IJCSI

International Journal of Computer Science Issues

Volume 10, Issue 1, No 1, January 2013
ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

www.IJCSI.org

IJCSI proceedings are currently indexed by:

Cornell University Library

Cogprints

Google scholar

.docstoc
find and share professional documents

ScientificCommons

View my documents on Scribd

BASE
Bielefeld Academic Search Engine

SCIRUS
search engine for science

SciRate.com

CiteSeerX beta

dblp.uni-trier.de
Computer Science Bibliography

Q·Sensei BETA

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

EBSCO HOST

ProQuest

www.IJCSI.org

IJCSI Publicity Board 2013

IJCSI Editorial Board 2013

EDITORIAL

In this first edition of 2013, we bring forward issues from various dynamic computer science fields ranging from system performance, computer vision, artificial intelligence, software engineering, multimedia, pattern recognition, information retrieval, databases, security and networking among others.

Considering the growing interest of academics worldwide to publish in IJCSI, we invite universities and institutions to partner with us to further encourage open-access publications.

As always we thank all our reviewers for providing constructive comments on papers sent to them for review. This helps enormously in improving the quality of papers published in this issue.

Google Scholar reported a large amount of cited papers published in IJCSI. We will continue to encourage the readers, authors and reviewers and the computer science scientific community and interested authors to continue citing papers published by the journal.

It was with pleasure and a sense of satisfaction that we announced in mid March 2011 our 2-year Impact Factor which is evaluated at 0.242. For more information about this please see the 3rd question in the FAQ section of the journal.

Apart from availability of the full-texts from the journal website, all published papers are deposited in open-access repositories to make access easier and ensure continuous availability of its proceedings free of charge for all researchers.

We are pleased to present IJCSI Volume 10, Issue 1, No 1, January 2013 (IJCSI Vol. 10, Issue 1, No 1). The acceptance rate for this issue is 33.09%.

IJCSI Editorial Board
January 2013 Issue
ISSN (Online): 1694-0814
© IJCSI Publications
www.IJCSI.org

IJCSI Reviewers Committee 2013

TABLE OF CONTENTS

Application of SVM Optimization Based on GA in Electronic Sphygmomanometer Data Fusion

Fengmei Gao[1] and Tao Lin[2,3]

**[1] School of Life Scienes and Technology, Xinxiang Medical University
Xinxiang, Henan, 453003,China**

**[2] School of Automation, Chongqing University
Chongqing, 400044, China**

**[3] School of Applied Electronics, Chongqing College of Electronic Engineering
Chongqing, 401331, China**

Abstract

If the proper kernel function parameter σ is chosen, using of the multi-sensor data fusion method based on SVM, the influence of cross sensitive disturbance variables including the temperature T and the power supply current I , can be significantly suppressed and the stability of the pressure sensor can be improved in the electronic sphygmomanometer. While kernel function parameter σ is difficult to ascertain after repeated test. GA(Genetic Algorithm) with powerful global searching for optimal solutions is able to meet the requirement of optimization for kernel function parameter σ of SVM(Support Vector Machine).

Keywords: *SVM, GA, Kernel Function Parameter, Multi-sensor Data Fusion.*

1. Introduction

In regard to human body, blood pressure usually refers to the surface arterial pressure of brachial artery, medically known as noninvasive blood pressure [1]. Blood pressure is one of important comprehensive physiological medical parameters, of which non-invasive detection methods include Korotkoff souna method, oscillometric method, double-cuff method, ultra-sound method, tension location survey method ,constant volume method, etc. Electronic sphygmomanometer, which has been treated specially about collection and filtering signal, can achieve blood pressure measurement in those regions which is not limited to only the upper-arm, but also the lower-arm and the wrist, even the finger. Oscillometric method is preferred in most electronic sphygmomanometer designs, because of its insensitivity from subjective measurement factors, conventional treatment, small equipment investment and good murmur resistance [1], [2].

The blood pressure measurement precision of the electronic sphygmomanometer based on oscillometric method is heavily conditioned, because whose pressure sensor performance is mainly influenced by temperature and power supply changes.With the development of the informatization and communalization of medical institutions, the electronic sphygmomanometer based on the systematic platform should satisfy patients' self-measurement, the diversity of adapter interface and working environment, and so on. Multi-sensor data fusion is one of the effective methods for improving reliability and measurement accuracy of the electronic sphygmomanometer [3], [11].

With SVM (Support Vector Machine) approach, the inverse model can be built to eliminate the effect of cross sensitive disturbance variables on the pressure sensor from the ambient temperature and the constant current power supply current in the electronic sphygmomanometer, so that the stability of the pressure sensor can be improved using the multi-sensor data fusion method of suppressing cross sensitive disturbance variables [3], [4]. In the training process of SVM, lots of trial are often needed to select kernel function parameter because of the complexity and nonlinear level of systems, but it is exactly these results that are at risk.GA(Genetic Algorithm) with powerful global searching for optimal solutions is able to meet the requirement of optimization for kernel function parameter of SVM, in order to modify the inverse model used for offseting the disturbance of cross sensitive disturbance variables, and to improve the measuring precision of blood pressure [4], [7].

2. Suppressing Cross Sensitive Disturbance Variables by SVM

Unlike multiple regression analysis, this method using of the multi-sensor data fusion based on SVM to suppress cross sensitive disturbance variables and improve the stability of the pressure sensor, don't have to establish the analytic function which has the untargeted parameters to be eliminated, but turn it into a convex quadratic optimization problem which is used to get theoretically the global optimization result by researching the estimation and prediction to small sample size according to VC dimension and structural risk minimization in statistical learning theory. Using kernel function by SVM, the sample points $\{(\boldsymbol{x_i},\ y_i)\}_{i=1}^{N^+}$ in input space $\boldsymbol{X}$ are mapped to the training points $(\boldsymbol{\varphi}(\boldsymbol{x_i}),\ y_i)$ in the higher dimensional Hilbert space $\boldsymbol{F}$, and the mapped training set $D=\{(\boldsymbol{\varphi}(\boldsymbol{x_i}),\ y_i)\}_{i=1}^{N^+}$ can catch regressions that is adopted to propose linear discriminant function in the Hilbert space $\boldsymbol{F}$ [5], [6]. It is this property that guarantees tthe generalization of the inverse model, can avoid the curse of dimensionality, becomes an irrelevance between the complexity of algorithm and the sample dimension, therefore can be suitable for the multi-sensor data fusion.

Let sample set be $\{(\boldsymbol{x_i},\ y_i)\}_{i=1}^{N^+}$, where $\boldsymbol{x_i}\in\boldsymbol{R}^d$ is input vector, y_i is corresponding expected value. The dual problem that corresponds to the constrained convex quadratic optimization problem can be expressed as

$$\arg\max_{\boldsymbol{\alpha}} \boldsymbol{\omega}(\boldsymbol{\alpha})=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}y_i y_j\alpha_i\alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)$$

$$s.t.\sum_{i=1}^{N}\alpha_i y_i=0\ ;\ 0\le\alpha_i\le C,\ i=1,2,\cdots,N \tag{1}$$

Where α_i is the Largrange multiplier, $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ is the kernel function.

Set $\boldsymbol{\alpha}^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_{N^+}^*)$ be the solutions for formula (1), only some of which is in general nonzero. The corresponding sample input $\boldsymbol{x_i}$ of nonzero solutions is used as support vector which a decision boundary decisions depends on. The purpose of data fusion based on SVM is fitting the relationship between the input $\boldsymbol{x}$ and output y .The Relationship expression is as follows [7], [9].

$$y(\boldsymbol{x})=\boldsymbol{\omega}^T\boldsymbol{x}+b=\sum_{i=1}^{s}\alpha_i K(\boldsymbol{x},\boldsymbol{x}_i)\ +b \tag{2}$$

In formula (2), $\boldsymbol{x_i}$ is support vector; s is the number of support vectors; $\boldsymbol{x}$ is the measured input; b is the offset of SVM; $\boldsymbol{\omega}$ is the weight coefficient of SVM,whose number is the same as the number of support vectors. The Gaussian radial basic function , which meets the Mercer criteria,is chosen and is as follows.

$$K(\boldsymbol{x},\boldsymbol{x}_i)\ =\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|}{2\sigma^2}\right) \tag{3}$$

Where σ is kernel function parameter, Adjustment of which can improve SVM predictive accuracy.

3. Measuring Blood Pressure Based on Oscillometric Method

3.1 Principle of Measurement

Blood pressure value parameters include systolic blood pressure and diastolic blood pressure during a heart pulsating cycle. To get blood pressure value, the oscillation envelope of pulses produced when blood flow strike vessel wall with oscillometric method, must be to detect and analyse [1], [3], [7]. Usually, oscillometric method can be divided into amplitude coefficient method and waveform feature method.

The measuring system of an electronic sphygmomanometer based on amplitude coefficient method is composed mainly of inflatable cuff, miniature electric air pump, electromagnetic gas valve, pressure sensor, temperature sensor, electric current sensor, microcontroller, etc. In many practical design based on oscillometric method, blood pressure can be measured in the course of inflation or deflation. Electromagnetic gas valve and miniature electric air pump under the control of microcontroller inject air into inflatable cuff at the rate of 5mmHg/s; The oscillation signal of surface arterial pressure pulses detected pressure sensor, is extracted alternating component through 0.8Hz second-order high-pass filtering and 300 times amplification, and gets A/D conversion through 38Hz second-order low-pass filtering to remove cuff frictional noise and power noise. The peak value $V_{\max}$ of oscillation envelope of pulses across the inflation cycle, and the concrete values of systolic blood pressure and diastolic blood pressure according to amplitude coefficient, are calculated by the microcontroller. As shown in Figure 1, the intersection V_{sp} of cuff pressure line and oscillation envelope of pulses is the amplitude value of systolic blood pressure, $V_{sp}/V_{\max}=k_{sp}$, k_{sp} always lies between 0.4～

0.65. In the process of an actual running, it is so critical to improve the stability of the pressure sensor for the accuracy of blood pressure measurement, because each pressure amplitude value of the sampling point obtained by pressure sensor will have an effect on the decision for the amplitude value of systolic blood pressure [1], [2].

In the process of blood pressure measurement, the ambient temperature and the constant current power supply current of pressure sensor is respectively under surveillance of temperature sensor and electric current sensor. This surveillance data willbe used to fuse the data of pressure sensor by the inverse model to eliminate the effect of cross sensitive disturbance variables.

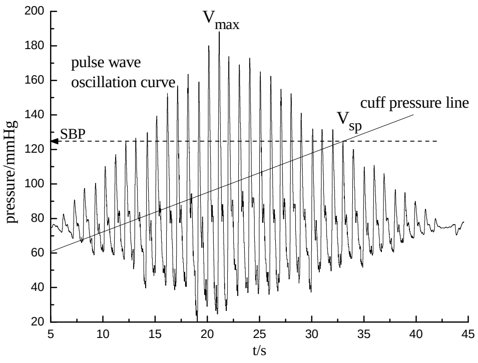


Fig. 1 Measuring systolic pressure based on oscillometric method.

3.2 Data sample preparation

For measuring systolic blood pressure, when SVM is trained to built the inverse model to eliminate the effect of cross sensitive disturbance variables, the three variables of pressure P , temperature T and electric current I are selected in three dimensional calibration experiment. In experimenting, the sensor choosing include the SC0073 dynamic micro pressure sensor, the JLB-11 electromagnetic balancing electric current sensor and the DS18B20 embedded miniature digital temperature sensor. The output voltage of temperature sensor used for detecting temperature disturbance variable T is U_T ; The output voltage of electric current sensor used for detecting electric current disturbance variable I is U_I ; The output voltage of pressure sensor used for detecting output variable P is U_P .

The number of total sample data pairs $(\boldsymbol{x}_i,\ y_i)\ (i=1,2,\cdots,\ N)$ is $N = N_p + N_t$ in calibration experimenting , where N_p is the number of training samples (N_p account for about 1/2～2/3 of the number of total sample), N_t is the number of testing samples [11], [14]. The three dimensional calibration experiment data of the SC0073 dynamic micro pressure sensor, as shown in Table 1.

Table 1: The three dimensional calibration experiment data of the SC0073 dynamic micro pressure sensor

SN	P /mmHg	I /mA	T /℃	U_I /V	U_T /V	U_P /V
1	80	5	-5	5.32	0.5	0.296
2	81	5	-5	5.32	0.5	0.384
3	82	5	-5	5.32	0.5	0.483
4	83	5	-5	5.32	0.5	0.536
5	84	5	-5	5.32	0.5	0.608
6	85	5	-5	5.32	0.5	0.713
7	86	5	-5	5.32	0.5	0.771
8	87	5	-5	5.32	0.5	0.884
9	88	5	-5	5.32	0.5	0.996
10	89	5	-5	5.32	0.5	1.08
……	……	……	……	……	……	……
282	115	11	45	13.32	1.1	5.509
283	116	11	45	13.32	1.1	5.642
284	117	11	45	13.32	1.1	5.756
285	118	11	45	13.32	1.1	5.895
286	119	11	45	13.32	1.1	6.002
287	120	11	45	13.32	1.1	6.249

4．GA Optimization

GA is a kind of simulated evolutionary algorithm, which imitate biological evolution law and encode the parameters of a problem to be solved into chromosomes whose information across group will be exchanged by operation including selection, crossover, mutation, etc., and will finally be able to develop by iteratively a globally optimal chromosome [4], [7].

Trained SVM with training samples is tested, as measured by the standard deviation MSETD between predicted output values and the pressure calibration value of testing samples, to reduce reliance of parameter choices on training samples. It is found through experiments that, these learning parameters of SVM including boundary of lagrange multiplier C , the condition parameter of convex quadratic optimization λ and ε -neighborhood parameter around solutions ε , have little impact on the output, while kernel function parameter σ is difficult to ascertain after repeated test [5], [9], [10]. Taking the standard deviation MSETD between predicted output values and the pressure calibration value of testing samples as objective function, kernel function parameter σ is optimized with global search performance of GA for optimal solutions, and then the proper offset b and the proper weight coefficient $\boldsymbol{\omega}$ are finded, so that output results are optimal or suboptimal to meet the precision and accuracy of system measurement.

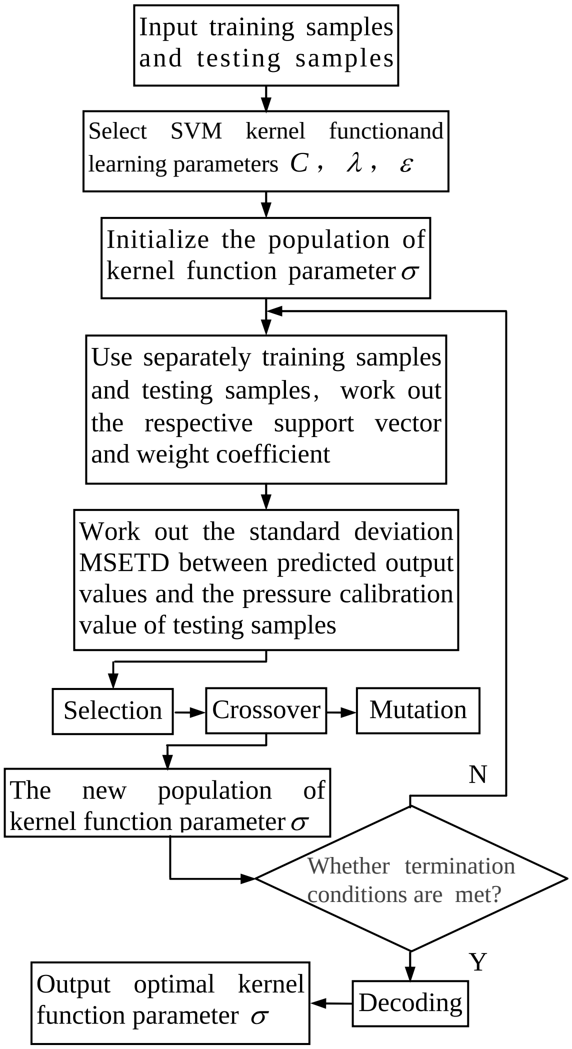

Input training samples and testing samples
Select SVM kernel functionand learning parameters C , λ , ε
Initialize the population of kernel function parameter σ
Use separately training samples and testing samples, work out the respective support vector and weight coefficient
Work out the standard deviation MSETD between predicted output values and the pressure calibration value of testing samples
Selection
Crossover
Mutation
The new population of kernel function parameter σ
N
Whether termination conditions are met?
Y
Decoding
Output optimal kernel function parameter σ

Fig. 2 GA optimizing kernel function parameter σ flow diagram.

GA optimizing kernel function parameter σ flow diagram , as shown in Figure 2.

4.1 Population Initialization

In the course of GA, the parameters in problem space aren't calculated directly, but the feasibility solutions of the problem to be solved are first expresses as Chromosomes or individuals in genetic space by encoding. Here, the population of kernel function parameter σ is initialized.

4.2 Fitness Function

Fitness function, which is used to distinguish the quality of individuals in population and is the only guide of natural selection, is usually derived from objective function. Here, the reciprocal of the standard deviation MSETD minimum between predicted output values and the pressure calibration value of testing samples is chosen as the fitness function value. The larger the fitness function value is, and the better the quality of individual is [3], [7].

4.3 Selection, Crossover and Mutation

Rely on selection , excellent individuals can be found from old population, and new population can be build to reproduce the next generation individual. The larger the fitness function value of individual is, the higher the probability of being selected is. Here, the roulette wheel method which is an selection method according to fitness Proportion, is used in selection, and the probability of being selected to the individual i is

$$p_i = \frac{F_i}{\sum_{j=1}^{n} F_j} \tag{4}$$

In formula (4), F_i is the fitness function value of the individual i , n is population size.

Crossover is used to randomly find out two individuals from the present population, whose chromosome information is exchanged and combined for each other to pass outstanding characteristics of father string down to son string, in order to reproduce the new excellent individuals. The real crossing method is adopted in crossover because all individuals are encoded using of real.

$$\begin{aligned} a_{kj} &= a_{ij}(1-d) + a_{li}d \\ a_{lj} &= a_{lj}(1-d) + a_{ki}d \end{aligned} \tag{5}$$

In formula (5), d is a random number on interval $[0, 1]$.

Mutation is used to reproduce a better individual, which has been randomly found from the present population and mutated slightly. Mutation aims at maintaining the diversity of population [3], [7]. The operation of mutation to the jth gene of the ith individual is as follows.

$$a_{ij} = \begin{cases} a_{ij} + (a_{ij} - a_{\max}) * f(g)\,, & r \geq 0.5 \\ a_{ij} + (a_{\min} - a_{ij}) * f(g)\,, & r < 0.5 \end{cases} \tag{6}$$

In formula (6), $a_{\max}$ and $a_{\min}$ are upper and lower bounds of gene a_{ij} . r is a random number on interval $[0\,,1]$; $f(g) = r' * (1 - g/G_{\max})^2$, r' is a random number, g is the present number of iterations, $G_{\max}$ is the maximum evolutionary generation.

4.4 Operation Results of GA

Here, population size is set to 40, binary digit capacity of variable is set to 20. Crossover probability is set to 0.7, mutation probability is set to 0.01, the maximum evolutionary generation is set to 40. After these learning parameters of SVM are set: boundary of lagrange multiplier C is 500, the condition parameter of convex quadratic optimization λ is 1e-10, ε - neighborhood parameter around solutions ε is 1e-6, genetic manipulation is implemented to kernel function parameter σ on interval $[0\,,10]$. Evolutionary process, as shown in Figure 3. Optimum value per generation, as shown in Figure 4.

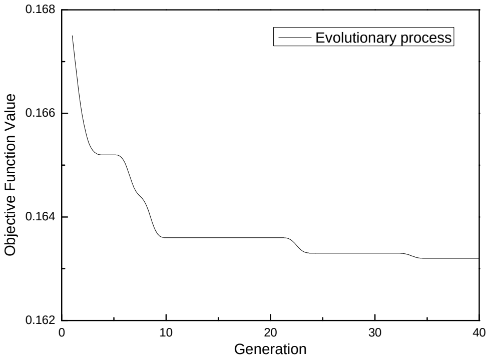


Fig. 3 Evolutionary process.

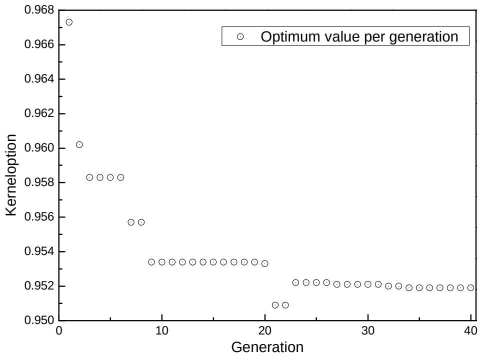


Fig. 3 Optimum value per generation.

According to the operation results of GA, kernel function parameter σ is 0.95187, the standard deviation MSETD between predicted output values and the pressure calibration value of testing samples is accordingly 0.16321.

5. Conclusions

When $\sigma =$ 0.95187, over the range of $\Box T =$ 50 ℃ and$\Box I = $6mA, the maximum fusion deviation of initial point $\left|\Box P'_{0m}\right| =$ 0.4016 mmHg; The full scale pressure $P_{FS} =$ 120 mmHg , of which the maximum fusion deviation $\left|\Box P'_{m}\right| =$ 0.6691 mmHg [12], [13]. So, the initial point temperature coefficient is

$$\alpha_0 = \frac{\left|\Box P'_{0m}\right|}{P_{FS}} \Box \frac{1}{\Box T} \tag{7}$$

the sensitivity temperature coefficient is

$$\alpha_s = \frac{\left|\Box P'_{m}\right|}{P_{FS} \Box \Box T} \tag{8}$$

the current impact coefficient is

$$\alpha_I = \frac{\left|\Box P'_{m}\right|}{P_{FS} \Box \Box I} \,. \tag{9}$$

The parameters contrast between before and after the fusion, as shown in Table 2.

From this it can be derived that , the influence of cross sensitive disturbance variables, the temperature T and the power supply current I in the electronic sphygmomanometer, can be significantly suppressed and the stability of the pressure sensor can be improved, using of the multi-sensor data fusion method based on SVM if the proper kernel function parameter σ is chosen.

Table 2: The parameters contrast between before and after the fusion

Evaluation Parameters	$\boldsymbol{\alpha_0}$ /℃	$\boldsymbol{\alpha_s}$ /℃	$\boldsymbol{\alpha_I}$ **/mA**
Before the fusion	6.296e-3	9.991e-2	8.326e-1
$\sigma = 0.95187$, Gaussian RBF kernel fusion	6.693e-5	1.115e-4	9.293e-4

Acknowledgments

This work has been supported through research grants from the Chongqing Natural Science Fund (cstc2012jjA10129).

References

[1] Yong P, Geedes LA. A surrogate arm for evaluating the accuracy of instruments for indirect measurement of blood pressur. Biomed Instrum & Technol, 24(7):130-135 (1990)

[2] Antti Jula, Pauli Puukka. Multiple clinic and blood pressure measurement versus ambulatoryblood pressure monitoring, Hypertension, 34(5):261-266 (1999)

[3] Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, S.W., Furey, T.S., Ares Jr., M., Haussler, D.: Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. U S A. 97(1), 262–267 (2000)

[4] Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S., Golub, T.R.: Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA. 98(26), 15149–15154 (2001)

[5] Vapnik, V.N.: Statistical Learning Theory. Wiley N.Y., Chichester (1998)

[6] Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification (2003)

[7] Houck, C.R., Joines, J., Kay, M.: A Genetic Algorithm for Function Optimization: A Matlab Implementation, NCSU-IE TR, vol. 95 (1995)

[8] Smith D., Singh S.:Approaches to multisensor data fusion in target tracking. A survey. IEEE Trans Knowl Data Eng 18(2):1696–1710(2006)

[9] Lv, Y., Wang, H., Sun, J.: A multi-sensor data fusion simulation platform design. Optics & Control 11(1), 22–24 (2004)

[10] Liu, Junhua.: Intelligent Sensor System (second edition). Xidian University Press, Xi’an (2010)

[11] Coue, C., Fraichard, T., Bessiere, P., Mazer, E.: Multi-sensor data fusion using Bayesian programming: an automotive application. In: Proceedings of IEEE Conference on Intelligent Robots and Systems, 30 September–5 October 2002, pp. 141–146 (2002)

[12] Vu T.N., Nguyen Q.T.: Application of Some Retrieved Information Method on Internet. International Journal of Computer Science Issues, vol.7, Issue 6, November 2010,ISSN (Online):1694-0814 (2010)

[13] Ren C. Luo, Min H. Lin and R. S. Scherp “Dynamic Multi-Sensor Data Fusion System for Intelligent Robots”, IEEE Journal of Robotics and Automation vol 4, number 4 (1988)

[14] Kokar M, Kim K H. Review of multi-sensor data fusion architectures and techniques. Proceedings of the 1993 IEEE International symposium on Intelligent Control, 261-266 (1993)

Fengmei Gao is a lecturer of Xinxiang Medical University, received the M.S. degree in electric machines and electric apparatus from Zhengzhou University of Light Industry, Zhengzhou, China, in 2006. Her research interests include biomedical engineering and intelligent control.

Tao Lin is a lecturer of Chongqing College of Electronic Engineering, received the M.S. degree in measurement technology and instruments from Chongqing Institute of Technology, Chongqing, Chia, in 2008. He is currently pursuing the Ph.D. degree at School of Automation, Chongqing University, Chongqing. His research interests include multi-sensor data fusion and intelligent control.

Cloud Computing and Agricultural Development of China: Theory and Practice

Yanxin Zhu [1,2], Di Wu[2] and Sujian Li1 [2]

[1]School of Mechanical Engineering, University of Science &Technology Beijing, Beijing, China；

[2]School of Business, Shijiazhuang University of Economics, Shijiazhuang, China

Abstract

Cloud computing technology has brought great opportunities to the development of China's agriculture；however it is also facing unprecedented challenges. According to the advantages of cloud computing, based on the status quo of China's agricultural development, the paper first discussed the impacts of cloud computing for China's agricultural development; and analyzed the field and the prospects of its possible applications in agriculture; then presented the application and promotion of cloud computing technology is a long-term system works, not only need to build the data center, integrate resources, enhance service capabilities, and also need to make information security .

Keywords: *Cloud Computing; China's Agriculture; Agricultural Informationization; Agricultural Modernization*

1. Introduction

With the continuous development of computer technology and network technology, various areas of the world have been undergoing enormous changes. The application of information technology will not only change the way of information interaction to shorten the distance of the world, but also conducive to social and economic development, improvement of production efficiency. Especially with the emergence and application of cloud computing technology, the resurgence of the climax of the national information construction, being seen as the third IT wave following the computer technology and Internet technology.

Currently, the countries in the world for the study of cloud computing technology is not very mature, Research in developed countries started earlier, and has made outstanding achievements in the basic framework, technical support, platform building. Major world-class IT companies, such as Microsoft, HP, Google, IBM, Oracle, and so on, have deeply realized the huge market potential and business opportunities in the field of "cloud computing", and all have been engaging in these studies (Zhang, 2010). Now, cloud computing has been used and promoted in the field of medicine and medical, manufacturing, financial services, energy, communication and other key areas, which will play an important role for improving the efficient use of resources, information sharing and integration. In China, Cloud computing applications in agriculture are in the phase of theoretical research, and lack mature cases. Therefore, this technology is great significant to improve management level in the weak field of agriculture information construction, the combination of agricultural informationization and modernization.

2. Cloud Computing Technology Overview

Cloud computing is a distributed computing technology, through a computer network the huge computing handler will be split and analyzed by a number of separate servers, then ultra millions or even hundreds of millions of information services will be available within seconds, so the users not only can get super computing capabilities but also can reduce resource inputs and waste. This is a paid service usage model, with ready access to demand unlimited expansion metering pay features, including IaaS (Infrastructure-as-a-Service), PaaS (Platform-as-a-Service), SaaS (Software-as-a-Service) and three levels of service(Chen & Deng,2009). Thus, cloud computing means computing power can be used as a commodity or service to be circulated and consumed through the Internet. Cloud computing technology application system framework is shown in Fig.1.

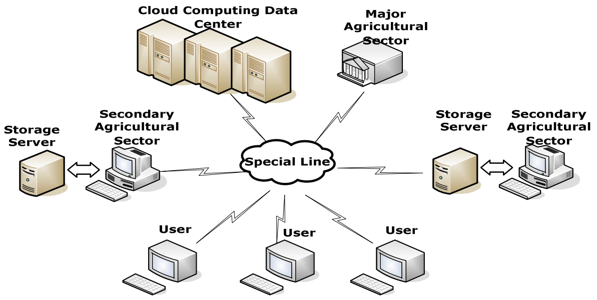

Cloud Computing Data Center
Major Agricultural Sector
Storage Server
Secondary Agricultural Sector
Special Line
Storage Server
Secondary Agricultural Sector
User
User
User

Fig. 1 Cloud computing technology application system framework

Cloud computing technology is an emerging hot technology appeared in recent years; it is very similar with utility computing and grid computing, and is considered as combined product with the computer technology such as grid computing, utility computing, distributed computing, network storage, load balancing, and network technology (Zhang & Gu, 2010) The evolution of this technology mainly experienced power plant mode phase of the pursuit of economies scale, utility computing phase, grid computing based on distributed processing and distributed cloud computing four stages(Chen,2009).

With Amazon launched Elastic Compute Cloud service, Google CEO Eric Schmidt first proposed the concept of "cloud computing", more and more IT companies such as IBM, Yahoo, HP, Dell have recognized the huge market potential for the technology advantages, and have started to research and promote in a number of areas. This technology also attracted the attention of the Governments of the United States, Britain, Japan and other developed countries, and all have begun to deploy a national cloud computing infrastructure, provide technical support for the development of information technology in areas such as government, economy, people's livelihood.

Cloud computing industry in China is still in the import phase; China Mobile, China Telecom, China Unicom, Huawei, Lenovo and other famous enterprises jointly established the "China Cloud Computing Technology Industry Alliance”, and began to explore the road of cloud computing technology. In the same time, all over the country have launched the cloud computing development plan, such as the Beijing "Xiangyun plan, the Shanghai "sea of clouds Plan", the Chongqing "cloud plans", Guangzhou days cloud plan”, shows China attaches great importance to the development of cloud computing technology; it also shows the determination to strengthen industry applications and promote economic development.

3. the relationship between Cloud computing and agricultural development

Although China has achieved fruitful results in crop cultivation, animal and plant breeding, agricultural production is still decentralized operation, low level information technology, coupled with farmers limitations constraints, the speed of agricultural modernization resulting is slow. Therefore, it is often the obvious contradiction between supply and demand in agricultural products; it not only hurt the enthusiasm of farmers engaged in agricultural production, reducing farmers' income, but also hindered the rapid socio-economic development. The applications of cloud computing technology in agriculture can solve the bottleneck problem of agricultural modernization and agricultural information, and can also break agricultural producers’ limitations in knowledge or technology, reduce duplication, improve utilization of existing resources to make up for dispersed, small-scale, regional differences agricultural production and the strong dependence on the natural climate vulnerability of agricultural production..

3.1Agriculture modernization needs cloud computing

Modernization of agriculture include three aspects: ① Widely use modern agriculture production equipment, agricultural machinery. ② Extensively use modern agricultural planting and breeding technology, Dohi technology, weather observation and forecasting; ③ Use modern forms of production organization and management methods, etc.

Europe, the United States or other developed countries have been the basic realization of agricultural modernization as early as the middle of the last century, but the level of agricultural development in China is still relatively backward, and still in the stage to forward the agricultural modernization. Seen from the development of agriculture, the agricultural mechanization goal has been basically achieved, but there are still many outstanding issues in technology and management, such as fewer agricultural technology service organizations and personnel, less necessary technical guidance, especially in the breeding, pollution-free crop cultivation and livestock breeding, soil testing, fertilizer, irrigation and soil improvement, meteorological observations and weather forecast were not enough technical support, most of farmers are in a state of blind conformity.

Organizational form of production in agriculture is relatively simple, backward, and a low degree of specialization of agricultural production areas, it is difficult to achieve Integrating Agriculture. In addition, due to the limitations of the farmers at market forecasting, business decision-making, information gathering and logistics management capacity is more lacking; it often leads to a mismatch between the supply and demand, not only damages the farmers' own interests, have also hindered the healthy development of the market supply and demand .

Therefore, to resolve these outstanding issues can not be separated from the IT technology application in the field of agriculture, especially cloud computing technologies play in the integration of resources, information sharing, online services differentiated advantages, will provide strong support for the realization of agricultural modernization.

3.2 Cloud computing role in promoting agricultural development

Cloud computing applications in agriculture makes agricultural producers do not need too much hardware and software investment, do not need to master advanced knowledge of computer and network technology; they can enjoy a more professional and more comprehensive services. The client just need to send the request to the cloud, then resources dispatch center will analysis and handle dynamically, and finally the corresponding processing results will be passed back to the client. For this calculation, the user does not need to know the calculation principle and process, simply according to the amount to pay. Agricultural producers can get planting and breeding techniques, pest control knowledge, and can also track and monitor the whole process of animals and plants from production, circulation to consumption, to achieve the scientific method in market forecasting, business decision-making, information collection and logistics.

Cloud computing application and implementation will play the following role:

(1) Agricultural Informatization

Agricultural information construction in China is relatively weak, compared with developed countries is still lagging behind. Some local government investment in the information construction is very inadequate, and producers can’t pay enough attention to the information, so low degree of information sharing hampered the process of the construction of agricultural information seriously (Qian, 2012). At present, import cloud computing technologies into agricultural industry, establish information network services platform, the level of Agricultural informatization will be a qualitative upgrading.

(2)Efficient use of agricultural resources

Decentralized management of agricultural production leads to low utilization of agricultural resources. However, cloud computing can integrate isolated production facilities, technical equipment, information services and other resources effectively; this form of paid services like as easy to buy hydropower (Cui, 2011).

(3) Promote the circulation of agricultural products

Currently, agricultural producers’ facing a prominent difficulty is the problem of sales of agricultural products. In China, farmers and consumers at both ends of the supply chain are difficult to derive much benefit because of small proportion direct sales, long distribution chain and complex link. Cloud computing will establish a bridge of communication between farmers and consumers; it is not only beneficial to the farmers to produce marketable products, as well as conducive to the realization of the value-added of the agricultural products.

4. Cloud computing applications in agriculture

4.1 High integration and sharing of agriculture information

During the transformation from China's traditional agriculture to modern and digital agriculture, increasing but disorderly information brings tremendous problems. Cloud computing offers a new management mechanism, which can integrate information resources in different regions and departments, build information sharing space and share infrastructure(Cao,2012) . In the ‘Agriculture Information Resources Cloud (AIRC)’, the agricultural sector and farmers can be real-time access to a full range of agricultural information that satisfies users extremely, and greatly reduces operating costs while substantially increase the efficiency of information haring. Meanwhile, the cloud computing technology has a powerful wireless access function. Users are able to get agricultural information through a variety of terminal not just the computer, which promotes the information sharing significantly.

4.2 Real-time monitoring and guidance in agricultural production

Application of cloud computing technology in agricultural production can be reflected in two aspects: production process monitoring and controlling, experiment simulation and support.

Currently, cloud computing technology already achieves real-time visual monitoring of crop growth(Zhang, 2011), not only able to quickly get the surface information, such as leaf area, leaf perimeter, stem diameter, stem height, etc, but also be able to detect the water and fertilizer content in the soil. Meanwhile, the crops information received from the cloud platform intelligent processing can automatically trigger corresponding improvement measures. For example: open the spray device when water content reaches the minimum threshold, alert to farmers when crops are ripe, identify weeds from crops and spray weed herbicide precisely.

Cloud computing technology also can be applied to the study of agricultural science. Particularly for some time consuming and high-cost experiments, or some experiments which are difficult to implement because of conditions limitations, the simulation can be great help to obtain the experimental results.

4.3 Providing agricultural science and technology service

As an important supporting technology of digital agriculture, cloud computing technology offers advanced information technology services, and realizes digitizing and visualizing expression, controlling, design and management of all the agriculture involving objects and the whole process. Agricultural extension, education and scientific research achieve trinity in the cloud computing environment. In addition, the cloud computing technology can be used to build precision agriculture technology and equipment systems, which make use of advanced agricultural production information and professional geographic information software to gain organic links among agricultural production and operating procedures. The system is able to optimize the investment in agricultural materials and improve material utilization, to achieve the purpose of reducing costs and increasing efficiency, and at the same time, it is able to effectively reduce the environmental pollution and realize sustainable agriculture development.

4.4 Construction and improvement of the agricultural products supply chain

Agricultural products have strong seasonal and regional features as necessities of life, which is prone to hoarding phenomenon. The convenience, breadth and popularity of the AIRC help farmers or agricultural enterprises understand the market information, the cloud platform facilitates the information exchange and communication between farmers and agricultural enterprises, it has very important significance for constructing and improving agricultural products supply chain, ameliorating agricultural products sales, and increasing farmers' profits. The agricultural products supply chain based on cloud computing technology is shown in Fig.2 (Qiu, 2010).

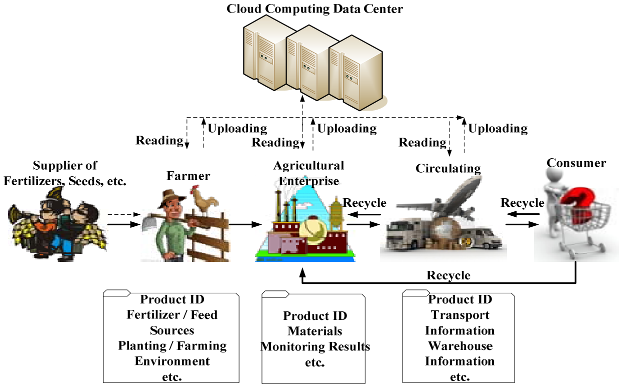

Cloud Computing Data Center
Reading
Uploading
Reading
Uploading
Reading
Uploading
Supplier of Fertilizers, Seeds, etc.
Farmer
Agricultural Enterprise
Circulating
Consumer
Recycle
Recycle
Recycle
Product ID Fertilizer / Feed Sources Planting / Farming Environment etc.
Product ID Materials Monitoring Results etc.
Product ID Transport Information Warehouse Information etc.

Fig.2. Agricultural Products Supply Chain Based on Cloud Computing Technology

4.5 Tracking and monitoring of the agricultural products quality

In the cloud computing platform, the animal husbandry can take advantage of advanced computer imaging technology to evaluate the animal meat, select and cultivate superior varieties, establish the magneto-therapy database and animal nutrition demand model, optimize feed formulation, to meet a number of animals nutritional needs indicators and exert the maximum production potential of livestock and poultry. In addition, tracking and monitoring of agricultural products quality and safety can be fully realized in the cloud computing platform. The cloud computing technology has been integrated into the scientific research, raw materials access, production and processing, storage and transportation, marketing, quality traceability and information services, inspection and quarantine, supervision and administration, etc.

5. Implementation of cloud computing technology in agriculture

Promotion and application of cloud computing technology is an inevitable choice to achieve the modernization and informatization of agriculture, is also an inevitable trend in the Internet technology popularization. But cloud computing application is still in its infancy stage and lacks references of success cases, therefore needs long-term exploration and step-by-step implementation. Meanwhile, it is more needed to raise awareness of the abundant agricultural producers and all-level government departments to improve their own qualities and enhance management capabilities for safeguarding the smooth implementation of cloud computing technology.

5.1 Build cloud computing data center

In order to operate and implement of the agricultural cloud computing technology better, first we must determine what constitutes the cloud computing data center and how to achieve the functions. The underlying of cloud computing data center is constituted by a large number of servers connected through the network and various types of controllers. Load balancing and computing virtualization are used for balancing the computing power of underlying server, and then dynamically deploy computing resources to agriculture-related personnel. Storage virtualization and cloud distributed file system are used to provide cross-server file storage service, automatically migrating information from the full server to get high utilization of storage resources (Cao, 2012). Application layer provides applied service for the agriculture-related personnel, users can select their desired landings when accessing to the cloud computing data center, but the underlying computing and storage details can not been seen. Cloud computing data center system is shown in Fig.3 (Cao, 2012).

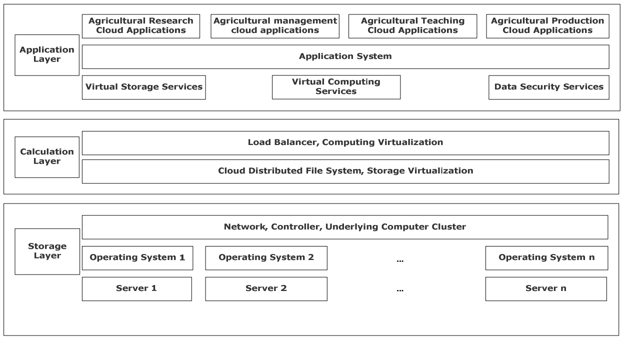

Application Layer
Agricultural Research Cloud Applications
Agricultural management cloud applications
Agricultural Teaching Cloud Applications
Agricultural Production Cloud Applications
Application System
Virtual Storage Services
Virtual Computing Services
Data Security Services
Calculation Layer
Load Balancer, Computing Virtualization
Cloud Distributed File System, Storage Virtualization
Storage Layer
Network, Controller, Underlying Computer Cluster
Operating System 1
Operating System 2
...
Operating System n
Server 1
Server 2
...
Server n

Fig. 3 Cloud Computing Data Center System.

5.2 Integrate agricultural resources

Repetition phenomenon is very serious in China's agricultural information construction and all-level agricultural sectors use their self-built information management system independently. Non-uniform informatization standards cause less networking to merge process relevant business between departments, plus low information shared degree, forming "islands of information". Therefore, in order to achieve cloud computing, first, we must establish the agricultural informatization standards complying with the law of agricultural production, and measure and reflect the characteristics and differences of the various components in agricultural management information system comprehensively, for facilitating establishment of national information platform (M.2010). Second, assess the utilization and applicability of the regional facilities and equipment resources, guide agricultural producers to rationally use local resources, and provide basis for implementation of cloud computing technology. Third, integrate technical resources of all enterprises, institutions and research institutes, strengthen the establishment of cooperation mechanisms in technology research and development, marketing, consulting, etc, and improve utilization and market-oriented operation of agricultural technology.

5.3 Improve the information service capacity of the agricultural sector

The agricultural sector will be the main force to promote the use of cloud computing technology in the agriculture field, and is also a direct participant in public cloud building. Service capacity and quality level of the agricultural sector will be directly related to the application results of this technology. Therefore, on the one hand, the knowledge level and technical ability of the agricultural sector personnel should be improved, being familiar with computer technology and network technology, to provide technical support for building and applying cloud computing platform. On the other hand, we should increase the service awareness, starting from the needs of agricultural producers, and eliminate bureaucracy and unrealistic blind construction. For the relevant government departments, they should be out of the misunderstanding that cloud computing is to build a data center, buy equipments and hardware. They not only should include the cloud infrastructure into the overall national plan for unified construction, but also should actively organize various types of research and development efforts to execute the research and development, pilot demonstration and promotion applications of the cloud application software, according to the regionalism, dispersion and farmers' ability to accept in agricultural production (Peng, 2011).

5.4 Pay attention to the agricultural information security

Cloud computing data center has strong openness and complicated business types, plus uncertain access source, will inevitably facing lots of threats and risks, therefore it is particularly important to better the data center information security. First, improve the security of data storage. Cloud computing data is divided into sub-modules and stores in different servers dispersing even cross-sector and cross-region, which is subject to many threats, so it is particularly important to backup the various blocks and update data to prevent the malicious tampering and attack. Second, implement the identification and classification of user rights to ensure the reliability, preventing the data destruction causing by illegal operation and intrusion. Third, strengthen network monitoring and maintenance efforts, and monitor network attacks situation and type in three-dimensional way in case of possible troubles (Wu, 2011).

6. Conclusions

Cloud computing technology attracts more and more attentions of countries and enterprises with its powerful advantages and market potential, the feasibility and applicability of whose application are also exploring in various industries. This technology will bring greater opportunities to the agricultural development in China, and also be the inevitable choice to achieve modernization and informatization in agriculture. However, all-level governments should be fully aware that its implementation will be a long exploring process, especially in the weak infrastructure and information construction agriculture area, where the application difficulties are hard to imagine. Therefore, government departments and institutions should pay full attention to the implementation of this technology, raise awareness, and especially provide strong support in platform construction, resource integration and service capabilities. We believe that with the supporting of modern information technology and network technology, China's agriculture is bound to usher in a rapid and healthy development period.

References

[1]Jianxun Zhang, Zhimin Gu,and Chao Zheng, " A Summary of Research Progress on Cloud Computing", Application Research of Computers, Vol. 27, No. 2, 2010, 429-433.

[2]Quan Chen, and Qianni Deng,"Cloud Computing and Its Key Technologies", Journal of ComputerApplications, Vol. 29, No. 9, 2009, 2562.

[3]Kun Qian, "The Application of Cloud Computing in Agricultural Management Information System", Hubei Agricultural Sciences，Vol.5, No.1, 2012, 159-162.

[4]Wenshun Cui, "Application and Developing Prospect of Cloud Computation in the Agricultural Informationization", Agricultural Engineering, Vol.2, No. 1, 2011, 40-43

[5]Liying Cao, Xiaoxian Zhang, and Yueling Zhao, "Application of Cloud Computing in Agricultural Information Resources Integration Mode",Chinese Agricultural Mechanization, No.3, 2012, 141–144.

[6]Mao Zhang, "Application of Computer Technology in Modern Agriculture", Agricultural Engineering, Vol.1, No.4, 2011, 26–28.

[7]Zhuqiang Qiu, Fei Wang, and Zhiyong Zhang, "Research on Self-constructed Traceability System Based on Agri-food Supply Chain Management",Guangdong Agricultural Sciences, No.4, 2010, 246–250.

[8]Mitsuyoshi Hori, Eiji Kawashima, and Tomihiro Yamazaki, "Application of Cloud Computing to Agriculture and Prospects in Other Fields", Fujitsu Scientific and Technical Journal, Vol.46, No .4, 2010

[9]Xiuyuan Peng, Xi Wang, Chuang Lu, and Kai Xuan, "Application of Cloud Computation in the Agriculture", Agriculture Network Information,No.2, 2011,8－10.

[10]Danhua Wu, Zhigang Huang, and Yongxian Liu, "The Prospect of Cloud Computing in the Application of Agricultural Information", South China Agriculture, Vol.5, No .9, 2011,61-63.

Yanxin Zhu, (1979-), Ph.D. postgraduate of University of Science &Technology Beijing; lecturer of Shijiazhuang University of Economics .She is engaged in the development of Logistics Engineering and Supply Chain Management.
Postal address: School of Business, Shijiazhuang University of Economics, No.136, Huaian East Road, Yuhua District Shijiazhuang City, Hebei Prov. China 050031.

Di Wu, (1985-), Ph.D. postgraduate of University of Science &Technology Beijing. She is engaged in the development of Logistics Engineering.
Postal address: School of Mechanical Engineering, University of Science &Technology Beijing, No.30, Xueyuan Road, Haidian District Beijing City, China 100083.

Sujian Li,(1959-). Professor of the Beijing University of Science and Technology, PhD Supervisor. He is engaged in the development of Logistics Engineering and Information technology.
Postal address: School of Mechanical Engineering, University of Science &Technology Beijing, No.30, Xueyuan Road, Haidian District Beijing City, China 100083.

Exploration on Big Data Oriented Data Analyzing and Processing Technology

Authors' Names and Addresses: XIAO DAWEI, No.31 TieShan West Road, Economic and Technological Development Zone Dalian, China,116600
XIAO DAWEI [1], AO LEI [2]

[1]Department of computer engineering, city institute, Dalian university of technology, Dalian, China

Abstract
At present, enterprises have urgent needs to conduct an effective and stable statistical analysis on big data. With a view to solving the issue of analysis and processing of big data in respect of enterprise business, this essay proposes a hybrid structure mode based on the MapReduce technology and the parallel database technology, discusses the principle on which the mode is used to realize the analysis and processing of big data and its advantages, and analyzes, expounds and proves the hybrid structure and provides a practical plan on big data processing. It is expected that this study has certain reference value in related researches. 103

Keywords: *big data, MapReduce, parallel database, hybrid structure*

1. Introduction

With the continuous improvement of informatization construction, most enterprises have completed the deployment of informatization system and the research and development, design and promotion of new products and new services have been greatly improved in respect of the operation efficiency of compared to traditional enterprises. However, although the enterprises have realized a high-efficiency and elaborate management, masses of business data are concurrently accumulated. Furthermore, there are categories of data and the requirement of being real-time is quite high. This is so-called "big data". All of the internet of things, cloud computing, mobile Internet, large-scale e-commerce, mobile phone, tablet computer and the various sensors spread all over each corner of the earth are data sources or the carrying way. At present, as a relatively new concept, big data is not proposed directly as the proper noun to give policy support by the Chinese government. However, in December 8, 2011, the Ministry of Industry and Information Technology of People's Republic of China issued the Internet of things " Twelfth Five-Year Plan", put forward as the information processing technology is one of the 4 key technical innovation projects, including the analysis of massive data storage, data mining, image and video intelligent analysis, which are important parts of big data. The other 3 key technical innovation projects including information technology, information technology, information security technology, are closely related with the large data. Large data has four characteristics: first, a huge amount of data, using PB as the unit; second, various data types, including network log, video, pictures, geographic location information and other data; third, the low density of value, with video as an example, the valuable data may be only one or two seconds in the continuous monitoring process; fourth, fast processing speed, this point is essentially different with the traditional data mining. Big data contains masses of valuable mode and information. For example, Wal-Mart, a global retail giant, will mine geographic locations, sales performance and social information of stores in big data to improve customers' understanding. For enterprise organizations, the value of big data is reflected in the two aspects: analysis and use and secondary development. The enterprises need to rely upon the technical platform on informatization system to mine in big data any needed important information and conduct intensive analysis and processing of these big data so as to construct a data warehouse and application and analysis platform based on important data information. By analyzing and processing the masses of the data information of enterprises, it will provide a correct guidance and policy-making for the existence and development of enterprises. Figure 1 illustrates big data management system structure. To design a system platform based on big data analysis, this essay proposes a solution which is big data oriented and integrates storage, management, analysis, processing and application. This proposal is very practical upon being tested and analyzed.

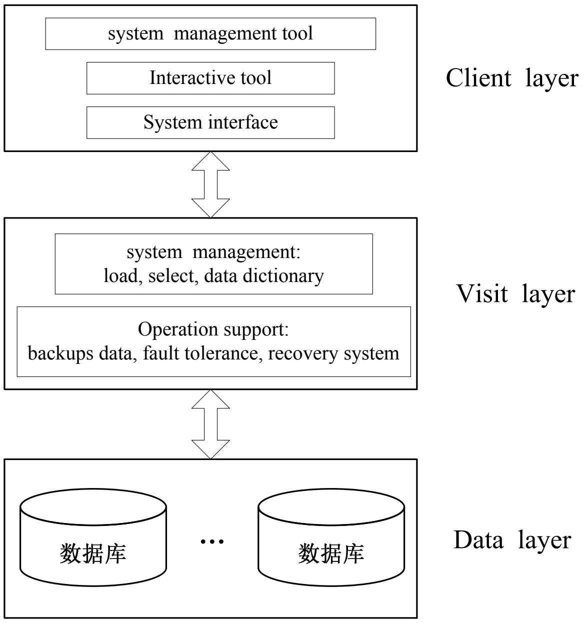


Fig.1 big data management system structure

2. Big Data Processing Technology

Theoretically, there are no limits to the improvement of the processing function of big data. At present, in the practical application on the processing of big data, the most often used and mainstream realization technology is the MapReduce technology, parallel database technology and the hybrid structure technology based on MapReduce technology and parallel database technology.

2.1 MapReduce Technology

In 2004, MapReduce was proposed by Google, it is an object-oriented programming model to deal with the large data, primarily used for processing internet data, such as document capture, inverted index construction. But because MapReduce has a simple and powerful data processing interface, and it hides many details on massively parallel execution, fault tolerance and load balance implementation, so the technology has been widely applied in the field of machine learning, data mining, data analysis, text tokenization, indexing research, creation of other kinds of data structures(e.g., graphs).

In a slide presentation, Google offers the following applications of MapReduce: distributed grep, distributed sort, web link-graph reversal, term-vector per host, web access log stats, inverted index construction, document clustering, machine learning, statistical machine translation.

The MapReduce technology realizes the abstract processing of complicated business logics involved in the parallel programming. It realizes complicated the computing process, provides simple and easily used interactive interface, and conceals the specific realization process for the parallel computing, processing, fault tolerance, data analysis and load balancing used in complicated businesses. The MapReduce technology includes two basic operation conceptions: Map(Mapping) and Reduce(Simplication). The Map technology mainly processes a group of input data record and distributes data to several servers and operation systems. Its means of processing data is a strategy based on the key/value. The Reduce technology mainly occupies itself in summarizing and processing the result after processing the above key/value. Issues and tasks in the real world may be modeled and described by means of this simple means of processing. Programs realized by this means will be distributed cluster. The data processing algorithm based on issues will be distributed to the distributed system formed by ordinary computers and then executed. The system will then solve the problem on the details in relation to the input of big data. Then the algorithm programs crossing computer cluster by means of the center server will be dispatched and managed: the management not only relates to the condition of each processing machine but to interactive communication request between the computers. Using such computing mode can help realize the mode of processing and computing of big data based on the distributed system structure for large-scale enterprises, without the need to grasp the process and details of the parallel processing. It will easily cause the realization of unified dispatch, management, storage, analysis and processing of the scattered resource information of the integration enterprise. It will realize the high-efficiency analysis and utilization of big data of enterprises by using the business data information of each branch of the enterprise.

MapReduce is designed for mass composed of low-end computer cluster, its excellent scalability has been fully verified in industry. MapReduce has low requirement to hardware, MapReduce allows to build cluster using inexpensive hardware. As a free open source system, MapReduce can store data in any format, can achieve a variety of complex data processing function. Analysis based on the MapReduce platform, without the need of complex data preprocessing and writing in the database process, can be directly analysed based on the flat file, and the calculation model which its use is mobile computing instead of moving data, therefore the analysis delay can be minimization. But the utility software based on MapReduce is relatively little, many data analysis function requires users to develop their own, which will lead to increasing cost. Because the MapReduce does not want to become a database system, so it does not provide SQL interface.

2.2 Parallel Database Technology

During the present phase, the popularizing and applying of relational databases feature a widest scope and they are at a mainstream position in the whole database system field. The original design object of relational database is to realize the application of the large-scale machines based on the “Host Computer – Terminal Computer” mode; however, its application scope is very limited. With the popularity and application of the “Client - Service”, the relational database system brings about an application era of “Client - Service” and is widely developed and applied. However, with the popularizing of the Internet technologies, the Internet information resources begin increasingly complicated, and the relational database begins to become unable to apply to complicated Internet application and cannot be used to express and administer each type of complicated document type and multi-media resource information. Therefore, the relational database system is improved and adjusted on this regard, such as adding the support function on the database system that is object oriented, at the same time, adding the function on handling each complicated information data.

Database processing technology based on parallel computing is a technology blended with the parallel computing mode and database processing technology. It originated in the seventies of the 20th century, mainly studies the parallelism of the relational algebra operations and the hardware design for implementation of relation operation, hope to realize some function of relational database operation by hardware. Unfortunately, the study failed. In the late 80s of the 20th century, the research direction of parallel database technology turned gradually to the general parallel machine, and the research was focused on the physical organization of parallel database, operative algorithms, optimization and scheduling policy. From the 90s until now, with the development of the basic techniques for processor technology, storage technology, network technology, the parallel database technology rise to a new level, the focus of the research is also transferred to the time and spatial parallelism of the data operation.

In the processing and analysis of the big data, data parallel processing manner is essential. Because the processing strategy of "divide and rule" provides unlimited reverie to extend system performance.

With the fast development and application of parallel processing and computing technology, people get to know that the processing of a parallel mode can be realized by means of the conceptions of time or space so that the processing efficiency of system tasks can be improved. After the entry of enterprises’ business information data into the big data era, this parallel computing mode can help well solve the data processing issue for large-scale enterprises’ business data system. Parallel computing includes two aspects: data parallel processing and task parallel processing. In terms of the data parallel processing means, a large-scale task to be solved can be dissembled into various system sub-tasks with the same scale and then each sub-task will be processed. As such, compared to the whole task, it will be easy to process. Adopting the task-paralleling processing mode might cause the disposal of tasks and coordination of relationships overly complicated. Using the parallel database technology is a means for realizing the parallel processing of data information. Parallel database support standard SQL language, through the SQL to provide data access service, SQ L is widely used because it is simple and easy to apply. But in big data analysis, the SQL interface is facing great challenges. The advantage of SQL comes from packaging the underlying data access, but the packaging affects its openness to a certain extent. User-defined functions which provided by parallel database is mostly based on the design of a single database instance, and therefore they cannot be executed in parallel cluster, it means that the traditional way is not suitable for the processing and analysis of big data. Moreover, the user-defined functions often need to pass complex system interaction in parallel database, and is familiar with the database structure and system calls, so it is difficult to use. Parallel database design is based on high-end hardware, software fault tolerance ability is poor, so the augmentability of parallel database is limited, the traditional data warehouse based on parallel database usually completed data preprocessing and analysis show with the help of external tools, so the data processing and analysis process involves a lot of data migration and calculation, of course, the analysis delay is often higher.

3. Constructing the Pattern on Big Data Processing

For the analysis and processing of big data, using a system based on the parallel database and data warehouse is not an ideal plan for the analysis and processing of big data. Using the combination of MapReduce and parallel database combines the advantages of the two means. After the analysis and comparison of the two means, the advantages and disadvantages of the two means can be seen. For such reasons, constructing a big data processing pattern based on MapReduce combined parallel data can on the one hand make up for their own disadvantages and can on the other hand improve the reliability of the system.

3.1 The hybrid structure mode of associating MapReduce with Parallel Database

SQL, as a universal relationship database system scripting language, has been widely used in the field of relational database, and SQL can be applied in the parallel database. Therefore, the SQL scripting language can act as a entry point for the combination of the two.

MapReduce defines a self-defined interface function for SQL scripting sentence and provides the same grammatical form as common SQL scripting sentence. Within the self-defined function realizes the data processing based on parallel computing. At the same time, such processing mode based on paralleling is applicable to the enterprise distributed system and can be used to hundreds of servers. These parallel processing machines and their computing of services conceal the realization details to users and are transparent for users, as a result of which big data processing can be realized by using MapReduce modules through SQL scripting sentence. The interface function of such MapReduce can normally operate under the database system circumstance, and its returned result sets remain a usable data table. Figure 2 illustrates the pattern on big data processing based on the MapReduce framework.

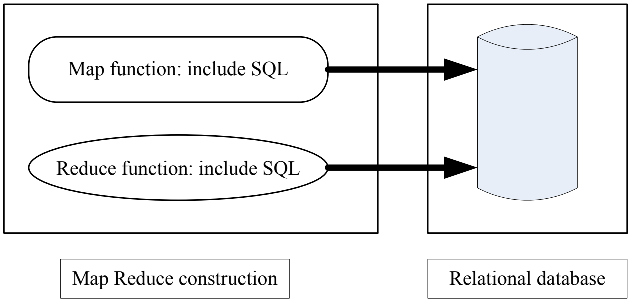


Fig.2 Pattern on Big Data Processing Based on the MapReduce Framework

3.2 Data Analysis

The loading, analyzing and processing of a large amount of data supported by data warehouse engines based on MapReduce can satisfy the need for the construction and mining needed for enterprise big data and can apply to the need of the performances and functions of the present and future information system. Below is an example on a Hotel Chain as the data analysis of solutions.

3.2.1 System Environment

1) Database Servers
Branch 1: DELL Power Edge R910.
Branch 2: DELL Power Edge R910.
Branch 3: DELL Power Edge R910.
Master Server: Dell 2950.
Data Storage Equipment: 5 MD1000 Direct Connection Storage.

2) Operation System
Two branches and the master server are respectively fixed FreeBSD Linux 9.0.

3) Network Environment
One switchboard for Gigabit LAN optical network, No Blocking Switch mode and Gigabit LAN optical network.

4) Figure 3 illustrates the Service Architecture of the Information System Clusters of a Hotel Chain.

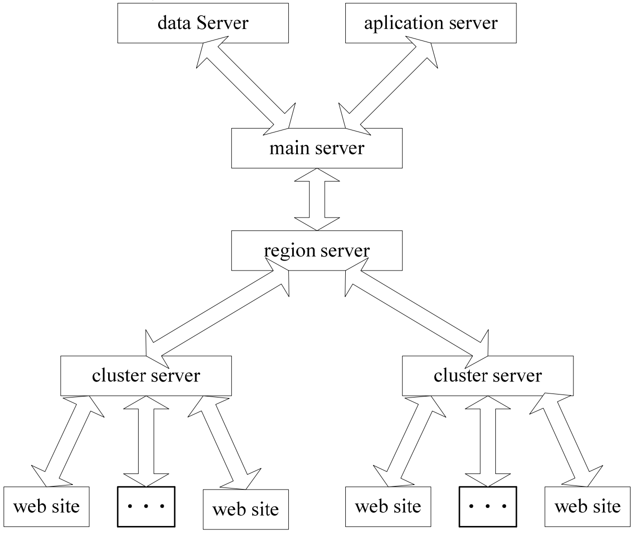


Fig.3 Tthe Service Architecture of the Information System Clusters of a Hotel Chain

3.2.2 Performance Test of System

1) Table 1 illustrates Loading,Rate of Engine Data

Table 1: Loading Rate of Engine Data

Filename	Description of File	Size of File	Time
Cinfo	Customer Information Form	90,134,500	3s
Userinfo	Customer Information Form	70,104,400	5s
Cjyinfo	Client Dealing Information Form	802,325,126,105	60s
Syslog	Journal Information Form	6,427,634,543	30s
Jysum	Dealing Summary Form	23,898,856,777,988	104s

2) Statistical Analysis of Data

The same business data will be loaded to Greenplum, Oracle and separate SQLSERVER database, the same SQL sentences will be executed by MapReduce. The time for the execution of sentences will be taken notes of. The systematic structure based on MapReduce is found to be highly efficient through test comparison. Table 2 illustrates Data Statistic Analysis Result.

Table 2: Data Statistic Analysis Result

Type of Operation	Record Number Of Source Forms	Time Consumed for Loading Map Reduce	Oracle	SQL SERVER	ImProVeMent
Customer Information Registration	1,014,997,563	5s	12s	16s	9
Statistics on Customer Dealings, polymerized computing	6,476,668,896,313	24s	46s	68s	28
Statistics on Customer Dealings, polymerized computing	35,478,993,448	9s	16s	22s	10
Information retrieval	22,015,202,412,569	36s	65s	101s	40

4. Conclusion

This essay introduces the mode of big data analysis and processing based on combined and parallel by MapReduce technology into the processing of big data of a hotel chain's information system, draws the loading rate of engine data, and makes comparison as to the execution time for loading the same data as such databases as Oracle and SQL SERVER. The result shows that a combined structure mode based on the combination of MapReduce technology and parallel database technology can improve the disposal efficiency of big data processing.

5. Research status and prospect

5.1 Research status

In recent years, the industry has design a variety of data analysis and processing platform through a lot of research, the following will introduct three typical platforms.

Greenplum MapReduce, MapReduce has been proven as a technique for high-scale data analysis by Internet leaders such as Google and Yahoo. gives enterprises the best of both worlds- MapReduce for programmers and SQL for DBAs- and will execute both MapReduce and SQL directly within Greenplum's parallel dataflow engine, which is at the heart of the Greenplum Database. Greenplum MapReduce enables programmers to run analytics against petabyte-scale datasets stored in and outside of the Greenplum Database. Greenplum MapReduce brings the benefits of a growing standard programming model to the reliability and familiarity of the relational database. The new capability expands the Greenplum Database to support MapReduce programs. Greenplum use MapReduce to improve data processing function of parallel database, but the scalability and fault tolerance of parallel database does not change.

Hive, defines a simple SQL-like query language, called QL, that enables users familiar with SQL to query the data. At the same time, this language also allows programmers who are familiar with the MapReduce framework to be able to plug in their custom mappers and reducers to perform more sophisticated analysis that may not be supported by the built-in capabilities of the language. QL can also be extended with custom scalar functions (UDF's), aggregations (UDAF's), and table functions (UDTF's).

HadoopDB, an open source parallel database. It is a hybrid that combines parallel databases with scalable and fault-tolerant Hadoop/MapReduce systems. HadoopDB is comprised of Postgres on each node (database layer), Hadoop/MapReduce as a communication layer that coordinates the multiple nodes each running Postgres, and Hive as the translation layer. The result is a shared-nothing parallel database, that business analysts can interact with using a SQL-like language. HadoopDB can't still be pushed down to the database layer for the complex connection operation (such as ring connection), so it didn't solve the performance problem fundamentally.

Despite there are a lot of big data processing platform, and they all have their own advantages, but they can't still solve the fundamental problems.

5.2 Research prospect

Simple function integration can’t effectively solve the problem of big data processing, so the research on hybrid architecture also need further.

There is a distance between the scalability of parallel database and the demand of big data analysis, so it is a very challenging task to improve the scalability of parallel database.

Despite the performance of MapReduce is increasing rapidly, there is still a large promoted space on some hands such as the parallelization of multiple analysis, complex analysis operation display, data compression efficiency.

References

[1] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters/ / Proceedings of the 6th Symposium on Operating System Design and Implementation(OSDΓ 04) .San Francisco, California, USA, 2004: 137-150.

[2] Xu Zipei , The Big Data Revolution. Guilin: Guangxi Normal University Press, 2012.

[3] Peng Hong, and Du Nan, " Research of parallel technology in massive commerce data management system ", Application Research of Computers, Vol. 26, No. 2, 2009, pp. 614-616.

[4] Wang Guiqiang, and Lu Chaojun, “Probing Parallel Technique-Based Statistical Analysis For Enormous Data”, Computer Applications and Software, Vol. 28, No. 3, 2011, pp. 162-165.

[5] Yu Chuli, Xiao Yingyuan, and Yin Bo, “A parallel algorithm for mining frequent item sets on Hadoop”,Journal of Tianjin University ofTechnology, Vol. 27, No. 1, 2011, pp. 25-28.

[6] Wang Min, Zhang Hong, and Yan Peng, “Parallel Technique Analysis for Effective Reducing Test Cost”, Computer and Digital Engineering, Vol. 38, No. 9, 2010, pp. 13-15.

[7] Das S, Sismanis Y, Beyer K S, Gemulla R, Haas P J, McPherson J, Ricardo: Integrating R and Hadoop. Proceedings of the SIGMOD. Indianapolis, the United States, 2010: 188-190.

[8] Zhang Li, SQL Server Database Principle and Application. Beijing: Tsinghua University Press, 2009.

[9] Li Huazhi, Database Solution, Beijing: Publishing House of Electronics Industry, 2010.

[10] Wang Shan, Wang Huiju, and Qin Xiongpai, “Architecting Big Data: Challenges, Studies and Forecasts”, Chinese Journal of Computers , Vol. 34, No. 10, 2011, pp. 1741-1752.

[11] http://www.dbms2.com/2008/08/26/known applications of mapreduce/

[12] http://www.asterdata.com/product/mapreduce.php

[13] http://www.greenplum.com/technology/mapreduce/

[14] https://cwiki.apache.org/confluence/display/Hive/Home

[15]http://strata.oreilly.com/2009/07/hadoopdb-an-open-source-parallel-database.html

First Author Xiao Dawei, Dalian, China. Birthdate: October,1978, received her master degree in computer application from Shenyang Jianzhu University of China. She is currently working as a full-time lecture in city institute of Dalian university of technology. She has published a work named “Computer Composition and Design” and 6 journal papers. She received excellent guide teacher award in 2011 in National Undergraduate Electronic Design Contest. Her research field is computer composition principle, database theory, MCU(microcomputer unit) design.

Second Author Ao Le**i**, Dalian, China. Birthdate: February,1979, is a master of Computer Science, graduated from the software engineering of Northeastern University in 2008. He is a lecturer of Computer science and technology in City Institute of Dalian University of Technology. He has published “The study of the course content and teaching method on Neural Network”, “Computer network educational reform and practice”, “The Design of the network module in the Embedded Database Management System”, “The Design and Implementation of the Embedded Database Management System Based on VxWorks”, “Research of computer network training course content”, “The Design of RACK Laboratory Network” , “computer network experiment course”, “Comprehensive training of computer network Course” from 2009 to 2012. His research direction is Embeded System Development and network engineering.

Computer Network-based Multimodal Teaching of British and American Literature

Liu Xucai
Foreign Languages School, Changchun Normal University,
Changchn, Jilin, China

Abstract

Guided under the multimodal teaching theory, the researchers mainly discuss the practice of the computer network-based multimodal teaching of British and American literature. This article emphasizes that the teachers take advantage of pictures, network screenshots, movies, video, sound, and other resources in the classroom lectures to mobilize the coordination operation of the auditory, visual and tactile senses, to enhance the students' impression of the writer and his works. By doing so, the students have a good understanding of the original work, instead of having the boring sense of the single text-mode teaching and the distress of comprehension of literary works, and have no difficulties in comprehending the text and then the classroom teaching is improved.

Keywords: computer network, teaching of British and American literature, multimodality, image mode, sound mode

Introduction

In the information age, the computer network provides people with various information which include both excellent pictures and texts audio-visually in a quick and convenient way. People no longer rely on a single text to get the information in the form of books, newspapers, etc. Students no longer obtain knowledge only through books and classroom attendance, while they can acquire knowledge, update knowledge and accumulate knowledge through the computer network. The interactive learning environment with friendly interface and intuitive image as well as a rich source of information provided by the computer network helps students to get more knowledge, help to stimulate students' interest in learning and collaborative learning and help students to take the initiative in finding and exploring. Therefore, the mode of teacher-centered classroom teaching can not meet the individual needs of students.

Because of this, the computer network or multimedia has been successfully applied to modern English teaching process, presenting unprecedented multimodal forms of the English teaching. Through information and communication technology (ICT), curriculum resources are rationally used and actively developed. Abundant teaching resources on the Internet are fully used. Students are provided with curriculum resources which can present students with actual life, and whose content is rich and instructive. The ways of learning and using English are also expanded. And teachers can enrich teaching content and forms to improve the effect of English teaching by using a variety of audio-visual resources and network resources. At the same time, teachers can take advantage of the computer and multimedia teaching software to explore new teaching modes to promote personalized learning. Therefore, the computer network technology changed the original single and monotonous books teaching, and made multimodal teaching feasible.

2. The Multimodal Teaching Theory

Multimodality refers to the combination of different semiotic modes in a finished communication or communication activities. Semiotic modes are these different systems for meaning-making, or possible "channels" such as speech, writing, images, etc. Semiotic modes can include visual, verbal, written, gestural and musical resources for communication. They also include various "multimodal" ensembles of any of these modes (Kress and van Leeuwen, 2001). Multimodality, the multiple presentations of modality, is the most important factor in teaching. Multimodal educational and pedagogical approach promotes not only the teaching through the means and principles of ICT education, e-learning and modern technology, but also the reforming of traditional culture(Kanari& Potamias,2011).

This theory, mainly in the perspective of social semiotics and based on the theory of Halliday's functional grammar, holds that the traditional paralinguistic images, colors, music and action symbols are no longer in the auxiliary position in modern social communication, but form a broader semiotic resources along with text symbols to make meaning.

Multimodal teaching advocates that teachers should take advantage of more channels and more teaching methods such as websits, pictures, role-plays, etc, to mobilize students' multiple senses engaging in language teaching. According to Kress et al., (2001), teachers often use gestures together with speech to draw attention to images and other references within the classroom. In particular, they maintain: various modes are interacting and interplaying: gestures, drawings, speech, objects. Each mode contributes to meaning construction: speech to create a difference, an image on the blackboard to get a visual background, manipulation of an object to locate the discussion in the physical setting, action to make clear the dynamic nature of the concept, the image in the textbook to do a stable summary, cohesion is achieved through repetition, synchronization, similarity and contrast.

3. Characteristics of the British and American literature course and the feasibility of multimodal teaching

3.1 Characteristics of the British and American literature course

3.1.1 Situational

The course of British and American literature includes British and American literary history, writers, and selected reading materials. All these aim at depicting beautiful pictures for readers. While reading, the readers can imagine seeing the Paradise that God has built for Adam and Eve and the enormous marlin that the old man Santiago catches, entering the palace where Prince Hamlet lives, landing the island where Robinson Crusoe lives lonely. The sense of these pictures is the manifestation of the literature class scenarios.

3.1.2 Episodic

The writers, works and the characters in the works in the British and American literature are all related to some moving stories. While reading works, the readers can "experiece" the plots, accomplishing the missions and overcoming the innumerable hazards and hardships, such as Bewulf's fighting closely with the mosters, Hamlet's revenging, Adam and Eve's being expelled from the Garden of Eden, Tess's being hanged, Rip Van Winkle's long sleeping, etc. The characters and their stories in the works contain the complex plots.

3.2 The feasibility of multimodal teaching under computer network-based conditions

Moreno and Mayer (2007) believe that

multimodal learning environments use different modes to represent content knowledge, for example verbal and non-verbal, where the non-verbal mode is the pictorial mode including static and dynamic graphics. These different presentation modes (verbal and non-verbal) are used to attract students' different sensory modalities (visual, auditory and tactile). Once the stories of a literay work and the pictures depicted in it are combined with the languages, the work is alive, which make the readers a sense of reality. Thus, every literary text is an integration of various modes, instead of a single text mode. Moreover, the computer internet happens to be able to provide the multimodal sounds, texts, images, or even different colors and fonts for English and American literature teaching, fully facilitating students' auditory, visual, tactile and other senses, which make this course organic integration of many means of languages, images, sounds, texts and symbolic resources.

4.Multimodality teaching construction of British and American literature course

The British and American literature teaching under computer network conditions, is a multimodal auxiliary teaching mode based on its own curriculum characteristics, the multimodal teaching theory and on the use of computer network platform.

Nowadays, the network culture is very popular, and it is necessary for teachers to offer the students who are addicted to internet culture some English and American Literature learning websites, and guide students to visit and browse them, learn something and finish a certain amount of job. Therefore, the teaching mode of British and American literature can be transformed from the mode of single teacher-centered lectures to the teaching mode based on classroom teaching supplemented by students' online study, and teaching presents three main modes, whose positions in the teaching process are as follows: text mode, image mode and sound mode.

4.1 text mode

In the course of British and American literature teaching, text mode is always playing a leading role. It includes the text in students' book, the text in the courseware, the text in the reading materials which students complete reading online as a task, the writing text which includes the literary essay writing assigned, the adaption of literary screenplays, etc.

4.2 image mode

The image mode includes all kinds of image materials that are displayed for literature learning, such as the course videos which the teachers upload to the LAN, network ppt courseware including the pictures, drawings and screenshots in the courseware, literary movies adapted from literary works, and even literary drama performances with living images. This mode makes text mode more vivid and intuitive.

4.3 sound mode

The sound mode regardes the sound as the carrier, including the ppt courseware for classroom teaching, the network ppt courseware, course videos, adapted movies, drama recording, which makes text mode and image mode more vivid.

Among the three modes, text is the main mode, in charge of providing key information, while image and sound are auxiliary modes, in charge of providing the background information. Although they emphasize different aspects, they are used to achieve the same macro purposes as is to enhance communication purposes of literature teaching content.

5.Multimodal teaching practice of British and American literature course

In the course of multimodal teaching, text mode plays the leading role, but they interact to and reinforce one another, build the same discourse meaning of British and American literature. In one case, image mode and sound mode make the features of text mode prominent; in the other case, text mode also makes the features of image mode and sound mode obvious.

We will take the teaching of Shakespeare and one of his masterpiece Romeo and Juliet for example in the following to present the whole process of the three modes.

5.1 text mode

When we introduce Shakespeare's life story and the four periods of his works, we mainly use the text mode, supplemented by ppt courseware. The text mode in the courseware always plays the leading role. For example, Shakespeare's writing career has been often divided into four periods. In the first period (1590-1594), he wrote mainly comedies influenced by Roman and Italian models and four history plays in the popular chronicle tradition. His second period began in 1595 with the tragedy Romeo and Juliet and ended with the tragedy of Julius Caesar in 1599. From about 1600 to about 1608, his third "tragic period", Shakespeare wrote mostly tragedies, and from about 1608 to 1613, the fourth period of mainly tragicomedies, also called romances. All these are presented by text mode, because only the text mode can express the boundaries of time and classification of the works. We use the text to express the time and classification, supplemented by pictures, screenshots, videos, etc. Because of its strong visual impact，the image mode，as the background of the teaching, emphasizes the literal meaning of the expression, and highlights the text mode. Meanwhile, we add sounds to the same page of ppt courseware using a custom animation effect, which causes the combination of the clear text, the obvious image and the striking sound. For example, when we add the "typing sound" to the page together with the voice in the video shots, we will make the student's auditory senses involved in the text mode. At this time, the auxiliary role of image mode and sound mode will deepen the students' impression of the text mode.

In addition, extracurricular literary reading is also based on text mode. After the teachers' guiding in class, the students can read the original works, or read the materials downloaded in the the designated campus network platform to expand literary background. Accordingly, the students must read the original works and download "film and television scripts" from the network platform to perform it in the later literary practice. Moreover, the teacher will teach the students how to analyze the literary works, the films adapted from the lietray works, and how to write critical literary essays, which is also based on the text mode.

5.2 image mode

Image mode is the auxiliary mode in the classroom lectures or the display of text mode. However, the image is the main mode in the image-based pages of some parts of the ppt courseware, or the course videos, online videos, screenshots, while text mode and sound mode are the auxiliary modes. For example, when the teacher uses the pictures to illustrate the story of Romeo and Juliet, the picture will take most of the entire ppt page, while the text is at the bottom of the picture, only as the role of the caption.

Likewise, the image is the main mode in film videos, or screenshots, while text mode is the auxiliary mode as the form of the caption at the bottom of the picture and sound mode is the auxiliary mode synchronizing with the screen character language. At this point, the text mode highlights the image mode for its function of interpretation, and the sound mode makes the image mode vivid for its synchronization, which also has a strengthening effect.

The primary modal role of the image is also reflected in the student network learning. Through the course network platform, students can watch the course videos, browse the network courseware to consolidate the content of school textbooks. According to the problems set by the teacher, after reading a work, the studnets can also see the movie adapted by the original work on the websites recommended by the teacher. For instance, after reading the work of Romeo and Juliet, students are encouraged to see the movie of different editions, such as 1936 edition, 1954 edition, 1968 edition, and 1996 edition. The students can understand the interpretations of the work by different directors and also form their own understanding of the work. The movie image mode is the interpretation of text mode of the work, and the sound mode synchronized with it enhances the image mode.

Another manifestation of the image mode is the students' drama performances based on the adaptation of literary works. In the play, according to the text description, the students set the classroom and the small stage, and play the roles based on the pre-recorded lines. Although the performances are based on the text mode, and occasionally, the switching of the scenes also needs the notice board, the image mode (live image) is the major mode before the audience. The text mode provides the performing situations for the image mode, and the sound mode – pre-recorded sound provides a clue for the performances.

5.3 Sound mode

Sound mode is always responsible for providing background information. In the ppt courseware, a text or picture to enter or exit needs the hints of the sound background to emphasize the input of a new piece of information. In the course videos, movie videos, sound is the guarantee to make the characters of the pictures lifelike.

However, when the students make drama performances after they record beforehand in accordance with the need of the scenes, tasks, background, drama performances recording – the sound mode becomes the main mode. Still take Romeo and Julia for example: The students download the related script from the teaching platform, adapt it according to their own understanding, pre-record the lines in accordance with the role assignment, soundtrack according to the story, make the mp3 format of the play. While playing the roles, the studnets just perform and converse based on the recording of the musical situation. At this moment, the playing sound dominates every performer on the stage, and also the development of the story. Thus, it becomes the major mode. On the contrary, the notice board (text image) of the sub-scene for live performances and the whole scene (image mode) play a supporting role.

6. Conclusion

Compared with the text mode of the past teaching material, the computer network-based multimodal teaching makes every mode interdepend and promote mutually in the use of the computer network technology. The different modes (verbal and non-verbal) are used to appeal to students' different sensory modalities (visual, auditory and tactile,etc). Moreover, multimodal courses allow instructional events or elements to be presented in more than one sensory mode (multiple representations), and then have been used to further facilitate student's learning (Shah & Freedman, 2003). Based on this, British and American literature classes become more vivid, lively and effective.

In the multimodal teaching, the text mode runs throughout all aspects of teaching, and always bears the irreplaceable role in the interpretation of the other modes; at the same time, the text mode, along with the sound mode, image mode, complement each other, which makes the text vital. The visual and auditory impacts of the sound mode and image mode supplement the lack of sense of pictures in the original works, fill students' sensory gaps, fully mobilize students multiple senses such as hearing, vision, strengthen the significance of the original works, as well as largely eliminate the difficulties in the students' comprehending the text.

The quick pace of change from text-based to more modes of presentations of information involves a quick response from language teachers to take advantage of multimodality to engage learners in meaningful cognitive, critical understandings. More close attention to the meaning-making potential of the multimodal teaching and learning can help language teachers and learners to cope more efficiently as they face new modes of information presentation.

References

[1] Abbas Pourhossein Gilakjani (2011). The Effect of Multimodal Learning Models on Language Teaching and Learning. *Theory and Practice in Language Studies*, 10, 1321-1327.

[2] Kress, G. R., & van Leeuwen, T. (2001). Multimodal discourse: The modes and media of contemporary communication. London: Edward Arnold.

[3] Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. *Educational Psycholog y Review*, 19, 309-326.

[4] P.Kanari,G. Potamias(2011). 4th International Conference of Education, Research and Innovations. http://library.iated.org/view/KANARI2011MUL, 2011, 2805-2810.

[5] Shah, P., & Freedman, E. G. (2003). Visuospatial cognition in electronic learning. *Journal of Educational Computing Research*, 29 (3), 315-24.

First Author Liu Xucai, female, born in Suangyang County, Jilin Province, 1972, is an associate professor in Foreign Languages School, Changchun Normal University, majoring in British and American Literature.

The Research and Implementation of the Key Techniques on Post-graduate Degree-granting Online Information Collection System

Ying-lai HUANG[1],Meng Ga[2],Chun-Ying Li[3],Jing Chen[4]

[1] Information and Computer Engineering College, Northeast Forestry University,
Harbin, 150040, China

[2] Information and Computer Engineering College, Northeast Forestry University,
Harbin, 150040, China

[3] Graduate Schools, Northeast Forestry University,
Harbin, 150040, China

[4] Graduate Schools, Northeast Forestry University,
Harbin, 150040, China

Abstract

On the premise of meeting the basic requirement of the post-graduate degree-granting online information collection, this paper discusses the key technologies and system optimization idea as well as its realization method involved in the system design and development process, including the SSH framework technology, the data initialization technology on template, the dynamic forms filled on AJAX technology and the generation technology of DBF data report forms, etc, through the combination of optimizing idea and modern information technologies to achieve the win-win performance of improving the system applications satisfaction and optimizing the work quality.

Keywords: *information collection, SSH, dynamic forms, system optimization*

1. Introduction

Along with the fast development of information technology and networks, it has become the latest trend by combining the computer hardware, software with office concept fully integrated with advanced management ideas in office area, and it attains the goal of improving office efficiency, alleviating the burden of work, and strengthening the quality of work by cooperating with each other. To the domain of education office, especially the work of degree information collection, its characteristics of large quantity of management objects, great refinement of the categorization, data tightly knit and tight restriction of time node determine that the information collected work is difficult and the statistical task is heavy. Therefore, it has a very important significance to apply modern information technologies in graduate degree-granting information collection work in order to improve the office level of university. The "China Academic Degrees and Graduate Education Information" oriented to all the universities and post-graduate of the country has been erected by the Academic Degrees and Graduate Education Development Center of the Ministry of Education, universities submit the graduate degree information in DBF format to the system every year for making copies and facilitate future queries use. However, the way of information collected and statistical work currently is not the same among the universities, most of them still use paper-based media to pass data repeatedly among students and staff, although it has achieved a certain degree of automation, it is only limited to the use of office software applications, which can't attain the goal of digital management completely. This paper establishes a whole post-graduate degree-granted online information collection system, with the analysis and implementation of the key technologies; it has attained the goal of digitization management of the whole process arranging from students' form filling to staffs' submitting.

2. System Analyses

2.1 System outline

The goal of post-graduate degree-granting online information collection system is to integrate students, college secretaries and university graduate academy into a unified platform, carrying out the management work based on the graduate degree-granting information data. With the digital management during the whole data collection and statistical process, it will reduce the use of paper-based media and improve office efficiency, ensure the quality of data and finally achieve the goal of low-carbon environment.

The system consists of six modules: (1).Personal information management module, which contains personal information view and maintain.(2).Degree information management module, which contains audit degree information, qualified degree information and unqualified degree information.(3)Statistical analysis module. (4)User information management module, which contains the college user management and Graduate Academy user management.(5)Code tables management module, which contains the import of code tables and the maintenance code tables.(6)Mailbox management module, which contains writing mails, mail inbox and outbox.

2.2 System Flow

The flow of Graduate degree granted online information collection system begins with the underlying data initialization and ends with degree-granting information submission. Firstly, the administrators import the basic data of responsible person of the University Graduate Academy and the colleges of related colleges; afterwards, each college is responsible in importing the students' basic information, including name, ID number and contact information; then, students use ID number to login in the system to further complete the personal information and submission; colleges review the corresponding students information submitted, if an information is qualified, then submit to the Graduate Academy for the second audit, otherwise, give a feedback to the student for re-modify; finally, the Graduate Academy makes the second audit to students information submitted from colleges, if one information is not qualified, gives a feedback the college for modifying, otherwise , the confirmed degree information will be exported in DBF style and submitted to the national degree office. The system workflow shows in Figure 1.

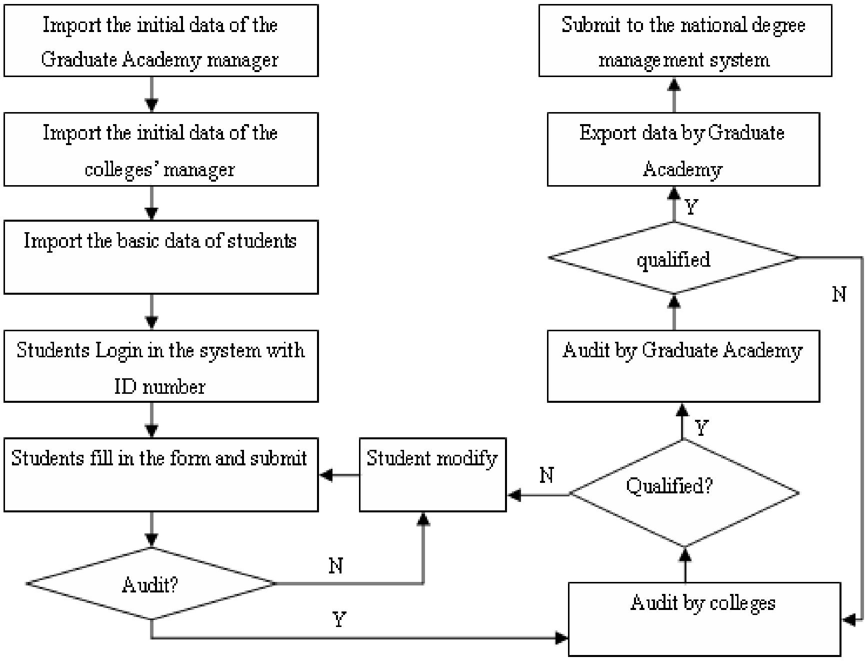

Import the initial data of the Graduate Academy manager
Import the initial data of the colleges' manager
Import the basic data of students
Students Login in the system with ID number
Students fill in the form and submit
Audit?
N
Y
Student modify
N
Qualified?
Y
Audit by colleges
Audit by Graduate Academy
qualified
Y
N
Export data by Graduate Academy
Submit to the national degree management system

Figure.1 System work flow diagram.

2.3 System Network Architecture

The design and implementation of the system is based on Browser/Server architecture, which enables users can login in the system to do management and maintenance work anywhere with networking. Almost all of the National Universities have the Network Information Center, System application and database are separately deployed in the network center, and use firewall and other safeguards to protect the system from the attacks of unauthorized users. The Legitimate users can access to the system through a browser on any PC with networking. System network architecture shows in Figure 2.

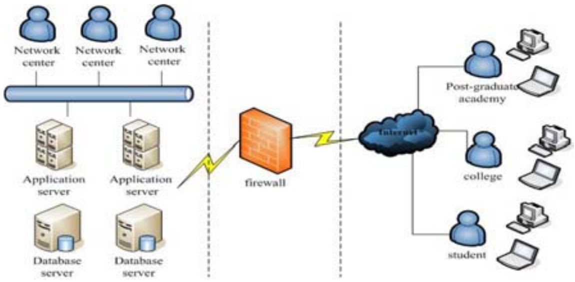


Figure.2 System Network Architecture figure.

3. Analyzes and implementation of key techniques

3.1 Struts+Hibernate framework technology

Generally, JSP+Servlet technology portfolio is more likely used in the development of traditional web applications, and in this paper it plays as a basic technology of the Struts framework. Integrate the two techniques with the tags library to form a unified framework, and separate the view layer from the complex business logic layer which can improve the hierarchy and the maintainability of the code [1], as well as greatly facilitates developers' coding process and the maintenance personnel's debugging work. Based on all the techniques above, introduce the technology of Hibernate as a further step, separate database management logic from business logic, then form a data persistence layer for maintaining the mapping relationship between Java entity classes and the database tables. Through operating the database by object-oriented method [2], encapsulate the details of database accessing in order to improve the scalability and maintainability of the program. Therefore, the application based on Struts+Hibernate framework has a strong flexibility. The application framework based on the combined techniques shows in figure 3.

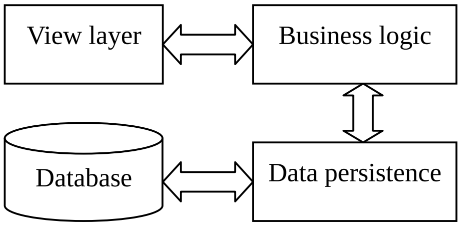


Figure.3 System Application Frame figure.

3.2 Data initialization Technology based on template

According to the results of investigation and analysis, there exist several problems about the users' needs: (1) As the fact that it is a stability job for faculty secretary and those who in charge of graduates' degree information collection, so, there will be almost no such situation like changing user's name frequently. (2) Students can not login in the system until the system already has their basic information. However, the number of students is large and it will surely increase the burden of work if faculty secretary records the basic information of students one by one. (3) The colleges have already retained large amount of previous degree information data for many years' before this system comes into use. In order to achieve the goal of unified management, the workload must be great if recoding these data one after another.

Therefore, this paper proposes the idea of using a template-based system data initialization method. For more precise, put the information into an excel template which designed by system in advance, then initialize the system data through importing the template in one time which can solve the problem in inputting data one by one. This system has three types of templates, including the system user template, students' information template and degree information template. User template consists of responsible person's name, faculty, contact; students' information template consists of name, ID number, contact; degree information template consists of 20 workbooks, the former 4 workbooks are designed according to different degree types which required different degree information, the later 16 workbooks provide the needed code information during the degree information filling for referencing. Degree information template shows in figure 4.

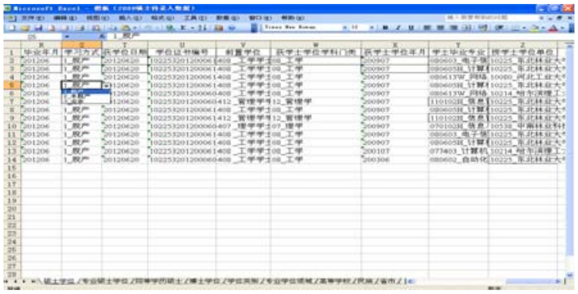


Figure.4 Degree information template.

During the template importing, the system automatically searches the data in the template and matches with the corresponding field of the database to finally insert them into the database. The processes above can be realized through Apache POI which supports HSSFWorkbook Class, HSSFSheet Class, HSSFRow Class, HSSFCell Class and corresponding methods include getSheetAt(int number)、 getRow(int number)、 getCell(int number)、 getStringCellValue() to achieve the goal of reading the value of ranks in the EXCEL document. After getting the values of cells in the document, the related fields in the database can be reset automatically.

3.3 Dynamic form filling technique using AJAX

The types of the graduated degree include academic master's degree, professional master's degree, the equivalent master's degree, PHD, and so on; each sort of the degrees requires different kind of information. After combining the information attributes according to these degrees in the database, there are more than 92 fields in the degrees totally, each of the degree generally has more than 30 fields and 52 at most. The filling scope of most attributes are limited in a reasonable range, such as the style of the study is limited in one of the these three kinds including full-time, part-time and amateur, the same as political landscape, pre-degree, pre-qualifications, the property of the job and the topic source. When the system is designed, in order to avoid making some unknown mistakes, the filling style needs to be supply as a form of drop-down list. So, it is necessary to import all kinds of the data tables into the system firstly, and then the system can automatically get the relative muster from the database and fills the relative drop-down list for users to choose when filling the form. In this way, the system finishes the initiation of the drop down list at once when the users of the system click the form filling interface. Because over half of the information are got from the database, in order to finish the reset of the list, users have to wait for the long time of page initiation to fill the list, which has greatly weaken users' experience.

Therefore, considering dynamical filling technology based on AJAX to promote this kind of experience. System just supplies simple JSP interfaces including different kinds of html labels when users logging in the form filling interface, which makes users preview the list intuitively as soon as possible. The list is filled only when it need to chose the property value of this list by trigging the onchange event of the drop-down list and finish the filling through getting the relative muster from the database by AJAX. By this way, the pressure can be dispersed effectively and it can also promise the user experience as well and finally improve the suitability of the system. The keywords of the AJAX technique for filling the list are listed below:

```
<select name="xxfs" style="width: 150px" onchange="selectchange();">//trigger onchange event when filling forms
function selectchange(){
var url="findsel.jsp?db=xxfs";//point out the jsp file for database operation
xmlHttpRequest=createXmlHttpRequest();//generate xmlHttpRequest
xmlHttpRequest.onreadystatechange=callback;//point out the method for accepting return values
xmlHttpRequest.open("GET",url,true);//point out the format of sending request
xmlHttpRequest.send(null);//send requests
}
function callback(){
if(xmlHttpRequest.readyState==4&&xmlHttpRequest.status==200)
{var result=xmlHttpRequest.responseText;//accept the results set of database operation
  var array=result.split(",");//split the results set
  var ids=array[0].split("|");//split code values set
  var values=array[1].split("|");//split data values set
  var len=ids.length;//get the total numbers of options in drop-down lists
  for(i=1;i<len;i++){//fill the drop-down list calculatedly

document.form1.xxfs.options[document.form1.xxfs.options.length] = new Option(ids[i], values[i]);
}
```

3.4 DBF Report Generation Technology

Currently, Academic Degrees and Graduate Education Development Center of the Ministry of Education requires universities to submit the information about graduate degree-granting information in DBF style in a set time. In order to connect with the its system, the system need to provide data exporting method for graduate academy users according to the regular format, JAVA offers the whole classes for generating and reading DBF documents. The keywords of generating the DBF report are listed as:

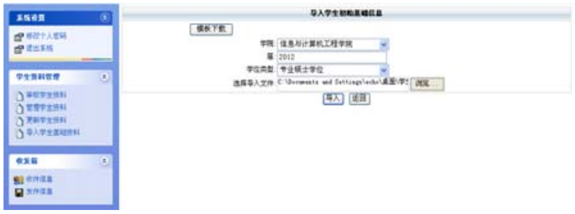
```
DBFField fields[] = new DBFField[56]; //generate DBF data columns
fields[0] = new DBFField(); //generate first column
fields[0].setName( "Xm"); //name the first column
fields[0].setDataType( DBFField.FIELD_TYPE_C);//define the format of the first column
fields[0].setFieldLength(40);//define the data length of the first column
…….//generate other data columns
DBFWriter writer = new DBFWriter(); //define write file object
writer.setCharactersetName("gbk"); //define the coding format of writing files
writer.setFields( fields);//write the data columns into files
Object rowData[] = new Object[56]; //generate object arrays for save data columns values
rowData[0] = u.getXm();//assign values for data columns
…….//assign values for other data columns
```

writer.addRecord(rowData);//write the data columns values into files

4. System running instance

List some examples of critical parts of the system according to the analysis and implementation of the key techniques above.

When initialing the student information, secretaries of each college click the baton of "Inserting the base information" to login into the starting interface, as shown in Figure 5. Choose the classes and degree type of students for importing, then choose a finished template, click the baton of "Importing", the importing results are shown in Figure 6:



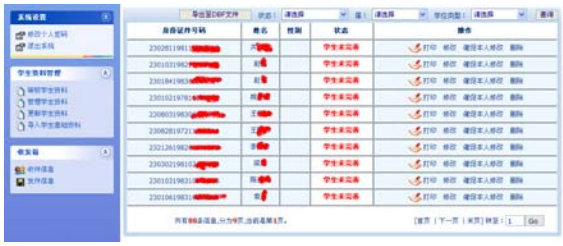
导入学生初始基础信息

模板下载

学院: 信息与计算机工程学院
届: 2012
学位类型: 专业硕士学位
选择导入文件: C:\Documents and Settings\echo\桌面\学 浏览...

导入 返回

Figure.5 Initial student information importing page.



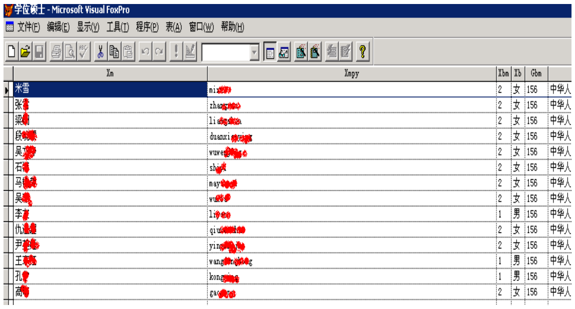
导出至DBF文件 学院: 请选择 届: 请选择 学位类型: 请选择 查询

身份证件号码	姓名	性别	状态	操作
2302811981[redacted]	[redacted]		学生未完善	打印 修改 设置本人修改 删除
2301031982[redacted]	[redacted]		学生未完善	打印 修改 设置本人修改 删除
2301841983[redacted]	[redacted]		学生未完善	打印 修改 设置本人修改 删除
2301021978[redacted]	[redacted]		学生未完善	打印 修改 设置本人修改 删除
2308031983[redacted]	王[redacted]		学生未完善	打印 修改 设置本人修改 删除
2308281972[redacted]	王[redacted]		学生未完善	打印 修改 设置本人修改 删除
2321261982[redacted]	[redacted]		学生未完善	打印 修改 设置本人修改 删除
2303021981[redacted]	[redacted]		学生未完善	打印 修改 设置本人修改 删除
2301031983[redacted]	[redacted]		学生未完善	打印 修改 设置本人修改 删除
2301061983[redacted]	[redacted]		学生未完善	打印 修改 设置本人修改 删除

共有88条信息,分为9页,当前是第1页。 [首页 | 下一页 | 末页] 转至: 1 Go

Figure.6 Initial student information importing successfully page.

After the double check of student information by college secretaries and managers of the graduate academy, they can click the button of "Student Information Management" to login in the student degree information list interface, in which they can search the information according to the college name of students, current audit status and degree type, click the button "Export to DBF file" to get the searching results. Figure 7 shows the process of exporting the information of the master's degree in the 2012.


文件(F) 编辑(E) 显示(V) 工具(T) 程序(P) 表(A) 窗口(W) 帮助(H)

学位硕士 - Microsoft Visual FoxPro

Xm	Xmpy	Xbm	Xb	Gbm	
米雪	mix[redacted]	2	女	156	中华人
张[redacted]	zhang[redacted]	2	女	156	中华人
梁[redacted]	lia[redacted]	2	女	156	中华人
段[redacted]	duanxi[redacted]	2	女	156	中华人
吴[redacted]	wuwe[redacted]	2	女	156	中华人
石[redacted]	sh[redacted]	2	女	156	中华人
马[redacted]	may[redacted]	2	女	156	中华人
吴[redacted]	wu[redacted]	2	女	156	中华人
李[redacted]	li[redacted]	1	男	156	中华人
仇[redacted]	qiu[redacted]	2	女	156	中华人
尹[redacted]	yin[redacted]	2	女	156	中华人
王[redacted]	wang[redacted]	1	男	156	中华人
孔[redacted]	kong[redacted]	1	男	156	中华人
高[redacted]	gao[redacted]	2	女	156	中华人

Figure.7 DBF exporting page.

5. Conclusion

Through analysis of the current state of post-graduate degree information collection and statistics, makes it clear that the significance of applying modern information technology in this kind of work. In addition, according to the common and special problems during the process of the work, studies the implementation of the key technologies and realization methods in solving these problems. It makes the collection of graduate students information more reasonable and humanity by combining different kinds of methods, improve the work efficiency of staffs and optimize the performance of the system at the same time.

Acknowledgments

funded by Innovation Project of Guangxi Graduate Education(JGXM_HLJ_2012008)、Supported by the Fundamental Research Funds for the Central Universities(DL11BB04).

References

[1] Juanjuan Yan， Bo Chen， Xiu-e Gao，etc. Research of Structure Integration based on Struts and Hibernate [J]. 2009 World Congress on Computer Science and Information Engineering，2009:530-534

[2]Yongchang Ren，Deyi Jiang，Tao Xing，ect. Research on software development platform based on SSH framework structure [J]. Procedia Engineering，2011(15):3078-3082

[3] http://www.iteye.com/topic/759437 [OL]. 2010-09-09

[4] Jie Xiao， Xiang Chen， Jianghai He,etc. Design and implementation of Web application based on AJAX and

Struts[J]. Computer Engineering and Design，2009，30(8):1934-1937

[5] Paula Montoto, Author Vitae, Alberto Pan Author Vitae，etc. Automated browsing in AJAX websites [J]. Data & Knowledge Engineering，2011，70(3):269-283

First Author Huang Ying-lai,,male, born in October 1978, PhD, lecturer. Acquired bachelor degree from Northeast Forestry University in July, 2003. Acquired professional degree from Northeast Forestry University in July, 2006.Now is studying in wood science and technology, Northeast Forestry University, PhD. Mainly engaged in signal processing, computer intelligent processing direction, has published 10 papers, hosting, participate in national, provincial project 10 items.

Second Author Gao meng, female, 1989, PhD, research specialty: forestry information engineering

Second Author Li chun-ying, female, 1970, vice research fellow.

Fourth Author Chen jing, female, 1976, lecturer.

Numerical simulation of an amphibious vehicle sailing resistance

Zhangxia Guo [1], Yutian Pan [1], Haiyan Zhang [2], Yongcun Wang [3]

[1]College Of Mechatronic Engineering, North University Of China, Taiyuan 030051,PRC

[2]China North Vehicle Research Institute, Beijing 100072 , PRC

[3]Northwest Institute of Mechanical & Elect rical Engineering,Xianyang,712099，PRC

Abstract

In order to evaluate the waterborne performance of amphibious vehicle, based on Fluid Dynamics and principle of marine mechanics related knowledge, the resistances and viscous flow field of amphibious vehicle in different headway were numericaly simulated by solving Navier-Stokes equatlons with the $k-\varepsilon$ turbulence model. we obtained the result of frictiona resistance coefficient 、 residual resistance coefficient and running resistance coefficient,thus we can calculate its total resistances. the reliability of computing methed was validated by comparing the calculation results with the test data.

Keywords: *Numerical Simulation, Amphibious Vehicle, Sailing Resistance, Viscous Flow Field.*

1. Introduction

The speediness of the amphibious vehicle is one of the most important qualities, for the amphibious vehicle, the speediness is closely connected to improve the combat effectiveness and survivability. So, forecasting and optimizing design the speediness of the weapons is one of the critical technology for the amphibious vehicle's design. The speediness of the amphibious vehicle contains two sides, resistance and propulsion, then, researching on the resistance during the vehicle running plays an important role in ameliorating the speediness.

The determination of the resistance of the amphibious vehicle has been mostly a perpetual remain on pool model resistance test, not only wasted a lot of manpower and material resources but also the flow filed around the vehicle can't be accurately described. Therefore, in the vehicle design stage ,there is no doubt that the CFD method is the ideal choice.

The amphibious vehicle are different obviously from ship structure, its features contain short length, small surface and shape change highlights, the vehicle road wheel, track and other transmission devices becomes the part of the body. Therefore, it's not applicable for the amphibious vehicle to analysis the theory of ship sailing resistance, in the amphibious vehicle's total resistance, sailing resistance component ratio and value are different from ship's[1].

The sailing resistance of the amphibious vehicle consist of friction drag, form drag and wave drag, the friction drag is relate to the viscosity of water, the form drag is relate to the pressure of water and the wave drag is relate to the speed of car[2]. For the towing tank test, it can't achieve that let it and real vehicle satisfy the Reynolds number and Froude number at the same time. So, tissue the test when the Froude number is equal to each other. In order to get the real resistance from the towing tank test, Froude get the following assumption, the sailing resistance of the amphibious vehicle have two parts, one part is friction drag which can be calculated according to the 1957 ITTC fairly flat formula and only associated with Reynolds number; The other part is called the residual resistance coefficient(form drag and wave drag included) consistent with the Froude similarity criterion ,that's to say, the corresponding dimensionless resistance coefficient and Froude number are equal, only about[3].

The amphibious vehicle using Froude two dimensional method conversion of the total resistance can be in the following form :
Residual resistance coefficient $C_r = C_{tm} - C_{fm}$
Real vehicle total resistance coefficient
$C_{ts} = C_{fs} + C_r + \Delta C_F$
Total resistance $R_{ts} = C_{ts} \times (\rho S V^{\wedge} 2 /2)$

Where C_{ts} is the real vehicle total resistance coefficient; C_{fs} is vehicle friction coefficient; C_r is the residual resistance coefficient; ΔC_F for model and real vehicle conversion of resistance between the compensation value, generally take 0.004; C_{tm} is towing total resistance coefficient; C_{fm} for towing friction coefficient; Rts is the real vehicle total resistance; ρ is the density of water， while S and V are wet surface area and speed. So through CFD numerical simulation we can calculate the model of residual resistance, thereby obtaining the residual resistance coefficient C_r, through the corresponding experience formula can be derived C_{fs}, then obtains total resistance coefficient C_{ts}, the resistance coefficient and resistance relationship can calculate the total resistance.

2. Mathematical model

2.1 Incompressible fluid continuity equation and N-S equation[4]

$$\frac{\partial u_i}{\partial x_i}=0$$

$$\rho\frac{\partial\left(u_iu_j\right)}{\partial x_j}=-\frac{\partial P}{\partial x_i}+\rho g_i+\rho\frac{\partial}{\partial x_j}\left[v\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right)\right]-\frac{\partial\left(\rho\overline{u_i'u_j'}\right)}{\partial x_j}$$

where $u_i=\left(u,v,w\right)$ is the velocity component in $x_i=\left(x,y,z\right)$ direction, while P , ρ , v , g_i , $-\rho\overline{u_i'u_j'}$ are the static pressure,fluid density, fluid viscosity, gravitational acceleration weight and Reynolds stresses, respectively.

Free surface fluctuation is tracked by the use of VOF, the equation can be written as:

$$\frac{\partial a_q}{\partial t}+\frac{\partial\left(u_ia_q\right)}{\partial x_i}=0,\left(q=1,2\right)$$

$$a_1+a_2=1$$

Where a_1 、 a_2 are the air phase, phase volume fraction.

2.2 Turbulence models

In the article ,the sailing resistance of the amphibious vehicle is calculated by turbulence models. The turbulence kinetic energy equation is [5~6]:

$$\frac{\partial}{\partial t}(\rho k)+\frac{\partial}{\partial x_i}(\rho k u_i)=\frac{\partial}{\partial x_j}\left[\left(\mu+\frac{\mu_t}{\sigma_k}\right)\frac{\partial k}{\partial x_j}\right]+G_k+G_b-\rho\varepsilon-Y_M+S_k$$

The turbulent dissipation rate equation:

$$\frac{\partial}{\partial t}(\rho k)+\frac{\partial}{\partial x_i}(\rho\varepsilon u_i)=\frac{\partial}{\partial x_j}\left[\left(\mu+\frac{\mu_t}{\sigma_\varepsilon}\right)\frac{\partial\varepsilon}{\partial x_j}\right]+C_{1\varepsilon}\frac{\varepsilon}{k}(G_k+C_{3\varepsilon}G_b)-C_{2\varepsilon}\rho\frac{\varepsilon^2}{k}+S_\varepsilon$$

where G_k and G_b are turbulence kinetic energy by average speed grads, turbulence kinetic energy by buoyancy, respectively,and Y_M is the condensability turbulence pulsant expanding to total dissipation ration,and μ_{t} is the turbulence viscosity constance,and $C_{1\varepsilon}$ (=1.44), $C_{2\varepsilon}$ (=1.92),and $C_{3\varepsilon}$ (=0.09) are the turbulence model constants.

2.3 Boundary conditions

This is a gas, liquid two-phase flow problems, because the amphibious vehicle's higher part is air, the lower part is water,during the sailing, it must cause the interaction between water and air, and generate waves.Wave theory can be divided into liner wave and nonlinear wave.In the specific conditions, the wave height relative to the wave length(or relative to the depth of water) is generally limited, in this wave of finite amplitude,fluctuations of the free water surface caused by nonlinear effects must be taken into consideration, thus the actual ocean waves are studied based on the nonlinear wave theory. In this article ,using two order Stokes wave, incident boundary speed to satisfy the following conditions:

x direction speed

$$u=\frac{\pi H}{T}\frac{\cosh ks}{\sinh ks}\cos\theta+\frac{3}{4}\frac{\pi H}{T}\left(\frac{\pi H}{L}\right)\frac{\cosh 2ks}{\sinh^4 kd}\cos 2\theta$$

y direction speed

$$v=\frac{\pi H}{T}\frac{\sinh ks}{\sinh kd}\sin\theta+\frac{3}{4}\frac{\pi H}{T}\left(\frac{\pi H}{L}\right)\frac{\sinh 2ks}{\sinh^4 kd}\sinh 2\theta$$

Where H is the height of wave; θ is the phase angle ;L is the wavelength ;k is the wave number ;d is the height of wave which does not consider the wave surface; s is height with wave surface considered; T is the period. This article uses UDF technology as the incident boundary conditions coupled to the calculation equation, in order to achieve the unsteady wave simulation.

The computational domain boundary conditions consist of entrance,exit and wall,in the flow direction of the entrance boundary given the flow velocity, the air and water volume fraction; exit boundary is far from the amphibious vehicle and the flow had reached a steady state; then,flow direction parallel to the distant boundary set free outflow boundary; considering the viscous effects,the vehicle surface defined as not slip wall; the calculation region at the bottom of fixed boundary.

2.4 Numerical methods

The finite volume method is adopted to discrete momentum equation. Convection using two order upwind difference scheme, diffusion using a central difference scheme, the pressure velocity coupling using SIMPLE algorithm.

3.The calculation model and results analysis

3.1 The amphibious vehicle towing test

The experiment was done in Dalian science and engineering university ship mold pond.The pond's total lenth is 160 ms and breadth is 7 ms, water's deep is 3.7 ms;The trailer is the empty beam structure, speed scope is 0.01~8 ms/s.

The Principal particulars about model list on Table 1, the model adopted steel quality model that is jointed with the high-quality cold armor plate of 2 mms.The model reduced scale is 1:4. Model test data as shown Fig 1.

Table1: Principal performance about model

serial number	item	unit	value
1	total lenth	m	1.994
2	total high	m	0.40
3	total breadth	m	0.73
4	water line	m	0.299
5	tonnage	kg	328

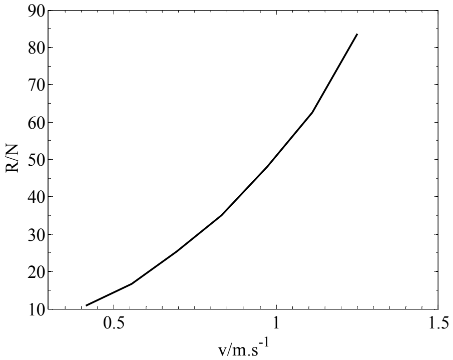


Fig. 1 Model test of the resistance curve

3.2 Establishment of calculation model

Practice has proved that, compared with structured grids,unstructured grids is more suitable for complex areas grid, its random data structure is more easy to be adaptive,so as to capture the physical characteristics of flow field,therefore,this calculation model averaging using unstructured grid.Calculated using the model are consistent with the model, numerical calculation flow field length,width and depth were 160m,7.m,3.7m(reference model experiment where the pool size).

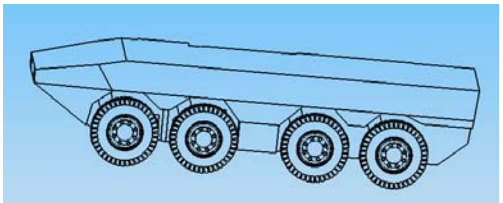


Fig. 2 Computed model

Fig 2 is an amphibious vehicle model diagram, In order to simplify the problem,tire using approximate cylinders instead,computational domain for the front 5 times vehicle length, 2 times body top commander .

Table.2: Computed resistance of the real vehicle

Number	V(m/s)	$C_r X10^{-2}$	$C_{fs} X10^{-3}$	$C_{ts} X10^{-2}$	R_{ts} (KN)
1	0.833	3.005	3.315	3.737	0.714
2	1.111	2.589	3.148	3.304	1.122
3	1.389	2.398	3.026	3.101	1.645
4	1.667	2.226	2.932	2.953	2.256
5	1.944	2.228	2.856	2.966	3.083
6	2.222	2.231	2.793	2.989	4.059
7	2.500	2.500	2.738	3.094	5.318

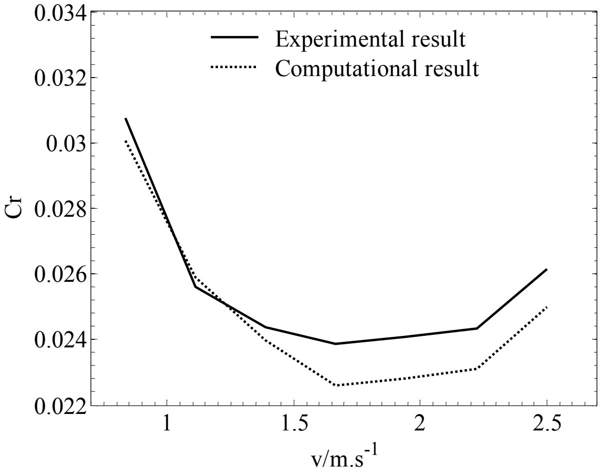


Fig. 3 Relationship between residual resistance coefficient and speed of the amphibious vehicle

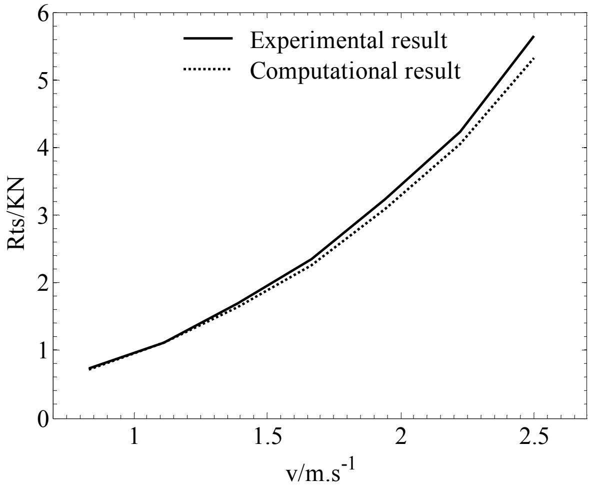


Fig. 4 Relationship between running resistance and speed of the total amphibious vehicle

3.3 Results of the computation

Through a series of the amphibious vehicle turbulent viscous flow theory and Froude two time method results of the computation shown in Table2, whereΔCF is 0.004, friction resistance curve, Fig 3 is the amphibious vehicle residual resistance coefficient with the sailing speed curve.Fig 4 is the amphibious vehicle total resistance with the sailing speed curve.

4 Conclusions

In this article ,using the theory of fluid mechanics for amphibious vehicle navigation performance was studied by numerical simulation,from the final results can be obtained the following conclusions:

1) In the turbulent viscosity theory,to the residual resistance coefficient comparison,numerical simulation of calculation and model experimental results are basically the same.In particular when the vehicle's speed is more than $1.667m/s$ the residual resistance coefficient varied little with the sailing speed changes, at the same time, when the vehicle's speed is less than $1.667m/s$ the rate of change is quick.

2) We can see from Figure 4 that the real vehicle resistance is more and more big along with speed increases, this is accorded with the resistance characteristic of the ships.At the same time it can be seen by comparison the calculated resistance curve and test curve is more consistent, especially at low speed,the calculation results are in good agreement with experimental results.It shows thatflow field numerical calculation of adapting turbulent model, calculation method and the boundary condition is reasonable.

3） In the current numerical calculation technology more mature circumstances,recommend the use of numerical calculation technique and test technology means combining forecasting of resistance for the amphibious vehicles.

Acknowledgments

Project supported by the National Nature Science of China (No. 51275489).

References

[1] Guoying Xu, Weiping Liu.Composition of water resistance to amphibious vehicle[J].Journal of Armored Force Engineering Insistute, 1998(3) : 59-63(in Chinese)

[2] Naijun Ju.Hydrodynamics analysis and simulation for amphibious vehicle(M). Beijing: The Publishing House of Ordnance Industry, 2005(in Chinese)

[3] Jianwei Yu. Research on Calculation and Prediction for Ship Resistance Based on CFD Theory [D]. Shanghai: Jiao Tong University, 2009(in Chinese)

[4]Garofallidis D. "Experimental and numerical investigation of the flow around ship model at various Froude numbers," Ph.D. thesis. Athens: Department of Naval Architecture and Marine Engineering, NTUA,1996.

[5]J.E.Choi,K.-S.Min,J.H.Kim,S.B.Lee andH.W.Seo,"Resistance and propulsion Characteristics of various commercial ships based on CFD results," in Ocean Engineering,vol. 37, pp. 550, February2010.

[6]. WU Song-ping, LIU Zhao-miao.Computaional fluid dynamics [M]. Beijing: Machine Industry Publishing Company, 2008(in Chinese)

Mr. Zhangxia Guo received the Master Degree in science from North University Of China, in 2005. Currently, she is an a lectorate at North University Of China, China. Her research interests include numerical simulation of vehicle.

Mr. Yutian Pan received the Bachelor's degree in science from North University Of China, in 1962. Currently, he is an Professor at North University Of China, China. His research interests include degign of vehicle.

Mr. Haiyan Zhang received the Master Degree in science from North University Of China, in 2005. Currently, she is an associate professor at China North Vehicle Research Institute, China. Her research interests include intelligent control and Fault diagnosis about vehicle

Telco Business Process Transformation using Agile Lean Six Sigma and Frameworx components: focus on the core engineering aspects with a case study

Mounire Benhima[1], Abdelaâli Himi [2], Camille Ameyao[3], and Edwige Ahonie Adou[4]

**[1] Eng., M.Eng, Member of TM Forum Trainers Panel / Business Transformation Subject Matter Expert, POWERACT Consulting
Casablanca, Morocco**

**[2] Eng., MSC. BPM Expert / Business Transformation Subject Matter Expert, Intellectus Consulting Services
Casablanca, Morocco**

**[3] Senior Manager MTN Business Service Delivery / Project Manager
Abidjan, Ivory Coast**

**[4] MSC., Business Sales and Reporting Expert MTN Business / Project Coordinator
Abidjan, Ivory Coast**

Abstract

The business transformation, a worldwide trend in many industries, is to improve the business performance in order to remain competitive in a challenging market. To accompany Telco industry in this trend, TM Forum issued Frameworx which addresses the key business aspects namely process, information, application and integration ones. The purpose of this paper is to present a methodology for Telco business process transformation harmonizing Frameworx and Lean Six Sigma (L6S) since it is well proven methodology supporting this transformation. This harmonization increases the L6S agility. The focus of the current paper is on the Define phase of L6S, with some highlights about other phases, supported by a case study. The Define phase activities all together are named “Core Engineering aspects”.

Keywords: Business Process design, Business Process reengineering, Business Process Optimization, Business Process Performance, Frameworx, eTOM, Business Metrics, Lean Six Sigma, Project Charter, Change Management, SIPOC, Business Process Flow, VSM, Voice of Customer, Critical To Quality, Root cause analysis, Fishbone, The Seven Wastes, Customer Experience, Operational Efficiency and Revenue & Margin.

1. The Business Transformation, De Facto solution for the challenges of the TELCO industry

In the dynamic of the TELCO industry, there are multiple types of actors including, to name some of them, content providers, terminals and equipment vendors, , integrators, social networks providers, service providers, TELCO products distributors and any TELCO provider or consumer. These types all together make the TELCO value chain.

This industry is today under the pressure of a changing and challenging business context (Ex.: Customer

Experience challenges and social media, quick changes with mergers and acquisitions, new unexpected entrants, new technologies, margins compression, operational efficiency, cost optimization) which threatens the competitiveness and even the survival of an actor.

Given such dynamics, TELCO industry players are more and more adopting the solution « Business Transformation» to overcome the challenges of such a business context. This increasing adoption is justified by is justified by its practical relevance and the very positive feedback from the market with several real case studies. Indeed, companies who adopted and implemented this solution have seen a return (ROI) and value (VOI) on Investment. In a subsequent section, a "Success Story" will be presented with the obtained ROI and VOI. The key elements subject to transformation, as presented in the Figure1, are the Business Model and Product Portfolio, Customer Experience, Business & Staff Culture, Business Processes, IT & systems and infrastructure.

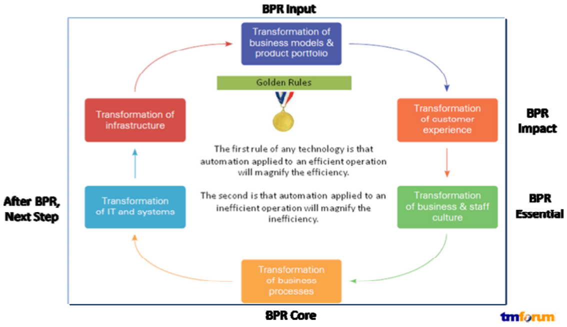

BPR Input
Transformation of business models & product portfolio
Golden Rules
Transformation of infrastructure
Transformation of customer experience
BPR Impact
The first rule of any technology is that automation applied to an efficient operation will magnify the efficiency.
The second is that automation applied to an inefficient operation will magnify the inefficiency.
After BPR, Next Step
Transformation of IT and systems
Transformation of business & staff culture
BPR Essential
Transformation of business processes
BPR Core
tmforum

Figure1: Key elements subject to Business Transformation

2. How to accelerate Business Transformation?

In a world characterized by responsiveness and speed, it is very natural to ask how to accelerate Business Transformation? The TELCO industry adopts massively TM Forum Frameworx as an accelerator for this transformation.

Today 8 of the Top 10 worldwide service providers use and adopt Frameworx. This adoption is justified, as previously explained, by the ROI and VOI obtained with the slogan «Do not start your transformation initiatives from scratch, use tmforum Frameworx as an efficient and reliable accelerator».

TM Forum Frameworx, as shown in the Figure2, and its components (called Frameworks) answer four key concerns related to the business transformation. First, The Business Process Framework (eTOM) answers the concern «How to improve / structure / define the business processes supporting the Business? ». Secondly, the Information Framework (SID) answers the concern «How to improve / structure / define the information manipulated by the business processes? ». Then, the Integration Framework (TNA) answers the concern «How to improve / structure / define the interaction between my business processes & the manipulated information and how to support the application development? ». Finally, the Application Framework (TAM) answers the concern «How to improve / structure / define the applications supporting my business processes? ».

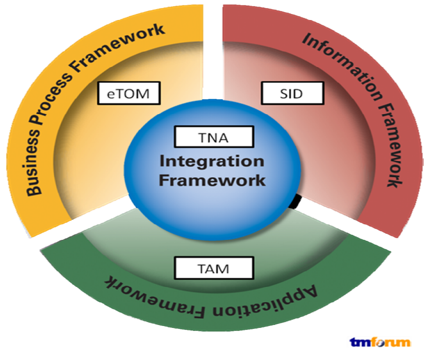


Figure2: Frameworx and its key components

And to measure the business performance, TM Forum issued the Business Metrics Framework. This framework is composed of three major performance domains with their related metrics. These three, the pillars of the balanced scorecard, are:

- The Customer Experience
  - *Metric example*: *% Orders Delivered By Committed Date (ID:F-CE-2c)*
- The Operational Efficiency and
  - *Metric example: Fulfillment Process Cost As % OpEx (ID: F-OE-1c)*
- Revenue & Margin.
  - *Metric example: OpEx / CapEx (ID: G-RM-2)*

3. How to start Business Transformation for inefficient operations?

The startup of the Business Transformation for inefficient operations, as mentioned in Figure1, is based on two golden rules. The first rule is that «automation applied to an efficient operations will magnify the efficiency» Bill Gates. The second rule is that «automation applied to an inefficient operations will magnify the inefficiency» Bill Gates. Based on these golden rules, the starting point to transform inefficient operations is the Business Process Reengineering (BPR) with the Slogan «Improvement before automation». BPR is being widely used jointly with Lean Six Sigma Methodology for better process optimization.

4. The BPR and process decomposition

In the engineering world, the architecture development of the work to be done is crucial and fundamental «I Plan my work first» and represents a fundamental principle. The BPR, part of this world, follows by syllogism this principle. Therefore, the business process architecture should be developed. This architecture breaks down and decomposes activities into several levels (0, 1, 2, etc).

In order to support the development of such a process architecture and decomposition, many key actors in the TELCO industry have chosen eTOM.

5. About eTOM

eTOM provides a standard process architecture, terminology, classification scheme and decomposition hierarchy. eTOM covers all business activities, with more detailed aspects than others, to meet the needs and interests of the TELCO industry with its major trends.

The eTOM suggested process architecture, as shown in Figure3, is composed of three Process areas. First, The Strategy, Infrastructure & Product Process Area includes processes that develop strategy, commit to the enterprise, build infrastructure, develop and manage products, and that develop and manage the Supply Chain. In the eTOM, infrastructure refers to more than just the IT and resource infrastructure that supports products and services. It includes the infrastructure required to support functional processes, e.g., Customer Relationship Management (CRM). These processes direct and enable the Operations processes.

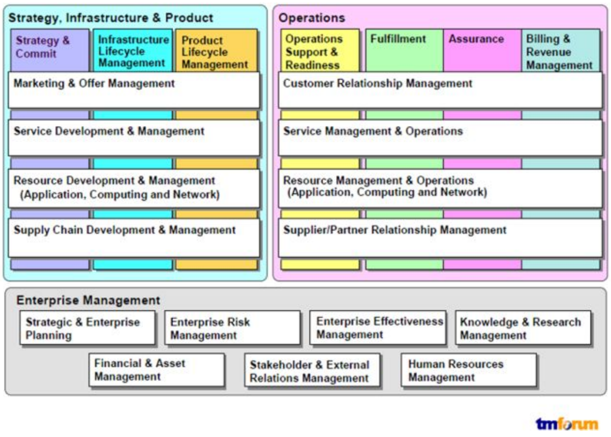

Strategy, Infrastructure & Product
Strategy & Commit
Infrastructure Lifecycle Management
Product Lifecycle Management
Marketing & Offer Management
Service Development & Management
Resource Development & Management (Application, Computing and Network)
Supply Chain Development & Management
Operations
Operations Support & Readiness
Fulfillment
Assurance
Billing & Revenue Management
Customer Relationship Management
Service Management & Operations
Resource Management & Operations (Application, Computing and Network)
Supplier/Partner Relationship Management
Enterprise Management
Strategic & Enterprise Planning
Enterprise Risk Management
Enterprise Effectiveness Management
Knowledge & Research Management
Financial & Asset Management
Stakeholder & External Relations Management
Human Resources Management
tmforum

Figure3: The Business Process Framework (eTOM)

Then, the Operations (OPS) process area contains the direct operations vertical end-end process groupings of Fulfillment, Assurance & Billing (the FAB process groupings), together with the Operations Support & Readiness process grouping. The FAB process groupings are sometimes referred to as Customer Operations processes.

Then, the Enterprise Management Process Area includes basic business processes required to run any business. These processes focus on Enterprise Level processes, goals and objectives. These processes have interfaces with almost every other process in the enterprise, whether operational, product or infrastructure processes. These are sometimes considered corporate functions and/or processes, e.g., Financial Management, Human Resources Management processes, etc.

The three process areas are further decomposed into further levels of decomposition which could be customized. Thus, eTOM is an open Framework allowing the specific needs of a given company to be integrated in it.

6. The BPR and process flows

The BPR starts by developing a process architecture and decomposition as previously explained. This architecture/decomposition is made of process elements which form a library. Initial process flows could be constructed by selecting the appropriate process elements from the library based on the business context to model. For detailed process flows, a detailed descriptive sheet will be elaborated for each process element. This sheet is simply named a Use Case. The key elements of a use case are the Name, Goal, Description, Actors, Pre-conditions, Triggers, Essential steps, Manipulated information, Post-conditions and Business rules.

All the use cases, within the scope of work, will form the Use Cases repository.

To build a process flow, a set of use cases suitable to the business context to model will be selected from the use cases repository. Then the process flow will be an orchestration, via a sequence, of the related use cases.

7. Lean Six Sigma (L6S) in a nutshell[1]

7.1 Introduction

Lean Six Sigma is a step by step methodology to optimize the process performance. The methodology steps, known as **DMAIC**, are:

- **D**efine: the main purpose is to define the project, the team and the process
- **M**easure: the main purpose is to validate the measurement system and collect process data
- **A**nalyze: the main purpose is to analyze the process data in order to identify the root causes for the problems and non performance
- **I**nnovate: the main purpose is to find, assess and launch the process improvements ideas
- **C**ontrol: the main purpose is to ensure a sustainable process improvement

7.2 Define Phase: Activities and Deliverables

The key activities related to the Define Phase are:

- Set project goal (s)
- Define the project objectives
- Define the project scope

[1] www.excellence-operationnelle.tv

- Elaborate the financial impact
- Set the project organization and planning
- Perform the Kick-off meeting
- Capture the process organization using SIPOC (**S**uppliers, **I**nputs, **P**rocess Steps, **O**utputs, **C**ustomers)
- Build the high level business process flow and identify/designate the process Owner (s)
- Capture the Voice Of Customer which highlights the customer needs
- Identify the Critical To Quality (CTQ) where measurable performance indicators are defined for the customer needs.
- Elaborate the change management essential charters which are:
    - Communication Charter
    - Leadership/Sponsorship Charter
    - Organization Optimization Charter
    - Sustainable Change Charter
- Identify the quick wins
- Perform a quality gate review for the Define Phase

The key deliverables related to the Define Phase are:

- Project Charter
- Kickoff meeting presentation
- SIOC
- High Level process flow
- Process accountabilities defined:
    - Owner (s) defined
    - Virtual Process Owner (s) for end to end processes identified if applicable
- Voice Of Customer
- Critical To Quality
- Change Management Essential Charters
- Pareto Analysis
- Quality gate review for the Define Phase

7.3 Measure Phase: Activities and Deliverables

The key activities related to the Measure Phase are:

- Build the Value Stream Map where process steps and their related performance data are captured
- Build the process performance Data Collection file which integrate data to calculate the measurable performance indicators identified in the CTQ
- Validate the measurement system
- Collect the process performance data
- Calculate the baseline process performance
- Perform the first process analysis in order to identify the trends to deepen during the Analysis Phase
- Identify the quick wins if any
- Refine the deliverables of the previous phase (s) where applicable

The key deliverables related to the Measure Phase are:

- Value Stream Map (VSM)
- Process performance data file designed and filled with a sample of process records
- Baseline process performance or capability calculated and communicated using suitable means (Ex. Graphs). The capability makes the link between what the process is capable to do and the customer needs.
- Quality gate review for the Measure Phase

7.4 Analyze Phase[1]: Activities and Deliverables

The key activities related to the Analyze Phase are:

- Generate a list of possible causes (Xs) for non process performance
- Prioritize the list of causes
- Verify the root causes of variations
- Assess the impact of each X on the performance indicators (Ys)
- Quantify the Gap/Opportunity:
    - Determine the performance gap

[1] http://www.isixsigma.com/new-to-six-sigma/dmaic/six-sigma-dmaic-quick-reference/

  - Display and communicate the gap/opportunity in financial terms
- Identify the quick wins if any
- Refine the deliverables of the previous phase (s) where applicable

The key deliverables related to the Analyze Phase are:

- Process performance graphs with the variation, Pareto Diagram, Results analysis
- Fishbone Diagram with the 5 Whys filled for the critical measures
- 7 wastes identified
- FMEA1 (Failure Mode and Effect Analysis) with the potential causes for the non process performance
- Correlations graphs between (Xs) and (Ys)
- Quality gate review for the Analyze Phase

7.5 Innovate Phase: Activities and Deliverables

The key activities related to the Innovate Phase are:

- Confirm the critical X (s)
- Generate the potential solutions
- Select the solution
- Optimize the solution
- Perform Risk Analysis
- Perform a pilot run of the solution
- Identify the quick wins if any
- Refine the deliverables of the previous phase (s) where applicable

The key deliverables related to the Innovate Phase are:

- FMEA2 (Failure Mode and Effect Analysis) with the action plan for the improvement areas
- The To-Be VSM
- The To-Be Process Flow (s)
- The To-Be Procedure (s) and Work Instruction (s)
- Solution Selection Matrix
- Pilot Project and simulation
- Implementation strategy
- Risk Management

7.6 Control Phase: Activities and Deliverables

The key activities related to the Control Phase are:

- Develop a Control Plan
- Implement the process changes, controls and documents
- Calculate the final financial and process measures
- Handover the project the process accountable (s)
- Identify the
- Identify the quick wins if any
- Refine the deliverables of the previous phase (s) where applicable

The key deliverables related to the Control Phase are:

- A finalized Action plan. This action plan could be used as a control plan for the remaining elements to master
- The To-Be procedure validated by the stakeholders
- The process users trained about the To-Be procedure
- Process Control Card (s) to track and maintain the characteristics influencing the process performance at a standard level

8. The L6S Project Streams

The suggested project streams are as follows:

- Business Architecture and Excellence Stream
  - The goal is to handle business architecture and lean six sigma activities
- eTOM/ITIL Standardization Stream
  - The goal is to handle the eTOM/ITIL alignment activities of the To-Be process design
- Essential Aspects Stream

  - The goal is to handle the activities related to change management and its sustainability
- QA and Deliverables Stream
  - The goal is to ensure that the deliverables are compliant with the methodology activities and deliverables

- Next Phase Readiness Stream
  - The goal is to ensure the readiness for the coming. This readiness is about collecting any required information and ensuring the availability of the customer stakeholders for the coming phase.

For this project Stream (s), RACI (Responsible, Accountable, Consulted, and Informed) is applied to identify the project accountabilities.

9. Methodology rules: Focus on the Brainstorming Workshops

While executing the methodology phases, many brainstorming workshops will be performed. As a best practice[1], the following rules have to be reminded at the beginning of each brainstorming workshop:

- Think freely
- All the ideas have the same value
- Do not criticize
- One idea = One Post-It
- Develop other's ideas
- Display all the ideas

10. L6S applied to Telco: Define Phase

The application of the Lean Six Sigma to Telco will be a harmonization between L6S and Frameworx with other tools. The focus of this harmonization will be on the L6S Deliverables.

The case study is about improving the Order To Cash/Payment related to the MTN Business operations in the Ivory Coast.

10.1 Project Charter

The project charter components are:

- Project Identity
- Project Description
- Key Project Performance Data
- Objectives/Performance Indicators
- Project Scope
- Financial Impact
- Project Organization
- Project Plan

Support Documents

The support Frameworx documents for "Project Charter" are:

- *GB921E*: it contains end to end business flows from different perspectives: Customer, Network and Product
- *QSP* for: Fulfillment, Assurance and New Services
- *Introductory Guide* (s) for PLM
- *Business Metrics*: Which contains the list of business metrics divided into three performance domains (Customer Experience, Operational Efficiency and Revenue&Margin)
- *GB921*: D and DX for eTOM decomposition and J for process flow design and context.
- RACI tool

Snapshots from Support Documents

Below are some snapshots, from the support documents, that can be used for this item.

[1] www.excellence-operationnelle.tv

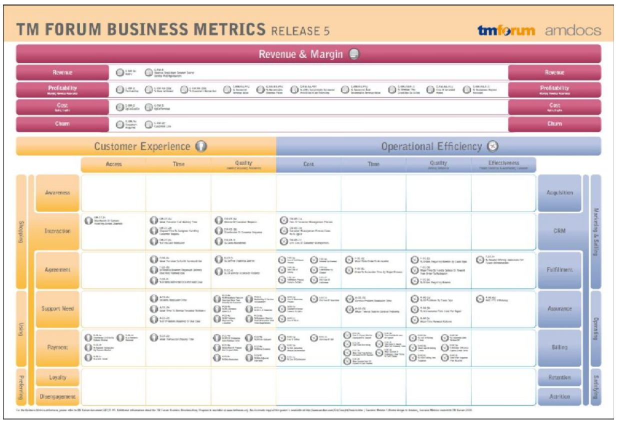

TM FORUM BUSINESS METRICS RELEASE 5
tmforum
amdocs
Revenue & Margin
Revenue
Profitability
Cost
Churn
Customer Experience
Operational Efficiency
Access
Time
Quality
Cost
Time
Quality
Effectiveness
Awareness
Interaction
Agreement
Support Need
Payment
Loyalty
Disengagement
Requisition
CRM
Fulfillment
Assurance
Billing
Retention
Attrition

Figure4: The Balanced Scorecard with the related metrics, Source : TM Forum

The above Figure shows the three performance domains making the balanced scorecard with their related metrics. These domains are Customer Experience (Metric example: % of Orders Delivered by Committed date), Operational Efficiency (Metric example: Mean Time Order To Activation) and Revenue & Margin (Metric Example : ARPU).

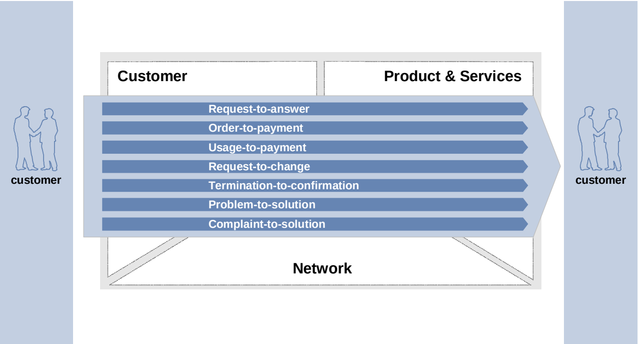
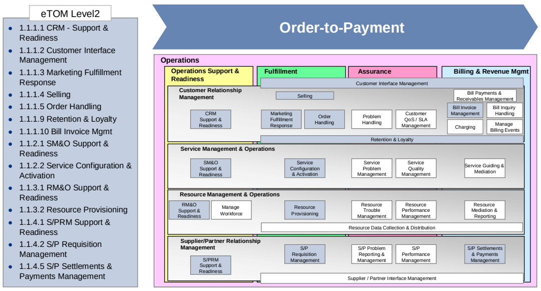


Figure5: Customer-Centric End-To-End Business Streams, Source : TM Forum

The above figure shows the Customer-Centric End-To-End business streams:

- Request-To-Answer
- Order-To-Payment
- Usage-To-Payment
- Request-To-Change
- Termination-To-Confirmation
- Problem-To-Solution
- Complaint-To-Solution

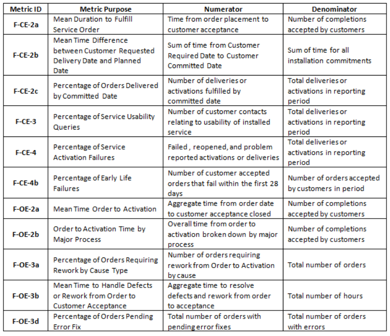

eTOM Level2

- 1.1.1.1 CRM - Support & Readiness
- 1.1.1.2 Customer Interface Management
- 1.1.1.3 Marketing Fulfillment Response
- 1.1.1.4 Selling
- 1.1.1.5 Order Handling
- 1.1.1.9 Retention & Loyalty
- 1.1.1.10 Bill Invoice Mgmt
- 1.1.2.1 SM&O Support & Readiness
- 1.1.2.2 Service Configuration & Activation
- 1.1.3.1 RM&O Support & Readiness
- 1.1.3.2 Resource Provisioning
- 1.1.4.1 S/PRM Support & Readiness
- 1.1.4.2 S/P Requisition Management
- 1.1.4.5 S/P Settlements & Payments Management

Order-to-Payment
Operations
Operations Support & Readiness
Fulfillment
Assurance
Billing & Revenue Mgmt
Customer Interface Management
Customer Relationship Management
Selling
Bill Payments & Receivables Management
CRM Support & Readiness
Marketing Fulfillment Response
Order Handling
Problem Handling
Customer QoS / SLA Management
Bill Invoice Management
Bill Inquiry Handling
Charging
Manage Billing Events
Retention & Loyalty
Service Management & Operations
SM&O Support & Readiness
Service Configuration & Activation
Service Problem Management
Service Quality Management
Service Guiding & Mediation
Resource Management & Operations
RM&O Support & Readiness
Manage Workforce
Resource Provisioning
Resource Trouble Management
Resource Performance Management
Resource Mediation & Reporting
Resource Data Collection & Distribution
Supplier/Partner Relationship Management
S/PRM Support & Readiness
S/P Requisition Management
S/P Problem Reporting & Management
S/P Performance Management
S/P Settlements & Payments Management
Supplier / Partner Interface Management

Figure6: Mapping of Order To Payment in eTOM, Source: TM Forum

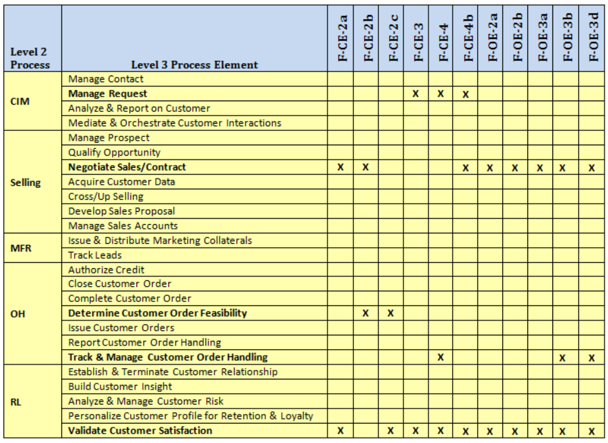
The above figure shows the mapping of the Business Stream "Order To Payment" with eTOM Level2 (s).



Metric ID	Metric Purpose	Numerator	Denominator
F-CE-2a	Mean Duration to Fulfill Service Order	Time from order placement to customer acceptance	Number of completions accepted by customers
F-CE-2b	Mean Time Difference between Customer Requested Delivery Date and Planned Date	Sum of time from Customer Required Date to Customer Committed Date	Sum of time for all installation commitments
F-CE-2c	Percentage of Orders Delivered by Committed Date	Number of deliveries or activations fulfilled by committed date	Total deliveries or activations in reporting period
F-CE-3	Percentage of Service Usability Queries	Number of customer contacts relating to usability of installed service	Total deliveries or activations in reporting period
F-CE-4	Percentage of Service Activation Failures	Failed , reopened, and problem reported activations or deliveries	Total deliveries or activations in reporting period
F-CE-4b	Percentage of Early Life Failures	Number of customer accepted orders that fail within the first 28 days	Number of orders accepted by customers in period
F-OE-2a	Mean Time Order to Activation	Aggregate time from order date to customer acceptance closed	Number of completions accepted by customers
F-OE-2b	Order to Activation Time by Major Process	Overall time from order to activation broken down by major process	Number of completions accepted by customers
F-OE-3a	Percentage of Orders Requiring Rework by Cause Type	Number of orders requiring rework from Order to Activation by cause	Total number of orders
F-OE-3b	Mean Time to Handle Defects or Rework from Order to Customer Acceptance	Aggregate time to resolve defects and rework from order to acceptance	Total number of hours
F-OE-3d	Percentage of Orders Pending Error Fix	Total number of orders with pending error fixes	Total number of orders with errors

Figure7: Metrics related to the business stream Order To Payment, Source : TM Forum

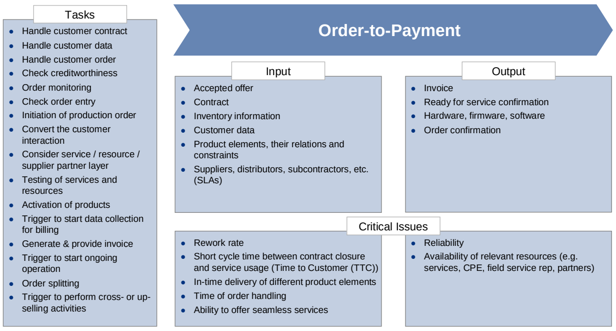
The above figure shows the metrics related to the business stream Order To Payment. These metrics are derived from the balanced scorecard previously presented.



Level 2 Process	Level 3 Process Element	F-CE-2a	F-CE-2b	F-CE-2 c	F-CE-3	F-CE-4	F-CE-4b	F-OE-2a	F-OE-2b	F-OE-3a	F-OE-3b	F-OE-3d
CIM	Manage Contact											
	Manage Request				X	X	X					
	Analyze & Report on Customer											
	Mediate & Orchestrate Customer Interactions											
Selling	Manage Prospect											
	Qualify Opportunity											
	Negotiate Sales/Contract	X	X				X	X	X	X	X	X
	Acquire Customer Data											
	Cross/Up Selling											
	Develop Sales Proposal											
	Manage Sales Accounts											
MFR	Issue & Distribute Marketing Collaterals											
	Track Leads											
OH	Authorize Credit											
	Close Customer Order											
	Complete Customer Order											
	Determine Customer Order Feasibility		X	X								
	Issue Customer Orders											
	Report Customer Order Handling											
	Track & Manage Customer Order Handling					X					X	X
RL	Establish & Terminate Customer Relationship											
	Build Customer Insight											
	Analyze & Manage Customer Risk											
	Personalize Customer Profile for Retention & Loyalty											
	Validate Customer Satisfaction	X		X	X	X	X	X	X	X	X	X

Figure8: Order-To-Payment eTOM processes with related Metrics (Part), Source : TM Forum

The above figure shows the eTOM process elements (Level2 and 3) involved in Order To Payment with their related metrics.



Order-to-Payment

Tasks
- Handle customer contract
- Handle customer data
- Handle customer order
- Check creditworthiness
- Order monitoring
- Check order entry
- Initiation of production order
- Convert the customer interaction
- Consider service / resource / supplier partner layer
- Testing of services and resources
- Activation of products
- Trigger to start data collection for billing
- Generate & provide invoice
- Trigger to start ongoing operation
- Order splitting
- Trigger to perform cross- or up-selling activities

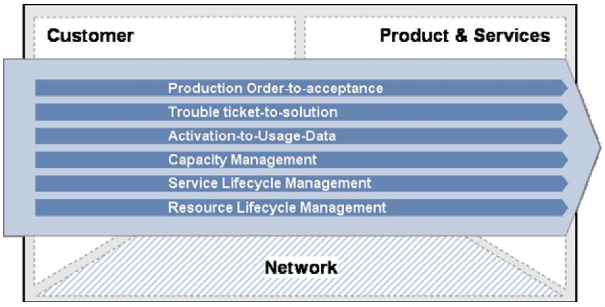
Input
- Accepted offer
- Contract
- Inventory information
- Customer data
- Product elements, their relations and constraints
- Suppliers, distributors, subcontractors, etc. (SLAs)

Output
- Invoice
- Ready for service confirmation
- Hardware, firmware, software
- Order confirmation

Critical Issues
- Rework rate
- Short cycle time between contract closure and service usage (Time to Customer (TTC))
- In-time delivery of different product elements
- Time of order handling
- Ability to offer seamless services
- Reliability
- Availability of relevant resources (e.g. services, CPE, field service rep, partners)

Figure9: Order To Payment Detailed information, Source: TM Forum

The above figure shows detailed information about the business stream Order To Payment.



Customer

Product & Services

Production Order-to-acceptance

Trouble ticket-to-solution

Activation-to-Usage-Data

Capacity Management

Service Lifecycle Management

Resource Lifecycle Management

Network

Figure10: Network-Centric End-To-End Business Streams, Source : TM Forum

The above figure shows the Network-Centric End-To-End business streams which, as stipulated by TM Forum, are:

- Production Order-to-acceptance
- Trouble ticket-to-solution
- Activation-to-Usage-Data
- Capacity Management
- Service Lifecycle Management
- Resource Lifecycle Management

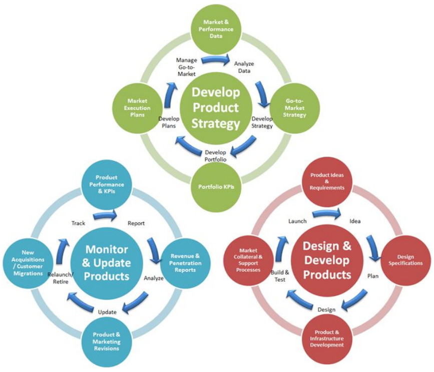

Market & Performance Data
Manage Go-to-Market
Analyze Data
Market Execution Plans
Develop Product Strategy
Go-to-Market Strategy
Develop Plans
Develop Strategy
Develop Portfolio
Portfolio KPIs
Product Performance & KPIs
Track
Report
New Acquisitions / Customer Migrations
Monitor & Update Products
Revenue & Penetration Reports
Relaunch/ Retire
Analyze
Update
Product & Marketing Revisions
Product Ideas & Requirements
Launch
Idea
Market Collateral & Support Processes
Design & Develop Products
Design Specifications
Build & Test
Plan
Design
Product & Infrastructure Development

Figure11: Product-Centric End-To-End Business Streams, Source : TM Forum, Tribold

The above figure shows the Product-Centric End-To-End business streams which, as stipulated by TM Forum, are:

- Develop Product Strategy
- Design & Develop Products
- Monitor & Update Products

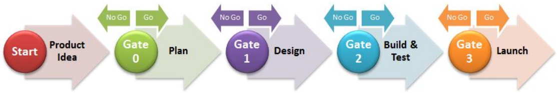


Figure12: Design & Develop Products Stages/Gates, Source: TM Forum, Tribold

The above figure shows the Stages related to the Business Stream "Design & Develop Products":

- Product idea-to-plan (idea is ready for planning)
- Product plan-to-design (Plan is ready or design)
- Product design-to-build&test (Design is ready for Build&Test)
- Product Build & Test-to-launch (Build is ready for launch)

Project Identity

A template, with an initial filled content, of this item could be as follows:

PROJECT Name	**Ex. : Improvement of the Order To Cash/Payment process**
Project Responsible	
Function	
Sponsor	
Project Coordinator	

How to use the support documents?

- Project Name could be derived from the Business Stream names and content
- Project Name could be derived from the Business Metrics names and content, or
- A combination of Business Stream and Business Metrics names and content

Example

- Project Name could be, as mentioned in the template, "Improvement of Order To Cash (Payment)". Order To Payment is one of the Business Streams previously presented.

Project Description

A template, with an initial filled content, of this item could be as follows:

Description du projet :	The Order To Cash, and end-to-end business stream, is a transverse process going from receiving the customer order till issuing the customer bill. This process needs to be more agile in order to improve customer experience, operational efficiency and revenue & margin.

How to use the support documents?

- The Project description could be derived from the Detailed information related to the Business Streams and Business Metrics.

Example

- The project description, as mentioned in the template, is derived from the Figures 4, 5, 6 and 7.

Key Project Performance Data

A template, with an initial filled content, of this item could be as follows:

Key Project Performance Data	**Revenue Breakdown**: • Customer Segment: B2B • Service : Optical Fiber **Average Monthly Fees per customer for optical fiber**: • **X** Euro (s)

How to use the support documents?

- The performance data fields could be derived from Business Metrics.

Example

- The key Project Performance Data field (s), as mentioned in the above template, was derived from the Metric Revenue Breakdown part of Business Metrics.

Objectives/Performance Indicators

A template, with an initial filled content, of this item could be as follows:

Objectives/ Performance Indicators	***) Increase** % of orders delivered by committed date *) **Decrease** Mean Time Order To Activation *) **Decrease** % of Orders requiring technical reworks

Concepts

- An objective is generally formulated as follows:
  - **Increase/Decrease/Stabilize** performance_metric

How to use the support documents?

- These objectives could be derived from Business Metrics, *QSP* (for: Fulfillment, Assurance and New Services) and Introductory Guide (s) for PLM.

Example

- The project description, as mentioned in the template, is derived from the Figures 4, 7 and 8.

Project Scope

A template, with an initial filled content, of this item could be as follows:

Project Scope	The Order To Cash process scope is: ***) Geography**: MTN Business, Ivory Coast; ***) Customer Segment**: B2B; ***) Products/Services**: Optical Fiber; ***) Execution Time**: Normal flow (Fast Track is out of scope); ***) Channel**: … ; ***) Triggers**: …

How to use the support documents?

- The project Scope could be derived from GB921E, QSP (For: Fulfillment, Assurance, New services) and Introductory guide (s) (Ex.: for PLM), GB921D, GB921DX, eTOM and

Business Metrics posters and GB921J for process context.

Example

- The project Scope key fields, as mentioned in the above template, was derived from GB921J.

<u>Financial Impact</u>

A template, with an initial filled content, of this item could be as follows:

Financial Impact			
Improvement area	**Baseline As-Is**	**Impact**	**Impact Cost (Euros)**
Increase % Orders delivered by committed date	The current performance	**<u>Win</u>** Days of service usage **<u>Avoid</u>** Delay in billing Revenue loss	Impact of one day lost = X (monthly fees per customer)/30 days Euros
…			

How to use the support documents?

- The financial impact improvement areas could be derived from the project objectives/indicators.

Example

- As mentioned in the template, Increase % Orders delivered by committed date is derived from the project objectives/indicators and followed by a brainstorming session to fill the impact and its cost for example the cost of one day lost (Euros) is equal to X (monthly fees per customer) divided by 30 days.

<u>Project Organization</u>

Support Documents

- The definition of project organization could be supported by:
    - The L6S Project Streams previously presented
    - RACI Tool

10.2 SIOC

The SIPOC is related to process and its components are:

- **S**uppliers
    - Who provide the inputs?
- **I**nputs
    - What are the inputs for the process
- **P**rocess Steps
    - What are the process steps?
- **O**utputs
    - What are the process results?
- **C**ustomers
    - To whom the results are intended to?

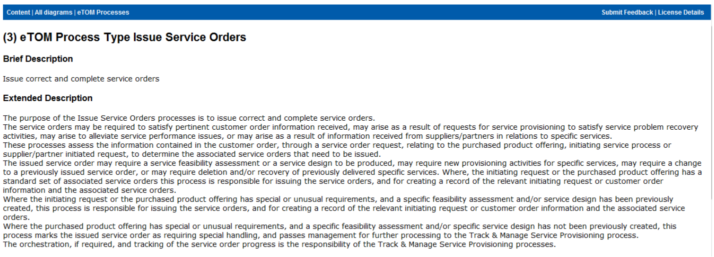
Support Documents

The support Frameworx documents for the SIPOC are:

- GB921D and DX for eTOM decomposition
- GB921E : for end to end business streams
- GB921F : for process flows examples
- QSP for fulfillment, Assurance and new services
- Introductory guide for PLM
- Business Metrics

Snapshots from Support Documents

Below are some snapshots, from the support documents, that can be used for this item.



Content | All diagrams | eTOM Processes
Submit Feedback | License Details

(3) eTOM Process Type Issue Service Orders

Brief Description

Issue correct and complete service orders

Extended Description

The purpose of the Issue Service Orders processes is to issue correct and complete service orders.
The service orders may be required to satisfy pertinent customer order information received, may arise as a result of requests for service provisioning to satisfy service problem recovery activities, may arise to alleviate service performance issues, or may arise as a result of information received from suppliers/partners in relations to specific services.
These processes assess the information contained in the customer order, through a service order request, relating to the purchased product offering, initiating service process or supplier/partner initiated request, to determine the associated service orders that need to be issued.
The issued service order may require a service feasibility assessment or a service design to be produced, may require new provisioning activities for specific services, may require a change to a previously issued service order, or may require deletion and/or recovery of previously delivered specific services. Where, the initiating request or the purchased product offering has a standard set of associated service orders this process is responsible for issuing the service orders, and for creating a record of the relevant initiating request or customer order information and the associated service orders.
Where the initiating request or the purchased product offering has special or unusual requirements, and a specific feasibility assessment and/or service design has been previously created, this process is responsible for issuing the service orders, and for creating a record of the relevant initiating request or customer order information and the associated service orders.
Where the purchased product offering has special or unusual requirements, and a specific feasibility assessment and/or specific service design has not been previously created, this process marks the issued service order as requiring special handling, and passes management for further processing to the Track & Manage Service Provisioning process.
The orchestration, if required, and tracking of the service order progress is the responsibility of the Track & Manage Service Provisioning processes.

Figure13: Issue Service Order details (GB921D)
Source: TM Forum

The above figure shows details about the process element (eTOM Level3) obtained from GB921D.

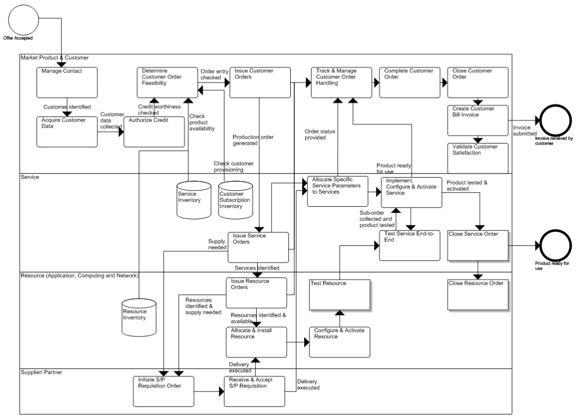

Offer Accepted
Market Product & Customer
Manage Contact
Customer identified
Acquire Customer Data
Customer data collected
Authorize Credit
Credit worthiness checked
Determine Customer Order Feasibility
Check product availability
Order entry checked
Issue Customer Orders
Production order generated
Check customer provisioning
Track & Manage Customer Order Handling
Order status provided
Complete Customer Order
Close Customer Order
Create Customer Bill Invoice
Invoice submitted
Invoice received by customer
Validate Customer Satisfaction
Product ready for use
Service
Service Inventory
Customer Subscription Inventory
Allocate Specific Service Parameters to Services
Implement, Configure & Activate Service
Product tested & activated
Sub-order collected and product tested
Test Service End-to-End
Close Service Order
Product ready for use
Supply needed
Issue Service Orders
Services identified
Resource (Application, Computing and Network)
Resource Inventory
Issue Resource Orders
Resources identified & supply needed
Resources identified & available
Allocate & Install Resource
Test Resource
Configure & Activate Resource
Close Resource Order
Delivery executed
Supplier/ Partner
Initiate S/P Requisition Order
Receive & Accept S/P Requisition
Delivery executed

Figure14: Order-to-Payment – Process-Flow (L3) BPMN Source: TM Forum

The above figure shows process flow related to Order To Payment. In the process flow, inputs and outputs are shown.

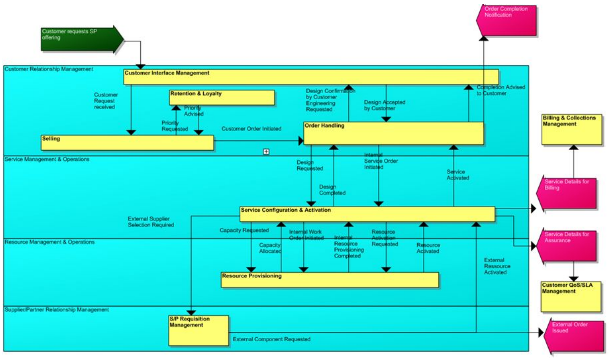

Order Completion Notification
Customer requests SP offering
Customer Relationship Management
Customer Interface Management
Customer Request received
Retention & Loyalty
Priority Advised
Priority Requested
Design Confirmation by Customer Engineering Requested
Design Accepted by Customer
Completion Advised to Customer
Customer Order Initiated
Selling
Order Handling
Billing & Collections Management
Service Management & Operations
Design Requested
Internal Service Order Initiated
Service Activated
Design Completed
Service Details for Billing
Service Configuration & Activation
External Supplier Selection Required
Capacity Requested
Internal Work Order Initiated
Internal Resource Provisioning Completed
Resource Activation Requested
Resource Activated
Capacity Allocated
Resource Management & Operations
External Ressource Activated
Service Details for Assurance
Resource Provisioning
Customer QoS/SLA Management
Supplier/Partner Relationship Management
S/P Requisition Management
External Component Requested
External Order Issued

Figure15: Fulfillment Flow - Level 2 Ordering Process Flow, Source: TM Forum

The above figure shows an example of Fulfillment Flow - Level 2 Ordering Process Flow where some inputs and outputs could be identified.

SIPOC

A template, with an initial filled content, of this deliverable could be as follows:

SIPOC: Order To Cash (Extract)

Supplier	Input	ProcessTask (s)	Output	Customer

Flow 1: ….

If …:

Supplier	Input	Task	Output	Customer
Customer	Customer Order	Receive Customer Order	Customer Order	Sales
……	……	……	……	……
Customer Care	Validated Customer Order	Issue Service Order	……	……
Customer	Customer Service configuration done	Test the service with customer	Delivery Slip	Customer
Customer	Signed Delivery Slip	Get Signed Delivery Slip	Signed Delivery Slip	Customer Care

How to use the support documents?

- SIPOC elements could be derived from the extended description related to eTOM process elements (GB921D and DX) and from process flows (GB921E, F …) where many inputs are shown.

Example

- As mentioned in the template, Validated Customer order Input, could be derived from Figure13.

10.3 High Level process flow

The high level process flow shows the interaction between the process elements involved in Order To Cash (Payment). The flow shows the sequence in which the process elements should be executed.

Support Documents

The support Frameworx documents for this Deliverable are:

- GB921E : for end to end business streams
- GB921F : for process flows examples
- QSP for fulfillment, Assurance and new services
- Introductory guide for PLM
- GB921J for process flow design guidelines and context

Example

- The Figure14 shows a process flow related to the Order To Cash.

10.4 Process accountabilities

To define process accountabilities, eTOM process elements could be used in combination with RACI tool.

Support Documents

The support Frameworx documents for this Deliverable are:

- GB921D, DX
- GB921E
- QSP for Fulfillment, Assurance and New services
- Introductory Guide (s) for PLM

Process accountabilities

A template, with an initial filled content, of this deliverable could be as follows:

eTOM process element	R (Responsible)	A (Accountable)	C (Consulted)	I (Informed)
Selling	Sales	Sales	Service Management …	Service Management …
…	…	…	…	…
Service Configuration & Activation	Service Management	Service Management	…	…
…	…	…	…	…

How to use the support documents?

- The support documents will provide the appropriate process elements within the project scope
- To these elements, RACI is applied.

Example

- The eTOM process elements in the Process accountabilities, as mentioned in the template, are derived from GB921E.

10.5 Voice Of Customer (VOC) and Critical To Quality (CTQ)

The voice of customer consists of identifying the key customer needs. To identify these needs, the touch points between the customer and the service provider, in the project scope, should be analyzed.

To prioritize customer needs, Pareto analysis could be used.

CTQ will define the measures/indicators to be associated with the key customer needs.

Support Documents

The support Frameworx documents for this deliverable are:

- GB921E
- QSP for Fulfillment, Assurance and New Services
- Introductory guide for PLM
- CEMGB962 : is about Introduction and Fundamentals related to the Customer Experience Management

Snapshots from Support Documents

Some selected snapshots related to these support documents are shown in the figures 5, 10, 11 and 12.

VOC

A template, with an initial filled content, of this deliverable could be as follows:

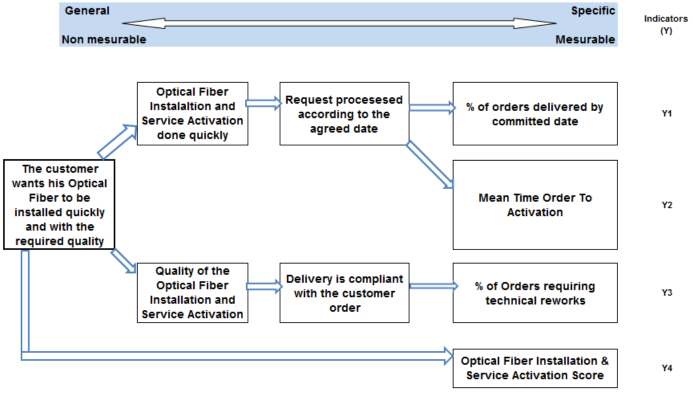

General
Specific
Non mesurable
Mesurable
Indicators (Y)
Optical Fiber Instalaltion and Service Activation done quickly
Request procesesed according to the agreed date
% of orders delivered by committed date
Y1
The customer wants his Optical Fiber to be installed quickly and with the required quality
Mean Time Order To Activation
Y2
Quality of the Optical Fiber Installation and Service Activation
Delivery is compliant with the customer order
% of Orders requiring technical reworks
Y3
Optical Fiber Installation & Service Activation Score
Y4

Figure15: Voice Of Customer and CTQ related to the case study

How to use the support documents?

To help identifying the customer needs:

- End to end business streams could be used
- Customer Experience (CE) performance domain, part of Business Metrics, could be used especially the topics related to CE.
- Customer Surveys (internal, via third parties)
- Pre-Sales
- Sales
- Customer Service

Example
As mentioned in the template, the voice of customer was identified:

- Using the three sub-domains related to the customer experience performance domain (Access, Time, Quality), and
- The feedback of the internal departments interacting with the customer.

And CTQ/Performance indicators were obtained using TM Forum Metrics which were customized to meet specific needs related to MTN Business Operations.

10.6 Change Management Essential Charters

The change management essential charters components are:

Communication Plan

The template for the communication plan is as follows:

Stakeholders	Communication type (Targeted, Global)	Purpose of the communication	Support Documents	Tool	Timing (Phase, …)	Owner
….	….	….	….	….	….	….

The main components of the communication plan are:

- Stakeholders: Target of the communication
- Communication Type : Targeted or Global
- Purpose of the communication
- Support documents
- Timing: when to do the communication
- Owner: Accountable for the communication

11. L6S applied to Telco: Analyze and Measure phases

The measure phase could be supported by Frameworx in multiple ways:

- While doing the VSM, eTOM end to end business streams could be used, jointly
- With Business Metrics while computing the process performance data to perform any customization regarding the process metrics definitions.

The analyze phase could be supported by Frameworx in multiple ways:

- While applying Fishbone technique to support the identification of causes, of non process performance, related to Methods and Inputs, eTOM end to end business streams could be used.
- While Applying 7 Wastes technique, eTOM business streams could be used as well. For example, to identify wastes related to the Waiting Time, we could have one between Order Handling and Service Configuration and Activation due to the non existence of an internal operational level agreement (OLA) between these 2 processes.

The core foundation of FMEA, which is composed of the process steps, is already defined during the Define Phase using the appropriate support documents as previously explained. The output of the FMEA is an action plan for which a steering committee has been defined.

12. L6S applied to Telco: Remaining phases briefly

Regarding the process design, To-Be VSM, To-Be Process Flow and To-Be Procedure was formalized and with a pilot run.

13. L6S applied to Telco: Benefits

The benefits could be seen from three different perspectives:

- Customer Experience
  - **Increase** in % of orders delivered by committed date by 67%
- Operational Efficiency
  - **Decrease** in the Mean Time Order To Activation by 36%
- Revenue & Margin
  - **Increase** of the Revenue Breakdown, by 55%, related to:
    - Customer Segment: B2B
    - Service : Optical Fiber

13. Conclusion

This paper suggests a methodology harmonizing Lean Six Sigma and Frameworx selected components with a focus on the engineering aspects related to the process. This harmonization, called L6S-Telco, allows quicker Transformation projects and getting standardization benefits since Frameworx components are used throughout the methodology where applicable as inputs for the design aspects. It does generate Six Sigma benefits as shown in the case study's methodology benefits. L6S-Telco is an approach to address business process transformation for Telco industry and might be applicable to many other service companies since many of the Frameworx components and Lean Six Sigma are too. The harmonization will optimize the CaPex and eventually OpEx since one single optimized project is executed to address two major needs namely process design and performance.

The current paper could be as input for the coming article (s) to apply L6S-Telco to other end to end business streams related to the customer, Network and Product. It can be also a basis to consider other layers of the enterprise architecture (Information, Application, Technology) to have an end to end business transformation.

Acknowledgments

Special thanks are addressed to my parents, my wife, my daughter and my son for their prayers, smiles, and support.

The authors would like to thank all those who participated in the project, basis of the case study, from MTN Ivory Coast corporate and MTN Business.

I would like to thank in advance anybody for any provided feedback or for any inspiration.

References

[1] M. Kelly, A. Kawecki, "Business Process Framework (The Business Process Framework), Addendum D: Process Decompositions and Descriptions", TM Forum July 2012

[2] M. Kelly, D. Freed, "Business Process Framework (The Business Process Framework), Addendum DX: Extended Process Decompositions and Descriptions", TM Forum April 2012

[3] J. Wilmes, A. Kawecki, " TM Forum Portfolio and Product Management Quick Start Pack ", TM Forum October 2012

[4] TM Forum, "Business Metrics and Benchmarking Program: Business Performance Management System, Level1 Metrics", TM Forum November 2011

[5] G. Vitt, M. Kelly, A. Kawecki, " Business Process Framework (The Business Process Framework), Addendum E: Application Note – End-To-End Business Flows", TM Forum April 2011

[6] G. Vitt, C. Dietze, M. Kelly, A. Kawecki, " Business Process Framework (The Business Process Framework), Addendum J: Application Note – Joining the Business Process Framework through to Process Flows", TM Forum October 2011

[7] M. Benhima, C. Ameyao, A. Ajaoui, A. Salhi, M. H. Jellouli, S. Garcia, "Transformation Business : Pour une meilleure optimization", TIC Magazine N°2 Janvier – Mars 2012

[8] M. Benhima, J. Reilly, H. Benhima, "Design of an enhanced Integrated Management System with Customer Experience Focus: The Business Process Framework (also known as eTOM) and ISO9001 together", IJCSI, Vol9, Issue 5, N° 2, Sept 2012

[9] K. Amin, S. Lynch, C. Michel, E. Obreja, K. Willets, C. Michel, Tribolod, Telecom New Zeland, A. Kawecki, "Product Lifecyle Management Introductory Guide", TM Forum November 2011

[10] K. Amin, S. Lynch, C. Michel, E. Obreja, K. Willets, C. Michel, Tribolod, Telecom New Zeland, A. Kawecki, "Product Lifecyle Management Introductory Guide", TM Forum November? 2011

[11] T. McCarty, M. Bremer, L. Daniels, P. Gupta, "Six Sigma Black Belt Handbook", Motorola University, 2004

[12] C. Frechet, "Mettre en oeuvre le Six Sigma", Eyrolles, 2005

Mounire Benhima is a Senior Consultant, specializing in Business Transformation with its implementation challenges. 15 years of experience in various industrial environments (mainly ICT) in many countries (Africa, America, Europe, Asia). He is distinguished to be the first one in the world to be certified The Business Process Framework Level4 (The highest level of Business Process Framework certification) with the highest score when it was an essay. He is also part of the international TM Forum Trainers Panel (21 members as per August 2012). Mounire has held various key consulting positions and has led numerous groups in major business transformation projects. As a result of his experience and background, Mounire has developed extensive operational and strategic skills in the Enterprise Architecture and Business Transformation with their impact on the organizational structure, quality management, business process engineering, information system governance, and change management.

Abdelaâli Himi Abdelaali HIMI is a Senior Consultant and Researcher at the Faccculté Science and Technology University Hassan I in Settat. He is an Engineer from ENSIAS since 1996. He has more than 16 years experience. Expert in management information systems and has accumulated many years of experience in project managemen, Business Process Reengineering, Enterprise Architecture and ERP. He is certified ITIL Intermediate (OSA), Cerified ISO27001 Lead Auditor and Lead Implementer and ISO2000 Lead Auditor and Certified Lean Six Sigma. He has published several papers on the management of services and the management of the business value chain.

Camille Ameyao Camille Ameyao is a Senior Manager of MTN Business Service Delivery He is in charge of Preseales, implementation and after sales support. He is an active member when introducing new technologies. For Business To Business, he manages Pre-Sales activities, resource management and operations activities, service management and operations activities and Technical support expertise to Business Sales. He is also assigned for special projects as acquisitions. He participated in many seminars delivering high quality speeches about challenges and new trends in the Telco industry.

Edwige Ahonie Adou Edwige Ahonie Adou is Business Analysis and Report Expert within MTN Business and Project Coordinator for many key projects related to MTN Business Business Transformation. For these key positions, she demonstrated high personal and project coordination skills which allowed her to be distinguished and assigned to key projects. She is an efficient communicator. Her main duties are implementing the integrated management systems related to processes, Designing new processes, generating the CxO Balanced Scorecard and ensuring Leadership during Business Transformation projects execution.

The Hybrid of Classification Tree and Extreme Learning Machine for Permeability Prediction in Oil Reservoir

Chandra Prasetyo Utomo

Faculty of Information Technology, Universitas YARSI
Jakarta 10510, Indonesia

Abstract

Permeability is an important parameter connected with oil reservoir. In the last two decades, artificial intelligence models have been used. The current best prediction model in permeability prediction is extreme learning machine (ELM). It produces fairly good results but a clear explanation of the model is hard to come by because it is so complex. The aim of this research is to propose a way out of this complexity through the design of a hybrid intelligent model. The model combines classification and regression. In order to handle the high range of the permeability value, a classification tree is utilized. ELM is used as a final predictor. Results demonstrate that this proposed model performs better when compared with support vector machines (SVM) and ELM in term of correlation coefficient. Moreover, the classification tree model potentially leads to better communication among petroleum engineers and has wider implications for oil reservoir management efficiency.

Keywords: *Permeability Prediction, Extreme Learning Machine, Classification Tree, Hybrid Intelligent Systems, Oil Reservoir, Regression Problem*

1. Introduction

Permeability is the flow capacity of fluid to be transmitted through a rock's pore space. According to the latest study in oil reservoir, millions of dollars can be saved or lost depending on the quality of permeability prediction. The information of permeability values in reservoirs is important because it is needed to find out the quantity of oil or gas exists in reservoirs, the quantity that can be retrieved, its flow rate, the prediction of future production, and the production facilities design. Based on that, correct knowledge of permeability is required for the whole reservoir management and development [1].

Conventional method used to obtain the permeability values is by taking rock samples in some depths then measuring its permeability in the laboratory. This method is very expensive, complex, and time consuming. In addition, laboratorial measurement is limited to the rock samples. So that, the continuous picture of permeability values can't be captured. Based on this reasons, a new method which is quite accurate, less expensive, simpler, faster, and able to deliver permeability distribution along the depth is highly needed.

A huge number of efforts have been carried out to obtain new method to predict permeability values from well log data. From 1927 to 1981, scientists had tried empirical models by delivering mathematical formulas to get permeability values. None of this formula gives satisfying result in general case. Since 1961, multiple variable regressions models had been applied. The distribution of predicted values gained from this model is still far from actual values. However, empirical and regression models gave hint about factors controlling permeability [2].

In the past two decades, computational intelligent techniques, such as artificial neural networks (ANN), have been utilized in permeability prediction. An ANN is a powerful and flexible tool for many applications including in petroleum area. This model is able to learn from previous data in order to predict values from new data. It gives better performance than previous models in predicting permeability from well logs in new wells [3]. Nevertheless, back propagation neural network suffers some drawbacks. It has some tuning parameters such as number of hidden neurons, learning rate, and momentum so it needs more efforts to find the best model. In addition, the gradient based learning algorithm used by ANN makes the training process becomes time consuming.

Many works have been tried to develop new ANN model to solve its weaknesses. In 2004, Huang [4] proposed new learning algorithm for single-hidden layer feed forward neural networks which is called extreme learning machine (ELM). Both in theory and experimental results, this learning algorithm gives better generalization performances and extremely faster learning speed than traditional popular gradient based learning algorithm [5]. Based on that, ELM has been highly exploited in many applications including in petroleum engineering area. In comparison with support vector machines (SVM) and conventional ANN for predicting permeability from well log data, ELM gives better generalization ability and faster speed [1]. This result stated that ELM is the current best single model in permeability prediction problem.

Although ELM gives fairly good results and faster speed, it still has some limitations. First, ELM can’t deal with high data distribution of permeability values. One of the main challenges in predicting permeability is high range of its values in each well [6]. Second, ELM can’t give knowledge representation of developed model. Because of its structure which is dense combination of simple computation, trained ELM is hard and complex to be written in mathematical formulas. As a result, it is impossible to produce understandable knowledge representation which is needed to communicate with expert for future study and research.

In this research, a new hybrid intelligent model which can manage high data distribution and give knowledge representation is proposed. To deal with high range data, a single model is not enough. The data should be classified into low permeability and high permeability then applied different models to predict the value.

This proposed hybrid model is basically combination of classification and regression models. Classification model is responsible to classify the data into low and high permeability. On the other hand, regression models are responsible to give final prediction value of its associated data. Classification tree is utilized as classification model since it can produce knowledge representation which is close to human intuition. ELM is used as regression model since it is currently the best single model in permeability prediction.

The rest of this paper is organized as the following. Section 2 is dedicated as previous works. In this section, review in permeability prediction and overview of ELM are presented. In Section 3, design of the proposed model and its implementation are explained. In Section 4, experiments, results, and analysis are provided. Finally, conclusions and future works are given in Section 5.

2. Previous Works

There are huge efforts from scientists and engineers in order to deliver best model to predict permeability values based on well logs data. This section describes previous works in permeability prediction which can be categorized into empirical models, multiple regression variable models, and artificial intelligence models.

2.1 Empirical Models

Empirical models are predicting permeability by defining mathematical formulas based on its correlation with some rock properties. Kozeny [2] introduced the first equation of permeability in 1927. He measured permeability as a function of empirical Kozeny constant, porosity, and surface area. Archie [7] established the concept of “formation resistivity factor” in 1941. His concept indirectly influenced the computation of permeability since it affected the way to calculate water saturation.

Tixier [8] proposed a formula in 1949 to determine permeability from resistivity gradients by using empirical correlation between resistivity and water saturation, water saturation and capillarity pressure, and capillarity pressure and permeability. In 1950, Wyllie & Rose [9] modified the formula proposed by Tixier. Their model is based on quantitative log interpretation theoretical analysis and some assumptions.

In 1956, Sheffield [10] delivered permeability formula based on Kozeny’s equation and formation of a correlation coefficient for some water well-known water-wet sands. However, he recommended his formula is suitable only for clean sands. In 1963, Prison [10] proposed formula which was determined by multiple correlation from relatively few data. For high gravity crudes (API > 40o) and for depths greater than 6500 ft, the formula must not be utilized.

Timur [11] generalized permeability equation based on the work of Kozeny and Willy & Rose. In 1974, Coates & Dumanoir [12] proposed an improved empirical permeability formula which is satisfied the condition of zero permeability at zero porosity and when irreducible water saturation is 100%. Coates and Denoo [13] simplified the previous proposed formulas and still satisfied the zero permeability condition. However, the formation must be at irreducible water saturation.

2.2 Multiple Variable Regression Models

Multiple variable regression models are expansions of the regression analysis that include extra independent variables in the equation. The model can be generalized as:

$$Y = C_0 + C_1X_1 + C_2X_2 + \cdots + C_nX_n + e \qquad (1)$$

where Y is the dependent variable, X_1, X_2,…, X_n are the independent variables, and e is a random error or residual. The regression coefficients C_1, C_2,…,C_n are the parameters to be approximated.

A general procedure of multiple variable regression for permeability prediction was established by Wendt and Sakurai [14] in 1986. The main drawback of using this model is the predicted permeability values is narrower than the actual values. Kendall and Stuart [15] enlightened above phenomena by stating this model gave the best prediction on the average. Weighting the high and low values are applied to improve the capability of regression model to predict outlier data. However, this may turn the predictor into unstable and statistically biased. Pereira [16] reported that density, derivative of density, gamma ray, and derivative of gamma ray are the best combination to be utilized as independent variables in multiple regression analyses.

2.3 Artificial Intelligence Models

Artificial Intelligence (AI) is set of models inspired by nature such as neural networks, fuzzy logic, and genetic algorithm. A lot of neural networks applications can be found in the petroleum industry, from exploration, drilling exploration, to reservoir and production engineering [17]. In predicting permeability, neural networks gave significant improvement [18-21]. This opened the door of others AI models to be applied in the petroleum industry area especially in the permeability prediction problem.

The combination of two or more AI models is called hybrid model. It complements the weaknesses of one model with the advantage of others. Since neural networks is one of the best AI model, most of published hybrid model are neural network based model. There are some proposed hybrid models in permeability prediction. Deni [22] proposed a hybrid of genetic algorithm and fuzzy/neural network inference system. Helmi [23] developed a hybrid of fuzzy logic, support vector machine, and functional network. Karimpouli [6] built up supervise committee machine neural network. Li [24] enhanced decision tree learning approach for neural decision tree model.

Although previous hybrid model gave better results than single model, it has some drawbacks due to the limitation of neural networks model. As a “black box” model, neural networks cannot give clear relationships among variables. Other limitations are it can fall into local minima, need to adjust too many parameters, and time consuming.

2.4 Extreme Learning Machines

A lot of works has been tried to resolve the drawbacks of ANN. Huang and Babri [25] proved that single hidden layer feedforward neural networks (SLFN) with at most *m* hidden nodes is able to approximate function for *m* distinct vectors in training dataset.

Let given *m* vectors in training dataset $\mathbf{D} = \{(\mathbf{x}^{(k)}, \mathbf{t}^{(k)}) \mid \mathbf{x}^{(k)} \in \mathbf{R}^n, \mathbf{t}^{(k)} \in \mathbf{R}^p, k = 1,..,m\}$ where $\mathbf{x}^{(k)} = [x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)}]^T$ and $\mathbf{t}^{(k)} = [t_1^{(k)}, t_2^{(k)}, ..., t_p^{(k)}]^T$. A SLFN with *M* hidden nodes, activation function $g(x)$ in hidden nodes, and linear activation function in output nodes is mathematically modeled as:

$$\sum_{i=1}^{M} \beta_i g_i(\mathbf{x}^{(k)}) = \sum_{i=1}^{M} \beta_i g(\mathbf{w}_i \cdot \mathbf{x}^{(k)} + b_i) = \mathbf{o}^{(k)}, \quad k = 1, \dots, m \tag{2}$$

where

$\mathbf{w}_i \in \mathbf{R}^n$ is the weights attached to the edge connecting input nodes and the *i*-th hidden node

$$\mathbf{w}_i = [w_{i1}, w_{i1}, \dots, w_{in}]^T, \tag{3}$$

$\boldsymbol{\beta}_i \in \mathbf{R}^p$ is the weights attached to the edge connecting the *i*-th hidden node and the output nodes

$$\boldsymbol{\beta}_i = [\beta_{i1}, \beta_{i2}, \dots, \beta_{ip}]^T, \tag{4}$$

$\mathbf{w}_i \cdot \mathbf{x}^{(k)}$ is the inner product of $\mathbf{w}_i$ and $\mathbf{x}^{(k)}$,
b_i is the bias of the *i*-th hidden node,
$\mathbf{o}^{(k)} \in \mathbf{R}^p$ is the output of neural network for *k*-th vector.

The meaning of SLFN can approximate *m* vectors is there are exist $\mathbf{w}_i$, $\boldsymbol{\beta}_i$, and b_i, such that:

$$\| \mathbf{o}^{(k)} - \mathbf{t}^{(k)} \| = 0 \tag{5}$$

$$\sum_{i=1}^{M} \beta_i g(\mathbf{w}_i \cdot \mathbf{x}^{(k)} + b_i) = \mathbf{t}^{(k)}, \quad k = 1, \dots, m \tag{6}$$

Those *m* equations can be written as:

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{T}, \tag{7}$$

where

$\mathbf{H} \in \mathbf{R}^{m \times M}$ is the hidden layer output matrix of the neural networks.

$$\mathbf{H} = \begin{bmatrix} g(w_1 \bullet x^{(1)} + b_1) & \cdots & g(w_M \bullet x^{(1)} + b_M) \\ \vdots & \ddots & \vdots \\ g(w_1 \bullet x^{(m)} + b_1) & \cdots & g(w_M \bullet x^{(m)} + b_M) \end{bmatrix} \tag{8}$$

$\boldsymbol{\beta} \in \mathbf{R}^{M \times p}$ is the weights connecting hidden layer and output layers

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1^{\;T} \\ \vdots \\ \beta_M^{\;T} \end{bmatrix}, \tag{9}$$

$\mathbf{T} \in \mathbf{R}^{m \times p}$ is the target values of *m* vectors in training dataset

$$\mathbf{T} = \begin{bmatrix} t^{(1)T} \\ \vdots \\ t^{(m)T} \end{bmatrix}, \tag{10}$$

In the conventional gradient descent based learning algorithm, weights $\mathbf{w}_i$ which is connecting the input layer and hidden layer and biases b_i in the hidden nodes are needed to be initialized and tuned in every iteration. This is the main factor which often makes training process of neural networks become time consuming and the trained model may not reach global minima.

Huang [5] proposed minimum norm least-squares solution of SLFN which doesn't need to tune those parameters. Training SLFN with fixed input weights $\mathbf{w}_i$ and the hidden layer biases b_i is similar to find a least square solution $\hat{\boldsymbol{\beta}}$ of the linear system $\mathbf{H}\boldsymbol{\beta} = \mathbf{T}$:

$$\left\|\mathbf{H}(\mathbf{w}_1, \dots, \mathbf{w}_M, \boldsymbol{b}_1, \dots, \boldsymbol{b}_M,)\hat{\boldsymbol{\beta}} - \mathbf{T}\right\| = \min_{\boldsymbol{\beta}} \left\|\mathbf{H}(\mathbf{w}_1, \dots, \mathbf{w}_M, \boldsymbol{b}_1, \dots, \boldsymbol{b}_M,)\boldsymbol{\beta} - \mathbf{T}\right\|. \tag{11}$$

The smallest norm least squares solution of the above linear system is

$$\hat{\boldsymbol{\beta}} = \mathbf{H}^{\dagger}\mathbf{T} \tag{12}$$

where $\mathbf{H}^{\dagger}$ is the *Moore-Penrose generalized inverse* of matrix H. This solution has three important properties which are minimum training error, smallest norm of weights, and unique solution which is $\hat{\boldsymbol{\beta}} = \mathbf{H}^{\dagger}\mathbf{T}$.

The above minimum norm least-square solution for SLFN is called extreme learning machine (ELM). Let given *m* vectors in training dataset $\mathbf{D} = \{(\mathbf{x}^{(k)}, \mathbf{t}^{(k)}) \mid \mathbf{x}^{(k)} \in \mathbf{R}^n, \mathbf{t}^{(k)} \in \mathbf{R}^p, k = 1,..,m\}$, activation function $g(x)$, and number of hidden node *M*. The training process of ELM is the the following:

Step (1) Randomly set input-hidden layer weights $\mathbf{w}_\mathrm{i}$ and bias b_i, $i = 1,\ldots,M$.

Step (2) Compute the matrix of hidden layer output **H**

Step (3) Compute the hidden-output layer weights $\hat{\boldsymbol{\beta}}$ for $\hat{\boldsymbol{\beta}} = \mathbf{H}^{\dagger}\mathbf{T}$ where $\mathbf{T} = [\mathbf{t}^{(1)},\ldots, \mathbf{t}^{(m)}]$.

The comparison between conventional widely used neural networks and ELM is summarized in the Table 1.

Table 1: The comparison between Back Propagation ANN and ELM

No.	Points of Comparison	Comparison
1.	Learning Algorithm	ANN: Gradient based learning ELM: Minimum least-squares
2.	Training Parameters	ANN: Need to tuning number of hidden nodes, learning rate, momentum, and termination criteria ELM: Simple tuning-free algorithm. The only one to be defined is number of hidden nodes
3.	Activation Function	ANN: Works only for differentiable functions ELM: Works for differentiable and many non-differentiable functions
4.	Speed	ANN: Very slow especially in the large dataset. All of weights are updated in every iteration. ELM: Extremely faster than BP ANN. Only three steps without any iteration
5.	Result	ANN: Get trained model which has minimum training error. There is possibility to finish in the local minima. ELM: Get trained model which has minimum training error and smallest norm of weight. Better generalization model and reach global minima.

3. Design and Implementation Model

The main challenge in permeability prediction is high range of permeability. A single model is not enough to deal with that. The data should be classified into low permeability and high permeability then applied different model to predict the value.

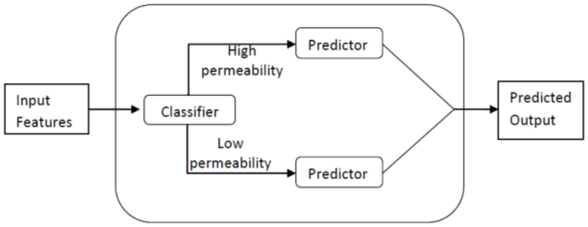


Fig. 1 Design of the proposed hybrid model

Hybrid model which is basically combination of classification and regression models is proposed. Classification model is responsible to classify the data into low and high permeability based on a threshold value. On the other hand, regression models are responsible to give final prediction value of its associated data. Design of this model can be seen in the Fig 1.

One of the objectives in this research is to propose new model which gives understandable knowledge representation. The best representation model which is close to human reasoning is classification tree. For this reason, Classification Tree model is used in the classification part. Since Classification and Regression Tree (CART) from Salford System [26] is one of the best tools for classification tree design, it is implemented in this proposed model.

As presented in previous section, ELM is the current best single model in permeability prediction. ELM developed by Huang [27] is implemented in this proposed model as final predictor.

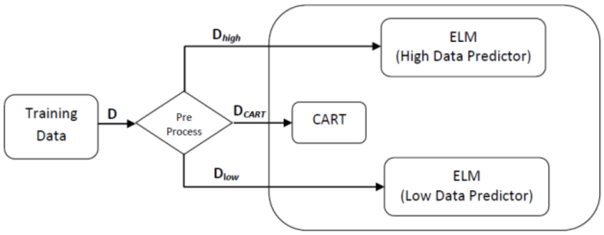


Fig. 2 Training procedure of proposed hybrid model

Let we have *m* vectors in training dataset **D**.

$$\mathbf{D} = \{(\mathbf{x}^{(k)}, \mathrm{t}^{(k)}) \mid \mathbf{x}^{(k)} \epsilon\, \mathbf{R}^n, k = 1,..,m.\}. \qquad (13)$$

The training algorithm of this hybrid model is designed as the following:

Step (1) Add Discretized Target

Discretize the target output $\mathrm{t}^{(k)}$ into two classes *"low"* and *"high"* based on selected threshold value. The new training dataset is $\mathbf{D}_{\mathit{1}} = \{(\mathbf{x}^{(k)}, \mathrm{t}^{(k)}, \mathrm{t_d}^{(k)}) \mid \mathbf{x}^{(k)} \epsilon\, \mathbf{R}^n, k = 1,..,m.\}$ with $\mathrm{t_d}^{(k)}$ is "low" if $\mathrm{t}^{(k)} \leq$ threshold, otherwise $\mathrm{t_d}^{(k)}$ is "high".

Step (2) Produce the Associated Training Data

In this step, three training dataset $\mathbf{D}_{\mathit{CART}}$, $\mathbf{D}_{\mathit{low}}$, $\mathbf{D}_{\mathit{high}}$ are produced. The training dataset for CART $\mathbf{D}_{\mathit{CART}}$ is $\mathbf{D}_{\mathit{1}}$ without original target value $\mathrm{t}^{(k)}$. The vector $(\mathbf{x}^{(k)}, \mathrm{t}^{(k)}, \mathrm{t_d}^{(k)})$ in $\mathbf{D}_{\mathit{1}}$ is putted into $\mathbf{D}_{\mathit{low}}$ if $\mathrm{t_d}^{(k)}$ = "low", otherwise it is putted into $\mathbf{D}_{\mathit{high}}$. The $\mathrm{t_d}^{(k)}$ element in the $\mathbf{D}_{\mathit{low}}$ and $\mathbf{D}_{\mathit{high}}$ are removed at the end of this step.

Step (3)Train the CART

Train the CART by training dataset

$$\mathbf{D}_{\mathit{CART}} = \{(\mathbf{x}^{(k)}, \mathrm{t_d}^{(k)}) \mid \mathbf{x}^{(k)} \epsilon\, \mathbf{R}^n, k = 1,..,m.\} \qquad (14)$$

Step (4) Train the ELMs

Train the low ELM by training dataset

$$\mathbf{D}_{\mathit{low}}\ \{(\mathbf{x}^{(l)}, \mathrm{t}^{(l)}) \mid \mathbf{x}^{(l)} \epsilon\, \mathbf{R}^n, l = 1,..,y.\} \qquad (15)$$

Train the high ELM by training dataset

$$\mathbf{D}_{\mathit{high}}\ \{(\mathbf{x}^{(h)}, \mathrm{t}^{(h)}) \mid \mathbf{x}^{(h)} \epsilon\, \mathbf{R}^n, h = 1,..,z.\} \qquad (16)$$

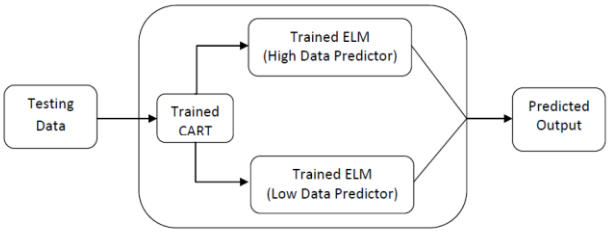

Testing Data
Trained CART
Trained ELM (High Data Predictor)
Trained ELM (Low Data Predictor)
Predicted Output

Fig. 3 Testing procedure of trained model

After finish four steps above, the trained hybrid model is produced and ready to predict permeability from new dataset. The illustration can be seen in Fig 3.

4. Experiments, Results, and Analysis

The data used in this experiment are 5 well logs data from Saudi Aramco. Data for Well 1 has 145 rows (vectors), for Well 2 has 141 rows, for Well 3 has 193 rows, for Well 4 has 147 rows, and for Well 5 has 141 rows. There are 5 input variables which are DT (sonic travel time), GR (Gamma Ray), PHIE (Effective Porosity), RHOB (Density), and SWT (Water Saturation). The target output to be predicted is PERM (Permeability).

Two kinds of experiments are conducted in this research. In the first experiment, one well is chosen as tested well and the rest wells are used to train the model. Because there are 5 wells, this experiment is repeated up to 5 times with different combination of training and testing wells. In the second experiment, all data are combined then divided randomly into training and testing data with ratio 80:20. The training data is used to train the model. Then, the trained model is tested by testing data to predict the permeability values.

The input features are normalized into [-1,1] and the output target is kept in the original value. The threshold used in this experiment is 1. This means, if the permeability value is less or equal than 1, then it is considered as low permeability. Otherwise, it is high permeability. A number of experiments had been tried to get the best parameters combination of CART such as in splitting criteria, stopping conditions, and thresholds.

Both classification and final prediction performance will be measured. The performance measurements for classification are Accuracy (ACC), True Positive Rate (TPR), and False Positive Rate (FPR). In order to measure the performance of the whole model, Root Mean Square Error (RMSE) and Correlation Coefficient (R) are used as performance criteria. The proposed model will be compared with SVM [28] and ELM based on this performance criteria.

ELM assigns randomly input weights and biases in the first step of execution. To reduce the influence of random generator, 10 sequences of executions are applied in each model and the average results are obtained.

Table 2: The performances of CART as classifier

Tested Well	TPR	FPR	ACC
1	0.8333	0.4220	0.6414
2	0.8333	0.0693	0.9148
3	0.5783	0.2818	0.6500
4	0.2727	0.0326	0.7075
5	0.4658	0.0735	0.6879

Table 3: The performances comparison of models

Tested Well	RMSE			R		
	SVM	ELM	Hybrid	SVM	ELM	Hybrid
1	7.87	9.77	12.24	0.55	0.44	0.44
2	19.47	14.48	13.98	0.67	0.77	0.73
3	16.82	15.52	15.29	0.38	0.39	0.42
4	9.38	8.514	9.51	0.40	0.44	0.35
5	10.40	8.42	9.60	0.38	0.47	0.47

The performances of CART as classifier to classify the high and low permeability data are shown in Table 2. These performances are obtained after tree pruning. When there is no pruning mechanism in classification tree induction, the classifier testing performances are bad and the final predictions of hybrid model are not reliable. Table 3 shows that the performances of proposed model are similar with current single best prediction model in permeability prediction.

The comparison of models based on RMSE can be clearly seen in Fig. 4. Except the models for tested Well 1, SVM models give the highest errors. The proposed models are better than ELMs in tested Wells 2 and Well 3.

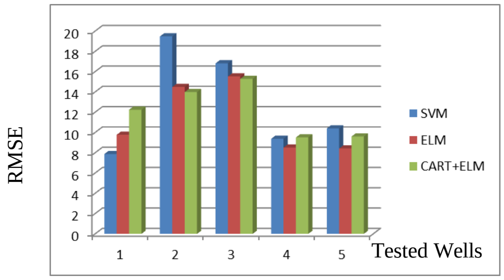


Fig. 4 The performances comparison based on RMSE

Fig. 5 shows the comparison of models based on Correlation Coefficient R. SVMs give the worst performances in tested Wells 2, 3, and 5. The proposed models are better than ELMs in tested Well 3 and 5, worse in tested Wells 2 and 4, and almost equal in tested Well 1.

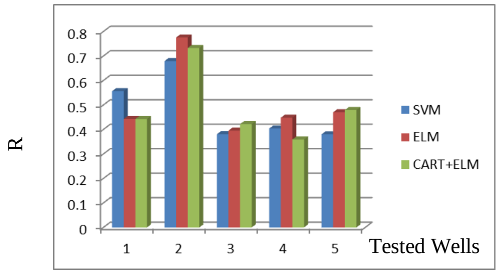


Fig. 5 The performances comparison based on Corr. Coefficient (R)

The performances results of the second experiment which is randomly divided data into training and testing data can be seen in table 4. In term of RMSE, the proposed model is worse than SVM and ELM. In term of R, the proposed model is better than SVM and ELM.

Table 4: The performances comparison of models in general Wells

Model	RMSE	R
SVM	12.88770	0.20670
ELM	12.49098	0.23453
CART + ELM	13.07807	0.26898

Another way to see differences of prediction is by looking the plot of actual and predicted values. Fig. 6 gives the permeability data plot of actual value and predicted value by ELM and proposed model. This figure shows that the proposed model can handle high distribution data and predict accurately the low permeability values. However, it is still not good enough to predict the high permeability values.

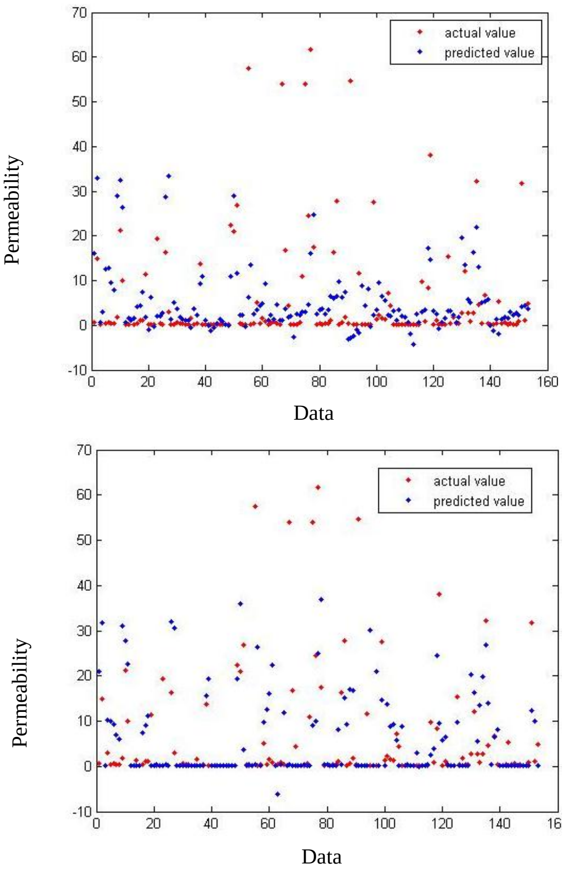


Fig. 6 Plotting permeability data of actual values and predicted values by ELM (top) and CART+ELM (bottom)

One of the most important objectives in this research is deliver knowledge representation. The classification tree is produced in the classification part. The classification tree produced in the second experiment can be seen in the Fig.7. This tree is simple and understandable. Some rules connected with relationship between permeability and the predictors can be drawn. It can be used to communicate with experts and researchers in domain problem.

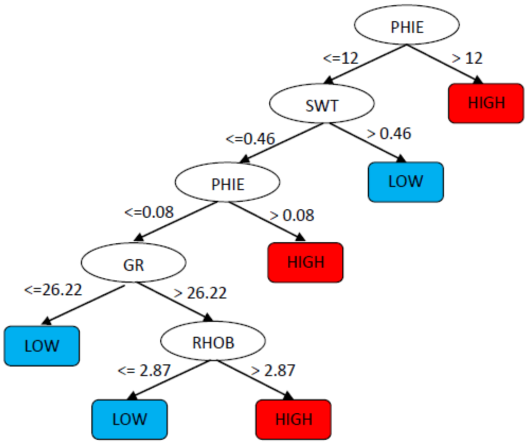


Fig. 7 Classification tree generated by CART in the classification part

5. Conclusions and Future Works

Based on the results and analysis of the experiments, some conclusions can be drawn. The proposed hybrid model, which is combination of Classification Tree as classifier and ELM as predictor, gives better performance than SVM and ELM in term of correlation coefficient in general Wells. The prediction in low permeability data is excellent but still not good enough in high permeability data.

The classification part plays important role in determining the prediction. The better accuracy of classifier, the better result in final prediction. The classification tree produced by this hybrid model is simple and understandable. This means, it will be promising tool to be widely used to communicate with domain expert. Although the proposed model just gave small improvement, it concludes that the use of hybrid model in this way is in the right direction.

The future work will be improvement in both classification and regression parts of this hybrid model. It is interesting to see how performance of classification tree with others induction tree algorithms. It is also necessary to investigate different possible hybrid models which combine classification tree with other regression models such as support vector regressions and fuzzy systems.

Acknowledgments

This paper is result from master thesis works by the author at the King Abdullah University of Science and Technology (KAUST). He would like to thank Prof. Mikhail Moshkov for supports and guidance. He would also like to thank Dr. Igor Chikalov for technical supports, Dr. Xiangliang Zhang for insight of using classification, and Dr. L. Ghouti for providing well logs data from Saudi Aramco Oil Reservoir. Many thanks also to Prof. Basem Shihada and Prof. Shuyu Sun for examining and approving this work as thesis committee member.

References

[1] S. O. Olatunji, *et al.*, "Modeling Permeability Prediction Using Extreme Learning Machines," in *Mathematical/Analytical Modelling and Computer Simulation (AMS), 2010 Fourth Asia International Conference on*, 2010, pp. 29-33.

[2] B. Balan, *et al.*, "State-Of-The-Art in Permeability Determination From Well Log Data: Part 1- A Comparative Study, Model Development," in *SPE Eastern Regional Conference & Exhibition*, Morgantown, West Virginia, USA, 1995.

[3] S. Mohaghegh, *et al.*, "State-Of-The-Art in Permeability Determination From Well Log Data: Part2- Verifiable, Accurate Permeability Prediction, The Touch-Stone of All Models," in *SPE Eastern Regional Conference & Exhibition*, Morgatown, West Virginia, USA, 1995.

[4] H. Guang-Bin, *et al.*, "Extreme learning machine: a new learning scheme of feedforward neural networks," in *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, 2004, pp. 985-990 vol.2.

[5] G.-B. Huang, *et al.*, "Extreme learning machine: Theory and applications," *Neurocomputing,* vol. 70, pp. 489-501, 2006.

[6] S. Karimpouli, *et al.*, "A new approach to improve neural networks' algorithm in permeability prediction of petroleum reservoirs using supervised committee machine neural network (SCMNN)," *Journal of Petroleum Science and Engineering,* vol. 73, pp. 227-232, 2010.

[7] G. E. Archie, "The Electrical Resistivity Log as an Aid in Determining Some Reservoir Charachteristics," *Trans., AIME,* vol. 146, pp. 54-62, 1942.

[8] M. P. Tixier, "Evaluation of Permeability From Electric-Log Resistivity Gradients," *Oil and Gas Journal,* p. 113, June 1949.

[9] M. R. J. Wyllie and W. D. Rose, "Some Theoretical Considerations Related to the Quantitative Evaluation of the Physical Characteristics of Reservoir Rock from Electric Log Data," *Trans., AIME,* vol. 189, p. 105, 1950.

[10] S. J. Pirson, *Handbook of Well Log Analysis*: Englewood Cliffs, N.J., Prentice-Hall, Inc., 1963.

[11] A. Timur, "An Investigation of Permeability, Porosity, and Residual Water Saturation Relationship for Sandstone Reservoirs," *The Log Analyst,* vol. 9, p. 8, July-August 1968.

[12] G. R. Coats and J. L. Dumanoir, "A New Approach to Improved Log-Derived Permeability," *The Log Analyst,* January-February 1974.

[13] SchlumbergerLtd., *Log Interpretation Charts*. Houston, Texas, 1987.

[14] W. A. Wendt, *et al.*, Eds., *Permeability Prediction From Well Logs Using Multiple Regression* (Reservoir Characterization. New York: Academic Press, 1986.

[15] N. R. Draper and H. Smith, *Applied Regression Analysis*: Wiley, 1981.

[16] J. L. L. Pereira, "Permeability prediction from well log data using multiple regression analysis," M.S.PNGE. 1424328, West Virginia University, United States -- West Virginia, 2004.

[17] M. P. McCormak, "Neural Networks in the Petroleum Industry," in *Society of Exploration Geophysicists 1991 Technical Program*, 1991, pp. 285-289.

[18] S. Mohaghegh, *et al.*, "A Methodological Approach for Reservoir Heterogeneity Characterization Using Artificial Neural Networks," presented at the SPE Annual Technical Conference, New Orleans, LA, September 1994.

[19] S. Mohaghegh, *et al.*, "Design and Development of An Artificial Neural Network for Estimation of Formation Permeability," presented at the SPE Petroleum Computer Conference, Dallas, Texas, August 1994.

[20] D. A. Osborne, "Permeability Estimation Using a Neural Network: A Case Study from The Roberts Unit, Wasson Field, Yoakum County, Texas," *AAPG South West Section Transactions,* pp. 125-132, 1992.

[21] J. M. Wiener, "Predicting Carbonate Permeabilities from Wireline Logs Using a Back-Propagation Neural Network," in *Society of Exploration Geophysicists 1991 Technical Program*, 1991.

[22] X. Deny, *et al.*, "Permeability Estimation Using Hybrid Genetic Programming and Fuzzy/Neural Inference Approach," presented at the Society of Petroleum Engineers Annual Technical Conference and Exhibition, Dallas, Texas, USA, 9-12 October 2005.

[23] T. Helmy, *et al.*, "Hybrid computational models for the characterization of oil and gas reservoirs," *Expert Systems with Applications,* vol. 37, pp. 5353-5363, 2010.

[24] X. Li and C. W. Chan, "Application of an enhanced decision tree learning approach for prediction of petroleum production," *Engineering Applications of Artificial Intelligence,* vol. 23, pp. 102-109, 2010.

[25] H. Guang-Bin and H. A. Babri, "Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions," *Neural Networks, IEEE Transactions on,* vol. 9, pp. 224-229, 1998.

[26] S. System, "Salford Predictive Modeler Builder v6.6," ed. San Diego, California: Salford System, 2011.

[27] G.-B. Huang, "ELM Source Codes: http://www.ntu.edu.sg/home/egbhuang/," ed, 2011.

[28] C.-C. Chang and C.-J. Lin, " LIBSVM : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology,* vol. 2, pp. 27:1--27:27, 2011.

C. P. Utomo received B.Sc in Computer Science from University of Indonesia and M.S in Computer Science from King Abdullah University of Science and Technology (KAUST) in 2009 and 2011, respectively. Currently, he is researcher and lecturer at the Faculty of Information Technology Universitas YARSI, Jakarta. His research interests are artificial intelligence, machine learning, and data mining. Specifically, he is eager in applying computational intelligent methods such as neural networks, extreme learning machine, and support vector machines in real world problems in the specific domain such as classification, regression, association rules, and clustering.

Traffic Demand Forecasting for EGCS with Grey Theory Based Multi- Model Method

Zhenshan Yang[1], Yunli Zhang[2]

**[1] College of Engineering, Bohai University
Jinzhou, 121013, China**

**[2] Center of Computer & Network, Liaoning Medical University
Jinzhou, 121002, China**

Abstract

Elevator traffic demand forecasting is the essential prerequisite for effectively implementing elevator group control system (EGCS). Considering that there exists lots of abnomal information in elevator traffic caused by subjectivity and occasionality in human behaviour and that observing traffic information continuously is costly and difficult, an improved grey forecasting based method using multi-model to forecast future elevator traffic demand of EGCS is proposed, the abnomal information which refers to outliers is processed, based on which a smoothing technique on original traffic data is conducted to transform the raw data into an increasing sequence, to further reduce the randomness of the observed traffic data and to make full use of regularity information. The proposed method not only avoid the theorrtical error of grey model per se, but also improved the forecasting accuracy, which is suitable for short period forecasting for elevator traffic demand. Simulation experiments show the validity of the proposed method.

Keywords: *Elevator Traffic Demand Forecasting, Elevator Group Control System (EGCS), Grey Model (GM), Abnormal Information, Multi-Model Forecasting, Smoothing Processing.*

1. Introduction

The increasing perfection of the function of modern high-rise buildings makes the elevator vertical transportation become more and more complicated. Elevator traffic demand forecasting deals mainly with the traffic conditions, evaluating the future traffic demand on real-time observed traffic data. Elevator traffic demand forecasting is the essential prerequisite for elevator traffic pattern recognition, selecting control strategy to effectively implementat EGCS [1]-[5].

Many studies have been done regarding the elevator traffic demand forecasting related problems based on, such as, neural network, exponential smoothing, fuzzy logic theory and so on, and significant progress has been made [6]-[9], but the problem to deal with both the abnormal traffic data according to the specified forecasting model and the the limited number of samples and the incomplete or insufficient information, has not been taken seriously. Therefore, the characteristics of elevator traffic demand make the applications of those methods not be effectively applied in some practical cases, and the satisfactory forecasting accuracy is hard to obtain. So, to develop effective forecasting method is of the utmost importance.

In this thesis, an improved grey prediction based method using multi-model to forecast the future elevator traffic demand of EGCS is proposed based on the following reasons.

1.1 Elevator Traffic Demand Possesses Obvious Grey Characteristics

For the elevator traffic system with EGCS, the passenger arrival process is a stochastic process [10], and the theory of grey system holds that any stochastic process is the grey variables changing in a certain amplitude range for a certain period of time [11]. The traffic flow of EGCS in the same time period of different working day may fluctuate differently to some extent, but stable in a long term.

Since obtaining the observed traffic data is costly and difficult as stated above, it is usually require that the traffic demand be forecasted based on less sample data, which provide opportunities for employing grey forecasting method.

1.2 The Essence of Grey Forecasting

Grey theory focuses on uncertain system with limited number of samples and amount of information, the valuable information is extracted by generating and developping operation on part of the known information to realize the correct understanding and effectively forecasting of the system operating regularity [12]. It can be simply stated that the development regularity of objects which contain incomplete information is forecasted based on the principle of grey system analysis. The core of which is Accumulating Generation Generators (AGO) whose aim is to increase the significance of system operating regularity by reducing the randomicity of the observed traffic samples, its forecasting technique is to set up grey forcasting mode extending from the past to the future on the basis of the past known or present unknown information employing small samples.

The remainder of this thesis is organized as follows. In Section 2, Problems description and modelling is discussed. In Section 3, the application of the proposed method to elevator traffic demand forecasting for EGCS is discussed in detail. In Section 4, simulation experiments and analysis are presented. And conclusions are drawn in Section 5.

2. Problems Description and Modelling

2.1 Description of Elevator Traffic Flow

The operation of EGCS is essentially to transport a certain amount of passengers in the building from their original floors to the ultimate destination floors in a timely manner. Traffic flow which indicates the traffic status of EGCS is expressed by the number of passengers, the period of passengers appearing, as well as the positions of passengers. It can be described by several kinds of data [13], but only part of the data which reflect the traffic characteristics inside the building are used in traffic analysis, which are the number of passengers entering and leaving the main terminal in specified time intervals, the total number of passengers within the building and the interfloor traffic conditions, etc. According to the characteristics of elevator traffic, the traffic flow is classified into four different traffic patterns: uppeak traffic; downpeak traffic; random interfloor traffic and lunch time traffic [1, 13]. The uppeak traffic pattern arises when all passengers are moving up from the main terminal floor, it occurs in considerable strength in the morning when prospective elevator passengers enter a building intent on travelling to destinations on the upper floors of the building. The downpeak traffic situation is observed when the dominant or only traffic flow is in a downward direction with all, or the majority of, passengers leaving the elevator system at the main terminal of the building. The random interfloor traffic is a characterization in which passengers are moving equally likely between floors, it exists for the majority of the working day in office buildings. Lunch time traffic occurs in the middle of the day and exhibits a dominant traffic flow to and from one or more specific floors, one of which may be the main terminal. Based on years of research, the 5 min. interval for traffic flow data collection has achieved general acceptance [13]-[15], on which the traffic data in a day are obtained at the main terminal every 5 min. interval along time axis under the two main traffic patrterns: uppeak traffic condition and downpeak traffic condition, and the traffic flow time series will be constituted, then the traffic demand forecasting models are constructed to forecast the traffic demand at the specified period of time in the future on the observed historical traffic data.

2.2 The Traditional Grey Model GM (1, 1)

The key technology of grey theory is the grey model which takes the grey generation function as its foundation, and differential fitting as the core. The grey theory shows that all the random variables are the gray variables and gray processes vary during a certain period of time within a certain range. Grey forecasting method neither emploies directly the raw data to model, nor finds the statistical laws and probability distribution of stochastic variables, but the generation operation on the raw sequence data is conducted to get new sequences with strong regularity in order to diminish the randomness and volatility of the raw data. Modeling and forecasting is based on the new sequences [14]. Utilizing the grey forecasting method to forecast elevator traffic demand needs neither to determine whether the passenger flow obeys a certain probability distribution, nor require a large number of observed samples.

1) Accumulating Generation Generator (AGO)

The sequence matrix of the raw data is constructed as follows.

$$X^{(0)} = \begin{bmatrix} X_1^{(0)} \\ X_2^{(0)} \\ \vdots \\ X_n^{(0)} \end{bmatrix} \tag{1}$$

Applying AGO to $X^{(0)}$ in (1) gives

$$X^{(1)} = \begin{bmatrix} X_1^{(1)} \\ X_2^{(1)} \\ \vdots \\ X_n^{(1)} \end{bmatrix} \tag{2}$$

where $X_i^{(0)} = \left\{ x_i^{(0)}(j) \middle| i = 1, 2, \dots, n, j = 1, 2, \dots m \right\}$, and $X_i^{(1)} = \left\{ x_i^{(1)}(j) \middle| i = 1, 2, \dots, n, j = 1, 2, \dots m \right\}$, $x_i^{(0)}(j)$ is the observed number of passengers who arrive at the main terminal and go to the upper floors or who come from the upper floors to the main terminal by elevators within the ith Δt for the jth day and

$$x_i^{(1)}(j) = \sum_{k=1}^{j} x_i^{(0)}(k),\ i = 1, 2, \dots, n,\ \ j = 1, 2, \dots, m.$$

2) Modelling of Grey Multi- Model Method

This model is a time series forecasting model. The generated adjacent neighbor mean sequence $Z^{(1)}$ of $X^{(1)}$ is defined as follows

$$Z^{(1)} = \left[z_i^{(1)}(j) \right]_{n\times m} \qquad (3)$$

where $z_i^{(1)}(j)$ is the mean value of adjacent data, i.e. $z_i^{(1)}(j) = \alpha x_i^{(1)}(j-1) + (1-\alpha) x_i^{(1)}(j)$, i=1,2,…n, $j = 2, 3, \dots, m$, $Z_i^{(1)} = \left\{ z_i^{(1)}(j) \middle| j = 2, 3, \dots m, i = 1, 2, \dots, n \right\}$, usually α is set as 0.5 [10]. $x_i^{(1)}(j)$ approximately follows the exponential rule, in terms of $x_i^{(1)}(j)$, the whitenization differential equationcan is constructed as follows

$$\frac{dX_i^{(1)}}{dt} + a_i X_i^{(1)} = u_i \qquad (4)$$

Discretizing Eq. (4) yields grey differential equation as shown below

$$X_i^{(1)} - a_i Z_i^{(1)} = u_i \qquad (5)$$

Then the sequence parameters to be identified are $\mathbf{a}_i = \left[a_i\ u_i\ \right]^T$, where a_i is developing coefficient, u_i is grey influencing coefficient.

3) Find Parameters a_i and u_i

Applying the least squares method yieldsis

$$\mathbf{a}_i = \left[a_i\ u_i\ \right]^T = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{Y}_n \qquad (6)$$

where

$$\mathbf{B} = \begin{bmatrix} -z_i^{(1)}(2) & 1 \\ -z_i^{(1)}(3) & 1 \\ \vdots & \\ -z_i^{(1)}(m) & 1 \end{bmatrix},\ \mathbf{Y}_n = \begin{bmatrix} x_i^{(1)}(2) \\ x_i^{(1)}(3) \\ \vdots \\ x_i^{(1)}(m) \end{bmatrix},\ i=1,2,\dots n.$$

4) Forecasting Results

According to Eq. (4), the solution of forecasted value $\hat{x}_i^{(1)}(j)$ of $x_i^{(1)}(j)$ is given as

$$\hat{x}_i^{(1)}(j+1) = \left[x_i^{(0)}(1) - \frac{u_i}{a_i} \right] \times e^{-a_i \times j} + \frac{u_i}{a_i} \qquad (7)$$

$$i = 0, 1,\ 2, \dots, n,\ j = 0, 1,\ 2, \dots, m$$

To obtain the forecasted value $\hat{x}_i^{(0)}(j)$ of the primitive data $x_i^{(0)}(j)$, the IAGO is used to establish the following grey model

$$\hat{x}_i^{(0)}\left(j+1\right) = \left(1 - e^{-a_i}\right) \times \left[x_i^{(0)}(1) - \frac{u_i}{a_i} \right] \times e^{-a_i \times j} \qquad (8)$$

$$i = 1,\ 2, \dots, n,\ j = 0, 1,\ 2, \dots, m$$

The above shows that each observation period Δt which is set 5min corresponds to a forecasting model, therfore the method proposed here is called multi-model method.

5) Limitations of the Traditional Grey Model in Practical Engineering Application

As stated earlier, grey forecasting is suitable for the situations in which the difficulty of incomplete or insufficient information is faced [16]. Since it can well reflecte the past state and future variation tendency, it has been widely used especially in short-term forecast. But like other methods, traditional grey forecasting method has its limitations [17]-[20], some of which are summarized as follows:

1. As the discrete degree of data increases, viz. the gray level increases, the forecasting accuracy becomes worse.

2. The forecasting result will be better, for a completely exponential growth sequence, and it is difficult to consider the situations where there exists outliers in the sequence, especially when the sample data growth deviates from

exponential rule, the forecasting accuracy will be getting worse.

3. Because the knowledge and rule are discovered on the raw sequence data taking no account or lacking of qualitative considerations and quantitative analysis for the objective factors which may impact the forecasting effect. Therefore, it is inferior in flexibility of application.

4. The growth rate of the raw data sequence $\lambda = \frac{dX_i^{(0)}}{dt} \Big/ X_i^{(0)} = \frac{d^2 X_i^{(1)}}{dt^2} = -a_i$ is a constant, but in general it is difficult to meet such implicit requirement. A slight change of a_i has little influence on forecasting accuracy, but a greater change of a_i will lead to the deterioration of forecasting accuracy.

5. Taking the primitive value $x^{(0)}(1)$ as the initial condition of the forecasting model casues that $\hat{x}^{(0)}(1)$ is irrelevant to the initial value of the raw data sequence. The information about $x^{(0)}(1)$ is overlooked, in this case, it is difficult to guarantee the minimum of the entire forecasting error.

Because of the reasons stated avove, to improve traditional grey forcasting method to increase forecasting accuracy concerning the specific engineering applications is of great significance.

3. Elevator Traffic Demand Forecasting for EGCS

3.1 The Processing of the Raw Data Sequence

The movement of people around a building is very complex, there is unpredictability in human behaviour. Due to some accidental or exceptional circumstances caused by the human occasionality and subjectivity, great random fluctuations of passenger traffic flow may occur during some specific time periods, which results in the outliers in the raw data sequence. Aside from the outliers processing, smoothing of the raw data sequence is needed for better forecasting results.

1) Outliers Discrimination

The outliers neither reflect nor represent the overall traffic characteristics, and will make the forecasting accuracy seriously deteriorate. Such outliers must be modified to refine the traffic models.

for all $\left\{x_i^{(0)}(j-1), x_i^{(0)}(i), x_i^{(0)}(j+1)\right\} \subset X_i^{(0)}$

If $x_i^{(0)}(i) > \gamma_1 \times \left(x_i^{(0)}(j-1) + x_i^{(0)}(j+1)\right)$

or $x_i^{(0)}(j) < \gamma_2 \times \left(x_i^{(0)}(j-1) + x_i^{(0)}(j+1)\right)$

Then $x_i^{(0)}(j)$ is called outliers, where $i = 1, 2, \ldots, n,$ $j = 2, 3 \ldots m,$ γ_1, γ_2 satisfies $\gamma_1 > 1$, $0 < \gamma_2 < 1,$ the value of γ_1, γ_2 are determined by the specific circumstance.

2) Outliers Modification

The outliers should firstly be removed, and the related data point in $X_i^{(0)}$ will be empty, then empty data positions in $X_i^{(0)}$ will be updated with the new data which are generated by non-adjacent neighbor mean generation method given as follows.

$$x_i^{(0)}(j) = 0.5 \times x_k^{(0)}(j-1) + 0.5 \times x_i^{(0)}(j+1) \qquad (9)$$

3) Smoothing Processing of the Raw Data Sequence

If the randomness of the raw data sequence is somehow smoothed, it will be easier to model and to forecast the expected performance of the system. Smoothing processing is to transform the raw data into an increasing sequence, which is intended for futher smoothing the randomness of the raw data sequence based on the above step 2).
for $2 \le j \le m-1$

$$x_i'^{(0)}(i) = 0.25 \times \left[x_i^{(0)}(j-1) + 2x_i^{(0)}(j) + x_i^{(0)}(j+1)\right] \qquad (10)$$

and for $j = 1,\ m$

$$\begin{cases} x_i'^{(0)}(1) = 0.25 \times \left[3x_i^{(0)}(1) + x_i^{(0)}(2)\right] \\ x_i'^{(0)}(m) = 0.25 \times \left[x_i^{(0)}(m-1) + 3x_i^{(0)}(m)\right] \end{cases} \qquad (11)$$

The above processing not only increases the weights of the current data, but avoid the value fluctuating excessively, and slowing the changing rate of the primitive sequence which tends to grew quickly to make the randomness of the new data sequence weaker than that of the raw data sequence. In this way, the application scope of the traditional grey model is expanded.

4) Elevator Traffic Demand Forcasting Results

Let the forcasting output of the accumulated sequence by the processed raw data sequence be

$$\hat{X}^{(1)} = \left\{\hat{x}_i^{(1)}(m+1),\ \hat{x}_i^{(1)}(m+2), \dots \ \middle|\ i = 1, 2, \dots, n\right\} \qquad (12)$$

where $\hat{x}_i^{(1)}(m+1), \hat{x}_i^{(1)}(m+2), i = 1, 2, \dots, n$, are respectively the forcasting output of the accumulated sequenc of the (m+1)th day, (m+2)th day, …. Employing IAGO to $\hat{X}^{(1)}$ yields the elevator traffic demand forcasting results

$$\hat{x}_i^{(0)}(m+1) = \hat{x}_i^{(1)}(m+1) - \hat{x}_i^{(1)}(m) \qquad (13)$$

5) Model Accuracy Testing

The model accuracy is tested by error of residuals to check whether the the relative error meets the given requirements. Let the residual sequence be

$$E_i^{(0)} = \left\{e_i^{(0)}(j) \middle| i = 1, 2, \dots, n,\ j = 1, 2, \dots m\right\} \qquad (14)$$

where $e_i^{(0)}(j) = x_i^{(0)}(j) - \hat{x}_i^{(0)}(j)$

Define $\overline{\varepsilon}_i = \frac{1}{m}\sum_{j=1}^{m}\left|e_i^{(0)}(j)/x_i^{(0)}(j)\right| \times 100\%$ as the average relative error of model *i*, and $p^0 = (1-\overline{\varepsilon}_i)\times 100\%$ is referred to as the accuracy of model *i*, which is the generally accepted criteria for evaluating the model evaluation grade whose values are set as shown in Table 1.

Table 1: Criteria for evaluating the model

Accuracy of model p^0	***Model evaluation grade***
$p^0 \leqslant 90\%$	excellent
$80\% \leqslant p^0 < 90\%$	good
$p^0 < 80\%$	poor or unacceptable

4. Simulation Experiments and Analysis

The simulation data are from an office block with a dining-hall in the first floor, the legal working time of the people in the block is from 7 :00 in the morning to 19:00 in the evening from Mon. to Fri. The time interval for observing traffic is Δt =5min, observations were made during 5 days from Mon. to Fri. between 7:00 and 19:00, 144 data were collected in a day. The observed data from Mon. to Thur. form the raw data sequence, and the Fri.'s observed data are taken as the comparison sample. i.e. using the raw data sequence from Mon. to Thur. to forecast the Fri.'s elevator traffic demand. Here γ_1, γ_2 are set as 2, 0.5 in up traffic and 2.4, 0.6 for down traffic, respectively. The accuracies of the models of the proposed method are shown in Table 2 indicating that both under up- and down traffic all the models meet the requirement of gry forcasting [10], where there are 132 models are with "excellent" level and 12 with "good" level for up traffic, and 134 with "excellent" level, and 10 with "good" level for down traffic.

Table 2: Model accuracy

Traffic patterns	***Model accuracy***	***Number of models***	***Model evaluation***
Up traffic	[0.8, 0.9)	12	Good
	[0.9, 1)	132	Excellent
Down traffic	[0.8, 0.9)	10	Good
	[0.9, 1)	134	Excellent

The simulation results are shown in Figure 1~ Figure 6, where Figure 1 and Figure 2, show the elevator traffic demand forecasting results of Fri., applying the raw data sequence from Mon. to Thur., compared with the obsvered data sequence on Fri. under up and down traffic patterns expectively. In order to compare the proposed method in this thesis with the traditional method, two types of error which are the mean square relative error (MSRE) and the mean relative error (MRE) are suggested. The forcasting err curves shown in Figure 3 and Figure 4 demonstrate that the proposed method performs better results than the traditional gry forcasting method under both up and down traffic in terms of forecasting errs as shown in Figure 5, Figure 6, whose MSRE and MRE are given by Table 2. The results in Table 2 suggest that both MSRE and MRE of the proposed method are much less than that of the traditional method, the reason for which is that there are some outliers in the raw data sequence created by human subjectivity and occasionality, which make the raw data sequence seriously deviate from the exponential growth rule resulting greater errors. After smoothing the randomness, the raw data become smoother, consequently, the forecasting accuracy is improved.

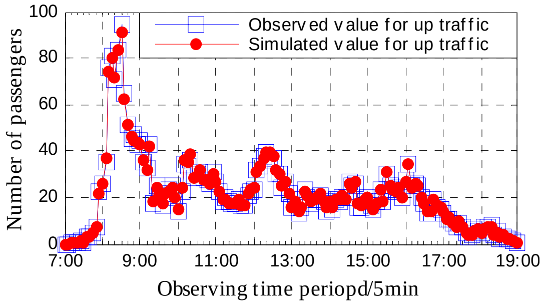


Fig.1. Simulated result for up traffic on friday by improved method

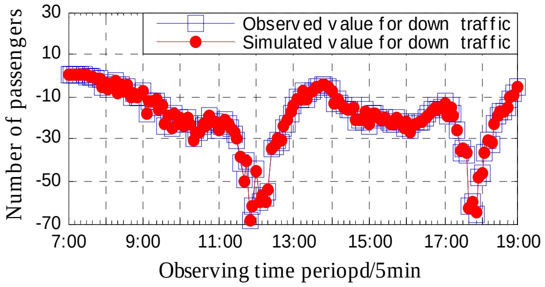


Fig.2. Simulated result for down traffic on friday by proposed method

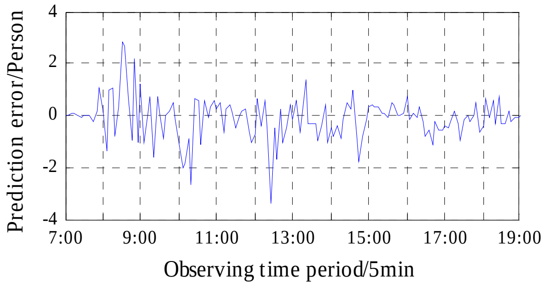


Fig.3. Forecasting error for up traffic by improved method

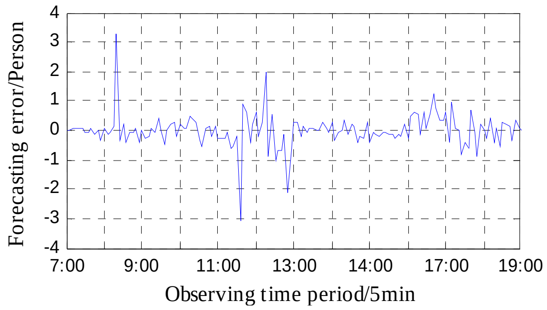


Fig.4. Forecasting error for down traffic by improved method

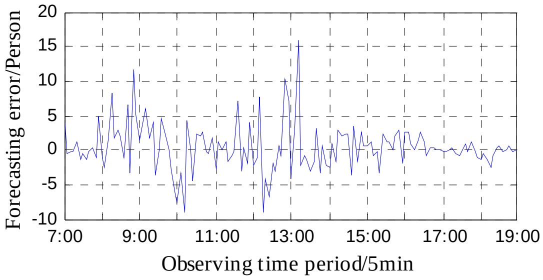


Fig.5. Forecasting error for up traffic by traditional method

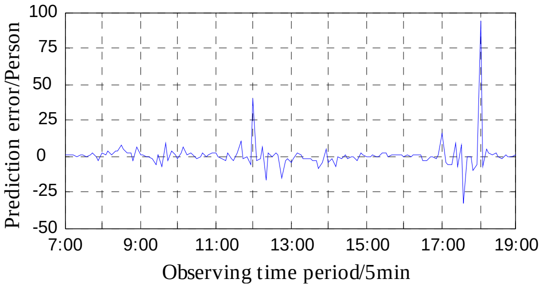


Fig.6. Forecasting error for down traffic by traditional method

5. Conclusions

In reguard to the strong grey characteristics of elevator traffic demand, a grey thoery based multi- model method to forecast the elevator traffic demand for EGCS is proposed in this thesis. Taking full advantage of the grey forecasting, the method is suited to the EGCS with limited number of samples and the incomplete information, on the other hand, for futher smoothing the randomness of the raw data sequence and weakening the influence of abnormal information caused by human subjectivity and occasionality, the raw data sequence is processed by comparing real time traffic data with the historical ones to modify the outliers to refine the traffic models, which effectively decrease the model error of the gry forcasting to reduce the impact of randomness on accuracy of modeling shch that the forecasting errs can be controled within the acceptable range for engineering problems, and consequently, making up for the deficiency of the grey forecasting method in theory. The simulation experiment indicates that the proposed method performs better results than the traditional gry forcasting method in forecasting auucracy, quite proximating to practical situation. The proposed method needs limited number of samples, especialy, it is suitable for short-term prediction of elevator traffic demand, providing the essential prerequisite for effectively implementating the EGCS.

Acknowledgments

This research was partially funded by the National Science Foundation of China (69874026) and the Doctoral Scientific Research Starting Foundation of Bohai university (BSQD2010-005)

References

[1] G. C. Barney, Elevator Traffic Handbook, London, Spon Press, 2003.

[2] Zhenshan Yang, Cheng Shao, and Guizhi Li, “Multi-Objective Optimization for EGCS Using Improved PSO Algorithm”, in the American Control Conference (ACC’07), 2007, Vol. 10, pp. 5059-5063.

[3] Zhenshan Yang, and Zunli Zhang, “A Simulation Based Verification Method for Elevator Traffic Planning”, in the IEEE International Conference on Computer Application and System Modeling (ICCASM 2010), 2010, pp. 572-576.

[4] [4] Anna Shang, “Elevator Traffic Flow Prediction Methods”, Hunan Agricultural Machinery (in Chinese), Vol. 37, No. 3, 2010, pp. 18-19.

[5] Yunli Zhang, and Zhenshan Yang, “Fuzzy logic controller for elevator traffic scheduling”, Lift Report, No. 1, 2009, pp. 64, 66-68.

[6] [6] Lixia Ji, Xiaogang Fu, “The Application of LS-SVM to the Prediction of The Elevator Transportation Flow”, Journal of Shanghai Dianji University, Vol. 9. No. 3, 2006, pp. 62-64.

[7] Shuguo Li, Lei Zhang, and Yang Bai, “Elevator System Applying Volume Estimating Method Based on The Data Mergence”, China Elevator, Vol. 22. No. 13, 2006, pp. 19-31.

[8] Haiyan Tang, Deliang Yu, Bao Ding, et al., Time Series Prediction of Elevator Traffic Flow Based on SVR”, Control Engineering of China, Vol. 18, No. 5, 2011, pp 723-726,792.

[9] Zhenshan Yang, “Traffic Regularity Evaluation for Elevator Vertical Traffic System”, Advanced Materials Research, Vol. 374-377, 2012, pp 2301-2304.

[10] Sifeng Liu and Yi Lin, Grey Systems, Theory and Applications. Berlin Heidelberg , Germany: Springer-Verlag Press, 2010.

[11] Xiaoxuan Zhang, “The Essential of GM(1, 1) Model”, Journal of Grey System, Vol. 10. No. 2, 2007, pp. 81- 87.

[12] FR Johnston, JE Boylan, and EA Shale, “An Emamination of the Size of Orders From Customers, Their Characterisation and The Implications for Inventory Control of Slow Moving Items”, Journal of the Operational Research Society, Vol. 54, No. 8, 2003, pp. 833-837.

[13] G. C. Barney, S M. Dos Santos, Elevator traffic analysis：Design and control. London：Peter Peregrinus Ltd, 1985.

[14] Meng, Lu, Kees Wevers, “Grey system theory and applications: A way forward”, Journal of Grey System, Vol. 10. No. 1, 2007, pp. 47- 53.

[15] CIBSE Guide D: Transportation Systems in Buildings, Norwich Norfork: Page Bros, 2005.

[16] Erdal Kayacan, Baris Ulutas, and Okyay Kaynak, Grey system theory-based models in time series prediction, Expert Systems with Applications, Vol. 37, No. 2, 2010, pp. 1784–1789

[17] D. Zhang, S. Zhang, and K. Shi, “Theoretical Defect of Grey Prediction Formula and Its Improvemen”, Systems Engineering -Theory & Practice, Vol. 12, No. 8, 2002, pp, 140-142.

[18] Xin Zhang, Gang Wei, Min Zhou, and Yijuan Yang, “Application of grey theory in forecasting City annual electricity consumption”, Journal of Shanghai University of Electric Power, Vol. 18, No. 2, 2002, pp. 9-12.

[19] Xiping Wang, “Grey Prediction with Rolling Mechanism for Electricity Demand Forecasting of Shanghai”, in the IEEE International Conference on Grey Systems and Intelligent Services, 2007, pp. 689-692．

[20] Huaizhu Shu, Summarize on the Forecast models of road traffic accident based on grey theory, Road Traffic & Safety, Vol. 9, No. 6, 2009, pp. 25-27.

Zhenshan Yang received the B.S. degree in industrial electrical automation from Shenyang Jianzhu University, in 1987. He received the Ph.D. degree in control theory & control engineering from Dalian University of Technology, in 2008. He is a committee member of China Electrotechnology Society (CES), the Deputy Secretary-General of the Affiliated Society of CES of Liaoning Province and the editorial board member of the *International Journal of Urban Planning and Design Research* (*UPDR*). Currently, he is a professor at Bohai University. His research interests include intelligent building, elevator traffic analysis, design, and intelligent Control.

Yunli Zhang received the B.S. degree in applied mathematics from Northeastern University, in 1987, and the master's degree in computer technology and application from Dalian University of Technology. Currently, she is a Professor at Liaoning Medical University, Jinzhou, China. His interests are in data mining and its application in intelligent control systems.

The Research on Search Algorithms in the Machine Learning

Hui Liu[1],Yonghui Cao[2]
[1] School of Computer and Information Technology, Henan Normal University, Xin Xiang, 453007, China

[2] School of Economics & Management, Henan Institute of Science and Technology, Xin Xiang, 453003 ,China

Abstract

Machine learning is the estimation of the topology (links) of the network, it can be achieved by utilizing a search algorithm through the possible network structures, because it is finding the best network that fits the available data and is optimally complex. In this paper, a greater importance is given to the search algorithm because we have assumed that the data will be complete. We focus on Two search algorithms are introduced to learn the structure of a Bayesian network in the paper. The heuristic search algorithm is simple and explores a limited number of network structures. On the other hand, the exhaustive search algorithm is complex and explores many possible network structures.

Keywords: *Structural learning, Search Algorithms, Heuristic Search, Exhaustive Search*

1. Introduction

A Bayesian network, Bayes network, belief network or directed acyclic graphical model is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.

Formally, Bayesian networks are directed acyclic graphs whose nodes represent random variables in the Bayesian sense: they may be observable quantities, latent variables, unknown parameters or hypotheses. Edges represent conditional dependencies; nodes which are not connected represent variables which are conditionally independent of each other. Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and gives the probability of the variable represented by the node. For example, if the parents are m Boolean variables then the probability function could be represented by a table of 2m entries, one entry for each of the 2m possible combinations of its parents being true or false.

Efficient algorithms exist that perform inference and learning in Bayesian networks. Bayesian networks that model sequences of variables (e.g. speech signals or protein sequences) are called dynamic Bayesian networks. Generalizations of Bayesian networks that can represent and solve decision problems under uncertainty are called influence diagrams.

Because a Bayesian network is a complete model for the variables and their relationships, it can be used to answer probabilistic queries about them. For example, the network can be used to find out updated knowledge of the state of a subset of variables when other variables (the evidence variables) are observed. This process of computing the posterior distribution of variables given evidence is called probabilistic inference. The posterior gives a universal sufficient statistic for detection applications, when one wants to choose values for the variable subset which minimize some expected loss function, for instance the probability of decision error. A Bayesian network can thus be considered a mechanism for automatically applying Bayes' theorem to complex problems.

A Bayesian network is not allowed to have a cycle because of the computational difficulties. A cycle in a Bayesian network leads to a "circular reasoning" between the variables. For example, if the dependencies in above network are: $X_1 \rightarrow X_2$, $X_2 \rightarrow X_3$, and $X_2 \rightarrow X_3$, a cycle will be formed. If evidence is entered into the variable X_1, the Bayesian network will run the evidence to X_2, then to X_3 . Then, the evidence will travel to X_1 because X_1 depends on X_3 . The evidence may run in the network forever because all the variables depend on each other in a circular way.

A heuristic arc addition is employed not to have a cycle in the Bayesian network while generating the Bayesian structure. An exhaustive arc addition is also employed to explore more network possibilities without limitation. In the exhaustive arc addition algorithm, a cycle check is employed before and arc is added. The following section presents the details of heuristic and exhaustive search algorithms.

2. Heuristic Search

In the heuristic search algorithm, the variables of the system have to be ordered in a certain way to prevent cycles from being created. The decision variables should be in the last columns in the database; and, the first columns of the database should be filled with the variables without parents, independent variables. After placing the independent variables in the first columns, the children of the independent variables should be placed in the following columns. The rest of the columns are filled with the children of the previously placed variables. Ordering of the variables is necessary because the heuristic arc addition adds the arcs from the first variables to the last variables. Because of the ordering, we need to have some knowledge about the variables. This does not mean that we need to know the dependencies between the variables. For example, let B be a Bayesian network with three variables, $\{X_1, X_2, X_3\}$.If we know the variable X_1 is the first variable and the variable X_2 is the decision node. Then the column order will be $\{X_1, X_2, X_3\}$.

The heuristic search starts with adding and removing arcs from the each variable to the last variable. Let the network have n variables. After adding an arc, the algorithm calculates the network score, records the score in a list, and removes the arc. The algorithm finds the arc that gives the highest increase in the network score. Let us assume that the arc from the kth variable to the last variable, n, gives the highest increase in network score. Then, the algorithm adds the arc from the kth variable to the last variable. After the arc is added, the algorithm adds and removes arcs from the remaining variables to the last variable. Then, the algorithm chooses the arc with the highest score increase and adds the arc to the network. This continues until no increase in the network score can be obtained by adding an arc to the last variable. Then, the algorithm starts adding arcs from the variables $\{1, 2, \cdots n-2\}$ to the $(n-1)th$ node. The algorithm adds arcs to $(n-1)th$ node until there is no increase in the network score. The algorithm stops when it adds an arc from the first variable to the second variable. The following is the heuristic search algorithm used in this research.

(1)Collect data
(2)Define the variables from the available data
(3)Start with a network with no arc.
(4)Estimate the parameters (only independent probabilities) of the BN using the MLE method using initial data
(5)Add a new arc from the ith variable to the jth variable to generate a network candidate and remove the arc. Repeat the process with $i = \{1, 2, \cdots j-1\}$ and generate networks $(B_1, B_2, \cdots, B_{j-1})$. Start j from n and decrease j by 1.
(6)Calculate the scores of the candidate networks and record them in a list.
(7)Find the network (B) with the maximum score and keep it for the next step.
(8)Repeat the steps 5, 6, and 7 until there is no increase in the network score.
(9)If $j > 1$, then go to step 5.
(10)Update the network parameters along with new data
(11)Update the network structure:

If enough new data obtained, go to step 1 and generate a new network structure.

If no structural update is necessary go to step 10.

Consequently, the heuristic search algorithm adds arcs only in the forward direction because this protects the network from having cycles and complex network structure. On the other hand, there is a price of arranging the variables at the creation of the database in the heuristic algorithm. Since the agents will not have much knowledge about the environmental variables, it is hard to arrange the variables at the beginning. There is a need for a better search algorithm that explores more possibilities in the network. The following paragraph introduces another searching algorithm that eliminates the arranging the variables, namely exhaustive search.

3. Exhaustive Search

The exhaustive search algorithm explores all the possible arcs in the network during its execution. The algorithm starts adding arcs from the ith variable to the jth variable where $i = \{1, 2, \cdots, n\}$, $j = \{1, 2, \cdots, n\}$, $i \neq j$. This covers $n \cdot (n-1)$ arcs throughout the network. The algorithm calculates the network score for each arc addition. Then, it chooses the arc with the highest increase in the network score. The algorithm repeats the above steps until there is no increase in the network score.

There are two major drawbacks in the exhaustive search algorithm. First, the number of arcs to be tried might become intractable when the number of variables is large. Second, during the search, the algorithm might introduce cycles to the network because it can add an arc in any direction. An additional algorithm is incorporated to the search algorithm to keep track of cycles. Using the additional algorithm, the search algorithm checks whether the new arc introduces a cycle or not. If the arc introduces a cycle, the algorithm does not add the arc to the network. The following is the exhaustive search algorithm used in this research.

(1)Collect data
(2)Define the variables from the available data
(3)Start with an empty network
(4)Estimate the parameters (only independent probabilities) of the BN using the MLE method using initial data
(5)Add a new arc from the ith variable to the jth variable to create a candidate network and remove the arc. Repeat the process for every value of i and j where $i=\{1,2,\cdots,n\}$, $j=\{1,2,\cdots,n\}$, $i \neq j$ .This step creates m possible networks $(B_1,B_2,\cdots,B_m)$. Algorithm creates $m=n\cdot(n-1)$ networks in first visit to step 5.
(6)Remove the network with cycles from the candidate fist.
(7)Calculate the scores of the candidate networks and record it in a list.
(8)Find the network (B) with the maximum score and keep it for the next step.
(9)Do step 5 through 8 until there is no increase in the network score.
(10)Update the network parameters along with new data
(11)Update the network structure:
If enough new data obtained, go to step 1 and generate a new network structure.
If no structural update is necessary go to step 10.
The search algorithms are explained in detail. There is a need to analyze the complexity of the search algorithm before there are implemented. The following section gives the complexity analysis of both search algorithms.

4. Complexity Analysis for Search Algorithms

As stated earlier, the heuristic search algorithm needs prior knowledge about the variables in terms of their order in the database. On the other hand, the number of iterations in the heuristic search algorithm may be tractable. In the heuristic search, the algorithm tries $(n-1)$ arcs in the first trip from step 5 to step 7. The algorithm repeats steps 5 through 7 until there is no increase in the network score. Assuming the algorithm adds an arc in every trip, the number of arcs tried will be one less then the previous trip. Algorithm can repeat step 5 through 7 at most $(n-1)$ times. In $(n-1)$ trips, the algorithm generates $(n-1)+(n-2)+\cdots+1$ networks candidates. When the algorithm reaches step 8, the algorithm loops back to step 5 and repeats the same process for the variables $\{X_{n-1},X_{n-2},\cdots,X_2\}$ .Therefore, after the first loop, the algorithm generates $(n-1)+(n-2)+\cdots+1$ network candidates. The complexity of the heuristic search algorithm is denoted as C_h .

In the following complexity analysis, each loop shows the number of network candidates tried until the algorithm reaches to the step 8. Since the algorithm will repeat itself for $(n-1)$ variables, the analysis has $(n-1)$ loops as the following.

Loop1

$$(n-1)+(n-2)+\cdots+1=n(n-1)-(1+2+\cdots+(n-1))$$
$$=n(n-1)-\frac{n(n-1)}{2}=\frac{n(n-1)}{2}$$

Loop2

$$(n-2)+(n-3)+\cdots+1=\frac{(n-1)(n-2)}{2}$$

$\vdots$
$\vdots$

Loop (n-1) $\frac{(n-(n-1))(n-(n-2))}{2}=1$

If we add the number of candidate networks from each loop, the following can be obtained:

$$C_n=\frac{n(n-1)+(n-1)(n-2)+\cdots+(n-(n-1))(n-(n-2))}{2}$$
$$C_n=\frac{2(n-1)^2+2(n-3)^2+\cdots+2(n-(n-2))^2}{2}$$

Then, we can further modify the equation as follows:

$$C_n = (n-1)^2 + (n-3)^2 + \cdots + 2(n-(n-2))^2 \quad (1)$$

Since each element in C_n is less than n^2. We can state that $C_n < n^2(n-3) < n^3$ (2)

Equation (2) illustrates the complexity of the heuristic search. The following paragraphs will explore the complexity of the exhaustive search algorithm.

The exhaustive search algorithm tries every possible arc in the network during its first visit to step 5. In a graph with n nodes, there can be $n(n-1)$ possible directed edges. Therefore, the algorithm generates $n(n-1)$ network candidates and the complexity of the first visit is $n(n-1)$.Then the algorithm continues until it reaches to step 9 and loops back to step 5 until there is no increase in the network score.

After the first loop, the complexity decreases by 1 in each step because the algorithm will not try the arc added in the previous step. The following presents the complexity analysis of the exhaustive search algorithm. First, the complexity is calculated for each loop. Then, they are added to obtain the complexity of the algorithm.

Loop 1	$n(n-1)$
Loop 2	$n(n-1)-1$
⋮	
⋮	
Loop N	$n(n-1)-N+1$

The exhaustive search algorithm does not perform a certain number of loops. The algorithm will continue until there is no increase in the network score. Therefore, we will assume that the algorithm end after N loops for the complexity calculations. If we add the complexities of all the loops together, the complexity of the exhaustive search, C_e, becomes the following.

$$C_e = n(n-1)N - (1+2+\cdots+(N-1)) \quad (3)$$

$$C_e = n(n-1)N - \frac{N(N-1)}{2} \quad (4)$$

If the network has great number of arcs, then the complexity of the algorithm becomes large. For example, if the algorithm ends in step $N=n$, the complexity becomes

$$C_e = n^2(n-1) - \frac{n(n-1)}{2} = \frac{2n^2(n-1) - n(n-1)}{2} \quad (5)$$

$$C_e = \frac{(n-1)n(2n-1)}{2} \quad \text{for } n = N \quad (6)$$

In general, number of nodes in a Bayesian network, n is much larger than 1. Therefore, we can reevaluate the complexity by assuming $n\rangle\rangle 1$ .The following equation represents the computational complexity of the exhaustive search algorithm when the number of steps is equal to the number of variables.

$$C_e \cong \frac{n \cdot n \cdot 2n}{2} = \frac{2n^3}{2} \Rightarrow C_e \cong n^3 \quad (7)$$

As can be seen above, the complexity of the exhaustive algorithm is larger than the complexity of the heuristic algorithm when $N = n$.

For the networks with large number of variables (nodes), the algorithm does not stop when $N = n$ .Let us calculate the worst case scenario for the exhaustive algorithm. The algorithm might explore all possible arcs in the network, which is equal to $n(n-1)$ .This is true because a complete graph with n nodes has $n(n-1)$ possible directed edges [30]. Therefore, we will replace N with $n(n-1)$ in the complexity analysis. Then, the complexity of the exhaustive search algorithm becomes the following.

$$C_e = n(n-1)N - \frac{N(N-1)}{2} = n(n-1)n(n-1) - \frac{n(n-1)(n(n-1)-1)}{2}$$

(8)

$$C_e = \frac{2n^2(n-1)^2 - n^2(n-1)^2 - n(n-1)}{2} = \frac{n^2(n-1)^2 - n(n-1)}{2} \quad (9)$$

We can simplify the equation above by assuming n >>1. In this case, the complexity of the algorithm becomes the following.

$$C_e \cong \frac{n^2 \cdot n^2 - n^2}{2} = \frac{n^2(n-1)^2}{2} \Rightarrow C_e \cong \frac{n^4}{2} \quad (10)$$

5. CONCLUSION

Bayesian network is a complete model for the variables and their relationships.Two search algorithms are introduced to learn the structure of a Bayesian network.
The heuristic search algorithm is simple and explores a limited number of network structures. the heuristic search algorithm adds arcs only in the forward direction because this protects the network from having cycles and complex network structure. there is a price of arranging the variables at the creation of the database in the heuristic algorithm. Since the agents will not have much knowledge about the environmental variables. There is a need for a better search algorithm that explores more possibilities in the network.

The exhaustive search algorithm is complex and explores many possible network structures. The heuristic search algorithm needs prior knowledge about the variables in terms of their order in the database. If the network has great number of arcs, then the complexity of the algorithm becomes large. The complexity of the exhaustive algorithm is approximately n-fold larger than the complexity of the heuristic search algorithm.

Acknowledgments

This work is financially supported by Science and Technology research projects by the Education Department of Henan province (Project No. 12A520025). Thanks for the help.

References

[1]F.V. Jensen, an Introduction to Bayesian Networks. London, UK: University College London Press, 2012.

[2]Ali Jarrahi and Mohammad Reza Kangavari, "An Architecture for Context-Aware Knowledge Flow Management Systems," International Journal of Computer Science Issues, vol. 9,2012.

[3]Y.Shoham, "Agent-oriented programming," Artificial intelligence, vol. 60(1), pp. 51-92, 2012 .

[4]J.Pearl, "Bayesian networks", in M. Arbib (Ed.), Handbook of Brain Theory and Neural Networks, MTT Press, pp. 149-153, 2012

[5]J.Pearl, "Bayesian networks," Technical Report R-246, MTT Encyclopedia of the Cognitive Science, October ,2011.

[6]F.V. Jensen, "Bayesian network basics," AISB Quarterly, vol. 94, pp. 9-22, 2011.

[7]W.Lam and F. Bacchus, "Learning Bayesian belief networks: an approach based on the MDL principle," Computational Intelligence, vol. 10, pp. 269-293, 2011.

[8]N.Friedman, M. Goldszmidt, D. Heckerman, and S. Russell, "Challenge: Where is the impact of the Bayesian networks in learning?" In Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI), pp.10-15, 2011.

[9]N.Friedman, K. Murphy, and S. Russell, "Learning the structure of dynamic probabilistic networks," in G.F. Cooper and S. Moral (Eds.), Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98), San Francisco, CA: Morgan Kaufmann, 2010.

[10]Amir Mosavi, "Multiple Criteria Decision-Making Preprocessing Using Data Mining Tools," International Journal of Computer Science Issues, vol. 7,2010.

[11]B.Theisson, C. Meek, and D. M. Chickering, and D. Heckerman, "Learning mixtures of Bayesian networks," in G.F. Cooper and S. Moral (Eds.), Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98), San Francisco, CA: Morgan Kaufmann, 2010.

[12]N.Friedman, "The Bayesian structural EM algorithm," in G.F. Cooper and S. Moral (Eds.), Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98), San Francisco, CA: Morgan Kaufmann, 2010.

[13]D.Spiegelhalter, P. Dawid, S. L. Lauritzen, and R. Cowell, "Bayesian analysis in expert systems," Statistical Science, vol. 8, pp. 219-282, 2009.

[14]D.Heckerman, D. Gieger, and M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," Technical Report MSR-TR-94-09, Microsoft Research, Redmond, WA, 2009.

[15]C.Claus, "Dynamics of multi-agent reinforcement learning in Cooperative multi-agent systems," Ph.D. Dissertation, Univ. of British Colombia, Canada, 2009.

[16]S. Sen and M. Sekaran, "Multi-agent coordination with learning classifier systems," in Proceedings of the IJCAI Workshop on Adaptation and Learning in Multi-agent Systems, Montreal, pp. 84-89, 2009.

[17]C.Boutilier, "Planning, learning and coordination in multi-agent decision processes," in Sixth conference on Theoretical Aspects of Rationality and Knowledge (TARK'96), The Netherlands, 2008.

First Author Hui Liu received the MS degree in computer education from Henan Normal University in 2008.She is currently a teacher in Henan Normal University. Her research interest is in the areas of machine learning.

Second Author Yonghui Cao received the MS degree in business management from Zhejinag University in 2006. He is currently a doctorate candidate in Zhejiang University. His research interest is in the areas of machine learning.

Framework Design of Secure Cloud Transmission Protocol

Dinesha H A[1] **Prof.V.K Agrawal**[2]

[1] **Assistant Professor, Dept of ISE and CORI, PES Institute of Technology,**
100 Feet Ring Road, Banashankari 3rd Stage, Bangalore-560085, India

[2]**Professor ISE, Director CORI, PES Institute of Technology,**
100 Feet Ring Road, Banashankari 3rd Stage, Bangalore-560085, India

Abstract

Cloud computing technologies are in high demand because of several benefits. Many business organizations are looking into cloud computing services to reduce the cost and complexity of their business infrastructure and its preservation. However, there are certain security issues in cloud computing technologies. To overcome those security issues, we propose secure cloud transmission protocol design. This framework design details will help us in developing a secure protocol for the customers who are using cloud computing technologies over insecure internet. In this paper we discuss: i) Overview model of proposed secure cloud transmission system in internet ii) Security requirements iii) roles and responsibilities of secure transmission protocol in OSI and iv) Framework Design of secure cloud transmission.

Keywords: *Cloud Computing, Protocol, Security, Secure cloud transmission protocol.*

I. Introduction

SINCE entering into 21st century, there has been a rapid boom of computer network development, Information technology is now more and more blended into our daily life at the coming of electronic era. The concept of cloud computing was jointly proposed by Google and IBM in 2007 [1]. Today due to the advancement of technologies and high-speed internet facilities, it is possible to realize cloud computing. Many organizations around the world are providing cloud services. Cloud computing is an internet- based model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) [2]. In internet, cloud computing technology provides four major services such as: i) Software as a Service ii) Data Storage as a Service iii) Platform as a Service and iv) Infrastructure as a Service [3], as shown in Fig. 1. Cloud service activities are upgraded or improved by the cloud service provider based on the customer needs. Our objective is to ensure security while customers are using cloud service over insecure internet.

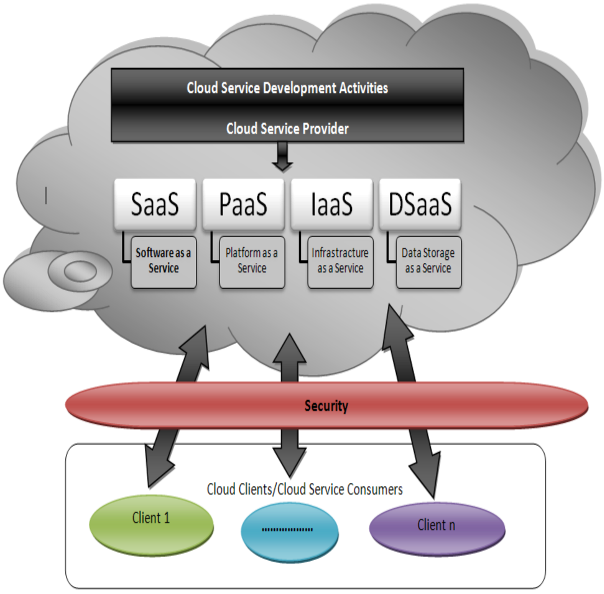

Cloud Service Development Activities
Cloud Service Provider
SaaS
PaaS
IaaS
DSaaS
Software as a Service
Platform as a Service
Infrastructure as a Service
Data Storage as a Service
Security
Cloud Clients/Cloud Service Consumers
Client 1
....
Client n

Fig. 1: Four major category services of cloud computing.

In internet services confidentiality, integrity and availability are the key challenges. The way to access any services over internet is through web browser. Web browsers typically use HTTP's protocols such as HTTP, HTTPS and S-HTTP. HTTP helps to communicate with web servers. In general, in HTTP, sending and receiving information between web server and web browser happens without encrypting messages [4]. However for sensitive transactions, such as Internet e-commerce or online access to financial accounts, the browser and server encrypt this information, referred as HTTPS [5]. HTTPS has been designed to withstand data hacking and provides data confidentiality [5]. HTTPS also facing some of the challenges such as : i) Complex encryption method[5] ii) Browser incompatibility in decrypting messages [5] iii) User needs to wait for long time to get session ends [5] iv) Man-in-the-middle attack [7] v) Eaves dropper attack [4]. Out of these drawbacks, only complex encryption has been addressed in Secured HTTP (SHTTP) [6]. Hence to overcome the rest of the security drawbacks and to establish secure channel, it is necessary to investigate a protocol which sits on top of HTTP and provide secured channel over an insecure internet for cloud transmission, known as secure cloud transmission protocol.

The paper is organized in the following manner: In section II, we brief about the cloud computing background, cloud services and its deployment types. In section III, we present the identified security issues, both in http protocols and cloud computing services. It also presents the requirements, roles in OSI architecture and framework of secure transmission protocol. Section IV concludes this paper along with the future work.

2. Cloud Computing

Cloud computing provides computation, software, data access, and storage services that do not require end-user knowledge of the physical location and configuration of the system that delivers the services [4]. Cloud computing services benefits in i) Reducing hardware installation & maintenance cost ii) Reducing infrastructure maintenance cost iii) E-waste minimization iv) On demand, anywhere, from any device services v) Efficient usage of electrical power vi) Flexibility and highly automated v) Virtual Business setup vi) Easier to replace and upgrade vii) Easier Maintenance and Management. Cloud computing helps customers by having its own intelligent features like i) Portability ii) Encrypted data storage iii) Fault Tolerance & Disaster Recovery iv) Elasticity vi) High Availability vii) Intelligent Management viii) Performance ix) On demand self services x) Service measurement xi) Resource pooling. In the following discussion, we briefly describe various services offered by cloud computing and deployment types of cloud services.

Cloud computing Services:

Cloud computing providers deliver applications via the internet, which are accessed from web browsers, desktop and mobile apps, while the business software and data are stored on servers at a remote location. As shown in fig. 2 Cloud Computing Technologies are grouped into 4 sections, they are SaaS, DSaaS, IaaS and PaaS[8] [3].

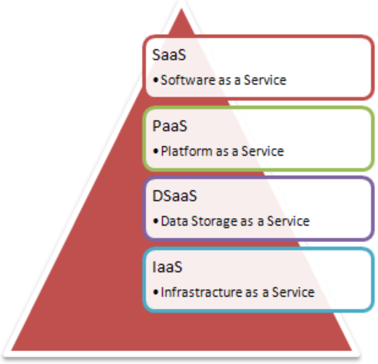


Fig. 2: cloud computing services

SaaS (Software as a Service) is on demand application service. It delivers software as a service over the Internet, eliminating the need to install and run the application on the customer's own computers [8] [3]. Fig. 3 shows that without installing, client can access the required application from cloud SaaS service provider over internet.

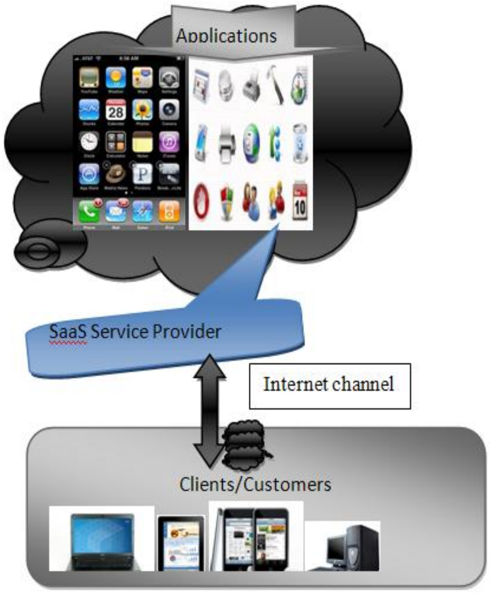


Fig. 3: Overview Model of Software as a Service

PaaS (Platform as a Service) is on demand platform service to host customer application. PaaS is delivery of computing platforms and/or solution stack as a service, often consuming cloud infrastructure and sustaining cloud applications. It facilitates deployment of applications without the cost and complexity of buying and managing the underlying hardware and software layers [8] [3]. Fig. 4 shows that the customer can access the required platforms remotely from PaaS service providers. It improves the flexibility in having multiple platforms in business environment.

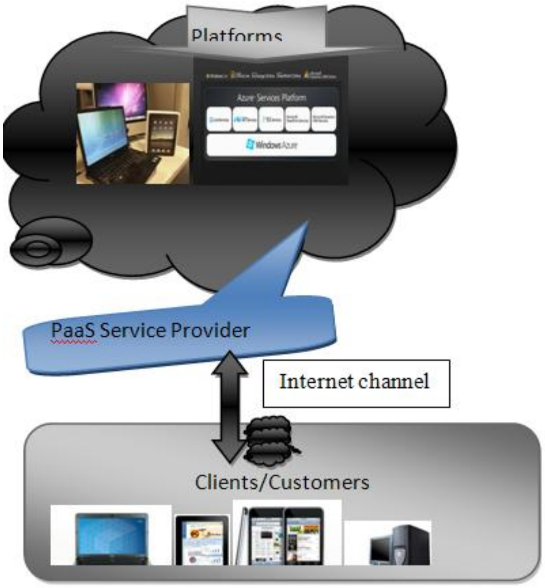


Fig. 4: Overview Model of Platform as a Service

DSaaS (Data Storage as Services) is on demand storage service. Cloud computing provides internet- based on demand back up storage services to customer [8] [3]. Fig. 5 shows the on demand accession of DsaaS services. In this service, customers can keep their data backup remotely over internet servers. These backup data maintenance is taken care by DsaaS service Provider. Cloud DsaaS service providers are responsible for customer data to keep confidentially. Here customers need not worry on setting up the large discs array to keep their huge amount of data.

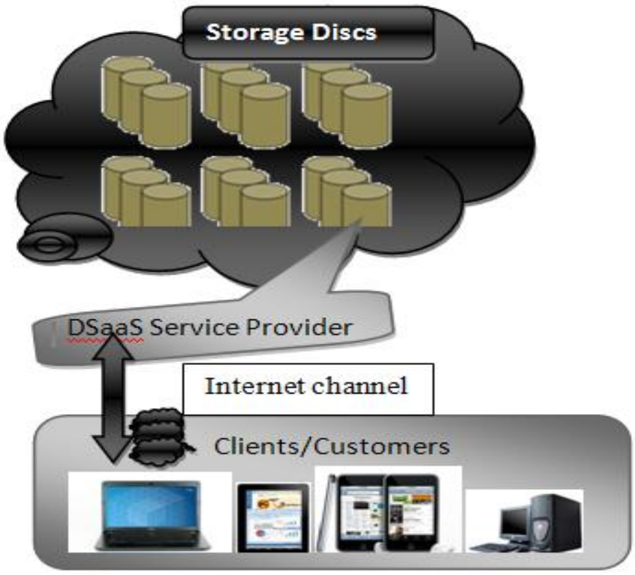


Fig. 5: Overview Model of Data Storage as a Service

IaaS (Infrastructure as a Service) is on demand infrastructure service. It delivers the computer infrastructure – typically a platform virtualization environment – as a service, along with raw (block) storage and networking. Rather than purchasing servers, software, data-center space or network equipment, clients can buy those resources as a fully outsourced service [8] [3]. Fig. 6 shows that customers can access infrastructure from IaaS service provider over internet.

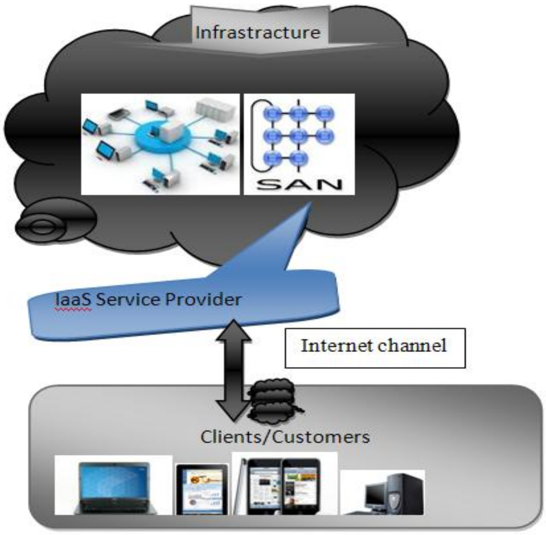


Fig. 6: Overview Model of Infrastructure as a Service

Deployment Types

Any organization can setup/use the cloud for its business maintenance purpose. There are four types of deployment that a customer can establish such as: Private, Public, Community and Hybrid [8].

Private cloud is a cloud service created with own/ rented resources. As shown in Fig. 7 the cloud infrastructure is owned or leased by a single organization and is operated solely for that organization

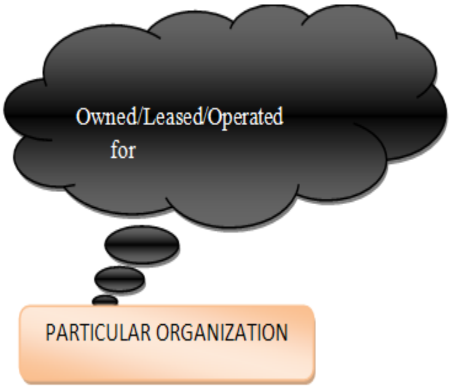


Fig. 7: private cloud

Community cloud: The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations).Fig 8 shows the community cloud created with similar group of customers with same set of resource requirements.

Public cloud: The cloud infrastructure is owned by an organization selling cloud services to the general public or to a large industry group. Fig. 9 shows the cloud infrastructure created with standard specification to any organization.

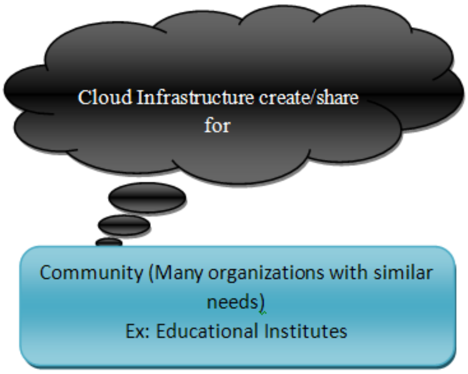

Cloud Infrastructure create/share
for
Community (Many organizations with similar needs)
Ex: Educational Institutes

Fig. 8: Community cloud

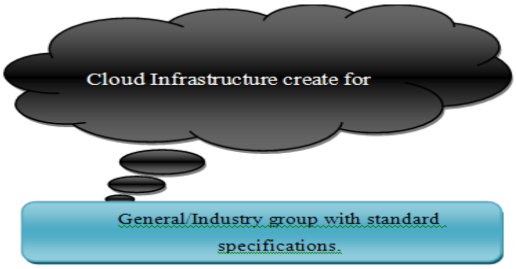

Cloud Infrastructure create for
General/Industry group with standard specifications.

Fig. 9: Public cloud

Hybrid cloud: The cloud infrastructure is a composition of two or more clouds (internal, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability. Fig 10 shows the hybrid cloud infrastructure created with any combination of public, private and community.

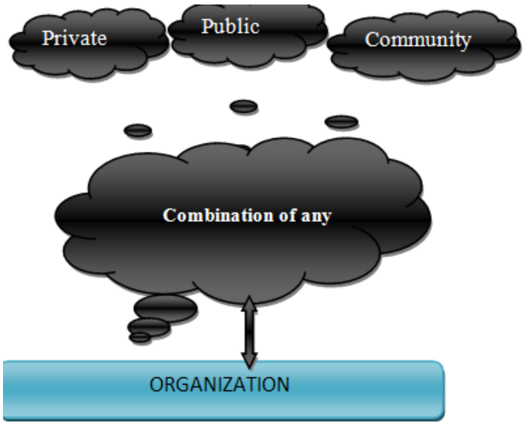


Fig. 10: Hybrid cloud

3. Secure Cloud Transmission Protocol (SCTP) Design

In internet, services are accessed through web browser using http's protocols such as HTTP, HTTPS & S-HTTP. However these protocols have some security issues which are discussed in [4] [5] [6]. HTTP is an application layer protocol which helps in sending and receiving the information. HTTP is not suitable for sensitive information transaction because it is not a secure protocol [4]. HTTPS is another protocol designed to provide security. This protocol works in presentation layer in encrypting the sensitive transaction. HTTPS is not effective because, along with message body it also encrypts the message header [5]. S-HTTP is designed in such a way that it encrypts only a message body [6]. These protocols do not help the security challenges such as man- in- middle attack, data integrity, strict authentication and authorized techniques and intruder detection [4][5][6]. Tables 1 discuss the security issues of HTTPs protocols. Based on the information given in Table 1, we suggest the security requirements such as secure channel, strict authentication and efficient cryptographic techniques.

Protocol Name	Description	Security Issues
HTTP	Application layer Request-response protocol with no security.	Data confidentiality, Integrity, Man-in-the-middle, Eavesdropping attacks.
HTTPS	I t is a Combination of HTTP and SSL/TLS. Performs the encryption to entire messages at Presentation layer. Provides authenticated public key certificate for web server	Man-in-the-middle attack Breakable by brute force technique, hackers can attack with login data (brute force technique). The encryption which performs at presentation layer (not efficient) happens to entire message. Browser dependability while encrypting and session transmission, hence long wait time.
S-HTTP	It's a HTTPS with efficient encryption	Man- in- middle attack.

Table 1: Security issues in existing protocol [4] [5] [6].

Another issue which needs to be handled by secure cloud transmission protocol is about cloud service. In SaaS, it needs to ensure user authentication with correct privileges checking before using any application [8] [9] [10]. In PaaS, before providing platform to launch customer application, it needs to ensure bug, vulnerability of platforms, Multi-tenanted application isolation, authentication privileges to particular user. In DSaaS, before using storage service, it needs to ensure Data Protection, Integrity, vulnerability and security from intruder. In IaaS, before taking infrastructure, it needs to ensure Physical Security, Privileged access rights, control and monitoring infrastructure, maintaining infrastructure, communication channel security, intruder detection, privileges to access the infrastructure and auditing techniques. The above issues are summarized in Table2 [8] [9] [10].

Name	**On demand service for**	**Control**	**Ensure security challenge**
SaaS	Application	No control on OS, H/W, N/W infrastructure.	Privileged access Authenticated access User Types
PaaS	Platforms (Hosting	Can control hosting	Bug, vulnerability

	Environment)	environment not on OS, H/W, and N/W infrastructure.	of platforms. Multi-tenanted application isolation, authentication privileges to particular user
DSaaS	Storage Area	No Control	Data Protection, Integrity, vulnerability and security from intruder
IaaS	Infrastructure : Computing Resources, Storage, Network or middleware	Can control OS, Storage. Applications not on cloud infrastructure	Physical Security, Privileged access rights, control and monitoring infrastructure, maintaining infrastructure, communication channel security, intruder detection , privileges to access the infrastructure, auditing techniques

Table 2: Cloud services security issues [8] [9] [10]

Overview of proposed protocol

To address the some of the important security challenges which are discussed in Table1 and Table 2, we are proposing secure cloud transmission protocol (SCTP). Expected objective of secure cloud transmission protocol is to provide secure channel over insecure internet independent of its devices, browsers and physical locations. As shown in Fig. 11, SCTP one of the features is to work independent of physical location, computation devices and browser types.

SCTP requirements and roles

Expected objectives of SCTP are to provide secured internet channel with effective authentication techniques and efficient cryptographic algorithms. An effective authentication technique is needed for ensuring strict user authentication and authorization. By considering Table 1 and Table 2, security issues, identified and analyzed SCTP requirements are:

- Strict Authentication (It applies strict techniques : Multilevel, multifactor password generation)
- Efficient Cryptographic Approach (Encryption and Decryption)
- Secure Channel (Fully protected media)
- Intrusion detection (Finding out attackers)

Fig. 12 proposes the SCTP roles in OSI Layers. We expect SCTP to perform the strict authenticated privilege access at application layer and efficient encryption at presentation layer.

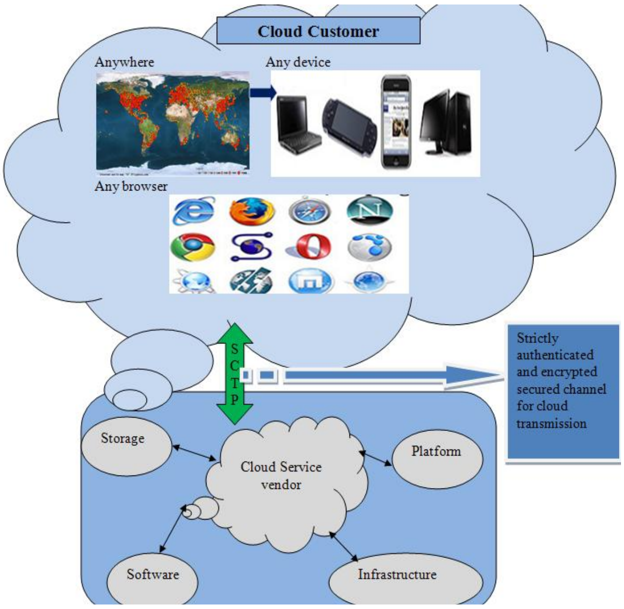
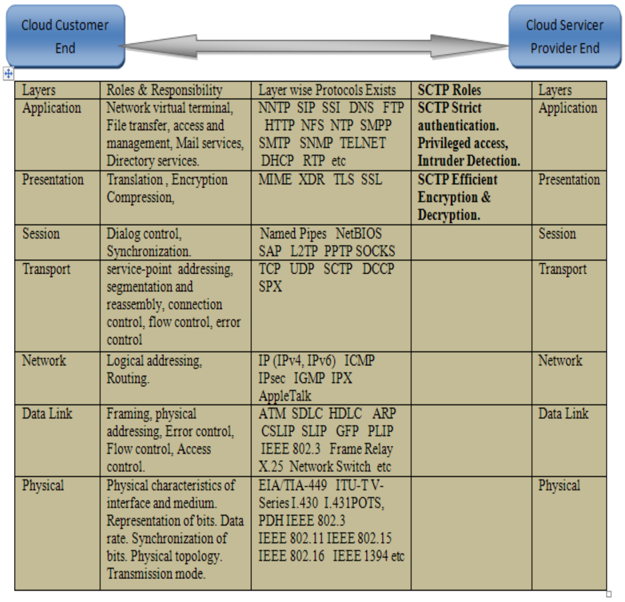


Fig. 11: Overview model of Secure Cloud Transmission Protocol (SCTP)


Cloud Customer End
Cloud Servicer Provider End

Layers	Roles & Responsibility	Layer wise Protocols Exists	**SCTP Roles**	Layers
Application	Network virtual terminal, File transfer, access and management, Mail services, Directory services.	NNTP SIP SSI DNS FTP HTTP NFS NTP SMPP SMTP SNMP TELNET DHCP RTP etc	**SCTP Strict authentication. Privileged access, Intruder Detection.**	Application
Presentation	Translation , Encryption Compression,	MIME XDR TLS SSL	**SCTP Efficient Encryption & Decryption.**	Presentation
Session	Dialog control, Synchronization.	Named Pipes NetBIOS SAP L2TP PPTP SOCKS		Session
Transport	service-point addressing, segmentation and reassembly, connection control, flow control, error control	TCP UDP SCTP DCCP SPX		Transport
Network	Logical addressing, Routing.	IP (IPv4, IPv6) ICMP IPsec IGMP IPX AppleTalk		Network
Data Link	Framing, physical addressing, Error control, Flow control, Access control.	ATM SDLC HDLC ARP CSLIP SLIP GFP PLIP IEEE 802.3 Frame Relay X.25 Network Switch etc		Data Link
Physical	Physical characteristics of interface and medium. Representation of bits. Data rate. Synchronization of bits. Physical topology. Transmission mode.	EIA/TIA-449 ITU-T V-Series I.430 I.431POTS, PDH IEEE 802.3 IEEE 802.11 IEEE 802.15 IEEE 802.16 IEEE 1394 etc		Physical

Fig. 12: SCTP roles & responsibilities with OSI layers roles.

SCTP Framework Design

After we analyzed the requirements of SCTP, In SCTP framework expected to involve: i) Strict Authentication

Techniques before using cloud services ii) efficient cryptographic approach to encrypt/decrypt the data over internet iii) ensuring the secure channel over insecure internet iv) intrusion detection. Fig. 13 is the functionalities framework diagram of SCTP which represents the identified functionalities of SCTP.

Actor documentation

Customer is an actor/setup who uses the cloud services for their business. **Cloud Service Provider** is the actor/setup who provides the cloud services over internet. **Intruder** can be person, tool and machine who do customized attacks against web applications, to identify and exploit all kinds of security vulnerabilities.

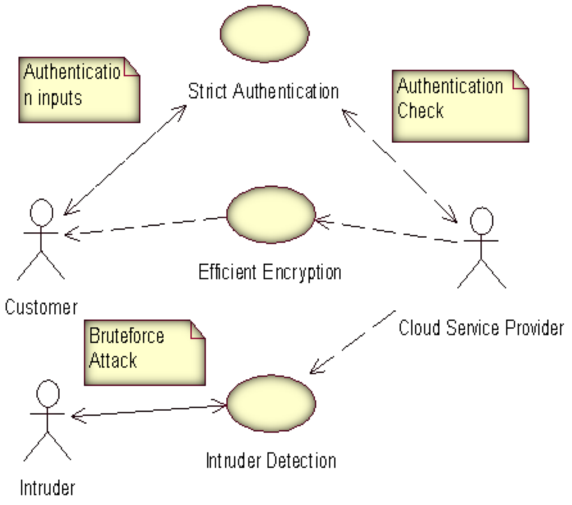


Fig. 13: SCTP framework for indentified functionalities

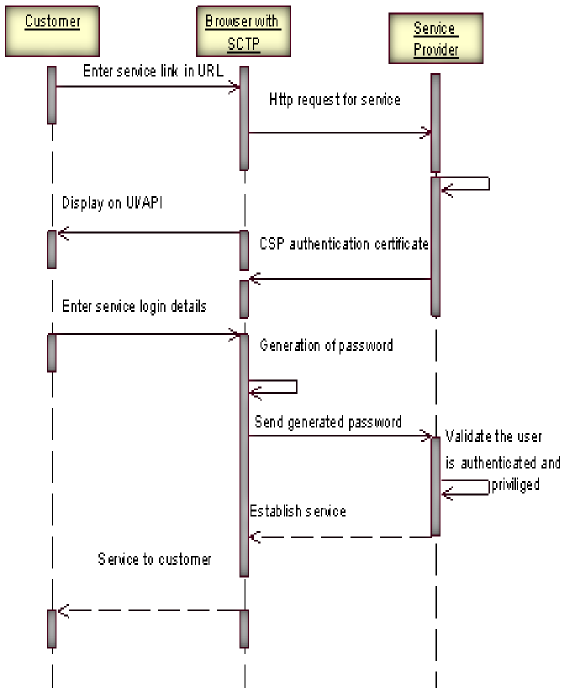


Fig. 14: SCTP Strict authentication

Functionalities:

Function: Strict Authentication

Description: This method is to ensure that the customer is authorized and authenticated before providing the service. This has multidimensional password generation, biometric and image- authorized techniques.

Flow of Events: (As Fig. 14 shown)

1. Send a service request
2. Enter login details
3. Generate password & authenticate the customer

Pre-Condition: Request for cloud service

Post-Condition: Find out whether authenticated customer or intruder

Function: Efficient Encryption

Description: This method is to ensure that the data transmission happens in encrypted form.

Flow of Events: (As Fig. 15 shown)

1. Enter encryption details
2. Generate key
3. Provide encrypted transmission

Pre-Condition: Should ask for data transmission

Post-Condition: Efficient encryption and data transferred over secure channel for authenticated customer

Function: Intruder detection

Description: This method is to find intruder.

Flow of Events: (As Fig. 16 shown)

1. Find if any unauthenticated customer is trying with brute force attack.
2. If intruder is suspected it then generate and report complaints to CSP.
3. Reject connection

Pre-Condition: Should be unauthenticated try

Post-Condition: Register complaints to CSP

Fig. 14, 15, 16 shows the sequence of operation in SCTP Strict authentication, efficient encryption/decryption and intruder detection respectively.

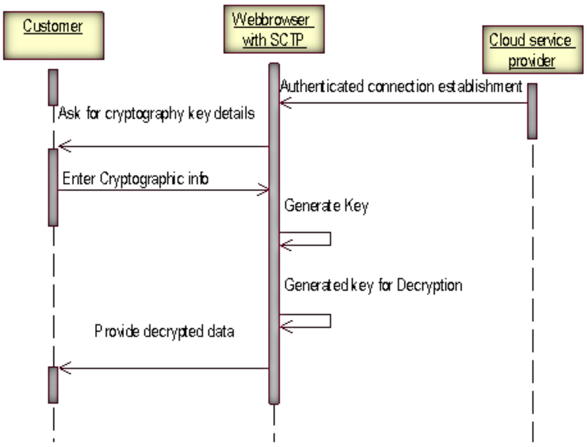


Fig. 15: SCTP encryption/decryption

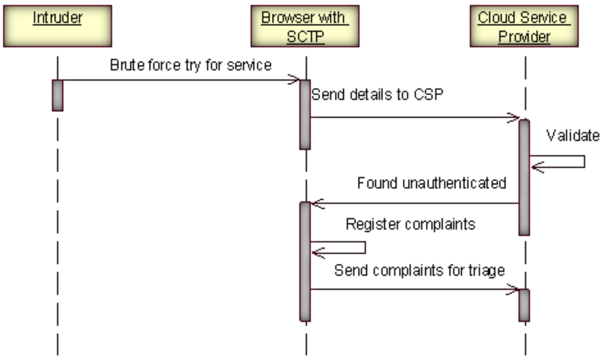


Fig. 16: Intruder detection system in SCTP

Secure cloud transmission protocol expected to works under many execution states such as connection establishment, authentication check, encryption/decryption, service use, measuring the service and connection end. Fig. 17 is a state chart diagram of SCTP which shows different states of SCTP during its execution.

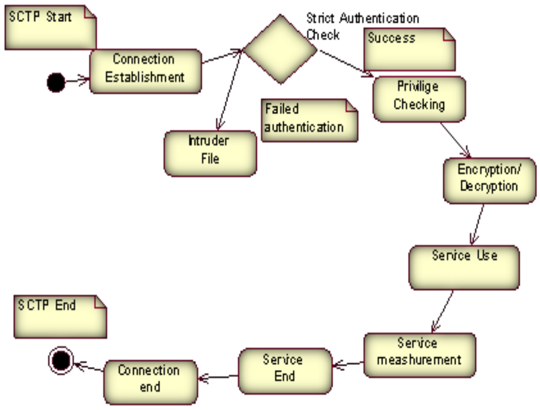


Fig. 17: SCTP State Chart Diagram

4. Conclusion and Future Enhancement

Cloud computing is a technology that uses the internet and central remote servers to maintain data, platforms, infrastructure and applications. Cloud computing allows consumers and businesses to use applications, platforms, infrastructure without installation and access their personal files at any computer with internet access. However there are internet security issues that need to be addressed. We propose a framework design of secure cloud transmission protocol (SCTP). SCTP is expected to create such secure channel over insecure internet. SCTP has proposed with strict authentication techniques and cryptographic approaches. SCTP design details which are presented in this paper may help us to fix the major security challenges which are identified in http protocols and cloud computing services. Our future plan is to carry out the SCTP detailed design along with its security proof.

Acknowledgment

Our sincere thanks to Prof. K N B Murthy, Principal and Prof. Shylaja S S, HOD, Department of Information Science and Engineering, PESIT, Bangalore, for their constant encouragement

References

[1]. Center Bo Wang, HongYu Xing "The Application of Cloud Computing in Education Informatization, Modern Educational Tech..." Computer Science and Service System (CSSS), 2011 International Conference on IEEE, 27-29 June 2011, 978-1-4244-9762-1, pp 2673 – 2676.

[2]. NIST Definition http://www.au.af.mil/au/awc/awcgate/nist/cloud-def-v15.doc

[3]. Cloud Computing services & comparisons http://www.thbs.com/pdfs/Comparison%20of%20Cloud%20computing%20services.pdf

[4]. Hyper Text Transmission Protocol: Communication Technology Proceedings-2003. ICCT 2003. International Conference on Study on conformance testing of hypertext transfer protocol by Xiaoli Yu; Jianping Wu; Xia Yin; Dept. of Comput. Sci., Tsinghua Univ., Beijing, China

[5]. hyper text transmission protocol with security: A Performance Analysis of Secure HTTP Protocol by Xubin He, Member, IEEE. http://en.wikipedia.org/wiki/HTTP_Secure. http://www.technolozy.net/difference-between-http-and-https-protocols.html.

[6]. S-HTTP: Secure Hypertext Transfer Protocol: http://www.javvin.com/protocolHTTPS.html. http://en.wikipedia.org/wiki/Secure_Hypertext_Transfer_Protocol

[7]. Man in the middle attack Moxie Marlinspike (2009). "HTTPS ... pwned !". http://blogs.orange-business.com/securite/2009/02/ssl-pwned.html. Retrieved 2011-06-20

[8]. A User Identity Management Protocol for Cloud Computing Paradigm Safiriyu Eludiora1, Olatunde Abiona2, Ayodeji Oluwatope1, Adeniran Oluwaranti1, Clement Onime3,Lawrence Kehinde apered in Int. J. Communications, Network and System Sciences, 2011, 4, 152-163

[9]. Cloud Computing Challenges and Related Security Issues: a survey project report on Cloud Computing Challenges and Related Security Issues by Traian Andrei and Prof. Raj Jain

[10]. Protocols for Secure Cloud Computing IBM Research – Zurich Christian Cachin April 2011

Author Biographies

First Author Dinesha H A was working with VMware pvt India ltd. Now he is with PES Institute of Technology, Assistant Professor in ISE & Research Scientist in CORI,100ft Ring Road, BSK III Stage, Bangalore -560085, Karnataka India (phone: +91-9945870006; FAX: 08026720886, email:sridini@gmail.com)

Second Author Dr V. K Agrawal was working in ISRO, now he is with the PES Institute of Technology, Professor in ISE & Director in CORI,100ft Ring Road, BSK III Stage, Bangalore -560085, Karnataka India (Ph: 080-26720783 FAX: 08026720886,email:vk.agrawal@pes.edu)

Design of Web Content Management System for Dental Laboratories

Reham Alabduljabbar and Samir El-Masri

Department of Information Systems, College of Computer and Information Sciences, King Saud University
Riyadh, Saudi Arabia

Abstract

Web Content Management system is a management tool for creating a dynamic website. It ensures logical structure of data organization and ease of content accessing and presenting. Dental laboratories need Web Content Management system (WCMS) to control their business. Maintaining a long-term relationship between dental laboratories and their customers (dental clinics and dentists) urges an active communication process between the two sides. The main contribution of this paper is to design a simple Web Content Management System for Dental Laboratories. The system adopts three layers of technical architecture. The paper will also discuss why there is a need to develop a standalone WCMS for Dental Laboratories whilst other open source WCMSs can be utilized such as Joomla, Drupal and WordPress.

Keywords: *Web content management system, Dental laboratory system, Three-tier architecture, Commercial content management system, Open source content management system, Joomla, Drupal.*

1. Introduction

Not all dental clinics have their own dental laboratories. Small to medium clinics send their patient lab cases to local, national or sometimes international dental laboratories. Communication is done through mailing handwritten forms or sending files and bills by email. Usually each dental clinic works with one dental laboratory. However, dental laboratories receive lab cases from different dental clinics. a long time is wasted at both sides on trying to track down missing lab case information over the phone or email, working out unreadable handwritten prescriptions, or following up on billing and payments.

Maintaining a long-term relationship between dental laboratories and their customers (dental clinics and dentists) urges active communication process between the two sides.

According to the Marketing Director at the Continental Dental Laboratories[1]: “communication—or a lack of it—will make or break the relationship between a laboratory and the dentist”. Until now, this communication process is done through a handwritten prescription and an impression that may or may not have been able to completely give the required information to the technician to meet the dentists’ expectations.

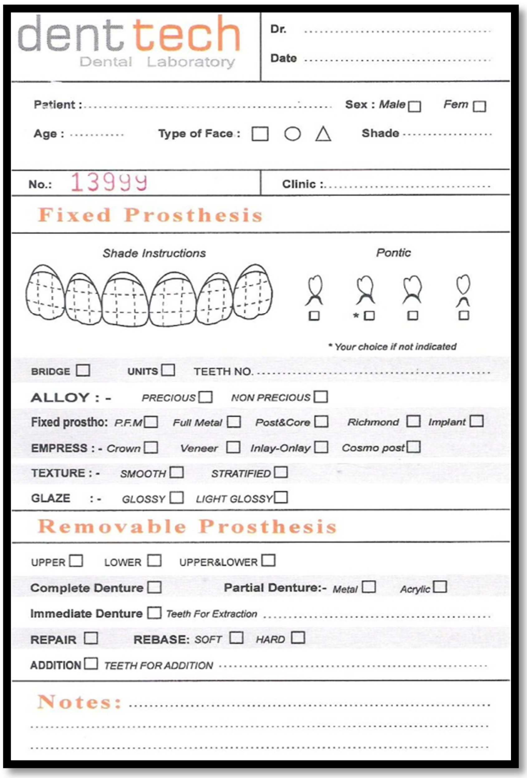
Vice President, Sales & Marketing at Trident Dental Laboratories[2] agrees that the fewer laboratories have to depend on verbal or written instructions, the better. The laboratory work depends on the prescription form received from the clinic, any simple mistake or incomplete information will result in loss in money and customers [4].

It is not only about lab case management, the dental laboratory, the dental technicians, and the laboratory owner have an obligation towards dentists to share their knowledge with them and to educate them regarding new products. Whether for product education or case management, communication between the dentist and the laboratory is the most important factor that has a great influence on the success of the relationship [4].

The market is crowded with desktop applications for managing dental laboratories operation; however, we are looking for a system that utilizes both case management and a relationship between dental laboratory and its customer. Many information need to be shared between dentists, dentist assistant, laboratory technicians, laboratory owner and others. With the development of the Internet technology, there is a need for an effective web application to mange such operation and to control accessing this shared information. Thus, the motivation of this paper is to design a simple Web Content Management System (WCMS) for creating dental laboratories websites.

[1] http://www.continentaldental.com

[2] http://www.tridentlab.com/



dent tech
Dental Laboratory

Dr. ....
Date ....

Patient : .... Sex : *Male* □ *Fem* □
Age : .... Type of Face : □ ○ △ Shade ....

No.: 13999 Clinic : ....

Fixed Prosthesis

Shade Instructions

Pontic
□ *□ □ □

** Your choice if not indicated*

BRIDGE □ UNITS □ TEETH NO. ....
ALLOY : - *PRECIOUS* □ *NON PRECIOUS* □
Fixed prostho: *P.F.M* □ *Full Metal* □ *Post&Core* □ *Richmond* □ *Implant* □
EMPRESS : - *Crown* □ *Veneer* □ *Inlay-Onlay* □ *Cosmo post* □
TEXTURE : - *SMOOTH* □ *STRATIFIED* □
GLAZE : - *GLOSSY* □ *LIGHT GLOSSY* □

Removable Prosthesis

UPPER □ LOWER □ UPPER&LOWER □
Complete Denture □ Partial Denture:- *Metal* □ *Acrylic* □
Immediate Denture □ *Teeth For Extraction* ....
REPAIR □ REBASE: *SOFT* □ *HARD* □
ADDITION □ *TEETH FOR ADDITION* ....

Notes: ....
....
....

Figure 1- Requesting Lab Case Form

The dental laboratory web content management system provides service to the dental clinics located in different locations. It would be used as a channel for dentists to technician, dental clinic-to-dental laboratory, to provide the long-term relationship and information sharing.

In daily services, a dentist in a certain clinic fills a form as in Figure 1 to order a lab case from a certain dental laboratory. Then, the dental assistant and other staff arrange with the laboratory for pickup, payment and delivery. This paper-based recording imposes several major drawbacks namely miscommunication between the laboratory and the clinic and lack of visual interactivity.

With dental laboratory WCMS, the medical data, pictures and patient records all can be accessible online. Dental laboratories can track and manage lab cases and payments online. Besides, dental clinics and dentists can access to track the lab cases, and they can be notified by email about the status of their lab cases. Both have an account and can view their lab case history, their current balance, and pay bills online. It aids in reducing paperwork and automating the approach to process lab cases.

Any dental laboratory can have its own instance of the system to manage its content and to ease communications with dental clinics its working with. The basic idea of web content management systems is to get organized and find a logical, consistent and easy way to place content on the web [8]. It allows non-technical users to create, edit, manage and control a large, dynamic collection of web material (HTML documents, images and video). WCMS involves a lifecycle starting from creation to destruction of content. The lifecycle includes reviewing the content before publishing it and it may include archiving before destroying. WCMS helps in keeping the site more consistent, ease the navigation, and most important it aids in controlling and tracking the content [13].

Figure 2 shows the framework for the dental laboratory WCMS. The core of this dental laboratory WCMS is the content, which is the patient lab case that is being sent from a certain dental clinic to the dental laboratory to be produced. The full content lifecycle starts from a dentist in a clinic submitting new patient lab case to the system and then, a laboratory technician is assigned to process this lab case. The content will be archived and later destroyed after delivering the patient lab case and receiving the payment.

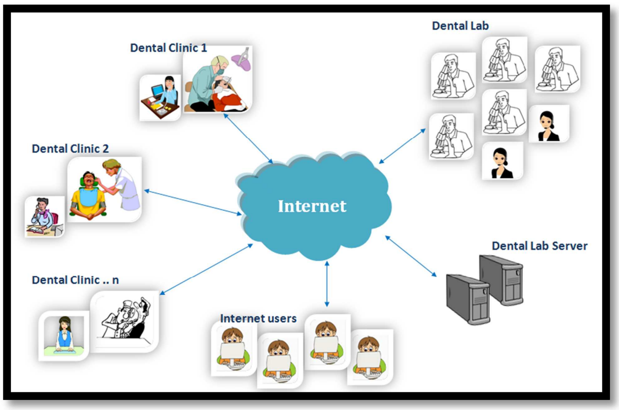


Figure 2- Dental Laboratory WCMS Framework

In the next section of this paper, a general review on commercial and open-source solutions for similar systems is given with a discussion whether there is a need to build a custom WCMS or the available open source WCMS solutions can be utilized. In the third section, the overall system analysis and design is proposed. Finally, the paper concludes with a summary of the key issues discussed in the paper.

2. Related Background

2.1 Commercial Dental Laboratories WCMS

Similar dental laboratory management systems envisaged is available on the international market in various formats, however the products available have several disadvantages. The greatest and most obvious disadvantages are:

- They are costly and usually unaffordable for small to medium dental clinics.

- They do not combine web management and content management.
- They normally do not have user friendly interface.
- They do not offer any kind of personalization or customization to their customers.
- They only offer dentist account; there is no dental clinic account.

We conduct a research and create a short list of systems in order to be further examined and narrowed or widened to fit the small to medium dental laboratories need. Table1 shows a comparison between some of the systems:

Table 1- Comparison between some available systems

	evident COMMUNICATE. CREATE. SHARE.	labnet	Precise
Web-based	√	Only via DDX [1]	X
On line scheduling	√	√	√
Rescheduling	√	Only via DDX	X
Patient lab case on line tracking	√	Internal only	√
Billing	√	√	√
Dentists Profile	√	√	√
Clinic Profile	X	X	X
Attaching files	√	Only via DDX	X
Reports	√	√	X
Customer Service	√	√	X
Chatting system	X	X	X
Personalization	X	X	X
Customization	X	X	X

2.2 Open-source WCMS Utilized by Dental Laboratories

In addition to theses commercial WCMS solutions, many open source solutions are utilized using Joomla, Drupal and WordPress. We have studied some websites which are utilized by different open-source solutions:

- Websites powered by Joomla:
  - Quest Dental Laboratory: http://questdental.us
  - A+ Dental Laboratory: http://www.a-plusdentallab.com/
- Websites powered by Drupal:
  - Vision Dental Lab: http://visiondentallaboratory.co.uk/
  - Mascol Dental Lab: http://mascoladentallab.com/
- Websites powered by WordPress:
  - ArrowHead Dental Lab: http://www.arrowheaddental.com/
  - Keller Laboratory: http://www.kellerlab.com/
  - Nik Dental Laboratory: http://www.nikdentallab.com/

Following are the common features of these websites:

- Sending a Lab Case: By filling handwritten form and schedule for pickup.
- Online Account: For scheduling a pickup.
- Lab Case Tracking: Not available.
- Shipment Tracking: Via carrier.
- Billing and Statements: Sent by mail.
- Product Education: promotes education for dentists through all the typical means, such as direct mail, journal ads, articles that they place in journals, and their Web site content.

Among all of the envisaged websites, none of them offer clinic and dentist account. Besides, they do not have a user-friendly interface. They are more likely static WebPages of mostly informational content with simple designs.

On the next section, a discussion is presented whether it is better to build a custom WCMS or to utilize an open-source solution.

2.3 Commercial vs. Open-source WCMS

As mentioned previously, there are many open-source solutions. Some of them are being created for many years, empowered by developers with technical background. A question may arise why we need standalone WCMSs for dental laboratories? Why dental laboratories do not utilize open source WCMSs solution? Although this is not the main goal of this research, it is worth it to bring up this discussion.

On the one hand, having a commercial WCMS specifically for dental laboratories will allow for full flexibility in developing [10]. Once it is available, many dental laboratories can utilize it instead of utilizing open-source solutions, this is because, the entire application is setup so it works exactly how needed by the dental laboratory. It will be faster to implement and associates a certain degree of safety as opposed to open-source. Moreover, it offers more support and stronger training documentation than open-source. Probably the most important concern regarding the commercial WCMS especially for small to medium dental laboratories is the cost.

On the other hand, for small to medium dental laboratories, open source WCMSs offer a low cost alternative to commercial solutions. Besides, Troubleshooting is made easier because of the technical support and online community. However, potential concern regarding the open-source solutions is the security. As the source code is available for public, attackers can use the source code to identify vulnerabilities. Thus, these systems raise significant security issues [9].

[1] DDX is a web based system that turns Labnet into a Web-enabled application.

According [1] in 2011, the best WCMSs today are Joomla and Drupal. Having studied them, Drupal is the best to utilize when developing large websites with hundreds of pages but smaller websites with lesser number of pages are better developed by Joomla [2]. Drupal is not very user-friendly, terms are confusing and the admin interface is relatively poor, whilst Joomla is more user-friendly with a more active developer and designer community. With Drupal, unlimited user permissions levels can be created, but Joomla offers only three user levels (Public, Registered and Special). Drupal does not support multimedia, photo galleries by default but Joomla supports multimedia by its default editor [3]. Of course both of their developers have overcome these issues by developing modules to extend their usability. However, modules not always free or easy to install.

To sum up the discussion, both approaches the commercial and the open-source have their advantages and disadvantages. It is depending on the requirements of the system, so there is no absolute answer to which is 'best'. However, open-source WCMSs does not fit the requirements of the system presented in this paper. We need many levels of permissions with user-friendly interface, plus securing websites to handle payments.

As we will see in the next section, the system and user requirements are not supported by the available dental laboratories commercial WCMS, thus, we are designing a new commercial WCMS for Dental Laboratories.

3. Overall System Analysis and Design

3.1 System Analysis

3.1.1Basic Requirements of WCMS

Following are the basic requirements of WCMS [5, 12, and 13]:

- Manage and deliver large amounts of unstructured material in multiple media and enable linking between the related materials.
- Provide a consistent and predictable information structure, user interface, and navigational mechanism.
- Support well-defined roles, responsibilities, and access control for various types of users e.g. retrieval only, editors, publishers, web manager, administrator etc.
- Enable workflow between authors, product managers, content administrators, editors, and system administrators.
- Provide a locking or concurrency control mechanism to prevent two people from simultaneously updating the same content.
- Enable searching and retrieval of content using the predefined business characteristics of products and services.
- Ability to archive data, and to output reports in digital and printed form.
- Providing a good scalability and portability, benefiting the extension of system function in future.

3.1.2 System Components and System Users for Dental Laboratory WCMS

The system has three major types of system engines (components) similar to the system proposed on [5, 6] in addition to the data warehouse repository:

- A Content Editorial Engine provides content and repository maintenance and approval functions for different levels of administrators in the dental laboratory.
- A Content Reception Engine collects content from external sources, and then delivers it to different parts of the system for approval and publication.
- A Content Publishing Engine stores approved content and send them to different parties via different channels (such as email, fax, and text messaging). It also serves as the Web storefront of the dental laboratory for user enquiries.

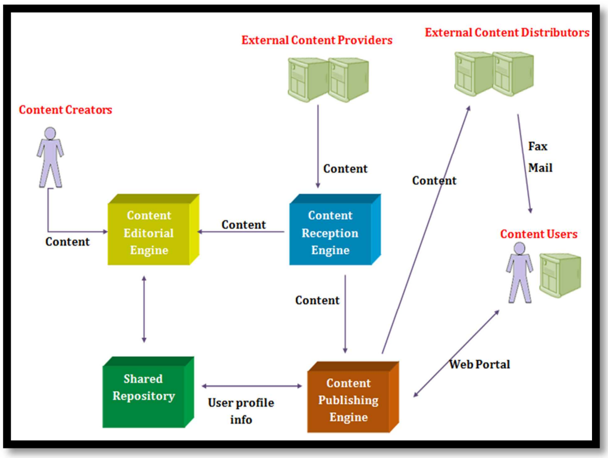


Figure 3-System users and components

Figure 3 depicts an overview of the Dental Laboratory WCMS highlighting the main system components and system users. A Dental Laboratory WCMS must be designed specifically to match the need and interest of each system user within and related to the dental laboratory.

Besides the management, there are four main types of system users involved, namely, Content Creators, Content Providers, Content Distributors, and Content Users:

1. Content Creators collectively refer to internal users who are involved in the content creation processes of the dental laboratory. The Dental Laboratory WCMS should be able to accommodate the different operational and administrative requirements of these different roles of internal users and to maintain appropriate security control. They interact mainly with Content Editorial Engines of the Dental Laboratory WCMS.
2. Content Providers are external sources (such as PayPal) providing content (such as payments) to the dental laboratory through a Content Reception Engine. To ensure timeliness, content from trusted sources are usually forwarded automatically to the Content Publishing Engine for immediate delivery.
3. Content Distributors are external service providers that render the content and deliver them to clients via different (traditional or electronic) channels, such as mass fax, mail, email, hardcopy delivery, and so on.
4. Content Users, who can be internal or external to the dental laboratory, are classified into three types in our case. Content Users obtain content access through a Content Publishing Engine. Based on their subscription data, the Content Publishing Engines also actively send appropriate content to the subscribed users. The three types are:
   - Public Visitors – Anonymous users are often allowed to access some limited amount of public content through a portal. This helps attract them to visit the dental laboratory's Web site.
   - Clients (dental clinics and dentists) – Customers who do basic business with the dental laboratory are allowed access and subscription to all unrestricted content. They have their own gateway where they can track their lab cases, pay and check their bills.
   - Internal Users – Internal staff can access "internal only" content related to them, as well as all the content for external users. They are also automatically subscribed to relevant content, according to their job functions, secretary, technician, driver, and so on.

3.1.3 Description of the Main System's Functions and Workflow

As mentioned earlier, the core of this dental laboratory WCMS is the content, which is the patient lab case that is being sent from dental clinic or a dentists to the dental laboratory to be processed. The full content lifecycle content creation, content editing, content approval and content publishing, which together consist of the core part of the website content management system. Following is description of each cycle:

1. Lab case creation: ability to create new lab case and submit it. In addition, content creation should provide basic editing methods and editing tools and be able to upload and download images in the content. Dentists can login using their accounts into the system and submit their cases online using an electronic form. In addition, the system will allow dentists to choose the technician if they wish. If the technician is available and able to accept the case and finish it by required time then the case will be assigned to that technician. Otherwise, the system will notify the dentist that the technician is not available and will give him/her the choice to choose someone else or just leave it for the laboratory staff to assign the lab case to a technician. Approximate pricing will be calculated after filling the form and indicating the due date for delivery. In addition, the system will offer the ability to attach files to lab cases. Dentists can attach images and any other file that is needed for better understanding of the lab case.
2. Lab case editing: editing should satisfy the requirement of submitted lab cases in terms of querying, previewing, modifying, deleting, submitting, etc. such basic operation and management, edit content items without affecting the published work and distribute the authority to approval submitted contents. In addition, lab case editing should track the process of lab case submission and approval status which has been rejected or in the process of approval. Thus, laboratory staff can view the lab cases as soon as they are submitted so they can arrange a pickup. When the lab case arrived, its status is updated as received.
3. Approval process: Approval process can add, modify, delete and manage the authority of the allocation roles and individuals. Provide a workflow that is configurable for users to allow different approval processes with varying case item status during the authoring, establish a variety of roles within a workflow process, and assign workflow to classes of content items as well as roles and individuals. The workflow in the dental laboratory will be managed by the ability to check the status of every lab case. In addition, the dental clinic will be notified by email about the status of the lab cases whether they have been received, processed or out for delivery, etc. Of course dentists can login to system and check the status as well.
4. Case approval: ability to approve the submitted lab case which is in the edit process, query and view the submitted lab cases for approval. There are two kinds of approval states: passed and not passed. The function that approval process should implement is to ensure each approval step can authorize only one approver.
5. Case publishing: The editorial content can be published after passing the approval process, in the process of content publishing, cancellation and republished function should be provided. Published content should apply static html pages as contents storage form, with the publication of the contents, the unpublished content and page module

can be updated as well, which also makes publishing Web pages convenient to manage and update. Thereby increase the speed of page browse and access. Laboratory staff will be alerted when lab cases are ready to deliver.

6. Website configuration and management: Ability to classify and manage the website columns of the publishing content, including the basic operation of the columns, such as columns add, modify, delete, as well as the template option of page modules and website columns. Besides, Website resource management can achieve the basic functions, such as upload and download files management, configuration and management of the published website parameters.
7. Rights authority: include allocation of operating authority in various sectors of system function module.

3.2 System Design

In the dental laboratory context, the initial structure that would be used in the WCMS had the following characteristics:

1. One central site for the dental laboratory (each dental laboratory would contain a unique domain).
2. The entire application is installed on a hosting web server for that dental laboratory.
3. A separate instance would be created for every dental clinic or dentists.

The proposed WCMS adopts Three-Tier Architecture to design the system as in [14]. Since there are many computers with different kinds of operating system will work in coordination in the system, platform independent system architecture is needed to adapt the change in future use.

3.2.1 Three-Tier Architecture

The system architecture is shown in Figure 4 and is composed of three layers [11]:

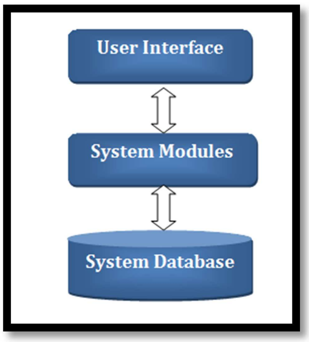


Figure 4-The system architecture

- The first (bottom) layer of the system is the database layer. It saves the system's content such as lab cases' data, dentists' data and images, etc. Content is frequently stored as XML since variety of data sources are used and each has its own characteristics [7]. XML is used to solve the incompatibility of different structures. XML facilitates reuse and enable flexible presentation options [12]. This layer mainly completes the local query, extracts and transforms distributed information from heterogeneous data sources. It uses Wrapper technology including the queries translate function and result translate function. It can translate the query result which gets from middle layer to local process. It extracts the query result and makes a XML document. Finally it returns the document to middle layer [13].
- The second (middle) layer is the system transaction layer, which is consisted of the system function modules, such as the lab case tracking. This layer mainly contains the query splitter and results Integrator two functions. In order to implement centre process, the system must use a common model which comes from different sources of information from a variety of data XML's characteristics determine that it is can describe a variety of data. It is means that it is a common data model. Heterogeneous data integration can solve this problem. It also enables the dynamic data release. Therefore, this system uses the XML model as a common model. It provides a unified query view by middleware layer on the client. It accepts the client's query command, split into various sub-queries and assigns to various data sources; and then integrates the results of its inquiries, sends to the client's browser displays for the user [13].
- The third (top) layer is the user interface layer, which included the client side of the system and displays the content to the users.

3.2.2 The System Functional Structure

A well-architected application is crafted into distinct layers, each of which encapsulates a particular role. There are different functions between the layers. Each layer maintains a clear separation to make them independent existence with different tasks. But interfaces are used to communicate between different layers, shielding the implementation of the internal detail. The specific function structure is discussed below [14]:

- Content integration: Allow users to find different forms of information from different systems, and the information includes documents, data, videos, and graphics and so on.
- Content intelligent: Content intelligent is the core function module of content management system, which is used for content classification archive and help users to locate required information quickly, the search methods can be divided into full-text search and context search.
- Content management: Content management system's main function modules, to support for content management

process, including content creating and editing, achieve the access to a variety of data, information, documents and procedures and provide collaboration tools to create content.

- Content distribution: Allow all users receive the published information at the same time, including the traditional way non-Web of graphical user interface GUI, web information portal and various other ways. The specific function structure chart is shown below as Figure5.

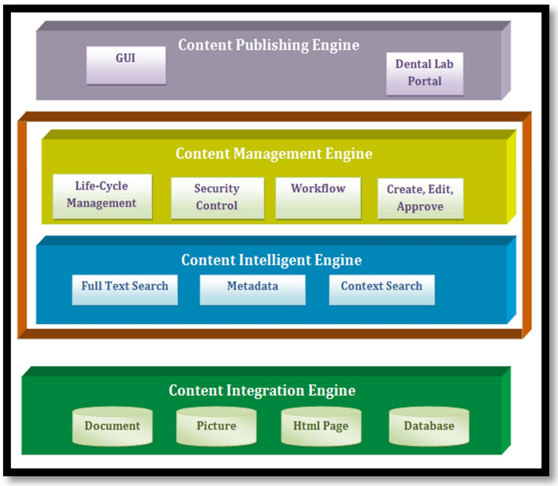


Figure 5-System Functional Structure

4. Conclusions and Future Work

The paper proposed a three-tier architecture design to develop a WCMS for dental laboratories based on a study of the requirements of dental laboratories. The proposed system was designed to enhance the clinical management level and to be used as a channel for dentists to technician, dental-clinic to dental-laboratory, to provide the long-term relationship and information sharing. The system was designed for this purpose which makes it more adaptive to the industry than any other existing industry WCMS.

It is expected that in the near future, the proposed system will be developed and implemented in several dental labs in Riyadh, Saudi Arabia. The system will significantly improve performance of dental laboratories and will assure long-relationship term between dental laboratories, dental clinics and dentists.

References

[1] Arah T., “The Best CMS: Joomla 1.6 vs Drupal 7.0” PC Pro blog, 2011, available at: http://www.pcpro.co.uk/blogs/2011/02/02/Joomla-1-6-vs-Drupal-7-0/ (accessed in June 2011).

[2] Bose S., “Drupal vs. Joomla: Advantages and Disadvantages”, Evon Technologies, 2011, available at: http://technology.ezinemark.com/Drupal-vs-Joomla-advantages-and-disadvantages-7d2d2b2f178c.html (accessed in May 2011).

[3] Burg S., “Joomla and Drupal - Which One is Right for You?”, December 2009, available at: http://www.alledia.com/blog/general-cms-issues/Joomla-and-Drupal-version-2/ (accessed in May 2011).

[4] DiMatteo, A. “Dental Labs: A Vital Key to Your Success”, Inside Dentistry published by AEGIS Communications, September 2009, Volume 5, Issue 8, available at: http://www.dentalaegis.com/id/2009/09/dental-labs-a-vital-key-to-your-success (accessed in May 2011).

[5] Gu, Y., Warren, J., Stanek, J. and Suthers, G. "A System Architecture Design for Knowledge Management (KM) in Medical Genetic Testing (MGT) Laboratories" Computer Supported Cooperative Work in Design CSCWD '06, 10th International Conference, Nanjing, China, May 2006, pp.1-6.

[6] Kwok, K. and Chiu, D. "A Web services implementation framework for financial enterprise content management" System Sciences, Proceedings of the 37th Annual Hawaii International Conference on 5-8 Jan. 2004, pp. 10.

[7] Liu Y., Yang L. and Zheng Y. "Research and design of open teaching and learning platform based on XML middleware" E-Health Networking, Digital Ecosystems and Technologies (EDT), 2010 International Conference, April 20, pp.356-359.

[8] McNay, H.E. "Enterprise content management: an overview," Professional Communication Conference, IPCC 2002 Proceedings IEEE International, Duluth, GA, USA, 2002, pp. 396-402.

[9] Meike, M., Sametinger, J. and Wiesauer, A. "Security in Open Source Web Content Management Systems" Journal of Security & Privacy, IEEE , (7:4), July-Aug. 2009, pp.44-51.

[10] Nakwaski M. and Zabierowski W. “Content Management System for Web Portal”, TCSET'2010, Lviv-Slavske, Ukraine, February 23-27, 2010.

[11] Preuner, G. and Schrefl, M. "A three-level schema architecture for the conceptual design of web-based information systems: from web-data management to integrated web-data and web-process management", Journal of World Wide Web, (3: 2), 2000, pp.125-138.

[12] Subrahmanyam J. “Future Trends Of Content Management Systems (CMS) for e-Learning: A Tool Based Database Oriented Approach“National Seminar on e-Learning and e-Learning Technologies ELELTECH, Hyderabad, India, 2005.

[13] Vidgen R., Goodwin S., and Barnes S. “Web Content Management”, e-Everything: e-Commerce, e-Government, e-Household, e-Democracy, 14th Bled Electronic Commerce Conference, Bled, Slovenia, 2001.

[14] Yu, J. "Distributed Data Processing Framework for Oral Health Care Information Management Based on CSCWD Technology," Information Science and Engineering (ICISE), 2009 1st International Conference, Nanjing, China: 26-28 Dec. 2009, pp. 2312-2315.

Reham Alabduljabbar Ms. Reham Alabduljabbar is a PhD Student enrolled at the Department of Information Systems, College of Computer and Information Sciences, King Saud University Riyadh, Kingdom of Saudi Arabia.

Dr Samir El-Masri completed his Electrical and Electronic Engineering degree at the College of Engineering, Lebanese University in 1993. Dr El-Masri has a Master's degree and PhD from the “Institut National Polytechnique de Grenoble” France in numerical simulations and software engineering in 1997. He worked for 3 years ended in 2001 at Hokkaido University, Japan as postdoctoral fellow and assistant professor. Dr El-Masri was a senior Lecturer/Associate Professor at the University of Western Sydney, University of Sydney, Central Queensland University and University of Southern Queensland, Australia from 2001 till 2006. He worked for at least 4 years in IT and software development industry at large Australian companies as a senior project manager/Program Manager and senior consultant in Australia. Dr El-Masri is currently an Associate Professor at King Saud University, Riyadh, Saudi Arabia since 2009, and he has published more than 80 research papers in international Journals, books and Conferences. Research Interests are Health Informatics. Dr El-Masri is a member of IEEE and IEEE Computer Society, Lebanese Order of Engineers, Saudi Association for Health Informatics and Australian Computer Society. He is an editorial member of the journal of engineering and technology research and the journal of health informatics in developing countries.

Red Blood Cell Recognition using Geometrical Features

Jameela Ali[1],AbdulRahim Ahmad[2], Loay E. George[3],Chen Soong Der[4],Sherna Aziz[5]
[1]College of Graduate Studies
[2,4]College of Information Technology
University Tenaga National -Malaysia
[3,5]Baghdad University-Iraq

Abstract

This paper presents the research on analysis and extraction of features of red blood cells for anemia recognition. Images from the blood samples collected at a hospital were used. Three geometrical features were used to distinguish between normal and anemic cells; Fourier descriptors, aspect ratio and moments. The City block distance measure was used as a criterion to determine the similarity degree between the tested samples and the established templates. Test results indicate that combination geometrical features gave high discriminative power approaching 98%.

*Keywords: Red blood cells, pattern recognition, Fourier descriptors, aspect ratio, moments***.**

1. Introduction

Pattern recognition is a field of research that examines the process and the design of systems to identify patterns in the data. Recognition system has emerged as a great challenge for computer vision. The longer term aim is to enable it to achieve near human level recognition for large number of categories under wide variety of conditions [1].

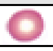
The red blood cell (RBC) recognition system can be used for educational purposes in medical schools and assist in the development of workers in the field of Hematology. RBCs come in a variety of shapes and textures, depending on the types of blood disease suffered by the patient. The variation is especially so in anaemia [2]. For example, the shape of a normal RBC is a biconcave disk, with 6 to 9 μm in diameter and 1.5 to 2.5 μm thick. In the peripheral smear of the sample slide, RBCs are a nucleate and contain predominantly haemoglobin that is distributed to form a dense outer rim with a paler centre that occupies approximately one third of the diameter of the cell[3]**,** The red color of blood is the result of a pigment called hemoglobin, which consists of iron and protein. Increase or decrease in the concentration of hemoglobin can result in different shapes, colors and sizes of the RBC and thus also can affect the textures [4][5].

Shape is one of the most important image features due to the fact that shape can effect human perception. Shape features has been extensively applied in RBC recognition to distinguish between normal cells and infected cells [6]. For infected cells, there are also many different shapes, four of which are; Sickle, Echnocyte, Teardrop and Ellipse which relate to four different types of anemia. Many shape representations and retrieval methods exists. However, most of those methods either do not well represent shape or are difficult to be normalized (making matching hard to do). One of the best methods is the one based on Fourier descriptors (FD). It achieved both representation and normalization well [10].

Textures another most widely used feature and has been an active topic in machine intelligence and pattern analysis since the 1950s. Texture features to discriminate different patterns of images by extracting the dependency of intensity between pixels and their neighboring pixels or by obtaining the variance of intensity across pixels [8]. Another three types of anemia are; Stomatocyte, Target and Hypochromic, also differentiated by different shapes of the cell. Moments-based texture analysis method has been introduced in medical images. Texture features are extracted by calculating moments in the texture pixels neighborhoods. Their capability to discriminate different textures has been verified by Wu [9]. In this research, we analyze and compare between eight different RBC shapes as show in table 1.

In this research, sample images of anemic and normal blood samples were collected and processed to obtain shape features (Fourier descriptors, moment and aspect ratio) to be used for training a recognition system. This paper is structured as follows; section II, discusses the proposed method. In the following section, the types of features are discussed followed by the discussion on experiments and the results in section IV. Finally, section V concludes.

Table1. Classification according to their shapes &textures

No.	Scientific cell name	Image cell
1.	Normal RBC	
2.	Teardrop RBC	
3.	Echnocyte RBC	
4.	Elliptocytes RBC	
5.	Sickle RBC	
6.	Target RBC	
7.	Stomatocytes RBC	
8.	Hypochromic RBC	

2. Methodology

The suggested RBC recognition scheme consists of three stages; (a) isolating target area, (b) determining geometrical features and finally to (c) recognizes the RBC as falling into one of the different types of anemia.
(a) The first stage involves image processing steps required to determine the background and target colors and to isolate the cell area (target) from the surrounding. Then, the external boundary pixels of the cell cut-out are traced.
(b) In the second stage the trace points are used to determine some adopted geometrical features like Fourier descriptors, aspect ratio and moments which have been used to describe the shapes of the blood cells.
(c) In the last stage, 100 different image samples for the normal and abnormal blood cells are used to train the recognizer to recognize 8 kinds of RBC, where 7are infected blood cells and one is of normal type. They are used as test materials to establish the template feature vector for each kind of blood cells. The City Block Distance (CBD) measure was used as a criterion to determine the similarity degree between the tested samples and the established templates that is found. This will permit the ordering of characteristic features and selection of the most useful features.

2.1 System layout

The proposed RBC recognition system is mainly designed to recognize anemia infected cells from normal cells, and recognize the types of anemia by using the extracted cell's features. The system is developed with the help and supervision of an expert (physician). It includes two main phases; training and recognition phase. Each phase involves three subs-stages; preprocessing, feature extraction and recognition. Fig. 1 shows the block diagrams of the training phase.

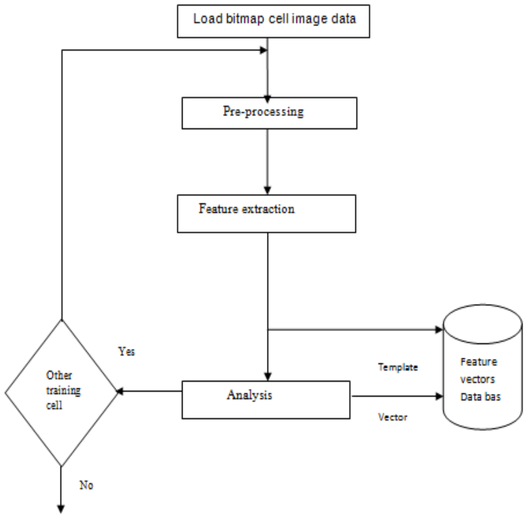


Fig (1) Block diagram of training phase

3. Feature Extraction

We focus on shape description because. Shape description cans be broadly categorized into two types, boundary based and region based. Boundary based methods use only the contour or border of the object shape and completely ignore its interior. Hence, these methods are also called external methods. The region based techniques take into account into account internal details (like holes etc) besides the boundary details. Recognition of a shape by its boundary is the process of comparing and recognizing shapes by analyzing the shapes 'boundaries [10];

3.1 Fourier Descriptors (FD):

Fourier Descriptors is most widely used in boundary based method [11]. The first set of features is Fourier

Descriptors(FD). In this technique, after retrieving RBCs binary shape images using Fourier Descriptors. Due to Fourier descriptors are used to describe the objects shape in terms of its spatial frequency content [12]

Fourier Descriptors based on the following Equation

$$c(n) = \frac{1}{m}\sum_{i=1}^{m-1}\left(\frac{\Delta xi}{\Delta yi} \cos \frac{2\pi ni}{L}\right)$$

$$s(n) = \frac{1}{m}\sum_{n=1}^{m-1}\left(a_n \frac{\Delta yi}{\Delta xi} \sin \frac{2\pi ni}{L}\right)$$

Where
m is the number of contour points
Δxi =xi-xi+1
Δyi =yi –yi+1
$\Delta\gamma i = \sqrt{(\Delta xi)^2 + (\Delta yi)^2}$

$L = \sum \Delta\gamma i$

(xi,yi) is the colume and row number of i thcountour point

Table (2) show the Fourier Descriptors of RBCs results.

3.1 MomentsInvariants(M):

Moments is one of complet of geometric in spatial domain in Region Based methods[13].moments is the second automated method for RBC image feature extraction .According to [14], Which was derived equations moments. A moments based on the following Equation

$$aMom(n) = \frac{1}{k}\left|\sum_{i \in c}\left|(x_i' + jy')^n\right|\right| \quad , \quad n = 1,2,3,4$$

Where,

$$x_i' = \frac{1}{L}\left[(x_i - \bar{x})cos\,\theta - (y_i - \bar{y})sin\,\theta\right]$$

$$y_i' = \frac{1}{L}\left[(x_i - \bar{x})sin\,\theta + (y_i - \bar{y})cos\,\theta\right]$$

$$\bar{x} = \frac{1}{m}\sum_{i \in c} x_i\,, \quad \bar{y} = \frac{1}{m}\sum_{i \in c} y_i$$

$$L = \min_{i \in c}\left(\left|x_i - \bar{x}\right|, \left|y_i - \bar{y}\right|\right)$$

Table (2) show a moment of RBC results

3.2 Aspect Ratio (AR).

The third and final method of automated feature extraction represents Aspect Ratio. Aspect Ratio is one of the most common examples of a Shape factor that represents quantities in shapes that have no dimensions used in image analysis[15].

$$ExtAspRat = \frac{(No\ of\ External\ Boundary\ Pixels)^2}{Total\ No\ of\ Cell\ Pixels}$$

Table (2) show Aspect Ratio of RBC results

Table (2) show the result of geometrics features

NO	Scientific cell name	Image cell	FD	AM	ER
1	Normal RBC		0.4971	0.01172	10.26
2	Teardrop RBC		0.3946	0.00268	19.11
3	Echnocyte RBC		0.3131	0.00802	41.19
4	Elliptocytes RBC		0.4111	0.00398	16.54
5	Sickle RBC		0.0540	0.00236	24.78
6	Target RBC		0.4955	0.01178	10.18
7	Stomatocytes RBC		0.5063	0.01101	11.86
8	Hypochromic RBC		0.4691	0.00687	13.12

4. City-Block Distance

The RBCs shape features extracted from the three methods above are presented to the City-Block Distance measurement for testing to make matching with the feature values in a reference database. In particular, City-Block Distance is a classifier that matches values of input features with values from features in a reference database. The use of City-Block Distance relies on four assumptions of distance function between points. For all points x, y, and z, a distance function ***D(x,y or z)*** satisfies the following properties: (a) Non-negativity: ***D(x, y) = 0***. (b) Reflexivity: ***D(x, y) = 0*** if and only if ***x = y***. (c) Symmetry: ***D(x, y) = D(y, x)*** and (d) Triangle inequality: ***D(x, y) + D(y, z) = D(x, z)* [16]**.

City-Block Distance achieves this classification and matching of image features by measuring the distance in between two connected pixels. Given two pixels at position ***(x_1,y_1)*** and ***(x_2,y_2)***, measurement by City-Block Distance can be expressed using function[16]

$$D_{12} = \left|x_1 - x_2\right| + \left|y_1 - y_2\right|$$

To assess the discrimination power of each adopted feature the following criteria was adopted

$$P_k = \frac{(n_c - 1)\sum_{j=1}^{n_c} \sigma(j,k)}{2\sum_{i=1}^{n_c - 1}\sum_{j=i+1}^{n_c}\left|\overline{F}(i,k) - \overline{F}(j,k)\right|}$$

Where

$$\overline{F}(j,k) = \frac{1}{n_j}\sum_{i=0}^{n_j-1} f(j,i,k)$$

$$\sigma(j,k) = \sqrt{\frac{1}{n_j}\sum_{i=0}^{n_j-1}(f(j,i,k) - \overline{F}(j,k))^2}$$

Where, nj is the number of training cell images belong to j-class, F (i,j, k) is the kth feature extracted from ith sample that belong to jth class, k=0,1,…….number of features. $\overline{F}(j,k)$ is the mean of kth feature for jth class. $\sigma(j,k)$ is the standard deviation of kth feature for jth class. Table (3) Discrimination power of adopted geometrical features.

Table (3) show discrimination poer of geometric features

NO	Featurs	discrimination power
1	MO	60.6%
2	AR	62%
3	FD	64%
4	FD,MO	86%
5	FD,AR	88%
6	MO,AR	86.5%
7	FD,MO,AR	98%

5. Experimental Results

The test results indicate that, for table (2) the cells divided into two groups to spherical and non-spherical according to (EAR) values. Where Target, Stomatocytes, and Hypochromic are spherical groups. Where value close to the normal cells value. Whilst Sickle, Echnocyte and Teardrop are non –spherical cells.

For table (3) first, discrimination power for FD better than AM and EAR respectively. Secondly collect together three features given high accuracy equal to98%.

6. Conclusion

Automated image-recognition systems provide significant benefits for medical test analysis. Since medical images are highly variants the development of reliable recognition processes is difficult. Blood recognition system is a difficult application in medical diagnoses, because the cells have several shapes, color and size. In this paper, we focus on three geometrical features (Fourier Descriptors, Moments and Aspect ratio) to extract features to 8 types of RBCs and used City-Block Distance method to distinguish between 8 different shapes. The results indicate that discrimination power of FD better than AM and EAR respectively and grouping the three features given high accuracy in discrimination power equal to98%.

References

[1] Woo Chaw Seng and Seyed Hadi Mirisaee, "A New Method for Fruits Recognition System", *MNCC Transactions on ICT*, 2009 Vol. 1, No. 1, June.

[2] Ferat Sahin, "A Radial Basis Function Approach to a Color Image Classification Problem in a Real Time Industrial Application", Master's thesis, Virginia polytechnic institute, Blacksburg, 1997.

[3] M. H. DE KEIJZER" Automated counting nucleated red blood cells in blood samples of newborns" Clin. Lab. Haem. 2002, 24, 343–345.

[4] F. Domino, Robert R. Sharp, Steven Lipper"NMR Chemistry Analysis of Red Blood Cell Constituents in Normal Subjects and Lithium-Treated Psychiatric Patients" BIOL PSYCHIATRY 1985; 20:1277-1283.

[5] Sveta Kabanova, Petra Klinbongard"Gene expression analysis of human red blood cells"International Journal of Medical Sciences, 2009,6(4):156-159.

[6] Vinay Saxena"Fourier descriptors under rotation, scaling, translation and various distortion forhand drew planar curves"Journal of Experimental Sciences, 20123(1): 05-07ISSN: 2218-1768.

[7] A.Karid, L.E.Nugroho"Acomprative experiment of several shape methods in recognizing blants"international journal of computer science & information technology, June 2011, (IJCSIT),Vol3, No3.

[8] Fabian Timm, Thomas Martinetz"Statistical Fourier Descriptors for Defect Image Classification" international conferanc of pattren recognition Ratzeburger Allee, 2011, 160, 23538 L¨ubeck, Germany.

[9] Ke Wu, Carole Garnier" A preliminary study of moment-based texture analysis for medical images"Conf Proc IEEE Eng Med Biol Soc. 2010: 5581–4.doi: 10.1109/IEMBS.2010.5626789.

[10] Zhang, D., & Guojun, L. A Comparative Study on Shape Representation Using Fourier Descriptors with Different Shape Signatures. Churchil: Australia 2009.

[11] Persoon, E., & Fu, K. Shape Discrimination Using Fourier Descriptors. IEEE Trans. On Sytems, Man andCybernetics, 7(3) 1995.

[12] Loay E. George,Shatha M. Noor"Handwritten Arabic (Indian) Numerals Recognition Using Fourier Descriptorand Structure Base Classifier"Journal of Al-Nahrain University Vol.14 (2), pp.215-224, June, 2011

[13] Eakins and M.Graham " contour-based image retrieval " JTAP, 1999, Generic,89.

[14] JAMEELA ALKRIMI," Medical Diagnosis by Analyses of Blood Cell Image" Master theses 2006.

[15] Tieng, M., Q., & Boles, W., W. Recognition of 2D Object Contours using the Wavelength Transform Zero-Crossing Representation IEEE Trans,1997, 19(8).

[16] International Association for Pattern Recognition. (2003), what is Pattern Recognition. IAPR Newsletter, 25(1), 1.

Spatial Filtering Applications from Medical Images to 2D Turbulence Using the Fourth-Order and Shock PDEs Methods in Complex Domain

Tamer Nabil^{1,2} and Waleed Abdel Kareem^{1,3}

^{1} Mathematics Department, Faculty of Science, King Khalid University, Abha, Saudi Arabia

^{2} Basic Science Department, Faculty of Computes and Informatics, Suez Canal University, Ismailia, Egypt

^{3} Mathematics Department, Faculty of Science, Suez University, Suez, Egypt

Abstract

The complex fourth-order as well as the complex shock partial differential equations (PDEs) is introduced for noise removal from medical images and 2D turbulent flow. The Lattice Boltzmann method (LBM) with a single relaxation model is used to obtain the velocity field of the turbulent flow. The two filtering methods are applied against the vorticity field of the flow. Comparisons between the results of the two methods for medical images and 2D turbulence are extensively studied. Investigation and identification of the filtering parameters are also considered. It is shown that the proposed filtering methods are effective for noise removal in both applications. Results indicate that the complex fourth-order PDE method extracts the coherent and incoherent parts more clearly compared with the shock method.

Keywords: *Complex fourth-order and shock PDEs filtering methods, medical images, 2D turbulence*

1. Introduction

Filtering can be considered as one of the most important problems in signal and image processing as well as for studies of turbulent flow. The main objective of a filtering method is to extract the original image from the noisy one. In turbulence studies, the filtering method divides each turbulent flow field into two parts: one is the organized coherent part and the second is the randomly distributed incoherent part.
A large number of filtering methods has been used for denoising, such as wavelet-based filtering that employs nonlinear thresholding [1-4], Curvelets [5, 6], total variation [7,8] and non-local mean filtering[9,10].

In recent years, PDEs start to play an important role for data filtering. Most PDEs filtering methods focused on parabolic equations. Osher and Rudian [11] proposed a hyperbolic equation, called shock filter that can serve as a stable deblurring algorithm approximating deconvolution. Alvarez and Mazorra [12] were the first to couple shock and diffusion for noise elimination and edge enhancement. Gilboa et al. [13] developed shock filter by adding a complex diffusion term to the shock equation. This new term is used to smooth out noise and indicate inflection points simultaneously. The imaginary value which is an approximated smooth second derivative that scaled by the time was used to control the process. They proved that, the results of this algorithm are robust on removing the signal noise. On the other hand, various approaches for filtering noise based on PDEs have been proposed and they are based on second order PDEs and scale space analysis. The methods include anisotropic diffusion equation [14] and a curve evolution equation that is based on geometric heat flow of the level sets of the data [15, 16]. You and Kaveh [17] found that these methods have been unable to achieve a good trade-off between noise removal and features preservation but they tend to cause the processed to look "blocky". They proposed a class of fourth-order PDEs to optimize the trade-off between noise removal and edge preservation.

The time evolution of the PDEs seeks to minimize a cost functional which is an increasing function of the absolute value of the Laplacian of the image intensity function. In a planar image, the Laplacian of the image tends to zero in its neighborhood and hence these PDEs can remove noise and preserve edges by approximating an observed image with a piecewise planar image.
Lysaker et al. [18] introduced a new method for image smoothing based on a fourth-order PDE model. The model is tested on a broad range of the real medical resonance image both in space and time, as well as on non medical synthesized images. This algorithm demonstrates good noise suppression without destruction of important

anatomical or functional details even at poor signal-to-noise ratio.
Rajan et al.[19] extended the second order nonlinear complex diffusion to fourth order complex PDE which produced a much better results.

There are two aims for this paper: one is to compare and study the performance of the complex shock filter and complex fourth-order PDE filter for filtering medical images. The second is to employ the two proposed methods for coherent vortex extraction from 2D homogeneous isotropic turbulence. To our knowledge, many literatures use wavelet theory analysis [20, 21] for filtering turbulent flow. In wavelet filtering, the total flow is divided into two parts, namely the coherent and incoherent parts. In this paper, we introduce the filtering methods using PDEs in spatial domain rather than transformation of the test data into the frequency domain. This spatial filtering may reduce the numerical errors that occur during the Fourier and wavelet transformations and their inverse. This paper is organized as follows. Sec. 2 discusses the proposed filtering methods and their implementations. Sec. 3 is devoted to discus the applications of the filtering methods to medical images and 2D turbulence. Results and discussion of the obtained results are introduced in sec.4 and finally sec.5 shows the conclusion of the results.

2. Proposed Filtering Methods

Let $Q(\vec{x})$ be a digital image and $Q_0(\vec{x})$ be its observation with random noise $\psi(\vec{x}), \forall x, y \in \Omega$. The noise is superimposed on the pixel intensity value by the formula

$$Q_0(\vec{x}) = Q(\vec{x}) + \psi(\vec{x}) \qquad (1)$$

Assume the noise level is approximately known, i.e.

$$\|Q_0(\vec{x}) - Q(\vec{x})\|^2_{L^1(\Omega)} = \int_\Omega (Q_0(\vec{x}) - Q(\vec{x}))^2 d\vec{x} = \sigma^2 \qquad (2)$$

Since noise can be recognized as fast oscillating signals over small areas, the important idea for denoising is to filter out high frequency signals while preserving the important features in the images.

2.1 The Complex-Shock Filtering Method

The complex shock filter for 2D data can be written as [13]

$$\frac{\partial Q}{\partial t} = -\frac{2}{\pi}\arctan(A\,\mathrm{Im}(\frac{Q}{\theta}))\|\nabla Q\| + \lambda Q_{\eta\eta} + \mu Q_{\xi\xi} \qquad (3)$$

With the initial condition

$$Q(\vec{x},0) = Q_0 \quad , \qquad (4)$$

and the boundary condition

$$\frac{\partial Q}{\partial \vec{N}} = 0 \quad \text{on} \quad (\partial\Omega) \quad , \qquad (5)$$

where the complex coefficient $\lambda = re^{i\theta}$ depends on the choice of the polar coordinates *r* and θ, μ is a real scale, *A* is a parameter that controls the sharpness of the slop near zero, $\vec{N}$ represents the perpendicular direction to the boundary $\partial\Omega$ of the image, $\eta = \eta(\vec{x})$ is the direction of ∇Q and $\eta = \frac{\nabla Q}{\|\nabla Q\|}$, ξ is the normal vector to η and $Q_{\eta\eta}, Q_{\xi\xi}$ are the second derivative in the direction of η and ξ respectively.

The condition $\frac{\partial Q}{\partial \vec{N}} = 0$ means that we minimize the boundary influence. The main properties and advantages of the shock filter for noise removal and edge enhancement can be found in [12]. The Numerical implementation of this model is based on the finite difference method. The finite difference method may be used to solve the PDE of the model by applying the iterative approach as follows.

Assuming Δt is the time step, and the space girds size $\Delta x = \Delta y = h = 1$ then

$$t = n\Delta t, \quad n = 0,1,2,...... \qquad (6)$$

$$x = ih, \quad i = 0,1,2,....., I \qquad (7)$$

and

$$y = jh, \quad j = 0,1,2,....., J \qquad (8)$$

where $I \times J$ is the size of the image.

$$Q^{n+1}_{i,j} = Q^n_{i,j} + $$
$$+ \Delta t(-\frac{2}{\pi}\arctan(A\,\mathrm{Im}(\frac{Q^n_{i,j}}{\theta}))\sqrt{\left\|\tilde{D}_x Q^n_{i,j}\right\|^2 + \left\|\tilde{D}_y Q^n_{i,j}\right\|^2} + $$
$$+ \lambda D^2_\eta Q^n_{i,j} + \mu D^2_\xi Q^n_{i,j}) \qquad (9)$$

Where $\tilde{D}_x, \tilde{D}_y$ are the first order symmetric approximations in *x* and *y* respectively. They are defined as,

$$\tilde{D}_x Q^n_{i,j} = \min(\left|Q^n_{i+1,j} - Q^n_{i,j}\right|, \left|Q^n_{i,j} - Q^n_{i-1,j}\right|) \qquad (10)$$

$$\tilde{D}_y Q^n_{i,j} = \min(\left|Q^n_{i,j+1} - Q^n_{i,j}\right|, \left|Q^n_{i,j} - Q^n_{i,j-1}\right|) \qquad (11)$$

and

$$D_{\eta}^{2}Q = Q_{\eta\eta} = \frac{Q_{xx}\|Q_x\|^2 + 2Q_{xy}Q_xQ_y + Q_{yy}\|Q_y\|^2}{\|Q_x\|^2 + \|Q_y\|^2 + \varepsilon} \quad (12)$$

where $\varepsilon = 10^{-6}$ is used to avoid division by zero.

$$D_{\xi}^{2}Q = Q_{\xi\xi} = \frac{Q_{xx}\|Q_x\|^2 - 2Q_{xy}Q_xQ_y + Q_{yy}\|Q_y\|^2}{\|Q_x\|^2 + \|Q_y\|^2 + \varepsilon} \quad (13)$$

where

$$Q_{xx}^{n} = Q_{i+1,j}^{n} - 2Q_{i,j}^{n} + Q_{i-1,j}^{n} \quad (14)$$

$$Q_{yy}^{n} = Q_{i,j+1}^{n} - 2Q_{i,j}^{n} + Q_{i,j-1}^{n} \quad (15)$$

$$Q_{xy}^{n} = \frac{(Q_{i,j}^{n})_x - (Q_{i,j-1}^{n})_x}{2} \quad (16)$$

with the symmetric boundary conditions

$$Q_{-1,j}^{n} = Q_{0,j}^{n}, \quad Q_{I+1,j}^{n} = Q_{I,j}^{n}, \quad j = 0,1,....., J \quad (17)$$

$$Q_{i,-1}^{n} = Q_{i,0}^{n}, \quad Q_{i,J+1}^{n} = Q_{i,J}^{n}, \quad i = 0,1,....., I \quad (18)$$

The scheme is convergent if $\Delta t \le 0.25\frac{\cos\theta}{r}$,[13]. For *r=1* the convergence condition becomes $\Delta t \le 0.25\cos\theta$. Gilboa [13] proved that $\theta = \frac{\pi}{1000}$ gives the best result according to experimental tests. In order to be more closely to analytic PDEs, a smaller time step may be used (according to convergent condition) at the beginning of the evolution and θ can be set to a very small value. The iterative scheme is ended if the following condition is satisfied

$$\|Q^{n+1} - Q^{n}\| \le 10^{-4} \quad (19)$$

2.2 The Complex Fourth-Order PDE Filtering Method

A Fourth –order PDE filtering method can be written as [17]

$$\frac{\partial Q}{\partial t} = -\nabla^2(c(\|\nabla^2 Q\|)\nabla^2 Q) \quad (20)$$

and the same initial and boundary conditions (4) and (5) are considered. Here the function *c(.)* is a positive and non-increasing function and it is defined by

$$c(\|\nabla^2 Q\|) = \frac{1}{1 + (\frac{\|\nabla^2 Q\|}{k})^2} \quad (21)$$

where *k* is a constant (sometimes called the flow constant or the soft threshold). The complex version of the this model is

$$\frac{\partial Q}{\partial t} = -\nabla^2(\tilde{c}(\mathrm{Im}(Q))\nabla^2 Q) \quad (22)$$

Where

$$\tilde{c}(\mathrm{Im}(Q)) = \frac{e^{i\theta}}{1 + (\frac{\|\mathrm{Im}(Q)\|}{k\theta})^2} \quad (23)$$

θ also represents the phase angle and Im(*Q*) is the imaginary part of the data. This partial differential equation can be solved numerically using the finite difference with the same methodology mentioned in Eqns.6-8 and 17-18. The scheme can be written as

$$\nabla^2 Q_{i,j}^{n} = Q_{i+1,j}^{n} + Q_{i-1,j}^{n} + Q_{i.j+1}^{n} + Q_{i,j-1}^{n} - 4Q_{i,j}^{n} \quad (24)$$

The numerical implementation of this method can be summarized and simplified as follows. First consider the function

$$g_{i,j}^{n} = (\tilde{c}(\mathrm{Im}(Q_{i,j}^{n}))\nabla^2 Q_{i,j}^{n}) \quad (25)$$

under the symmetric boundary condition

$$g_{-1,j}^{n} = g_{0,j}^{n}, \quad g_{I+1,j}^{n} = g_{I,j}^{n}, \quad j = 0,1,....., J \quad (26)$$

and

$$g_{i,-1}^{n} = g_{i,0}^{n}, \quad g_{i,J+1}^{n} = g_{i,J}^{n}, \quad i = 0,1,....., I \quad (27)$$

Finally, the numerical scheme can be finalized in the form

$$Q_{i,j}^{n+1} = Q_{i,j}^{n} - \Delta t \nabla^2(g_{i,j}^{n}) \quad (28)$$

3. Applications of the Filtering Methods

The applications of the two filtering methods are applied against 2D medical images as well as 2D homogeneous turbulence. First the methods are applied to 2D medical images and the important statistical parameters are estimated. Then the estimated parameters are examined in 2D turbulent data. In the following section, the parameters estimation will be considered then in the next section the 2D homogeneous turbulence data will be discussed.

3.1 Filtering Parameters estimation

The filtering methods are examined against several medical test images and here three test images are chosen with a size of 128^2 (Fig.1(a), Fig.2(a) and Fig.2(c)). A Gaussian noise with a standard deviation $\sigma = 20$ is additively considered in the original image (test image 1) as shown in Fig. 1(b). Also, Fig.2(b) and Fig.2(d) show the second and third test noisy images with standard deviations $\sigma = 10$ and $\sigma = 15$, respectively. Fig.3 shows another test data taken from 2D turbulent flow. The contours show the total vorticity of the turbulent flow.

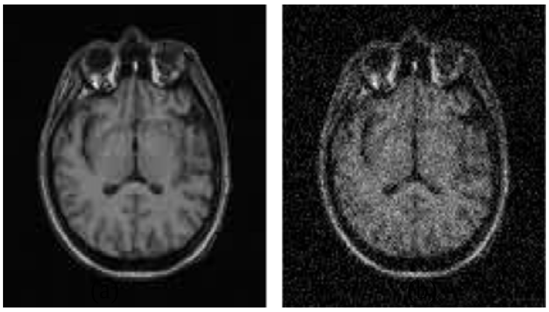
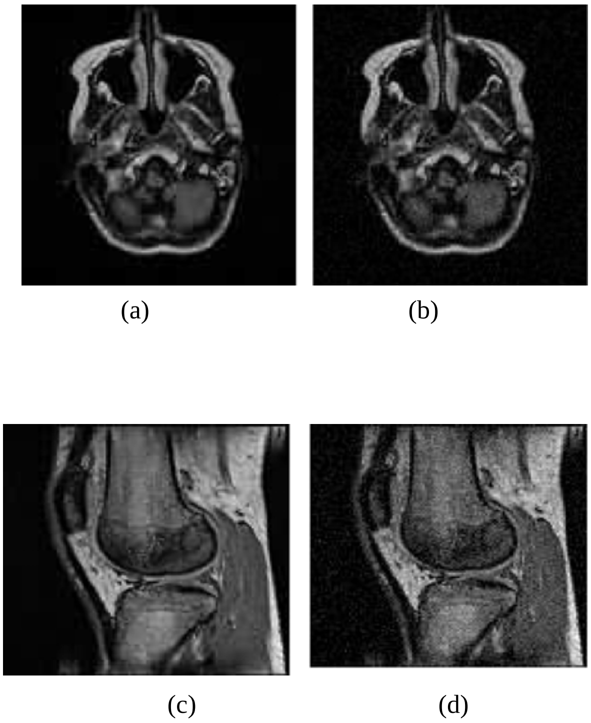


Figure 1: First test image (a) Original image (b) Noisy image



(a) (b)

(c) (d)

Fig.2: The second and third test image (a) Original second image (b) Noisy image ($\sigma = 10$) (c) Original third image (d) Noisy image ($\sigma = 15$)

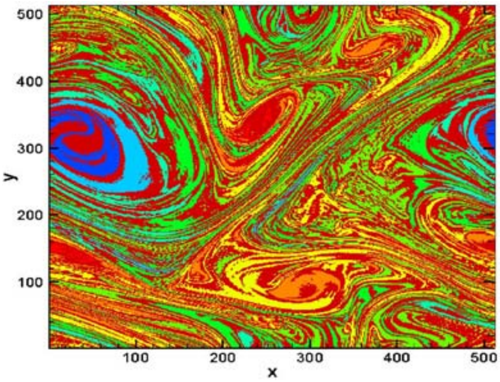


Fig.3: Total vorticity

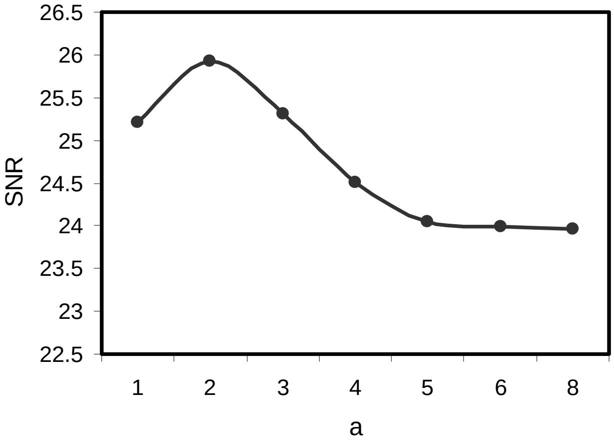


Fig. 4: Complex shock results for the parameter A-estimation

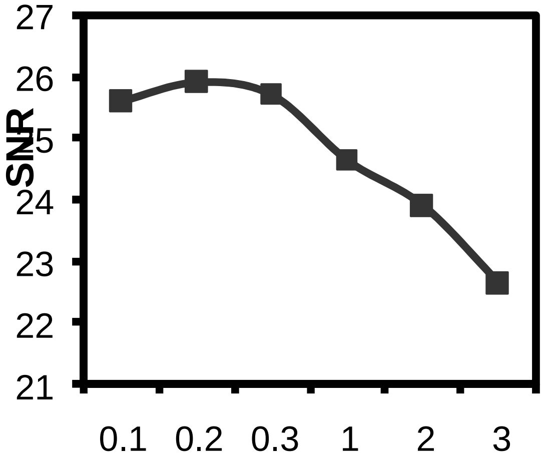


Fig.5: Complex shock results for the parameter µ-estimation

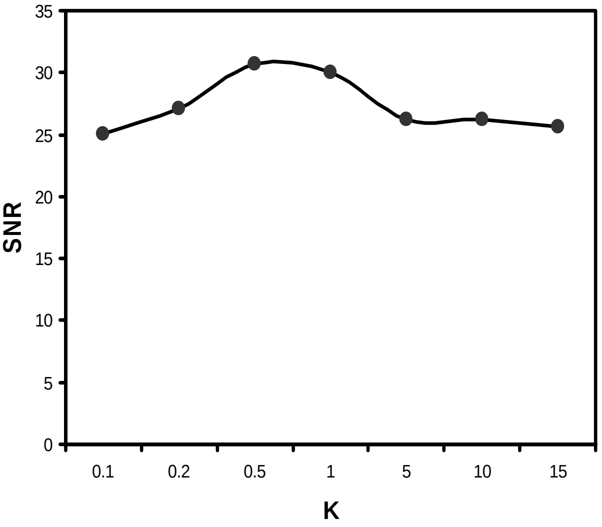


Fig.6: Complex fourth- order results for the parameter *K*-estimation.

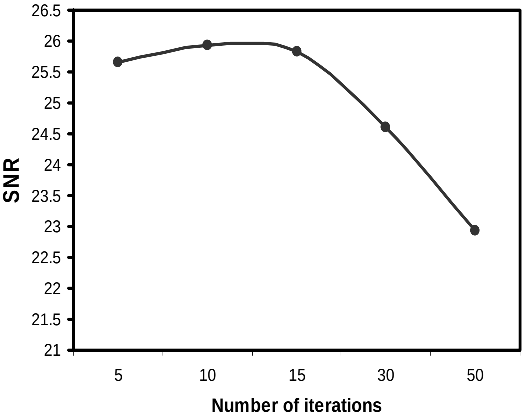


Fig.7: Complex shock results for the number of iterations *n*-convergence.

Several values of the parameters are tested in the filtering equations to reach the best values of the signal-to-noise-ratio (SNR) which is defined as

$$SNR = 10\log\left(\frac{\sum_{i=1}^{N} Q_i^2}{\sum_{i=1}^{N} (\hat{Q}_i - Q_i)^2}\right) \quad \textbf{(29)}$$

where *N* is the image size. The parameters used in the complex shock method are the phase angle θ, the polar radius *r*, the constants *A* and μ. The choices of

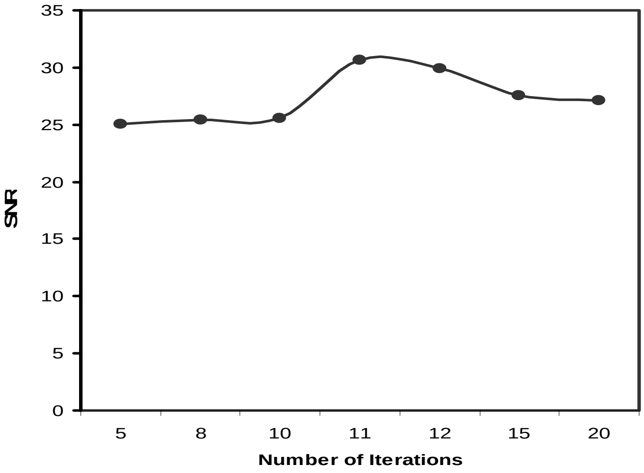


Fig.8: Complex fourth-order results for the number of iterations n convergence

$\theta = \dfrac{\pi}{1000}$ and *r=1* are based on the study introduced in [13]. The parameters *A* and μ and the corresponding SNR values are shown in Fig.4 and Fig.5, respectively. It is clear that *A=2* and $\mu = 0.2$ gives the best SNR values. Also, the soft threshold coefficient *k* in the complex fourth-order method is chosen as *k=0.5*. Fig.6 shows that at *k=0.5*, the best value of *SNR* can be reached. The number of iterations *n* is depicted in Figs. 7 and 8 against the *SNR* values for the complex shock and complex fourth-order methods, respectively. In the complex shock method the number of iterations *n=10* leads to the best estimation of SNR as well as it satisfies the convergence condition sated in Eq.19. In the case of the complex fourth-order the value is chosen as *n=11* for the same reasons. Finally, the time step is chosen as $\Delta t = 0.1$ in the two filtering methods for the stability condition [13].

3.2 2D Homogeneous Turbulence

The lattice Boltzmann method (LBM) is used for simulations of 2D and 3D decaying homogeneous turbulence (e.g. Xu et al.[22] and Abdel Kareem[23]). In this paper the LBM method is used to investigate the 2D forced turbulence. The velocity data u_x and u_y of a 2D turbulent flow are generated using the *D2Q9* single relaxation time (*SRT*) model where, the LBM equation can be written

$$f_\alpha(x+e_\alpha\delta t, t+\delta t) - f_\alpha(x,t) = -\frac{1}{\tau}\left(f_\alpha(x,t) - f^{eq}{}_\alpha(x,t)\right) + 3\rho\omega_\alpha(e_\alpha \bullet F), \tag{30}$$

where $f_\alpha^{eq}(x,t)$ is the equilibrium distribution function. The discrete velocity set and the respective weighting coefficients are e_α and ω_α , respectively. The force $F = F_x\hat{i} + F_y\hat{j}$ is defined as follows

$$F_x = A\sin(K_y y + \phi) \tag{31}$$

and

$$F_y = A\sin(K_x x + \phi) \tag{32}$$

Here ϕ is the random phase and ***A*** is the forcing amplitude. The equilibrium distribution function is defined as4

$$f^{eq}{}_\alpha(x,t) = \omega_\alpha\rho\left[1 + 3(e_\alpha \bullet u) + \frac{9}{2}(e_\alpha \bullet u)^2 - \frac{3}{2}(u \bullet u)\right], \tag{33}$$

The mathematical definitions of the discrete velocity set and the corresponding weighting coefficients are

$$e_\alpha = \begin{cases} (0,0,0), & \alpha = 0 \\ (1,0),(0,1),(-1,0),(0,-1) & \alpha = 1-4 \\ (1,1),(-1,1),(-1.-1),(1,-1) & \alpha = 5-8 \end{cases}$$

and

$$\omega_\alpha = \begin{cases} \frac{4}{9}, & \alpha = 0 \\ \frac{1}{9}, & \alpha = 1-4 \\ \frac{1}{36}, & \alpha = 5-8 \end{cases}$$

The simulations are done in a square computation domain with a resolution of 512×512 The vorticity is calculated from the velocity using the mathematical expression $\nabla \times \vec{u}$ and hence the vorticity can be calculated as $\omega = \frac{\partial u_x}{\partial y} - \frac{\partial u_y}{\partial x}$. The filtering methods are applied against the vorticity in the spatial domain. The number of iterations used in the two methods is fixed and it is chosen as *n=10*. This number of iterations leads to a good convergence.

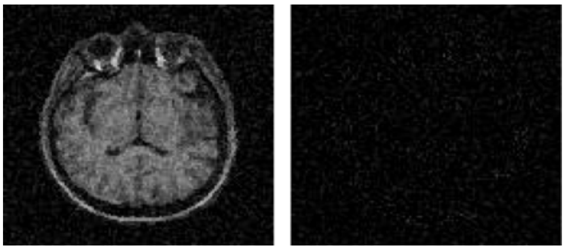
4. Results and Discussion

The shock and fourth-order PDEs filtering methods are applied to different types of data. The first data type is taken from image processing problems and the second type is taken from a study of turbulent flow. The applications of these methods to turbulent flow indicate that moving from image processing filtering to turbulence filtering can be achieved and lead to reasonable and important physical results. These physical results may help understanding important features of turbulence. The important results of the filtering methods can be discussed as follows.

4.1 Complex Shock and Fourth-Order PDEs Filtering Results

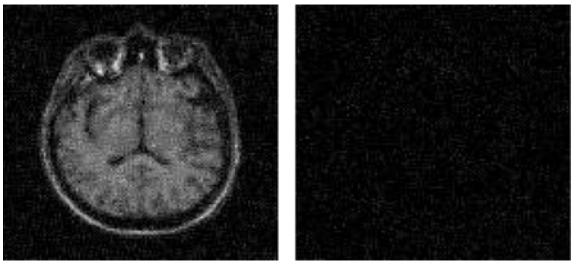
Fig.9 shows the denoised and noise image parts using the complex shock method for the first test image. Fig.10 shows the results extracted by the fourth-order method for the first test image, where the denoised and noise image parts are depicted. It can be observed that the extracted denoised image resemble the original image and the noise part doesn't contain any features from the original image. It can be also observed from the results that the method preserves the important features of the original image and isolates the noise from the image in a reasonable way. Figs. 11 and 12 show the filtering results for the second and third test images, respectively. Fig.11(a) shows the denoised image that extracted using the complex shock filtering method. Fig.11(c) shows the denoised image that extracted using the complex fourth-order method. Figs.11(b) and 11(d) show the corresponding extracted noisy parts. Also, Fig.12(a) and (c) show the denoised parts extracted by the complex shock and complex fourth-order, respectively. The removed noisy parts are shown in Figs.12(b) and 12(d), respectively. It is clear that the noisy parts don't include important features from the original images. For turbulent flow, Fig.13 shows the extracted coherent and incoherent parts of the flow field using the complex shock method. It can be observed that the coherent field is similar to the total vorticity field which is depicted in Fig.3. The incoherent part is smoothly distributed along the square region and no coherent regions can be observed.

For the complex fourth-order results that depicted in Fig.14, the coherent and incoherent parts of the flow field are also extracted smoothly. The coherent field is found similar to the total vorticity field and the incoherent part is smoothly distributed along the square region and no coherent regions can be observed.



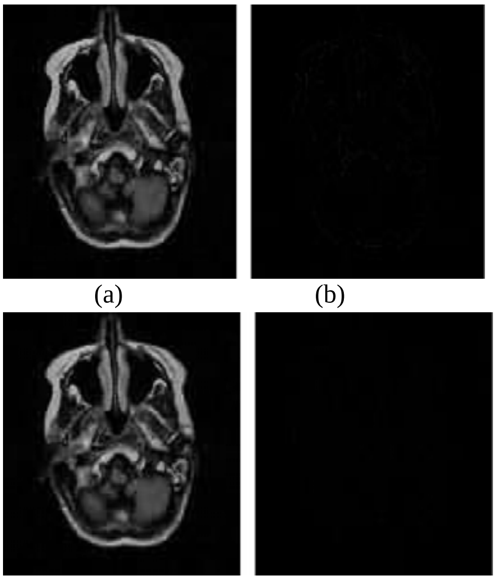
(a) (b)

Fig.9: Complex shock results for the test image: (a) denoised (b) noise



(a) (b)

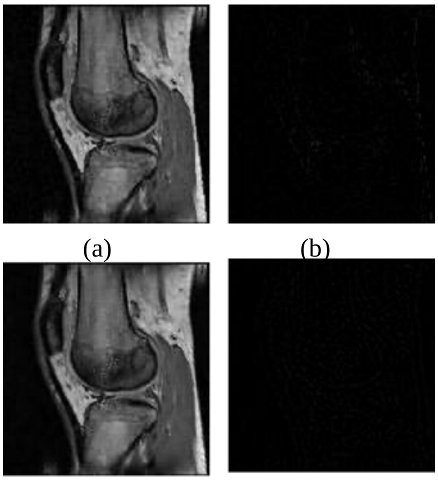
Fig. 10: complex fourth-order resu;ts for the test image: (a) denoised (b) noise.



(a) (b)

(c) (d)

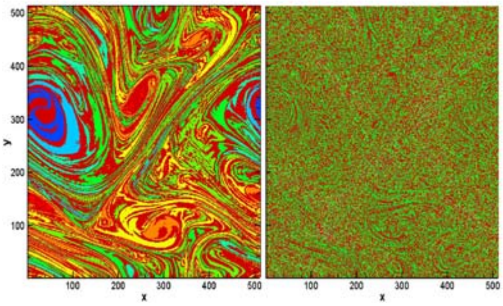
Fig.11: Results for the second test images: (a) denoised by complex shock (b) noise (c) denoised by complex fourth (d) noise



(a) (b)

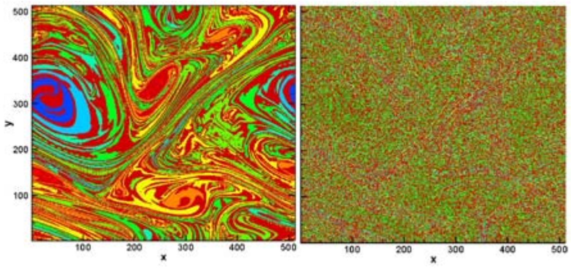
(c) (d)

Fig.12: Results for the third test images: (a) denoised by complex shock (b) noise part (c) denoised by complex fourth (d) noise



(a) (b)

Fig.13: Complex shock results for turbulent flow: (a) coherent part (b) incoherent part



(a) (b)

Fig.14: Complex fourth-order results for turbulent flow: (a) coherent part (b) incoherent part

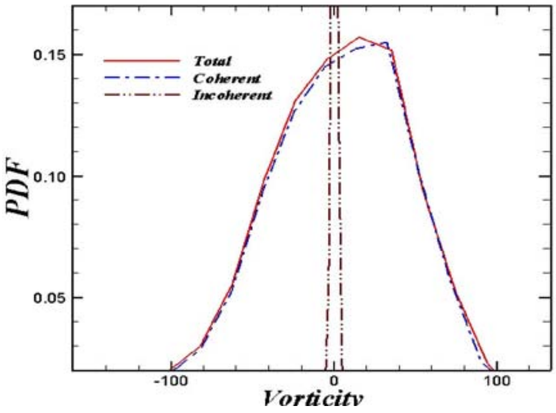


Fig.15: Complex shock: Probability density function (PDF) for the total vorticity, coherent and incoherent parts.

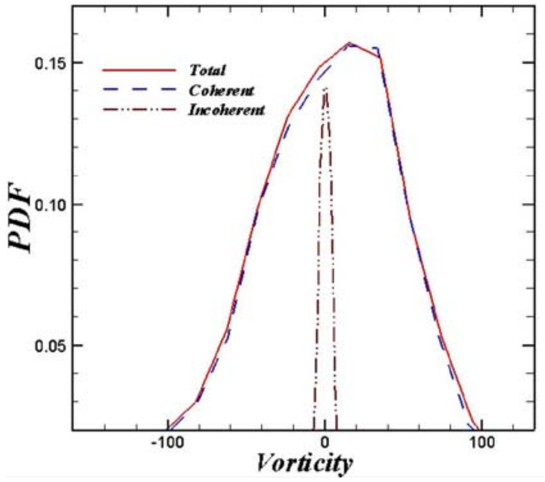


Fig. 16: Complex fourth order: Probability density function (PDF) for the total vorticity, coherent and incoherent parts.

4.2 Analysis and Comparisons of the Filtering Methods

The *SNR* in the medical image application is found as 25.927 and 30.6712 for the complex shock and complex fourth-order methods, respectively. The noisy image SNR was found as 21.135 which indicate that the two methods are succeeded in the noise removal process. The higher value of the complex-fourth-order *SNR* indicates that the complex fourth-order is more efficient than the complex shock method. The following tables show the *SNR* values estimated for the three test images at four different noisy standard deviations.

Table 1: *SNR* for the first test image

σ	*Noisy image*	*Complex shock*	*Complex fourth-order PDE*
5	35.9874	36.4274	36.9324
10	29.9818	30.8293	31.3778
15	26.2685	28.3086	29.514
20	21.135	25.927	30.6712

The results in the tables support that the fourth-order method is superior to the complex shock filtering method. The SNR values are found larger in all test image cases even at different values of the standard deviation.

Table 2: *SNR* for the second test image

σ	*Noisy image*	*Complex shock*	*Complex fourth-order PDE*
5	36.231	37.15	39.526
10	33.1873	35.4226	37.3295
15	26.151	27.0219	28.7613
20	23.2150	26.518	29.9814

For the turbulent flow, it can be observed that the complex fourth-order coherent part is more smoothly compared with the complex shock results. It can be observed that some coherent vortices are smoothly visualized using the complex fourth-order results.

Table 3: *SNR* for the third test image

σ	*Noisy image*	*Complex shock*	*Complex fourth-order PDE*
5	35.0352	36.9761	37.7521
10	32.3156	33.6213	34.8191
15	25.5096	27.0552	29.2892
20	22.5316	28.0532	31.9731

Also, the incoherent part in the complex fourth-order is distributed smoothly in the region. However, some vortical regions, though it is very small, can be observed in the complex shock filtering results. The probability density functions (PDFs) for the total vorticity, coherent and incoherent parts are shown in Figs. 15 and 16. Fig. 15 shows the PDF for the shock results and Fig. 16 shows the results for the fourth-order results. It can be observed that in both cases, the coherent part is almost similar to the total vorticity. The PDF for the incoherent part is Gaussian in both cases. There is a difference between the PDF for the incoherent parts, where in the shock-case a Gaussian PDF with very weak tails can be observed. The PDF for the fourth-order incoherent part is larger than the shock result because in the fourth-order case many noisy regions are extracted without affecting the vortical regions.

5. Conclusions

The complex shock and the complex fourth-order PDEs filtering methods are proposed to extract coherent and incoherent parts of some important experimental data. The test data are of two different types. One of the data set corresponds to medical applications and the second corresponds to turbulent flow. It was shown that the two methods can extract important features in both applications. It was also statistically shown that the complex fourth-order method is superior to the shock method. The characteristics of the extracted coherent and incoherent parts are found similar to previous efforts using the wavelet decompositions. The coherent part is found similar to the non-filtered field and the incoherent part is structuresless. The fourth-order method smoothly extracted the coherent vortices and removed the incoherent background without affecting the geometrical shapes of the vortices.

Acknowledgments

The authors thank the support of King Khalid university (KKU)-Project ID:KKU-SCI-011-008.

References

[1] D. Donoho and I. Johnston, "Ideal spatial adaptation via wavelet shrinkage ", Biometrika, Vol. 91, 1994, pp. 425-455.

[2] D. Donoho, "De-noising by soft thresholding ", IEEE Trans. Inform. Theory, Vol. 41, 1995, pp. 613-627.

[3] I.A. Ismail and T.Nabil, "Applying wavelet recursive translation invariant to low-pass filtered images", Int. J. Wavelets, Multiresolution Inf. Process, Vol. 2, 2004, pp. 99-110.

[4] Z. Liu, J. Tian, L. Chen and Y. Wang, " Wavelet-based image denoising variance field diffusion ", Optics Communications, Vol. 285, 2012, pp. 1744-1747.

[5] J. Starch, E.Candes and D. Donoho, " The curvelet transform for image denoising ", IEEE Trans. Image Process., Vol. 11, 2000, pp. 670-684.

[6] A.Ali, P.Swami and J. Singha, " Modified curvelet thresholding algorithm for image denoising ", J. Comput. Sci., Vol. 6, 2010, pp. 18-22.

[7] L.Rudin, S.Osher and E. Fatemi, " Nonlinear total variation based noise removal algorithms ", Physica D, Vol. 60, 1992, pp. 259-288.

[8] T. Goldsten and S. Osher, "The split Bregman method for L1 regularized Problems", SIAM J. Imaging Sci., Vol. 2, 2009, pp. 323-343.

[9] A. Buades, B. Coll and J. Morel, " A nonlocal algorithm for image denoising ", Proc. IEEE Int. Conf. On Computer Vision and Pattern Recognition, Vol. 2, 2005, pp. 60-65.

[10] M. Mahmoudi and G. Sapiro, " Fast image and video denoising via nonlocal mean of similar neighborhoods ", IEEE Signal Process. Letters, Vol. 12, 2005, pp. 839-842.

[11] S. Osher and L. Rudin, " Feature-oriented image enhancement using shock filters ", SIAM J. Num. Anal., Vol. 27, 1990, pp. 919-940.

[12] L. Alvarez and L. Mazorra, " Signal and image restoration using shock filters and anisotropic ", SIAM J. Num. Anal, Vol. 31, 1994, pp590-605.

[13] G. Gilboa, N. Sochen and Y. Zeevi, " Image enhancement and denoising by complex diffusion process ", IEEE Trans. On Pattern Anal. and Mach. Intell., Vol. 25, 2004, pp. 1020-1036.

[14] P. Perona and J. Malik, " Scale-space and detection using anisotropic diffusion ", IEEE Trans. On Pattern Anal. and Mach. Intell., Vol. 12, 1990, pp. 629-639.

[15] B. Kimia, A. Tannenbaum and S. Zucker, " On the evolution of curve via a function of curvature I ", J. Math. Anal. and Appl., Vol. 163, 1992, pp. 438-458.

[16] G. Sapiro and A. Tannenbaum, "Affine invariant scale-space" , Int. J. Comput. Vis., Vol.11, 1993, pp. 25 - 44.

[17]Y. You and M. Kaveh, "Fourth-order partial differential equations for noise removal", IEEE Trans. Image Process., Vol.9, 2000 , pp.1723 - 1730.

[18] M. Lysaker, A. Lundervold and X. Tai, "Noise removal using fourth-order partial differential equation with applications to medical magnetic resonance image in space and time ",IEEE Trans. Image Process., Vol.12, 2003 , pp.1579 - 1590.

[19] J. Rajan, B. .Jeurissen and J. Kannan, " Denoising magnetic resonance images using fourth order complex diffusion ", IEEE Int. Mach. Vis. and Image Process. Conf., 2009, PP.123 - 127.

[20]M. Farge, K. Schneider and N. Kevlahan, " Non-gaussianity and coherent simulation for two-dimensional turbulence using an adaptive orthonormal wavelet basis ", Phys. Fluids, Vol. 11,1999, pp.2187 - 2201.

[21] M. Farge and K. Schneider , "Coherent vortex simulation (CVS), a semi-deterministic turbulence model using wavelet ", Flow, Turbulence, and Combustion, Vol.66, 2001, PP.393 - 426.

[22] H. Xu, Y. Qian and W. Tao, “Revisiting two-dimensional turbulence by Lattice Boltzmann Method ", Progress in Computational Fluid Dynamics, Vol.9, 2009,pp.133 - 140.

[23] W. Abdel Kareem, “Tracking of vortical structures in three-dimensional decaying homogeneous turbulence ", Int. J. Modern Physics C, Vol.22, 2011, pp.1373 – 1391.

Tamer Nabil received the B.Sc. degree in Mathematics from Faculty of Science, Helwan University at Cairo, Egypt, in 1997, the M.Sc. degree in pure Mathematics from Helwan university at Cairo, Egypt, in 2000, and Ph.D. degree in Computational Mathematics from Suez Canal University at Ismailia, Egypt, in 2005. Currently, he is an Assistant Professor in Mathematics Department, faculty of Science, King Khalid University, Abha, Saudi Arabia. Also, his permanent work is Assistant Professor at Basic Science Department, Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt. His research interests are computational harmonic analysis, Numerical methods in Fluid mechanics, and Mathematical methods in image analysis.

Waleed Abdel Kareem received his B.Sc. (1996) and M.Sc. (2001) in Mathematics from Menoufiya University, Faculty of Science, Department of Mathematics, Egypt. In 2006, he received his PhD from Tohoku University, Graduate School of Engineering, Department of Mechanical Systems and Design, Sendai, Japan. Currently, he is Assistant Professor at King Khalid University, Faculty of Science, Department of Mathematics, Abha, Saudi Arabia. His permanent work is Assistant Professor at Suez University, Faculty of Science, Department of Mathematics, Suez, Egypt. His research interests are Lattice Boltzmann method, Turbulent flow and computational harmonic analysis.

.

Characterization of Physiological Glucose Concentration Using Electrical Impedance Spectroscopy

Quazi D. Hossain[1] and Sagar K. Dhar[1]

**[1]Dept. of Electrical and Electronic Engineering, Chittagong University of Engineering and Technology
Chittagong- 4349, Bangladesh**

Abstract

Non-invasive glucose monitoring is crucial for effective diabetes mellitus treatment while a sound correlation of a non-invasive parameter to glucose level variation is quite challenging. This paper presents characterization of glucose concentrations using Electrical Impedance Spectroscopy (EIS) in three different solutions: 1) 0.9% NaCl, 2) Saline (NaCl 1.3gm, KCl 0.75gm, $Na_3C_6H_5O_7$ 1.45gm, D-glucose 6.75gm in 500mL) and 3) Human Blood for every 25mg/dl change of glucose in total 150ml solution. A rectangular current pulse of 1.5s duration with 1mA peak is applied to the solutions and corresponding voltage is acquired across the solutions with Agilent InfiniiVision 7000B Series oscilloscope and Matlab R2011a Instrument Control Toolbox. The circuit proposed for current injection and voltage acquisition requires only two electrodes would reduce electrode polarization and skin irritation greatly which is a major concern in many previous works use generally four electrodes. Experimental results show sound correlation between EIS and blood glucose concentration. It is clearly found from the EIS that the DC impedance of solutions increases linearly with the increment in glucose concentrations.

Keywords: *Electrical Impedance, Impedance Spectroscopy, Non-invasive Glucose Monitoring, Diabetes Mellitus.*

1. Introduction

Electrical impedance spectroscopy (EIS) is getting more attention day by day as a mean of non-invasive physiological blood glucose monitoring. Although, there are different other means of non-invasive glucose measurement such as optical, photoacoustic and electromagnetic, it is always crucial to select a method that provides deterministic sensitivity of the measuring parameter with respect to the blood glucose variation. This work is aimed to investigate the variation in EIS with respect to the variation of glucose levels in human blood. To characterize this issue, EIS of three different solutions are compared and examined in this work.

Diabetes is one of the major ailments of concern in this 21st century. In 2012, more than 371 million people have diabetes which is increasing all over the world and is predicted to be 380 million by 2025 [1]. Diabetes is not only causing health concern but also have a significant socio-economic impact. In 2012, 4.8 million people died and 471 billion USD were spent due to diabetes [1]. However, the major fact of concern is that around half of the people with diabetes don’t know they have it [1] and get diagnosed only when serious condition arises. It may be due to the unavailability of technology, lack of consciousness and the invasive natures of the present diagnosis tools of glucose measurement. But People inherently may show reluctance of using invasive tools for diabetes measurement although having consciousness about the complications of diabetes and it can be predicted easily that if the diabetes can be measured in a non-invasive manner, the rate of using diagnosis tools will be increased which consequently will decrease the no. of undiagnosed people and hence the no. of death by diabetes. Besides, to avoid complexities due to diabetes, frequent glucose monitoring is necessary which is not possible with invasive glucose monitoring devices that requires blood collection at the present state. Moreover, the invasive techniques cause pain, high cost per measurement and potential risk of infection. All these issues inevitably lead to the necessity of a non-invasive glucose measurement system.

Non-invasive methods that are used for the determination of glucose so far can be categorized into two groups. First group includes the methods of near-infrared and mid-infrared absorption, optical rotation, Raman shifts and photo-acoustic absorption where measurements are based on the intrinsic properties of glucose molecules. On the other hand, second group measures the effect of glucose on the physical properties of blood and tissues, includes the methods of electrical impedance, electromagnetic and thermal technology. Among the different non-invasive techniques, infrared (IR) spectroscopy is the most tested and investigated one. [3-5] presents non-invasive glucose monitoring techniques based on IR spectroscopy faces problems cause of low absorption coefficient and non-specific scattering coefficient of glucose in the IR band [2]. Another technique of non-invasive glucose measurement, Raman spectroscopy, on the other hand suffers from the complexity of wavelength instability of laser and interference by other compounds [2]. Photo-acoustic method is based on the technique to excite the target with a laser and to get the acoustic response [6] heavily sensitive to the temperature and pressure change. On the other hand, [7-10] present electromagnetic sensors for measuring blood glucose are based on the fact that the change in the glucose concentration changes the dielectric parameters of blood. The feasibility of such electromagnetic sensors with in-vitro experiments is presented in [11] where it is shown that the antenna resonant frequency in the range of 1GHz to 10GHz increases when blood glucose increases. The major disadvantage in this method is the change in dielectric property of blood depends on several components other than glucose and will face inaccuracy when it will be tested in-vivo. In contrast, EIS although measures the dielectric property of blood can be improved significantly by proper electrical equivalent circuit modeling. However, the challenge in all these methods is to find a sound and deterministic relation between glucose and non-invasive parameters and to avoid the disturbance by skin-sensor interface and this paper, an experimental investigation of previous work [12], shows EIS as a firm solution in this regard.

The first complete non-invasive glucose monitoring device was approved in 2003 [13] in EU named Pendra by Pendragon Medical Ltd. was based on EIS presented in [14]. The device was based on the measurement of modulus of impedance at resonant frequency (minimum impedance) which was a variable of blood glucose concentration. But when it was in the market for home use, showed less precision as the correlation coefficient was only 0.64 with only 56% data in acceptable range [15]. This is because the resonant frequency can be changed in blood due to elements other than glucose which has not been compensated in Pendra. On the other hand, impedance spectroscopy using voltage-current pulse technique presented in this paper, directly measures the electrical nature of blood where the effect of other elements than blood can be compensated easily using proper circuit modeling.

Voltage and current pulse technique for realizing electrical impedance spectroscopy as a mean of non-invasive blood glucose monitoring are presented in [16, 17] but requires 4 electrodes for current injection and voltage detection. Alternatively, this paper presents a system with only two electrodes for both current injection and voltage detection which would reduce skin irritation due to electrode polarization. In-vitro experiments on three different solutions: 1) 0.9% NaCl, 2) Saline (NaCl 1.3gm, KCl 0.75gm, $Na_3C_6H_5O_7$ 1.45gm, D-glucose 6.75gm in 500mL) and 3) Human Blood are performed and found the linear variation in impedance modulus against different glucose concentrations.

2. Methods and Materials

2.1 Electrical Equivalent Circuit

Biological tissues are generally modeled by similar electrical circuits shown in Fig. 1 [18]. Here, the extracellular medium is modeled by Re which corresponds to the Plasma of blood. Then the cell membrane is modeled by Cm and Rm and the intra-cellular medium is modeled by Ri. In every case, subjected electrical current will flow through biological cells or extra-cellular medium and the current through the cells may be classified as the current across the trans-membrane ionic channel (shown as the path with R_m) or by the plasma membrane (shown as the path with Cm). Due to the very high trans-membrane channel resistance Rm, it may be ignored and the circuit can be simplified with a single extra-cellular medium resistance R_e, an intra-cellular medium resistance R_i and a cell membrane capacitance C_m.

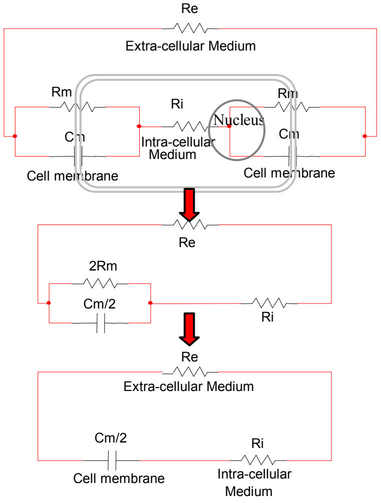


Fig 1. Electrical equivalent circuit of biological cells

When glucose concentration in blood is changed, it changes the ionic balance in the plasma and increases the extra cellular resistance Re. On the other hand, whenever, blood glucose in plasma is increased, water from intra-cellular medium is transferred to the plasma that changes the permittivity of cell membrane Cm and intra-cellular medium resistance Ri. Consequently, the change in blood glucose can be observed by any one parameter of Re, Ri and Cm or collectively by the impedance spectroscopy Z(jw) as shown in Eq. (1). However, the main complexities in blood glucose characterization is due to the change in EIS cause of elements other than glucose. According to Fig. 2 [19], it is evident that, the mostly varied component in blood is glucose althogh triglycerides (lipid) and urea also vary significantly. But the contribution of lipid is solely in Ri and Cm since it is not soluble in plasma. So, the effect of lipid can be avoided just calculating Re from EIS as presented in [12]. On the other hand, urea is highly soluble and can make significant effect on Re and on EIS. But change in urea is much more lower than change in glucose. So, the glucose variation could be effectively monitored by EIS and the effect of other elements except glucose can also be effectively filtered by observing Re only rather than the measurement based on EIS collectively.

$$Z(jw) = \frac{R_e + jwR_eR_iC_m}{1 + jw(R_e + R_i)C_m} \qquad (1)$$

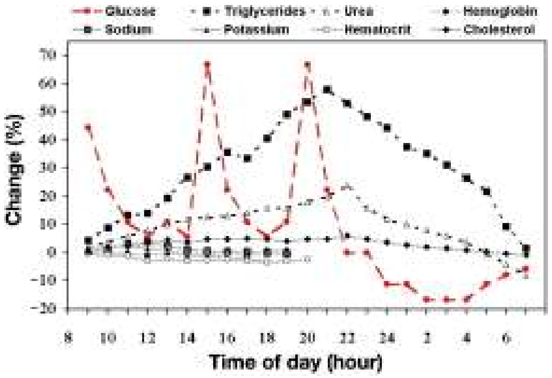


Fig 2. Diurnal blood component variation in blood. Meals are taken around 8:30, 13:30 and 18:30, denoting breakfast, lunch and dinner respectively [19,20].

2.2 Response to Current Pulse and EIS

EIS refers to the presentation of electrical impedance against frequency. One method can be applied for acquiring EIS is current pulse injection and voltage acquisition as shown in Fig. 3. When a current pulse is applied, the voltage developed can be expressed by Eq. (2) where h(t) represents the impulse response of bio-electrical circuit and '*' represents the convolution sum between h(t) and I(t). In this case, the impulse transform H(jw), the Fourier transform of h(t), equals the EIS i.e. Z(jw) which can be obtained by Eq. (3).

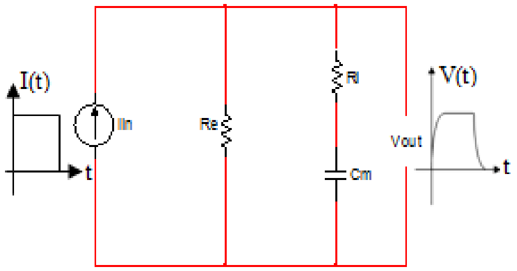


Fig 3. EIS by current injection and voltage detection

$$V(t) = I(t) * h(t) \qquad (2)$$

$$H(jw) = \frac{V(jw)}{I(jw)} = Z(jw) \qquad (3)$$

Therefore, from experimental data, we can calculate EIS by current pulse injection and voltage detection according to Eq. (3). But these EIS data correspond to the Eq. (1) by which we can estimate different EIS parameters separately that are Re, Ri and Cm [12]. However, this analysis is based on frequency domain and it is also possible to perform a time domain analysis to find out Re, Ri and Cm from estimating rise time, settling time and time constant of voltage response against current pulse. Eq. (4) and (5) represents the input current pulse I(t) and voltage response V(t) of bio-electrical circuit where T represents the duration of current pulse. Eq. (6-8) represent time constant, rise time and settling time respectively and solving these three relations, it is possible to find Re, Ri and Cm. Although, this time domain approach is faster and easier to calculate EIS parameters but erroneous compared to frequency domain analysis requires DFT algorithms to estimate impedance parameters. In this paper, frequency domain analysis is performed for better accuracy.

$$I(t) = u(t) - u(t - T) \qquad (4)$$

$$\begin{aligned} V(t) = u(t)&\left[R_e - \frac{R_e^2}{R_e + R_i} e^{-t/(R_e+R_i)C_m}\right] - \\ u(t-T)&\left[R_e - \frac{R_e^2}{R_e + R_i} e^{-(t-T)/(R_e+R_i)C_m}\right] \end{aligned} \qquad (5)$$

$$T_c = (R_e + R_i)C_m \qquad (6)$$

$$T_r = 2.2(R_e + R_i)C_m \qquad (7)$$

$$T_s = 4(R_e + R_i)C_m \qquad (8)$$

Once impedance data is retrieved from frequency domain analysis of voltage and current pulse, it is enough either to work with magnitude or phase of Z(jw) to estimate R_e, R_i and C_m. Eq. (9) and (10) present the magnitude and phase of Z(jw). But Re represents more accurate variation of glucose level compared to Ri and Cm and interestingly Re equals the DC impedance which refers to the $|Z(jw)|$ when w = 0 and can be found easily from Eq. (9).

$$|Z(jw)| = \frac{\sqrt{(R_e + R_eR_i(R_e + R_i)w^2{C_m}^2)^2 + w^2C_m^2(R_eR_i)^2}}{1 + (R_e + R_i)^2 w^2 C_m^2} \qquad (9)$$

$$\theta(w) = -\tan^{-1}\left(\frac{wC_m R_e^2}{1+(R_e+R_i)^2 w^2 C_m^2} \bigg/ \frac{R_e + R_e R_i (R_e+R_i) w^2 C^2}{1+(R_e+R_i)^2 w^2 C_m^2} \right) \quad (10)$$

2.3 Materials

Finding and setting up a sound correlation between glucose variation and a non-invasive parameter is crucial always. In this paper, to observe the co-relation between glucose level and EIS, three solutions: 1) 0.9% NaCl, 2) Saline (NaCl 1.3gm, KCl 0.75gm, $Na_3C_6H_5O_7$ 1.45gm, D-glucose 6.75gm in 500mL) and 3) Human Blood are tested. Solutions 1 and 2 are chosen because of their similar behavior with blood although lack cellular components. Since, the DC impedance Re is subjected to non-cellular components of blood, these two solutions are also expected to have similar EIS response against glucose compared to blood at w = 0 which is observed true in experimental results.

3. EIS Measurement System

The operation of EIS measuring system followed in this paper is presented in Fig. 4 basically based on four sections: 1) current injection, 2) voltage acquisition, 3) frequency domain conversion and 4) EIS estimation. The experimental setup of the system is shown in Fig. 5.

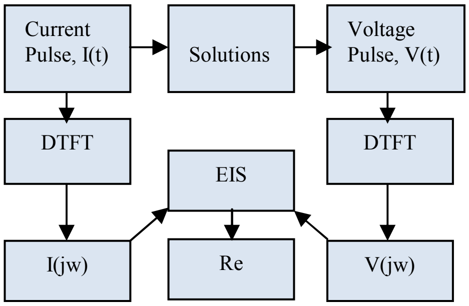


Fig 4. EIS measurement system block diagram

Fig 5. Experimental setup

The major part of the EIS measurement system is the current pulse source generation which is realized in this paper firstly generating a voltage pulse by monostable operation of a 555 timer IC and then converting the voltage pulse to a current pulse using a non-inverting amplifier as shown in Fig. 6 [12]. The input impedance of a 741 IC is 1MΩ and until the impedance of load exceeds 10kΩ the circuit in Fig. 6 will act as an ideal current source which is enough impedance range for human blood normally lies below 1kΩ.

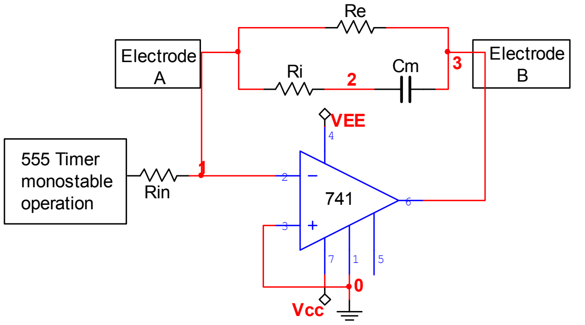


Fig 6. Current pulse generation

One of the major advantages of the system presented here for EIS measurement is using only two electrodes instead of four. The two electrode positions are shown in Fig. 6 as Electrode A and Electrode B. The parallel network of Re, Ri and Cm represent the solutions. In this work, the solutions are taken into a rectangular box of dimensions $2'' \times 2'' \times 2''$ where two sides are used as Electrode A and Electrode B made of Cu. D-glucose is added to manually vary the glucose concentration of solutions. The voltage across the solution is acquired using Agilent InfiniiVision 7000B Series oscilloscope and Matlab R2011a Instrument Control Toolbox after an instrumentation amplifier in Fig. 7 for noise rejection as shown in Fig. 8. The gain can also be possible to control by this instrumentation amplifier according to $A_{CL} = 1 + 2R / R_g$. Once voltage data are acquired, the EIS is estimated using Discrete Time Fourier Transform.

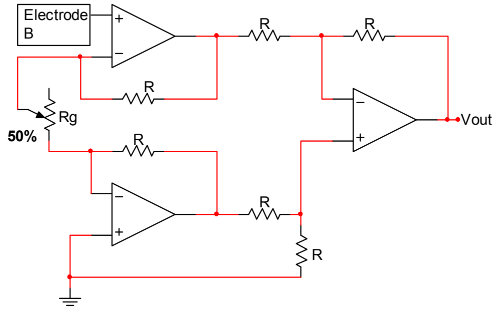


Fig 7. Instrumentation amplifier

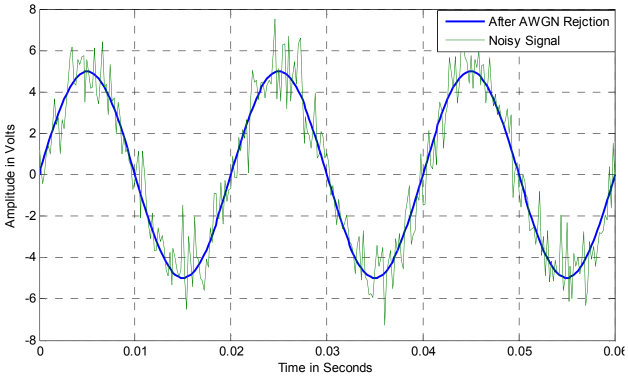


Fig 8. Noise rejection by instrumentation amplifier

4. Results and Discussions

EIS of three solutions: 1) 0.9% NaCl, 2) Saline and 3) Human blood is estimated against glucose variation every 25mg/dl. After taking every reading by applying a current pulse, the solutions are kept around 15-20 minutes to be relaxed. Fig. 9 shows the EIS variation of 0.9% NaCl solution. It is evident that EIS has the significant change for every 25mg/dl change in glucose concentration. Moreover, the DC impedance that is the value of EIS when w=0 encircled in Fig. 9 shows almost linear variation.

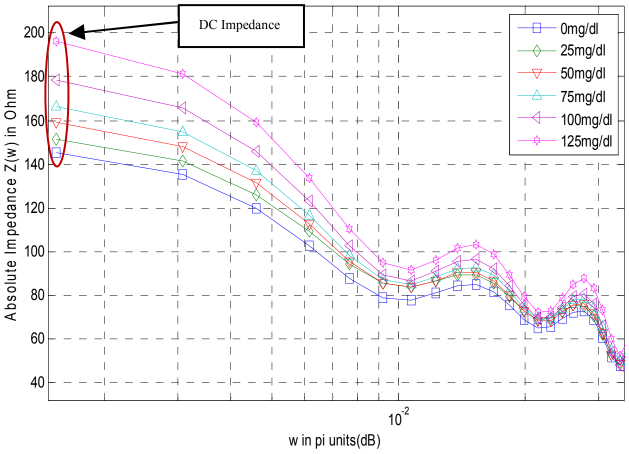


Fig 9. EIS of 0.9% NaCl solution

Again, Fig. 10 shows the variation in EIS for saline and linear variation in DC impedance is found here too. However, for the glucose variation of 0-125mg/dl, the DC impedance varies from about 140 to 200Ω in NaCl solution which is just 180-200Ω in saline for 0-225mg/dl glucose variation. That indicates the change in DC impedance is larger in 0.9% NaCl solution than saline. However, both solutions indicates the firm relation between glucose variation and EIS.

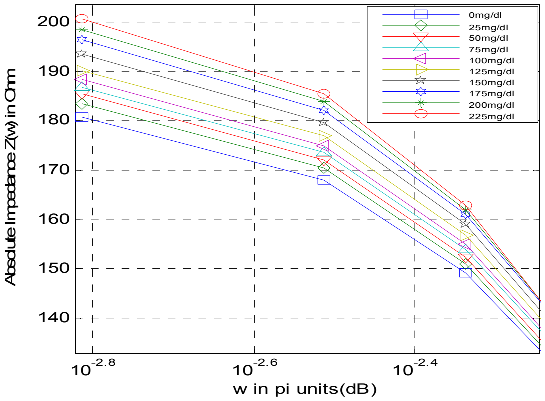


Fig 10. EIS variation of saline

Fig, 11-13 show the variation in EIS for three human blood samples for change in glucose concentration of every 25mg/dl. It is clear from three sample data that, when glucose concentration increases, the impedance of blood also increases. Moreover, the DC impedances of three blood samples also get changed linearly when glucose concentration varies as it was in saline and NaCl solution. However, the rate of change in DC impedance is not equal for three blood samples. For blood sample 1, although the DC impedance varies from around 200-295Ω for 0-250mg/dl change in gluocose concentration, it is around 325-360Ω for 0-125mg/dl for blood sample 2 and 320-345Ω for 0-250mg/dl. It is because of the variation of nature and ingredients in blood samples. For example, if the blood sample 1 has the lower amount of ionic compounds than blood sample 2 and 3, its impedance will be affected more by the variation of glucose concentration. Another factor may be the variation of blood cells in different samples. However, for a particular blood sample, EIS changes linearly in change of glucose concentration and sample specific callibration could be useful for glucose measurement from EIS.

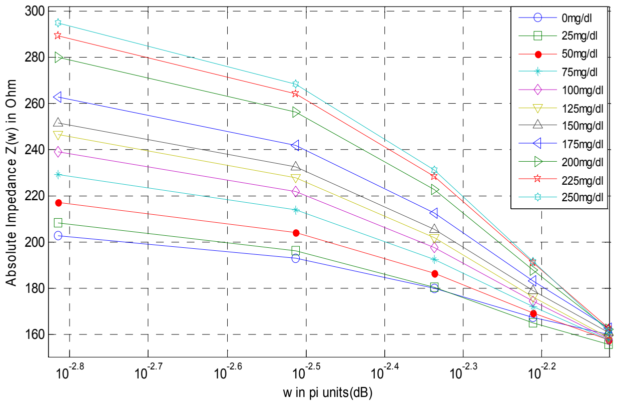


Fig 11. EIS variation in blood sample 1

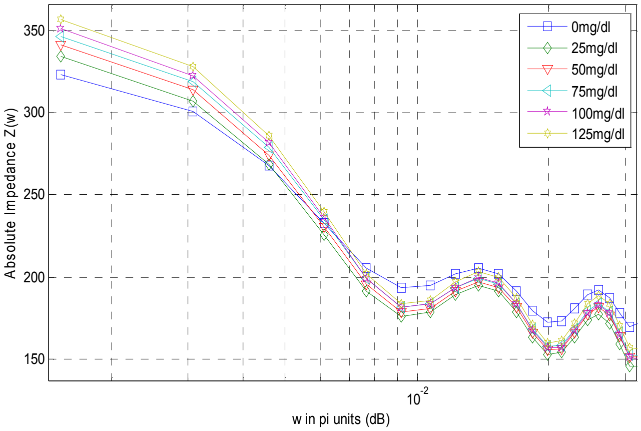


Fig 12: EIS variation in blood sample 2

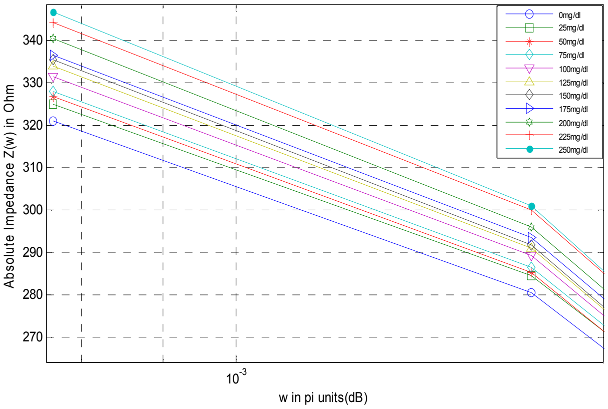


Fig 13. EIS variation of blood sample 3

5. Conclusion

Finding a sound correlation between blood glucose concentration and a non-invasive parameter is challenging always. This work measures EIS against different blood glucose concentration and found strong correlation between DC impedance of EIS and glucose concentrations. EIS increases linearly when glucose concentration increases. Although, EIS in different blood samples show different rate of change, initial calibration can be used to cope up this situation. Finally, glucose concentration in human blood can be measured accurately by EIS and since no blood collection or invasion would require in this method when tested in-vivo, non-invasive and continuous blood glucose monitoring would be possible accurately. Disturbances due to electrode-skin interference, skin and muscle characteristics can also be eliminated effectively using EIS when only DC impedance is get considered.

Acknowledgment

The authors would like to acknowledge the invaluable suggestions and guidelines by Dr. Sabuj Baran Dhar, Assistant Professor, Cox's Bazar Medical College Hospital and Tanzilur Rahman, Research Assistant, Tokyo University. We would also like to acknowledge gratefully the resources provided by Sandhani, Chittagong Medical College Hospital and Electronics Lab, Premier University.

References

[1] "The fifth edition of the IDF diabetes atlas", International Diabetes Federation, 2012.
[2] A. Tura, A. Maran and G. Pacini, "Non-invasive glucose monitoring: Assessment of technologies and devices according to quantitative criteria", Diabetes Research and Clinical Practice, vol. 77, pp.16-40, 2007.
[3] T. MoriKawa, F. Saiki, H. Ishizawa and E. Toba, "Non-invasive measurement of blood glucose based on optical sensing and internal standard method", IEEE Int. Joint Conf. on SICE-ICASE, pp. 1481-1484, Oct. 2006.
[4] H. Ishizawa, A. Muro, T. Takano, K. Honda and H. Kanai, "Non-invasive blood glucose measurement based on ATR infrared spectroscopy", Annual Conf. on SICE, pp. 321-324, Aug., 2008.
[5] S. Koyama, Y. Miyauchi, T. Horiguchi and H. Ishizawa, "Non-invasive measurement of blood glucose of diabetic based on IR spectroscopy", Annual Conf. of SICE, pp. 3425-3426, 2010.
[6] O. C. Kulkarni, P. Mandal, S. S. Das and S. Banerjee, "A feasibility study on non-invasive blood glucose measurement using photoacoustic method", IEEE Int. Conf. on Bioinformatics and Biomedical Engineering, pp. 1-4, 2010.
[7] B. R. Jean, E. C. Green and M. J. McClung, "A microwave frequency sensor for non-invasive blood-glucose measurement", IEEE Sensors Applications Symposium, pp. 4-7, 2008.
[8] M. Hofmann, T Fersch, R. Weigel, G. Fischer and D. Kissinger, "A novel approach to non-invasive blood glucose measurement based on RF transmission", IEEE Int. Workshop on Medical Measurements and Applications, pp. 39-42, 2011.
[9] V. V. Meriakri, E. E. Chigrai, I. P. Nikitin and M. P. Parkhomenko, "Dielectric properties of water solution with small content of glucose in the millimeter wave band and the determination of glucose in blood", 6th Int. Symp. On Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves, pp. 873-875, 2007.
[10] Y. Nikawa and D. Someya, "Non-invasive measurement of blood sugar level by millimeter waves", IEEE MTT-S Int. Microwave Symp. Digest, pp. 171-174, 2001.
[11] J. Venkataraman and B. Freer, "Feasibility of non-invasive blood glucose monitoring", IEEE Int. Symp. on Antenna and Propagation, pp. 603-606, 2011.
[12] S. K. Dhar and Q. D. Hossain, "Non-invasive bio-impedance measurement using voltage-current pulse technique", Int. Conf. on Electrical, Electronics and Biomedical Engineering, pp-70-74, May, 2012.

[13]“Unsafe and ineffective devices approved in the EU that were not approved in the US”, US Department of Health and Human Services, May 2012.

[14] A. Caduff, E. Hirt, Yu. Feldman, Z. Ali and L. Heinemann, “First human experiments with a novel non-invasive, non-optical continuous glucose monitoring system”, Journal of Biosensors and Bioelectronics, vol. 19, issue 3, pp. 209-217, Nov., 2003.

[15] C. M. Girardin, C. Huot, M. Gonthier and E. Delvin, “Continuous glucose monitoring: A review of biochemical perspectives and clinical use in type 1 diabetes”, Journal of Clinical Biochemistry, vol. 42, issue 3, pp. 136-142, 2009.

[16] A. M. Aguilar and R. P. Areny, “Electrical impedance measurement using voltage/current pulse excitation”, 19th IMEKO World Congress on Fundamental and Applied Metrology, pp. 662-667, Sept. 2009.

[17] T. Dai and A. Adler, “In vivo blood characterization from bioimpedance spectroscopy of blood pooling”, IEEE Transaction on Instrumentation and Measurement, vol. 58, no. 11, Nov. 2009.

[18] L. I. Kalakutskiy and S. A. Akulov, “Modeling of the bioelectrical impedance of blood by synthesis of the equivalent electrical circuits,” Proc. IFMBE 25/VII, pp. 575-577, 2009.

[19] SJ Pocock, D. Ashby, AG Shper, M Walker, G. Broughton, “Diurnal variations in serum biochemical and haematological measurements”, Journal of Clinical Pathology, vol.42, no.2, pp-172-179, Feb. 1989.

[20] I. H. Boehm, A. Gal, A. M. Raykhaman, J. D. Zahn, E. Naidis and Y. Mayzel, “Non-invasive glucose monitoring: a novel approach”, Journal of Diabetes Science and Technology, vol.3, issue 2, March 2009.

Quazi D. Hossain (MIEEE) was born in Bangladesh in 1976. He received his B.Sc degree in electrical and electronic engineering. from Chittagong University of Engineering and Technology (CUET), Chittagong, Bangladesh, in 2001; the Master of Engineering degree in semiconductor electronics and integration sciences from Hiroshima University, Hiroshima, Japan, in 2007; and the Ph.D. Degree in microelectronics from the University of Trento, Trento, Italy, in 2010. During his Ph.D. program, he also spent a period with the Smart Optical Sensors and Interfaces Group, Bruno Kessler Foundation, Trento, as a Postgraduate Researcher. From 2001 to 2007, he was with CUET as a Lecturer. In 2007, he became an Assistant Professor with the Faculty of Electrical and Computer Engineering, CUET. He is also the Life member of Institute of Engineers, Bangladesh. His research interests include image sensors and related readout circuit simulation, experimental characterization of semiconductor devices and biosensors.

Sagar K.Dhar received his B.Sc degree in electrical and electronic engineering from Chittagong University of Engineering and Technology (CUET), Chittagong, Bangladesh in 2008; Presently he is pursuing M.Sc Engineering degree in electrical and electronic engineering in the same institute as a part time student. He is also working as a Lecturer in the Department of EEE, Premier University, Chittagong, Bangladesh. He is a member of IEEE Solid State Circuit Society, Communication Society and Circuit & system society. His ongoing research interests: Time based communication circuits, Data converters and Biosensors.

Diagnosis of Switching Systems using Hybrid Bond Graph

Taher Mekki[1], Slim Triki[1] and Anas Kamoun[1]

**[1] Research Unit on Renewable Energies and Electric Vehicles (RELEV), University of Sfax
Sfax Engineering School (ENIS), Tunisia**

Abstract

Hybrid Bond Graph (HBG) is a Bond Graph-based modelling approach which provides an effective tool not only for dynamic modeling but also for fault detection and isolation (FDI) of switching systems. Bond graph (BG) has been proven useful for FDI for continuous systems. In addition, BG provides the causal relations between system's variables which allow FDI algorithms to be developed systematically from the graph. There are many methods that exploit structural relations and functional redundancy in the system model to find efficient solutions for the residual generation and residual evaluation steps in FDI of switching systems. This paper describes two different techniques, quantitative and qualitative, based on common modelling approach that employs HBG. In quantitative approach, global analytical redundancy relationships (GARRs) are derived from the HBG model with a specified causality assignment procedure. GARRs describe the system behaviour at all of its operating modes. In qualitative approach, functional redundancy can be captured by a Temporal Causal Graph (TCG), a directed graph that may include temporal information.

Keywords: *Hybrid Bond Graph, Global Analytical Redundancy Relation, Temporal Causal Graph.*

1. Introduction

Several physical systems with switching are nonlinear dynamic systems. When switching occurs, the system may change its mode of operation. If a system has *n* switching states, then it gives rise to 2^n possible operating modes. One way to represent mode switching is to generate 2^n sets of differential-algebraic equations (DAEs). Each set describes continuous behaviour of system in that particular mode. In practice, not all modes are practically realizable. Many recent researches on switching systems have been devoted to the synthesis of control laws that guarantee not only the stability but also good performances [1]. The control algorithms are generally developed considering that the system works in normal situation. Unfortunately, when failures occur, these algorithms become inefficient and even dangerous for the system itself or its environment. In order to reach higher performances and more rigorous security specifications, a Failure Detection and Isolation system has to be implemented. The literature in that field is abundant and different solutions have been proposed for continuous or discrete, linear and non linear systems. However, only few solutions have been proposed for switching systems.

Traditionally, two different communities: (1) the Systems Dynamics and Control Engineering (FDI) community (e.g., [2,3]), and (2) the Artificial Intelligence Diagnosis (DX) community (e.g., [4,5]), have developed model-based diagnosis approaches. The two communities have employed different kinds of models, and made different assumptions concerning robustness of the generated solution with regard to modeling errors, measurement noise, and disturbances. The general principle of all model-based FDI approaches is to compare the expected behavior of the system, given by model, with its actual behavior. The first step of a FDI procedure consists in generating a set of residuals. These residuals are special signals that reflect the discrepancy between the two behaviors. Analytic Redundancy Relation (ARR) methods are classically used for residual generation in the FDI community [6]. The DX community has developed methods such as possible conflicts [7,8], and analysis of temporal causal graphs [9,10] for diagnosis of continuous systems. These methods are based on the structural analysis of dynamic models, much like the ARR schemes developed by the FDI community. The two communities use different algorithms, but the overall framework for fault isolation is similar, defined by a two-step process: (i) residual generation, followed by (ii) residual evaluation [2,3].

In this work, we focus on the Hybrid Bond Graph as unified graphical method of modeling and diagnosis of switching systems. There are two main approaches while using HBGs: those who use switching elements with fixed causality [11,12], and those who use ideal switching elements which changing causality [13]. Therefore, we start with common modeling framework, hybrid bond graph, to describe and compare the ARR approach developed by [14] with temporal causal graph based diagnosis [15,16]. In particular, the residual generation and evaluation algorithms used by the two methods are presented and a discussion between the algorithms is established.

2. Quantitative hybrid bond graph-based fault detection and isolation

2.1 Analytical Redundancy Relations (ARRs) and Global ARRs

An ARR obtained from a physical law represents some conservation phenomena: Bernoulli equation etc. in hydraulic domain; Newton's law etc. in mechanical domain; and Kirchoff's law etc. in electrical domain. The ARRs for ordinary (non hybrid) system can be derived algorithmically from derivative causality bond graph model or Diagnostic Bond Graph (DBG) [17,18]. Whereas the bond graph model for hybrid system with discrete mode changes is called the Hybrid Bond Graph, and the GARRs are derived in similar manner from differentially causalled HBG or Diagnostic Hybrid Bond Graph (DHBG). A DHBG is a HBG which describes the behavior of a hybrid system at all modes based on unified set of causality assignment [19]. In other words, for a given DHBG, a reassignment of causality to the HBG is not required to describe the behavior of the system at different operating modes. This unique feature of DHBG implies that the causal paths of graph maintain the same structure, except that some of the sub-paths are eliminated due to the OFF states of the controlled junctions. Based on these uniform causal paths, a set of unified relations is derived to describe the hybrid system at all modes. This set of relations which is called Global ARRs (GARRs), are utilized as ARRs for the hybrid systems. Likewise for continuous system, the GARRs provide a convenient tool to deduce the fault detectability and isolability of a hybrid system. To illustrate how to derive the GARR for FDI application, consider the three coupled tanks depicted in Fig. 1. These tanks are connected by pipes which can be controlled by different valves. Water can be filled into the left and right tanks using two identical pumps. Measurements available from the process are the continuous water levels $h_i(t)$ of each tank. The connection pipe, with valve R_{12} (res. R_{23}), between the tank 1 and 2 (res. 2 and 3) is placed at a height of 0.5m (res. 0.7m).

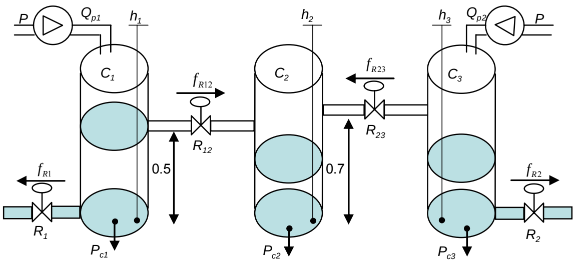


Fig. 1 Three Tank System.

Fig. 2 shows the hybrid bond graph model of this system. The two flow sources into tanks 1 and 3 are indicated by Sf_1 and Sf_2, respectively. The tank capacities are shown as C_1, C_2 and C_3. The pipes are modeled by resistances R_1, R_{12}, R_{23} and R_2. Pumps and valves are modeled by controlled junctions, which are shown in the figure as junctions with subscripts (1_1, 1_2, 1_5, and 1_6). The control signals for turning these junctions *on* and *off* are generated by the finite state automata in Fig. 2. For autonomous transitions in the system, also modeled by controlled junctions, the transition conditions computed from system variables (junctions 1_3 and 1_4). A mode in the system is defined by the state of the six controlled junctions in the hybrid bond graph model. Therefore, theoretically the system can be in 2^6 different modes.

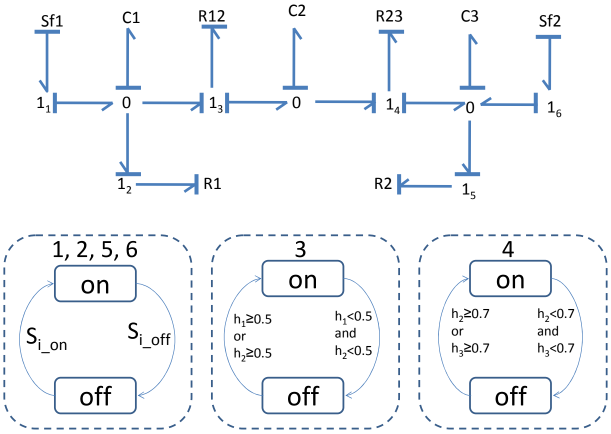

Sf1
C1
R12
C2
R23
C3
Sf2
1_1
0
1_3
0
1_4
0
1_6
1_2
R1
R2
1_5
1, 2, 5, 6
on
S_{i_on}
S_{i_off}
off
3
on
$h_1 \geq 0.5$ or $h_2 \geq 0.5$
$h_1 < 0.5$ and $h_2 < 0.5$
off
4
on
$h_2 \geq 0.7$ or $h_3 \geq 0.7$
$h_2 < 0.7$ and $h_3 < 0.7$
off

Fig. 2 HBG of the three tank system.

Fig. 3 shows the DHBG model of the hybrid tank system deduced from the three step procedure Hybrid Procedure Assignment Causality Sequential (SCAPH) and Model Approximation (MA) detailed in [19]. First step, a R_s component is added to every sensor junction; in this example a very-high resistive component is added to the junction 0_1, 0_2 and 0_3. The second step is to apply the SCAPH algorithm. In this step, all controlled junction are assigned with their preferred causality. As shown in Fig. 3, the output variables of the controlled-junctions 1_{c1} and 1_{c2} (e_5 and e_{10}, respectively) are assigned as inputs of the 1-port component (R_{12} and R_{23}, respectively). There is no source connected to the controlled-junction. Then the two sources Q_{p1} and Q_{p2} are assigned with their preferred causality. Since the DHBG is required to generate the GARR, the three storage components C_1, C_2 and C_3 are assigned with their preferred derivative causality. We extend the causal implication using the components constraints to remain bonds to complete the SCAPH algorithm. The final step is to eliminate every R_s that is in indifferent causality. In our case R_s added to the junction 0_1 and 0_3 are redundant and, therefore, are removed from the DHBG model of Fig. 3.

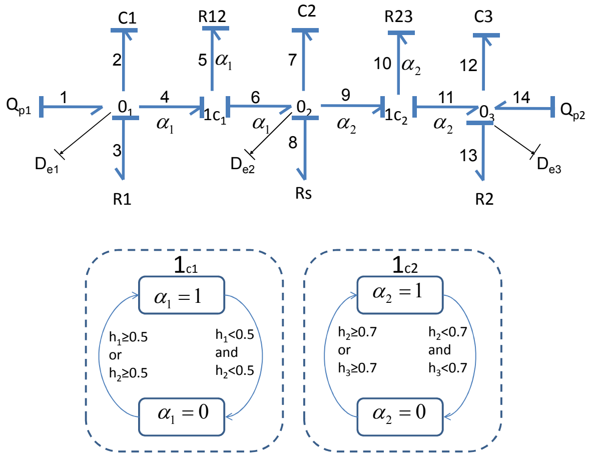


Fig. 3 DHBG of the three tank system.

The constitutive relation of the R_s component connected to the junction 0_2 is $f_8 = 0$. From the DHBG, the constitutive relations of junctions 0_1, 1_{c1}, 0_2, 1_{c2} and 0_3 are given by the following equations:

Junction 0_1

$$f_1 - f_2 - f_3 - \alpha_1 f_4 = 0 \tag{1}$$

Junction 1_{c1}

$$e_4 - e_5 - e_6 = 0 \tag{2}$$

Junction 0_2

$$\alpha_1 f_6 - f_7 - f_8 - \alpha_2 f_9 = 0 \tag{3}$$

Junction 1_{c2}

$$e_9 - e_{10} - e_{11} = 0 \tag{4}$$

Junction 0_3

$$\alpha_2 f_{11} + f_{14} - f_{12} - f_{13} = 0 \tag{5}$$

Where

$$f_4 = f_6 = \begin{cases} 0 & \text{if } h_1 < 0.5 \text{ and } h_2 < 0.5 \\ \frac{1}{R_{12}}(De_1 - 0.5) & \text{if } h_1 > 0.5 \text{ and } h_2 < 0.5 \\ -\frac{1}{R_{12}}(De_2 - 0.5) & \text{if } h_1 < 0.5 \text{ and } h_2 > 0.5 \\ \frac{1}{R_{12}} sign(De_1 - De_2) * |De_1 - De_2| & \text{if } h_1 > 0.5 \text{ and } h_2 > 0.5 \end{cases}$$

$$\Rightarrow f_4 = f_6 = sign(\max(De_1, 0.5) - \max(De_2, 0.5)) * (\max(De_1, 0.5) - \max(De_2, 0.5))$$

$$f_9 = f_{11} = \begin{cases} 0 & \text{if } h_2 < 0.7 \text{ and } h_3 < 0.7 \\ \frac{1}{R_{23}}(De_2 - 0.7) & \text{if } h_2 > 0.7 \text{ and } h_3 < 0.7 \\ -\frac{1}{R_{23}}(De_3 - 0.7) & \text{if } h_2 < 0.7 \text{ and } h_3 > 0.7 \\ \frac{1}{R_{23}} sign(De_2 - De_3) * |De_2 - De_3| & \text{if } h_2 > 0.7 \text{ and } h_3 > 0.7 \end{cases}$$

$$\Rightarrow f_9 = f_{11} = sign(\max(De_2, 0.7) - \max(De_3, 0.7)) * (\max(De_2, 0.7) - \max(De_3, 0.7))$$

Three structurally independent GARRs can be generated from (1), (3) and (5) after eliminating the unknown variables and obtained as follows:

$$GARR1 = Q_{p1} - C_1 \frac{d}{dt} De_1 - \frac{1}{R_1} De_1$$
$$- \frac{\alpha_1}{R_{12}} sign(\max(De_1, 0.5) - \max(De_2, 0.5)) * (\max(De_1, 0.5) - \max(De_2, 0.5)) \tag{6}$$

$$GARR2 = \frac{\alpha_1}{R_{12}} sign(\max(De_1, 0.5) - \max(De_2, 0.5)) * (\max(De_1, 0.5) - \max(De_2, 0.5))$$
$$- C_2 \frac{d}{dt} De_2 - \frac{\alpha_2}{R_{23}} sign(\max(De_2, 0.7) - \max(De_3, 0.7)) * (\max(De_2, 0.7) - \max(De_3, 0.7)) \tag{7}$$

$$GARR3 = \frac{\alpha_2}{R_{23}} sign(\max(De_2, 0.7) - \max(De_3, 0.7)) * (\max(De_2, 0.7) - \max(De_3, 0.7))$$
$$+ Q_{p2} - C_3 \frac{d}{dt} De_3 - \frac{1}{R_2} De_3 \tag{8}$$

2.2 Fault detectability and fault isolability

Unlike continuous systems, hybrid systems are multiple modes in nature. This feature suggests that the system's fault detectability and fault isolability are required to be evaluated at different operating modes for effective FDI analysis and designs. The unified characteristic of the GARRs provide a convenient way to generate the Fault Signature Matrix (FSM) of each operating mode. Table 1 shows the FSM of the three tank system and table 2 describes the modes. The fault detectability and fault isolability of each parameter is gained from the {D_b, I_b} values of the four FSMs. This information can be summarized in table 3.

Table 1: FSM for the three tank system

	GARR1	GARR2	GARR3	D_b	I_b
R_1	1	0	0	1	0
C_1	1	0	0	1	0
R_{12}	α_1	α_1	0	α_1	α_1
C_2	0	1	0	1	1
R_{23}	0	α_2	α_2	α_2	α_2
C_3	0	0	1	1	0
R_2	0	0	1	1	0

Table 2: Modes of the system

Modes	α_1	α_2
Mode 1	1	1
Mode 2	1	0
Mode 3	0	1
Mode 4	0	0

Table 3: Fault Detectability and fault Isolability of components

θ	Detectability	Isolability
R_1	all-mode	Nil
C_1	all-mode	Nil
R_{12}	Mode 1, 2	mode 1, 2
C_2	all-mode	all-mode
R_{23}	Mode 1, 3	mode 1, 3
C_3	all-mode	Nil
R_2	all-mode	Nil

In this work, MATLAB 7.0 is used to simulate the model of the tank system. The parameters of the system are fixed as follow; R_1= R_{12}= R_{23}= R_2=1m^{-1}s^{-1}, C_1= C_2= C_3=1kg^{-1}m^{4}s^2, and simulation time is fixed to 10s with sampling time t_s=0.01s. The inputs $Q_{p1}(t)$, $Q_{p2}(t)$ are presented in Fig. 4.

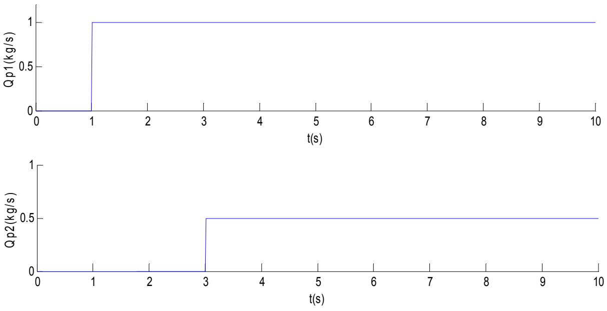


Fig. 4 Inputs on the system.

A fault is simulated in R_{12} component where its parametric value changes abruptly from 1 to 5 at t=1s. The fault profile is shown in Fig. 5. Figure 6 depicts the measured variables and the switches signals.

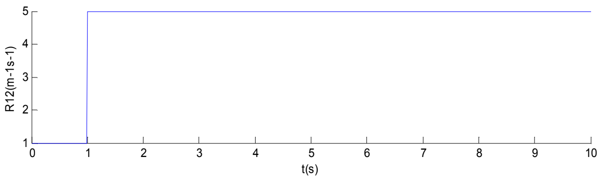


Fig. 5 Fault profile in R_{12} component.

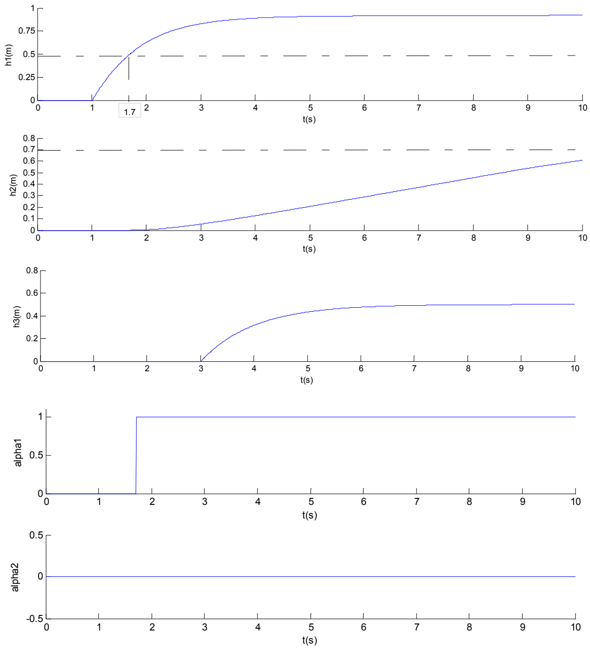


Fig. 6 Plot of measured variables and operating modes alpha 1 and 2.

The residuals GARR1, GARR2 and GARR3 due to the fault in R_{12} are shown in Fig. 7 where line denotes the thresholds re $\varepsilon_1 = \pm 0.02, \varepsilon_2 = \pm 0.01 \text{and} \, \varepsilon_3 = \pm 0.01$.

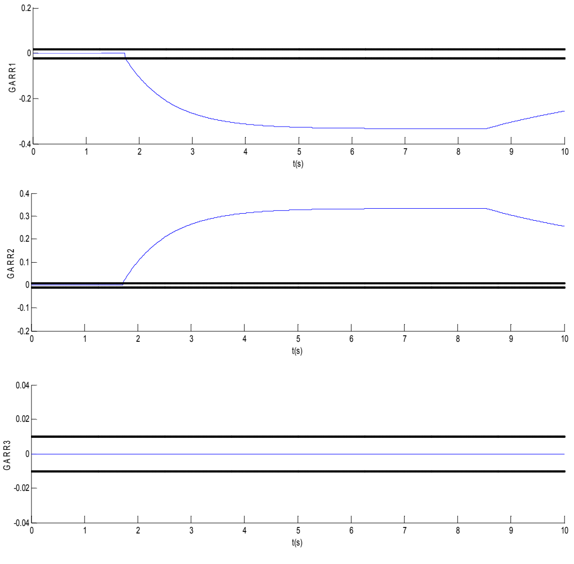


Fig. 7 Residuals responses for single fault in R_{12} component.

In general, if the residual exceeds the predetermined threshold, system is considered as faulty. According to the FSM tables, we can easy deduce that the faults in R_{12}

initiate at Mode 4, i.e. $\alpha_1 = 0, \alpha_2 = 0$, which is a non-observable mode. Hence the fault can not be detected until the system enters a mode in which fault is detectable, i.e. at time *t*=1.7s (see Fig. 6) when the system change to the Mode 2 ($\alpha_1 = 1, \alpha_2 = 0$). Fig. 7 reveals that GARR1 and GARR2 are sensitive to fault. From the FSM table at that mode, R_{12} has fault signature [1 1 0]. According to the FSM table 3, R_{12} is not detectable at Mode 3.

3. Qualitative hybrid bond graph-based detection and isolation

3.1 Temporal Causal Graph and Parametrized Causality

The DX community from the field of Artificial Intelligence, have developed a number of diagnosis algorithms based on consistency-based techniques [5]. There are many works in the literature that use the BG and HBG modeling language for this purpose. For example, [9] developed a diagnosis schema, which called TRANSCEND, based on the qualitative analysis of fault transient information for diagnosis of continuous systems. Since hybrid systems cannot be described by a single continuous or discrete event model, [20] extend the TRANSCEND system to Hybrid TRNASCEND where the diagnosis algorithms use a combination of qualitative and quantitative reasoning mechanisms. Authors in [21] propose a qualitative event-based approach to fault diagnosis of hybrid systems that extends the TRANSCEND and Hybrid TRANSCEND methodologies to incorporating discrete faults.

In all proposed frameworks, system uses a *Temporal Causal Graph* (TCG) model in order to analysis fault transients. Then, the diagnosis is based on this analysis, where observed deviations from nominal behavior expressed in qualitative form are compared against qualitative predictions of faulty behavior, i.e., *fault signature*, to isolate faults. For a particular mode, the TCG is constructed based on the system equations for that mode. Using the bond graph model this process is made easy. First, causality is assigned using SCAP [22], or, in the case of HBGs, may be reassigned based on the assignment of the previous mode using Hybrid SCAP [22,23,24]. After that, the bonds are converted to system variables and the bond graph elements are converted into labeled edges connecting the variables of their associated bonds (see Fig. 8). Signal links are converted into single edges, with the qualitative label corresponding to the qualitative relationship between the variables.

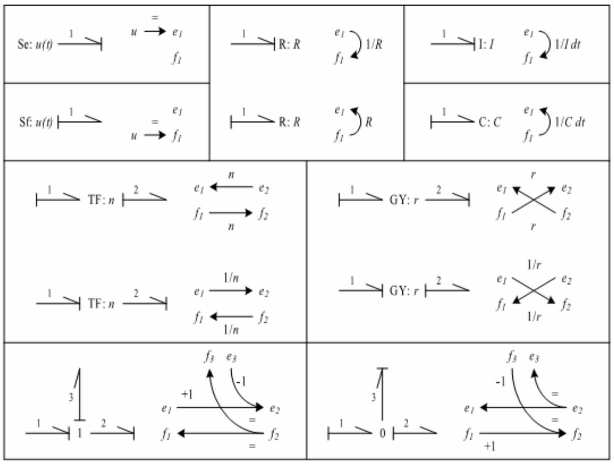


Fig. 8 Temporal Causal Graph transformations.

Fig. 9 shows a bond graph of the three-tank system and its corresponding TCG in the case when control junctions are both in mode OFF.

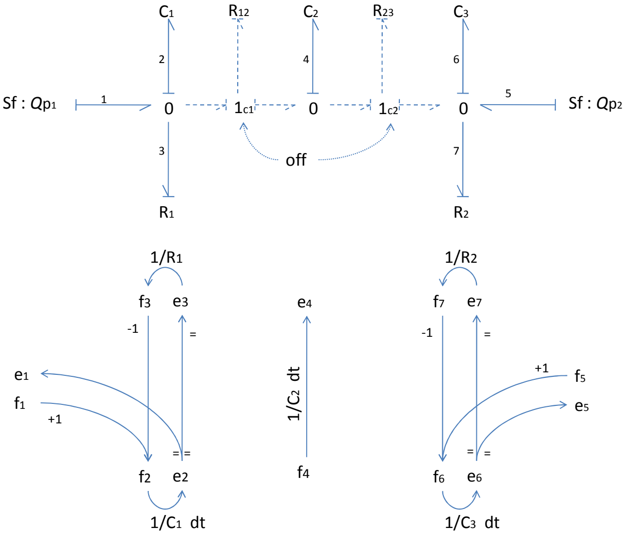


Fig. 9 Temporal Causal Graph of Three Tank System in mode 00.

The numbered bonds are converted to corresponding variables with subscript indicating the bond number. For example, bond 3 becomes e_3 and f_3. The resistor relates these two variables, i.e., $f_3 = \frac{1}{R_1} e_3$. The causality of the bond indicates that the 0-junction is imposing effort on R_1, and R_1 is imposing flow on the 0-junction. Therefore, the causal relationship is from e_3 to f_3. The label is R_1, which corresponds to the constituent equation of the resistor in the given causality. For the capacitance, the relationship between f_2 and e_2 is an integration, hence the *dt* label. Junctions sum one type of variable according to the bond signs, and set the other type equal. For the 0-junction, bond 2 is determining the effort of the junction, so e_1 and e_3 are set equal to e_2 . According to the bond signs, $f_2 = f_1 - f_3$. Since f_2 must be determined, f_1 and f_3 causally affect f_2 with labels +1 and -1, respectively.

In order to deal with the change of causality, the TCG can be derived for each possible system configuration or mode. However, in case of many locally acting switches, the combinatorial explosion quickly leads to an intractable problem. These problems can be mitigated to some extent by dynamically generating the TCG of each possible system mode in response to a failure. This may still result in a problem with large computational complexity which can be further reduced by measuring system variables that indicate specifically which local switches may have occurred and predictions for each of the variables that determine different causal assignments are required to be made and analyzed [25]. Once a set of possible TCGs is available, Gaussian decision techniques have been applied to compute the most likely mode of continuous behavior [9]. Others attention to hybrid diagnosis [26] concentrates on efficiently processing a set of TCGs. [27] describes how a hybrid model can be made amenable to the diagnosis algorithms that were developed in [28] by systematically generating one parametrized TCG. In this graph, the directed links are enabled by conditionals that correspond to the mode in which these links are present. The result is a set of predictions that are parametrized by the state of the local switches and the diagnosis problem then becomes one of constraint satisfaction [16]. The solution to this constraint satisfaction problem contains the possible parameter changes (i.e., the faults) and the effect on the system mode that this is required to have.

3.2 Parametrized Causality: Temporal Causal Matrix

To derive qualitative predictions, the system may be written as a directed graph that captures the causal (directed) relations between system variables [27]. Hence, the TCG can be represented by a weighted adjacency matrix where the columns are cause and rows are the effect variables and the entries capture the parameters on the graph edges. This is called the Temporal Causal Matrix (TCM). In the case of switched systems, modeled discontinuities result in causal changes. Therefore, the TCM may take several different forms and so do the corresponding predictions of future behavior, depending on whether a mode change occurs. To handle the change in TCM, the causal relations can be parametrized to make them dependent on the mode of operation. To this end, first the system is described in a noncausal form by using implicit equations. An implicit model of the three-tank (see Fig. 1) consists of the following equations:

$$\begin{cases} \alpha_1(-P_{c1}+P_{R12}+P_{c2})+(1-\alpha_1)f_{R12}=0 \\ \alpha_2(-P_{c2}+P_{R23}+P_{c3})+(1-\alpha_2)f_{R23}=0 \\ f_{c1}-Q_{p1}+f_{R1}+f_{R12}=0 \\ f_{c2}-f_{R12}+f_{R23}=0 \\ -f_{c3}+f_{R23}-f_{R2}+Q_{p2}=0 \\ P_{c1}-C_1\lambda^{-1}f_{c1}=0 \\ P_{R1}-R_1f_{R1}=0 \\ P_{R12}-R_{12}f_{R12}=0 \\ P_{c2}-C_2\lambda^{-1}f_{c2}=0 \\ P_{c3}-C_3\lambda^{-1}f_{c3}=0 \\ P_{R2}-R_2f_{R2}=0 \\ p_{R23}-R_{23}f_{R23}=0 \end{cases} \quad (9)$$

where λ represents the time differentiation operator and λ^{-1} indicates integration over time. From Eq. (9.1), i.e. $\alpha_1\left(-P_{c1}+P_{R12}+P_{C2}\right)+\left(1-\alpha_1\right)f_{R12}=0$, in case the water in tank 1 or tank 2 reach the level 0.5 the pipe R_{12} become conducting, $\alpha_1 = 1$, and $P_{c1}-P_{c2}=P_{R12}$, else, $\alpha_1=0$, and $f_{R12}=0$. The TCM for this system of equations contains the relations between each of the variables. For example, Eq. (9.6) embodies a temporal relation between P_{c1} and f_{c1} and Eq. (9.2) a relation between P_{c2} , P_{c3} and P_{R23} that is only active when α_2= 1.

In TRANSCEND framework, only the three values $-$, 0, $+$ are used to indicate values that are too low, normal, and too high, with respect to some nominal value, respectively. For example, a value of a model variable that is measured to be above its nominal value is marked $+$. In case the outflow through pipe line R_{12} of the tank system in Fig. 1 is too high, this is represented by f_{R12}^{+} . To find parameter deviations, a back propagation algorithm is used in TRANSCEND framework [27]. In qualitative matrix algebra this is equivalent to repeated multiplication of the initial deviation with the transpose TCM. Next, predictions of future system behavior are generated for each of the possible parameter deviations. To achieve a suffently high order prediction for the measured variable the initial deviation is repeatedly multiplied with the TCM. The TCM is derived from an implicit model formulation, Eq. (9), that includes mode selection parameters, α_i , to switch between equations. The possible causal assignments of higher relations are then made mutually exclusive by introducing selection parameters [27]. The TCG for three thank plant is shown in Fig. 10. Junctions and resistors define instantaneous magnitude relations, and capacitors and inertias define temporal effects on causal edges. For this example, to simplify the TCG structure, all = links and corresponding variables have been removed.

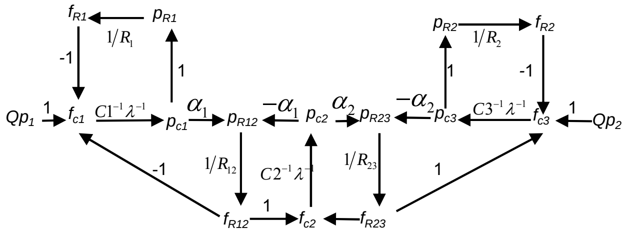


Fig. 10 Temporal Causal Graph of Three Tank System.

The resulting TCM is given by:

$$\begin{bmatrix} 1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \lambda^{-1} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \alpha_1 & 0 & 0 & 1 & 1 & 0 & -\alpha_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda^{-1} & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha_2 & 1 & 1 & 0 & -\alpha_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda^{-1} & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} f_{c1} \\ P_{c1} \\ P_{R1} \\ f_{R1} \\ P_{R12} \\ f_{R12} \\ f_{c2} \\ P_{c2} \\ P_{R23} \\ f_{R23} \\ f_{c3} \\ P_{c3} \\ P_{R2} \\ f_{R2} \end{bmatrix} \qquad (10)$$

The predictions of the TCM are parametrized by the active mode. This leads to more efficient diagnosis compared to the use of a bank of TCMs, which, in this case of two switches, would consist of four TCMs that need to be processed separately. The fault detectability and fault isolability of each parameter is gained from the {D_b, I_b} values of the four FSMs. This information can be summarized in tables 4, 5, 6, 7 and 8.

Table 4: FSM at mode 1 ($\alpha_1 = \alpha_2 = 1$)

Mode 1	P_{c1}	P_{c2}	P_{c3}	D_b	I_b
R_1^+	0+	00	00	1	1
C_1^-	+-	0+	00	1	1
R_{12}^+	0+	0-	00	1	1
C_2^-	0+	+-	0+	1	1
R_{23}^+	00	0+	0-	1	1
C_3^-	00	0+	+-	1	1
R_2^+	00	00	0+	1	1

Table 5: FSM at mode 2 ($\alpha_1 = 0, \alpha_2 = 1$)

Mode 2	P_{c1}	P_{c2}	P_{c3}	D_b	I_b
R_1^+	0+	00	00	1	1
C_1^-	+-	0+	00	1	1
R_{12}^+	0+	0-	00	1	1
C_2^-	0+	+-	00	1	1
R_{23}^+	00	00	00	0	0
C_3^-	00	00	+-	1	1
R_2^+	00	00	0+	1	1

Table 6: FSM at mode 3 ($\alpha_1 = 0, \alpha_2 = 1$)

Mode 3	P_{c1}	P_{c2}	P_{c3}	D_b	I_b
R_1^+	0+	00	00	1	1
C_1^-	+-	00	00	1	1
R_{12}^+	00	00	00	0	0
C_2^-	00	+-	0+	1	1
R_{23}^+	00	0+	0-	1	1
C_3^-	00	0+	+-	1	1
R_2^+	00	00	0+	1	1

Table 7: FSM at mode 4 ($\alpha_1 = \alpha_2 = 1$)

Mode 4	P_{c1}	P_{c2}	P_{c3}	D_b	I_b
R_1^+	0+	00	00	1	1
C_1^-	+-	00	00	1	1
R_{12}^+	00	00	00	0	0
C_2^-	00	+0	00	1	1
R_{23}^+	00	00	00	0	0
C_3^-	00	00	+-	1	1
R_2^+	00	00	0+	1	1

Table 8: Fault Detectability and fault Isolability of components

θ	Detectability	Isolability
R_1^+	all-mode	all-mode
C_1^-	all-mode	all-mode
R_{12}^+	mode 1, 2	mode 1, 2
C_2^-	all-mode	all-mode
R_{23}^+	mode 1, 3	mode 1, 3
C_3^-	all-mode	all-mode
R_2^+	all-mode	all-mode

4. Conclusions and discussion

In this paper, causal properties of bond graphs are used to generate the elimination schemes such that direct and deduced redundancies can be expressed only in terms of known process variables. First, a quantitative fault diagnosis framework for hybrid systems is developed basing on HBG and on a set of unified constraint relations called global analytical redundancy relations (GARRs). These relations can be derived systematically from the diagnostic hybrid bond graph (DHBG). The GARRs explicitly show the system component fault detectability and fault isolability and generate alarm signals for effective and efficient fault detection and isolation (FDI). The quantitative method treats sensor, actuator and parameter faults which are of three types abrupt, progressive and intermittent. Noise and robustness issues are considered in such method. The application of the quantitative diagnosis is constrained to both open and closed loop systems. However, the subset of the equations of complex models with implicit relations, complex non-linearities, and algebraic loops, etc., cannot be resolved.

In addition to the quantitative method, another qualitative approach dealing with temporal causal graph (TCG) is used. This method allows one to ameliorate fault isolation, it treats only faults parameter indeed, noise and robustness issues are not considered in such diagnosis and only abrupt faults are handled. The application of this method is constrained to only open-loop systems because it relies on temporal trends of systems evolution obtained from the measurements which may not show any deviation if they are controlled. It is well known that controllers try to hide the fault effects. The qualitative approach overcomes limitations of quantitative schemes, such as convergence and accuracy problems in dealing with complex non-linearities and lack of precision of parameter values in system models. The qualitative reasoning scheme is fast, but it has limited discriminatory ability.

As presented in this article, a single approach for diagnosis has limitations, and it does not satisfy all the requirements to perform a good diagnosis. Hence, several works can be found in literatures that combine diagnostic approaches. The objective is to find two or several approaches that complete each other, in a way that the qualities of one approach can overcome the drawback of another. In a future work, we will focus on complex models with implicit relations, complex non-linearities, and algebraic loops.

References

[1] D. De Scuter, “Optimal control of a class of linear hybrid systems with saturation”, SIAM Journal of Control Optimization, Vol. 39, No. 3, 2000, pp. 834-851.

[2] J. J. Gertler, “Fault detection and diagnosis in Engineering Systems”, Ph.D. Dissertation, George Mason University, Fairfax, Virginia, 1998.

[3] R. J. Patton, P. M. Frank and R. N. Clark, Issues in fault diagnosis for dynamic systems, Springer Verlag, New York, 2000.

[4] w. C. Hamascher, J. Decleeer and L. Console, “reading in model based diagnosis”, Morgan-kaufmann Pub, San mateo, 1992.

[5] R. Reiter, “A theory of diagnosis from first principles”, Artificial Intelligence, 1987, Vol. 32, pp. 57-95.

[6] M. Staroswiecki and G. Comtet-Verga, “Analytical redundancy relations for fault detection and isolation in algebraic dynamic systems, Automatica, 2001, pp. 687-699

[7] B. Pulido and C. Alonso-Gonzlez, “An alternative approach to dependency-recording engines in consistency-based diagnosis”, Atificial Intelligence, 2000, pp. 111-120.

[8] B. Pulido and C. Alonso-Gonzlez, “A compilation technique for consistency-based diagnosis”, IEEE Trans, 2004, pp. 2192-2206.

[9] P. J. Mosterman and G. Biswas, “Diagnosis of continuous valued systems in transient operating regions, IEEE Trans, 1999, pp. 554-565.

[10] E. J. Manders, S. Narasimhan, G. Biswas and P. J. Mosterman, “A combined qualitative/quantitative approach for fault isolation in continuous dynamic systems, IFAC, 2000, pp. 1074-1079.

[11] W. Borutzky, “”Representing discontinuities by means of sinks of fixed causality”, International Conference On Bond Graph Modeling, 1995, Vol. 27, pp. 65-72.

[12] P. J. Gawthrop, “Hybrid bond graph using switch I and C components”, International Conference On Bond Graph modeling, 1997.

[13] J. E. Stromberg, J. Top and U. Sodermann, “Variable causality in bond graphs by discrete effects, International Conference On Bond Graph modeling, 1993, pp. 115-119.

[14] C. B. Low, D. W. Wang, S. Arogeti and Z. J. Bing, “Causality assignment and model approximation for quantitative hybrid bond graph-based fault diagnosis, IFAC, 2008, pp. 10522-10527.

[15] J. Feenstra, P. J. Mosterman, G. Biswas and R. J. Breedveld, “Bond graph modeling procedures for fault detection and isolation of a complex flow processes, International Conference On Bond Graph modeling, 2001, Vol. 33, pp. 77-82.

[16] M. Torrens, R. weigel and B. Faltings, “Java Constraint library”, Workshop on Constraints and Agents, 1997.

[17] A. K. Samantary, S. K. Ghoshal, S. Chakraborty and A. Mukherjee, “Improvements to single fault isolation using estimated parameters”, Simulation: Transactions of the Society for Modeling and simulation International, 2005, Vol. 81, No. 12, pp. 827-845.

[18] A. K. Samantary, K. Medjaher, B. O. Bouamama, M. Staroswiecki and G. Dauphin-Tanguy, “Diagnostic bonf graphs for online fault detection and isolation”, Simulation Modeling Practice and Theory, 2006, Vol. 14, No. 3, pp. 237-262.

[19] C. B. Low, D. Wang, S. Arogeti and J. B. Zhang, “Causality Assignment and Model approximarion for Hybrid Bond Graph”, IEEE Transactions on Automation Science and Engineering, Vol. 7, No. 3, 2010, pp. 570-580.

[20] S. Narasimhan, “Model-based diagnosis of hybrid systems”, Ph. D. Dissertation, Vanderbilt University, 2002.

[21] M. J. Daigle, “A qualitative event-based approach to fault diagnosis of hybrid systems, Ph. D. Dissertation, Vanderbilt University, 2008.

[22] I. Roychoudhury, M. Daigle, G. Biswas, X. Koutskous and P. J. Mosterman, “A method for efficient simulation of hybrid bond graphs, International Conference On Bond Graph modeling, 2007, pp. 177-184.

[23] D. C. Karnopp, D. L. Margolis and R. C. Rosenberg, “Modeling and simulation of Mechatronic Systems, New York: John Wiley & Sons, 2000.

[24] M. Daigle, I. Roychoudhury, G. Biswas and X. Koutskous, “Efficient simulation of component-based hybrid modeld represented as hybrid bond graphs, Institute for Software Integrated Systems, Vanderbilt University, 2006.

[25] S. Narasimhan, G. Biswas, G. Karsai, T. Pasternak and F. Zhao, “Building observers to address fault isolation and control problems in hybrid dynamic systems, In proceeding of the IEEE International Conference on Systems, 2000, pp. 2393-2398.

[26] S. Mclraith, G. Biswas, D. Clancy and V. Gupta, “Hybrid systems diagnosis”, Springer Verlag, 2000.

[27] P. J. Mosterman, “Diagnosis of physical systems with hybrid models using parametrized causality, Institute of robotics and mechatronics , Germany, 2001.

[28] P. J. Mosterman, “A hybrid bond graph modeling paradigm and its application in diagnosis, Ph. D. Dissertation, Vanderbilt University, 1997.

A novel approach for modeling user's short-term interests, based on user queries

Albena Turnina

**Information Technology Department, Sofia University
Sofia, Bulgaria**

Abstract

In this paper is presented a novel approach for modeling user short-term interests, based on user queries. The proposed user model is represented by a weighted semantic network, composed of nodes and arcs. This semantic network can be used to express relations between query terms submitted by a user in his searches. Our approach is rooted on the idea that there are relations between topics of user's interests, which can be measured and used to provide a search context. We propose an approach for modeling user's interests, based on data taken from his previous queries. We aim to identify relations between topics of user's interests in a set of successive queries. The search personalization features of ShareTec digital library are also presented and the implementation of the proposed model is outlined.

Keywords: *User profile, short-term interests, adaptive query, search personalization.*

1. Introduction

In this work we propose a novel approach for modeling user short-term interests, based on data, taken from user's queries, intended to facilitate a personalized search. The proposed user model is represented by a weighted semantic network, composed of nodes and arcs. This semantic network can be used to express relations between query terms, inputted by a user in his searches. To the best of our knowledge, the same model does not exist, although some of the ideas behind the model have already been realized in existing systems for a personalized search. Our approach is based on the idea that there are relations between topics of user's interests, which can be measured and used to provide a search context. We propose an approach for user modeling, based on data taken from user's previous queries and an identification of relations between topics of interests in successive queries. The proposed model could be able to track dynamic changes in user's interests and to stay updated, representing actual topics of user's interests. We aim to investigate the possibilities for an achieving search personalization using solely the proposed model, which takes into account only usage data, taken implicitly from the queries. The intuition behind our model is that there are relations between different topics of user's interests, which are specific to a particular user and these relations are possible to be detected and used for delivery of personalized search results.

2. Theoretical background

There are three paradigms for information access to a web content in a hypertext environment - searching by browsing, searching by means of queries submitted to search engines and through recommendation systems [18]. The recommendation systems offer topics, analyzing what users with similar interests were chosen in the past. In searching by browsing, the users explore Web pages, one by one by following hyperlinks. The browsing is not a convenient way to find a certain and specific information. In this study we focus on the queries and investigate the approaches to provide a search personalization by adapting the queries.

The two main kind of personalization systems are Content-based systems and Collaborative-filtering systems. The personalization in the Content-based systems is based on the similarity between a user and the documents in a collection. The Collaborative-filtering systems exploit the idea that the users with similar interests are likely to prefer the same resources. There are other personalization techniques and systems, which are based on rules or explore demographic data as well as hybrid systems which combine several techniques together. A short overview of the existing personalization strategies and example systems is presented below.

- Content-based systems – encompasses systems which track user`s browsing behavior and recommend topics similar to those previously chosen by the user. This technique has some shortcomings, such as "over-specialization" and the lack of relations between topics [25], [20]. Example systems are WebWatcher and Letizia.

- Collaborative-filtering systems – includes systems that use cumulative experience of a group of users in order to facilitate searching experience of a certain user. The idea behind this strategy is that users with similar behavior/preferences are likely to have similar interests. Employed algorithms investigate similarities between the users and gives recommendations according to preferences of the most similar neighbor. Collaborative filtering approach is used in popular systems such as Yahoo, Excite, Microsoft Network and Amazon.com.

- Demographic-based systems – encompasses systems that group users on the base of their demographic similarities. In this kind of systems recommendations are given according to demographic characteristics of users, such as age and gender [23].

- Rule-based systems - in these systems the recommendation are given to the users according to their answers to a set of predefined questions [35]. Example system is Broadvision.

- Hybrid systems - combines two or more personalization strategies and approaches [2].

The search personalization techniques are used in systems that provide individualized collections of pages to users. This personalization is based on a user model which presents the user’s interests and search activities [18]. According to (Micarelli et al) the search personalization approaches are subdivided into several major types, which are as follows: Current Context, Search History, Rich User Models, Collaborative approaches, Result Clustering and Hypertextual Data [18].

The personalized information retrieval (PIR) systems can be classified according to different criteria, the main of which are: the type of the system and the personalization approach. The type of the system is based on the application domain or on the type of the service, provided by the system, like Web search systems, Multilanguage search systems, personalized news feeds, e-learning, etc. The personalization approaches are related to the way in which personalization is realized. The personalization in PIR systems can be realized by a query adaptation, by personalization of search results, or both. In multilingual systems the personalization process can include additional steps, needed for translation of the query and results in different languages [21], [22]. According to the chosen approach, the various personalization techniques can be implemented [14].

The query adaptation can be realized by modification of the query, by relaxation of the query or by substitution of the original query with one or more adapted queries. The modification of the query can be performed by wide variety of techniques which encompass: the substitution of the query terms with terms or concepts, taken from a reference vocabulary; the expansion of the query by terms, taken from a user profile; the change of weights of terms and/or relations between them and using the methods of pseudo-relevance feedback and relevance feedback. The modification of the query can be performed explicitly by the user as well.

The expansion terms can be derived from the user profile. These expansion terms are representative for the user’s preferences and serve to give a search context. The exact number of the expansion terms can be predefined or can be selected dynamically. (Chirita et al., 2007) argue that the number of the expansion terms can be tuned according to features of the query, such as the query length, scope of the query, etc. [10]. The process of a dynamic selection of the expansion terms is known as a “selective query expansion”.

The main advantage of the query adaptation approach is the lack of additional processing required in the other two approaches. However, the query adaptation is unlikely to influence significantly the returned results [14].

The personalization of the search results is the process of a selection of the most relevant results and their ranking according to a certain user, a community of users or all users of a system. This personalization can be realized by a number of techniques, which are as follows: pre-ordering of results, filtering of results and result scoring [18].

The techniques of pre-ordering of results and the filtering of results are performed after the initial results list has been extracted from the system. They are realized as an additional processing step over the retrieved set of documents. In contrast, the technique of a result scoring takes place in the process of initial selection of documents where the adaptivity is implemented as an integral part of the result scoring function.

The essential part of each personalization systems is the user model [7], [13], [19]. The user model represents the user’s characteristics and the user’s behavior in a system. There exist different approaches for modeling of a user according to purpose and scope of the system. For example, Web personalization systems use tracking algorithms to follow the user in his surfing, whereas Adaptive hypermedia systems build a user profile, based on the user interaction with the system. The user can be modeled according to his demographic characteristics and specific features such as: level of expertise in a particular domain, cognitive abilities, visual characteristics and etc. (Brusilovsky, 2001) classification of user's characteristics encompass the knowledge, purpose, background, interests, environment and experience with hypertext of a user [8]. (Gasparini et al., 2011) present a new approach for modeling of a user in a domain of e-learning, which takes into account contextual user aspects, such as technological, educational, personal, and cultural characteristics [12]. Related research on the implications of the cultural user's characteristics in the design of user interfaces are conducted by [9], [24].

The distinction between a user profile and a usage profile can be found in different sources. The former is related to user's characteristics and the usage profile is related to the behavior and activities of a user in a system. However, user profile can be found as a representative for both – a user and a usage profile. The user profile is considered to be essential and the most common part of the personalization systems. There are numbers of different techniques for the symbolic representation and the construction of the profile. The degree of a complexity of a user profile depends on purpose of the system. It varies from a simple explicit questionnaire to a complex dynamic structure, containing both explicit and implicit information. The important characteristic of a sophisticated user profile is its ability to be changed dynamically, reflecting the user changes and his shifts of interests. The user profile can take part in a personalization process of the Information Retrieval systems in three different ways - when takes place in the retrieval process, when is used for pre-ordering of the search results and when is used for a query adaptation [14].

The representation of users and documents can be based on various methods and techniques taken from different fields like Information Retrieval or Artificial intelligence. The vector-space model is popular and largely used model for documents and user profiles representation. In this model any document or profile is represented as a set of keywords or as an n-dimensional vector. This representation serves as a basis for comparison between a document and a profile and allows measurement of degree of similarity between them. The document is considered to be relevant to a user if the degree of similarity is over a predefined threshold value. The popularity of the vector-space model is mainly due to its simplicity and proven effectiveness. However, the vector-space model has some shortcomings such as potential loss of information when documents pass linguistic processing in which the keywords are separated. Because of this processing the embedded meaning in sentences and phrases can be lost. The disadvantage of the profile, constructed by means of keywords is that it requires a large amount of user's feedback. This feedback is needed for the selection of the exact words which present a given user. A largely known method for calculation of probability that the document is relevant to a user is by means of Bayesian probabilistic classifier. Other user profile representation techniques include semantic networks, associative concepts or rules although the last serve mainly in the field of a Web log mining [14].

The user profile, based on semantic network/s is composed of nodes (nod) and links (arcs) between them. The semantic network is a graphical notation, used for knowledge representation, based on interconnected nodes and arcs. The semantic networks can be used to represent the knowledge over which the inference can be done. According to (Sowa, 1992) the most popular semantic network types are as follows: Definitional networks, Assertional networks, Implicational networks, Executable networks, Learning networks and Hybrid networks [28]. The Learning network type is of a particular interest for this study because of its ability to extend itself when new knowledge is acquired. This new knowledge is able to modify the existing semantic network, to add new or to remove the existing nodes and links and to change their assigned weights. The semantic network can serves to represent symbolically the relations between terms and/or concepts and their mutual occurrences in texts or documents. Varies weighting schemas can be implemented to measure weights of nodes and links between them.

The semantic network as well as the ontology can be used to model the relation between a word and a concept. This is of a particular importance when the documents are expressed in natural language. The mapping of a word to a concept can be accomplished by means of reference vocabulary such as WordNet, by learning mechanism or manually.

The user profiles, based on semantic networks can have different level of sophistication and can encompass one or multiple networks, each one having its owl level of complexity. Amongst the systems, which implement a user profile, represented as semantic network/s are: WIFS, InfoWeb, SiteIF, ifWeb and PIN. The user profiles in these systems have different level of complexity and sophistication. In more complicated implementations, concepts are represented by nodes called Planets and the words are represented by nodes called Satellites, connected to Planets. In WIFS system the individual semantic network for each user's interest is deployed.

The user model can be represented by specially designed user ontology or by weighted domain ontology. A widely deployed ontology user model is an overlay model in which the user is modeled by domain ontology and his knowledge is represented as a part of an expert knowledge. The ontology allows inference over collected facts and drawn of new facts [11]. An interesting approach is the construction of a user profile, based on domain ontology with weights, assigned to concepts according to user's interests. These weights are updated through spreading activation algorithm [27]. In [15] the authors present a set of statistical methods for learning user ontology from domain ontology by spreading activation algorithm.

The user profile can be constructed by means of a diverse set of techniques emerged from various fields such as machine learning, Information retrieval and Artificial intelligence. Nevertheless not very popular, there are approaches based on the use of genetic algorithms and neural networks.

In this work we present a user model, intended to provide search personalization, based only on usage data taken from user's previous queries. We do not argue that this user model can serve as only source of personalization in a particular system, rather it is intended to complement long-term profiles, modeling user's characteristics. The details of our approach and characteristics of the proposed model are outlined in Section 4 of this work. Section 3 provides brief overview of existing systems, related to our work. The details on the planned implementation and brief presentation of the digital library ShareTec are shown in Section 5. Finally in Section 6 the conclusion and future directions are presented.

3. Overview of the existing systems

Our main focus of interest is directed towards personalization approaches which tackle the problem of a query adaptation. The central part of our work is the development of a novel approach for modeling user's short-term interests. Therefore we have been investigating a number of systems, which explore the problem of a user modeling. In this brief overview the systems and approaches, related to our work are presented. Firstly, we explore the systems that model the user's interests by means of semantic networks. Secondly, we focus on the systems that adapt the original user query in order to provide personalization.

IfWeb is a personalized information retrieval system in which a user profile is represented by a semantic network. The profile is constructed by extracted words which have the highest weight of a set of documents. Each of the extracted word in their approach is used to create a single node in the semantic network. Nodes are connected by links when the words they represent appear together in documents [1]. This approach is extended by SiteIF system where the extracted words are linked to concepts, taken from a dictionary WordNet. The similar approach can be observed in PIN system where the words, extracted from documents are nouns and concepts are learned by means of neural networks. In SiteIF system the user model is represented by weighted semantic network in which nodes are connected by links that have different weights. This system uses WordNet for finding semantic similarity (synonymy) between words [31]. InfoWeb is a filtering system for retrieval of documents in digital libraries. In this system the profile is represented by a semantic network which models the long-term interests of a user. Initially, the semantic network is represented by a collection of unconnected nodes as each node is a separate concept. These nodes called Planets contain single, weighted term that is representative for a given concept. In the process of gathering information about a particular user, his profile is enriched with weighted words, mapped to different concepts. Words are contained in nodes, called Satellites which are connected to conceptual nodes - Planets. The conceptual nodes – Planets, can be linked to each other's. The user model in system WIFS is partially based on semantic networks. The specific of WIFS system is explicitly provided a set of topics of interests by a user during his initial registration. The interests can be implicitly extended further by the system as a new data about a user is gathered in automatic manner. The set of topics, that a given user is interested in, are used to associate the user with a set of stereotypes. Stereotypes themselves are defined by people and experts to describe the knowledge domain of computer science. In this system the user profile models different aspects of the knowledge of the user, which includes personal information and a list of active stereotypes which are associated with the user. Each user's interest is modeled by a separate semantic network. Each semantic network contains a main node called Planet and many Satellites nodes connected to the main node [19].

The conceptual nodes in InfoWeb system are created by the technique of explicit feedback, whereas in WIFS system they are created by human experts. In MiSearch the user profile is represented by means of two alternative models, both based on a vector-space model. The aim of both models is to describe the long-term interests of the user. The first model consists of concepts derived from user's queries, and the second one from concepts derived from snippets of the selected (clicked) documents by the user. The models contain numbers of vectors. Each of them is representative for a particular user's interest. Both models use the ODP hierarchy to categorize the concepts. Mapping of retrieved documents to ODP categories is done by means of a text classifier [29]. (Zhou et al.2012) present a system where the user model is constructed by terms, extracted from the user's tags and bookmarks. They create a statistical model based on these bookmarks to represent the different topics. Their model can be used for an identification of topics in documents, based on user's tags and bookmarks. By means of this model, the most relevant terms for a particular user, can be identified and used to enrich and expand the user query to the system [36]. The query adaptation can be performed on the base of rules. (Koutrika and Ioannidis, 2004) proposed a method, based on rules for rewriting the query by means of which the movies database can by queried. In their approach, the original query is replaced by a number of queries as the process is managed by a set of rules, based on the individual preferences of the user. They propose to connect the different queries through a disjunctive logical operator "OR". For example, if a user likes a particular film genre and make a search for movies issued in a particular year, the system will issue the query that searches the specified film genre in addition to the original search [16]. (Stamou and Ntoulas, 2009) proposed a system in which personalization is performed by re-ordering of search results returned by Google. In their approach they model not only the long-term interests of a user but also short-term interests, taken from a current user query [30]. The long-time interests are modeled on the base of the information gathered from the past user queries as well as from the selected by a user results in the list of returned results. Document and query terms are linked to the concepts, derived from reference ontology. The current user's interests are identified by a current query submitted to the system before the results are presented. After sending the query, the system tries to identify whether that same query was issued in the system before. In such case, the results that were drawn before are presented to a user as a result of a current query. Otherwise, the system determines the degree of similarity between the concepts of the current query and the documents previously classified under different concepts of reference ontology. Thus, the authors model both long-term and short-term interests of a user and by combining both, calculates the user's interests. The relevance of the retrieved documents is a function of user's interests and is calculated as the sum of long-term interests, short-term interests and the weight of the document [30].

According to (Shen et al., 2005) the two main aspects of a personalized search are user's interests and a context of the search (understood as a disambiguation of query). In their system UCAIR they focus on modeling short-term interests of a user, through an approach called eager implicit feedback. In this approach, the current context of a query is output from the previous query within the same session, as well as from results from a previous query, identified as relevant by the user (clicked). In order to determine whether two consecutive queries are related, the system performs two searches - one for the previous query and one for the current query. The lists of results derived from both searches are then compared to determine the similarity between them. If the queries are semantically related, the current query is expanded with terms, taken from the previous query. The results retrieved due to the adapted query, are pre-ordered according to a user model. The user model updates itself dynamically when the user clicks on documents, presented in the result list [26].

A different approach for adapting query and results is presented by (Liu et al., 2004). They propose a system in which the user model is built as a vector consisted of conceptual terms. The conceptual interests are recorded on the base of Google Directory, which in own turn is based on ODP. In their approach, the query adaptation is performed by specifying the category of the query. Thus, the system is trying to identify the concepts, related to the query in order to provide the necessary context of the search [17].

4. Our approach – a short-term user model

4.1 Introduction

In this work is proposed a novel approach for modeling short-term user's interests, needed to provide a search personalization. Through the proposed model we can express in a formal way the relations between topics of user's interests, taken from a data derived from a user searches. Our aim is to provide a search personalization, based on a query adaptation, performed by enrichment of the current query with expansion terms taken from a user profile. The expansion terms are selected according to the weights of links between them and the current query terms as well as to their own weights. These weights serve to represent the importance of the given topics and significance of relations between the different topics to a particular user. The weights are assigned to nodes and links in the user profile according to proposed weighting schema. We argue that keywords (query terms) used by a user in his searches, reflect his information needs and interests. The user in his search activities in the system submits multitude of queries, some of which are related to each other. Each of these quires can be a specification of the previous one. It could be a series of queries that taken together express a user's search intention. Therefore, the discovery of such related terms in a series of queries, submitted by a user is essential. In our model we try to identify the query terms, which often appear together in the user searches, or often appear in the consecutive queries and to recognize the connected topics for a particular user. We believe that such relations exist and can reflect trends in user's interests, which can be useful for providing personalized search results. In our work we have been inspired by so-called method of "previous query" used in Google search engine, the meaning of which is that the previous queries issued by the user influence the search results of the current query. For example, if the user searches for information about the trip, but before that he had shown an interest in a particular country, the system will offer results in which the terms trip and that country appear together [34].

In our model we want to reflect the importance of some terms for the user which appear often in his queries as well as significance of connections between some terms. In order to measure relatedness of terms we use the term "distance" which reflects the distance of terms in a series of a predefined number of consecutive user queries. This "distance" serves as a way to assign weights to links between concept nodes in our model. These weights of the links will be used for determination of the terms, used for the expansion of the current query. The model is built by a predefined number of successive user queries as their exact number is still a subject of investigation. The proposed model is intended to be changed dynamically with acquisition of new knowledge in order to be able to represent the changeable user's information needs. The model is intended to operate in systems in which the searching is performed by means of free words. It has been shown that this is the easiest and most preferred way for users to perform a search, although it can present ambiguity. Through the presented model we proposed one solution to the problem of an identification of a search context which can help in avoiding ambiguity and can serve as an enabler for the search personalization.

4.2 The characteristics of the model

The proposed model is represented by a weighted semantic network composed of two kinds of nodes – one for presenting the concepts (concept nodes) and the other for presenting the query terms (term node). The term nodes can be associated to concept nodes according to semantic similarity between them. The concept nodes in the model can be related by weighted links. The concept nodes and the term nodes have their own weight, assign according to proposed methodology. The presented model has some common features with the user models of InfoWeb and WIFS systems but has its own specific characteristics as well. These characteristics are related to the way in which relatedness between the query terms are modeled and the weights to nodes and links are assigned. The weights of the nodes and the links in the proposed model can be changed dynamically, according to a user searches.

The query terms pass a linguistic processing - stemming before been added to a model as the term nodes. In order to associate the term node to the concept node, some vocabulary or reference ontology has to be used. One of the most viable solutions is WordNet vocabulary. WordNet is a collection of 100,000 words, nearly 80.000 organized in semantic clusters (synsets). But according to a certain implementation other vocabulary can be used as well. In the use case, presented below we propose to integrate the model in ShareTec digital library. In this particular case the education ontology TEO will be used as reference ontology in order to associate the term nodes to concepts. The weight of a term node reflects the frequency of occurrence of a term in a series of user queries. The weight of a concept node is obtained as a sum of weights of term nodes associated to it. Thus, the weight of one concept node is proportional to the number of the term nodes associated to it and their frequency in user queries. This weight reflects the relative importance of that concept for a particular user. At the same time, the weight of a concept node and the weight of a term node are inversely proportional to the time past since the last usage of a given term in the user queries. Thus, the weights of the nodes reflect the actuality of the user's interests. The weights of the links reflect the degree of relatedness of the connected concept nodes and are representative for the particular user's connected interests. The weight of a link is cumulative value which sums multiple occurrences together of terms in the same query, in the previous query or in the predefined number of queries. The connection/link between the concept nodes in the model is made when the term nodes associated to them are related in the user queries. This relation is measured as it is set out below.

After the user inputs a new query, the system first checks whether the terms, contained in the query, already exist in the user profile. If a particular term is new to a profile it is added as a term node and is associated to the most suitable concept node. At the same time the default weight is set to the term node and the weight of the respective concept node is increased. If the term node has already existed in the profile its weight as well as the weight of the concept node, to which it is associated, are increased with a default value. When two or more terms occur together in the same query, the system checks whether there is a link between the concept nodes, to which the terms are associated. If the concept nodes are not linked together in a profile, the new link is created and the default weight is set to it. If the link has already existed, its weight is increased with the predefined value. At the same time, the links are created (if not existed) between the concept nodes of the current query terms and the concept nodes of the previous query terms. But the weights of these links will be lower than the weights of links between concept nodes of a current query terms. This process can recursively be done to the pre-defined number of previous queries. In this process, the more distant backward the query is the lower weights of links between its concept nodes and current query concept nodes will be. The intuition behind the model is that the terms occurring together in the same query are the most semantically related and when taken together, they express the information needs of the user of the system. That is why we propose to assign the higher weights to the links between their concept nodes.

The current query often happens to be a specialization of the previous query and this is the reason why we also assign the weights to a links between the concept nodes of the terms in these queries but they will be lower. And so on as we proceed with a predefined number of previous queries. It appears that when a user inputs a query, the terms contained in it, will be related to each other and to the terms in a given number of previous queries. We propose the following: to assign the value of one (1.00) to a weights of the links between concept nodes of terms that appear in the same query; the value of 0.90 to the links between terms in consecutive queries, the value of 0.80 to the link between concept nodes of terms in current query and query before previous query and so on. At the moment we are not quite sure about certain weighting values because they have to be tuned by means of a set of experiments. Through the proposed model, we are able to express formally the idea that the relations between query terms exist and they can be measured in the range of queries. These relations will get weaker with the increase of the “distance" between the current query and the previous query. At the same time the model will be able to express the relative importance of some query terms for the user, measuring the frequency of their appearance in his searches. We argue that the proposed model could be able to gather information for creating patterns of a user search behavior. It still remains an open research question how many queries to be traced backward in order to build a functional and computationally reasonable profile. The proposed user model will have the ability to updates itself dynamically when new nodes and links are added and removed, and the weights are changed with accordance to the actual information, collected by the system. With each new query, issued by the user, the weights of existing nodes and links will be reduced proportionately. The exact amount of reduction of weights has been not established yet. But it will be in dependence of the number of the queries, tracked backward by the system. If the value of the weight of some node or link is reduced under predefined threshold, the respective node or link will be removed from the model. We believe that could keep the model current and actual by means of the proposed method.

In the process of personalization there is a risk that the personalization could be misdirected or unwanted. Adapting the query can result in increase of relevance of returned results if it is successful but if it is not, the relevance will decrease. In the process of enrichment of a query with terms, taken from a user profile the expectations are towards improving the relevance by providing missing search context. Thus taking into account, the expected benefit of a query adaptation against the risk of an improper personalization we propose the current query to be replaced by a series of queries. This series could include queries, adapted by terms derived from the user profile as well as the original query. The expansion terms are taken from a user profile. The selection of these terms is based on weights of links, connecting terms of a current query with terms from the previous queries and on weights of terms themselves. The last weight serves to express the importance of the terms for a particular user. In this work when we talk about expansion terms we mean the concept nodes which can be used for the expansion. In order to be considered as a valid expansion term, the concept node has to have its own weight greater than predefined threshold value. Moreover, the weight of a link between the concept node used for the expansion and the concept node in the current query also has to be over threshold value. If there are such multiple links, the few rounds of the query adaptations have to be performed. In the first round, the concept nodes with higher cumulative value of the links weights and their own weights are extracted from the model and used for the expansion. In the second round, the concept nodes with lower weights are used for the expansion and so on until the weight value falls under the predefined threshold. We expect that some heuristics will be helpful in the case of having too many expansion concepts.

The search results derived from the adapted queries as well as from the original query will be mixed in a result list. This approach is related to the so-called result diversification, known from the field of IR, which consciously diversified a set of results, returned by the system in response to a user query. The advantage of this method is that the user is offered results retrieved from the adapted query and from the original, formulated by the user query. Thus, the risk that personalization may be misdirected or unwanted is avoided. The user action - clicking on the document, extracted as a result of an adapted query is evidence that personalization was successful. We propose the user feedback to be used not only to validate the effectiveness of a personalization but also to update the user profile. The user action - clicking on documents, returned from an adapted query, will add extra weight to the link between the concept nodes – one to which the term from current query is associated and the other which serves for the expansion.

5. Implementation

The proposed in this work user model is not domain dependent and as such could be implemented in systems operating in different domains. We aim to integrate the model in e-learning environment in order to represent learner short-term interests and facilitate his searches.

The project Share.TEC: Sharing Digital Resources in the Teaching Education Community aims to support teachers and educators by providing access to digital resources, related to educational field. Share.TEC system includes a portal through which an individual user can access available resources [32] [33]. The portal provides a wide range of personalization features which are based on individual users' characteristics and on their national and cultural background. Three elements are crucial in the process of providing personalization: the adaptive approaches and techniques, a system interface and a user model. The main users activities, related to the adaptivity in digital libraries are: setting preferences and searching for digital content [3] [4] [5]. The user model in ShareTec includes three sets of data, by means of which the characteristics of individual user and his behavior in the system can be described and analyzed. They are as follows:

- Explicit preferences defined by the user himself
- Implicit user preferences collected through an analysis of a user behavior in an automated manner
- Summary of the behaviors of all users of the system

In the process of searching and filtering of search results, two clusters of data are created. The first cluster contains a list of search results, whereas the other cluster contains raw data including used in the searching keywords, comments, annotations and ratings. This raw data must be processed and analyzed before being able to describe a user behavior and preferences. In order to use that data, it must be aggregated and processed into a form that can model a user behavior. As stated above, this data has to be analyzed in order to be able to generate the implicit user's preferences from it. In this work we propose a user model, which can be used for modeling a part of a usage data, namely - the query terms. The proposed model aims to examine the effectiveness of personalization, obtained by means of solely information extracted from the user queries. But this model can as well be used together with another user models. In this case we expect an increase of search relevance of results. Realized adaptive features in ShareTec portal include automatic ranking of search results according to a user profile; recommendations to the user and individualized forms for assessments, bookmarks and annotations. The user profile in ShareTec is tightly integrated with the teacher ontology (TEO), developed in the frame of the same project. The search process in the system uses a user model and TEO ontology and is based on search engine Solr. The search component performs semantic query expansion by means of following techniques:

- Expansion based on explicit preference
- Expansion based on ontology - uses a parent-child relation
- Multilingual expansion
- Expansion for recommendation - uses the most concerned resource in accordance to user profile

The analysis of the adaptive, expansive query techniques show that in Share.TEC system does not exist a query adaptation, based on previous user's queries. Share.TEC recommendation system and a search engine use various sources of data like OMM metadata, teacher ontology TEO and implicit information taken from a user model. Solr search engine is configured to use in its indices the individual fields of OMM and TEO. In this work we propose to use TEO ontology concepts for an initialization of the conceptual nodes in presented user model. The intuition behind this is that the query terms used for searching in ShareTec are more likely to be related to concepts of teacher ontology, which models the educational domain. The practical implementation in ShareTec is done with the help of special tasks in Hadoop, working in asynchronous mode with a database in the portal. Our proposal follows the existing implementation of the system and the data processing will be implemented as an additional subtask in Hadoop.

6. Conclusions and future work

Future work is aimed to the practical aspects of the implementation of the proposed model and to answer the research questions, related to it. Amongst the questions that emerge, is the determination of the proper weigh values to links between concept nodes, described in the proposed model. Another research question that needs to be addressed is to be determined the exact numbers of the expansion terms, which can be used for a query adaptation. The proposed model is based on terms, taken from the previous user queries. The numbers of successive queries, which will be used to build the model, is not determined yet. But it appears that the proper number is essential for our model. On one hand, taking too many queries may result in a lack of accuracy and actuality of the model. On the other hand, taking not enough number of queries will lead to inefficiency and inability of the model to represent user's interests. And last but not least, the validation of the proposed model has to be conducted and the determination whether it presents a viable solution for a search personalization has to be provided.

Acknowledgments

"Acknowledge to the European funded Project UNITE (FP7 248583), namely its secondment programme coordinated by UNINOVA-GRIS, that supported the development of various ideas and concepts presented in this paper"
This work was partially supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014)

References

[1] Asnicar, F.A., Tasso, C.: ifWeb—a prototype of user model-based intelligent agent for document filtering and navigation in the World Wide Web. In: Adaptive Systems and User Modeling on the World Wide Web, Chia Laguna, Sardinia (1997)

[2] Balabanovic, M. (1998).Learning to Surf: Multiagent Systems for Adaptive Web Page Recommendation. PhD thesis, Department of Computer Science, Stanford University

[3] Boytchev, P., Grigorov, A., Earp, J., Stefanov, K., Georgiev, A. (2010) Adaptability Approaches in Digital Libraries, Second International Conference S3T, September 11-12, 2010, Varna, Bulgaria , pp. 6-13, ISBN 978-954-9526-71-4.

[4] Boytchev, P., Grigorov, A., Sarti, L., Georgiev, A., Stefanov K. and Chechev M. (2010), Recommender systems and repository search: the Share.TEC proposal, Sofia University e-Learning Journal, 2010/3, ISSN 1314-0086.

[5] Bozhilov, D., Stefanov, K., Stoyanov, S. (2009) The Effect of Adaptive Learning Style Scenarios on Learning Achievements, in IJCEELL V19 N4/5/6 2009, Special issue"Stimulating Personal Development and Knowledge Sharing", eds. R. Koper, K. Stefanov and D. Dicheva, pp.381-395.

[6] Brajnik, G., Guida, G., Tasso, C.: User modeling in intelligent information retrieval. Inf. Process.Manag. 23, 305–320 (1987)

[7] Brusilovsky, P., Tasso, C.: Preface to special issue on user modeling for Web information retrieval. User Model. User-Adapt. Interact. 14, 147–157 (2004)

[8] Brusilovsky P. (2001). User Modeling and User-Adapted Interaction, 11: 87-110

[9] Callahan, E. (2005). Cultural similarities and differences in the design of university websites. Journal of Computer-Mediated Communication, 11(1)

[10]Chirita, P.-A., Firan, C.S., Nejdl, W.: Personalized query expansion for the Web. In: 30th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007), pp. 7–14. ACM, Amsterdam (2007)

[11]Dolog, P., Nejdl. W,. Semantic Web Technologies for the Adaptive Web. The Adaptive web. Lecture Notes in Computer Science, 2007, Volume 4321/2007, 697-719, DOI: 10.1007/978-3-540-72079-9_23. p.697-719

[12]Gasparini, I., Weitzel, L., Pimenta, M.S. & Oliveira, J.P.M.d. (2011). Adaptive e-learning for all: integrating cultural-awareness as context in user modeling. In T. Bastiaens & M. Ebner (Eds.), Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2011, pp. 1321-1326.

[13]Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A. : User profiles for personalized information access. In: Brusilovsky, P., Kobsa, A., Nejdl,W. (eds.) The AdaptiveWeb, 1 edn, pp. 54–89. Springer, Berlin (2007)

[14]Ghorab, M., Zhou, D., O'Connor , A., Wade, V., Received:Personalised Information Retrieval: survey and classification < http://link.springer.com/article/10.1007%2Fs11257-012-9124-1>

[15]Jiang, X., Tan, A., (2009). Learning and inferencing in user ontology for personalizedSemantic Web search. Information Sciences 179 (2009) 2794–2808. p. 2794 – 2808

[16]Koutrika, G., Ioannidis, Y.: Rule-based query personalization in digital libraries. Int. J. Digit. Libr. 4, 60–63 (2004)

[17]Liu, F., Yu, C., Meng, W.: Personalized Web search for improving retrieval effectiveness. IEEE Trans. Knowl. Data Eng. 16, 28–40 (2004)

[18]Micarelli, A., Gasparetti, F., Sciarrone, F., Gauch, S., : Personalized Search on theWorld Wide Web < http://citeseer.uark.edu/projects/citeseerX/papers/personalized%20search.pdf>

[19]Micarelli, A., Sciarrone, F.: Anatomy and empirical evaluation of an adaptive Web-based information filtering system. User Model. User-Adapt. Interact. 14, 159–200 (2004)

[20]Mobasher B., DaiH., Luo T.,Nakagawa M., and Wiltshire J. (2002). Discovery of aggregate usage profiles for Web personalization. Data Mining and Knowledge Discovery, Vol. 6 (1), pp. 61–82.

[21]Oard, D.W.: Multilingual information access. In: Encyclopedia of Library and Information Sciences, 3rd edn, Taylor & Francis, Oxford, UK, pp. 3682–3687 (2010)

[22]Oard, D.W., Diekema, A.R.: Cross-language information retrieval. In: Williams M. (ed.) Annual Review of Information Science (ARIST), pp. 472–483. Information Today Inc., Medford (1998)

[23]Pazzani J. M. (2005). A framework for collaborative, content-based and demographic filtering. Artificial Intelligence Review, December 1999, vol. 13, no. 5-6, pp. 393-408(16).

[24]Reinecke, K.; Schenkel, S.; Bernstein, A. (2010) Modeling a User's Culture. In: The Handbook of Research in Culturally-Aware Information Technology: Perspectives and Models, IGI Global.

[25]Shahabi C. and Chen Y. (2003). Web Information Personalization: Challenges and Approaches, In the 3nd International Workshop on Databases in Networked Information Systems (DNIS 2003), Aizu-Wakamatsu, Japan.

[26]Shen, X., Tan, B., Zhai, C.: Implicit user modeling for personalized search. In: 14th ACM InternationalConference on Information and Knowledge Management (CIKM 2005), pp. 824–831. ACM, Bremen (2005)

[27]Sieg, A., Mobasher, B., Burke, R., Learning Ontology-Based User Profiles: A Semantic Approach to Personalized Web Search < http://www.comp.hkbu.edu.hk/~iib/2007/Nov/iib_vol8no1_article1.pdf >

[28]Sowa, J., (1992) : Buiding a semantic network, < http://www.jfsowa.com/pubs/semnet.htm>

[29]Speretta, M., Gauch, S.: Personalized search based on user search histories. In: IEEE/WIC/ACM International Conference onWeb Intelligence (WI 2005), pp. 622–628. Compiegne University of Technology, Compiegne (2005)

[30]Stamou, S., Ntoulas, A.: Search personalization through query and page topical analysis. User Model. User-Adapt. Interact. 19, 5–33 (2009)

[31]Stefani, A., Strapparava, C.: Personalizing access to Web sites: the SiteIF project. In: 2nd Workshop on Adaptive Hypertext and Hypermedia, Pittsburgh, Pennsylvania, USA (1998)

[32]Stefanov, K., Nikolov, R., Boytchev, P., Stefanova, E., Georgiev, A., Koychev, I., Nikolova, N., Grigorov, A. (2011) Emerging Models and e-Infrastructures for Teacher Education, 2011 International Conference on Information Technology Based Higher Education and Training ITHET 2011, paper 33, IEEE Catalog Number: CFP11578-CDR, ISBN: 978-1-4577-1671-3.

[33]Stefanov, K., Boytchev, P., Grigorov, A., Georgiev, A., Petrov, M., Gachev, G., Peltekov, M. (2009), Share.TEC System Architecture, In Proceedings of First International Conference on Software, Services & Semantic Technologies (S3T, 29-29 October 2009), Eds.: D. Dicheva, R. Nikolov and E. Stefanova, Sofia, Bulgaria, 2009, pp. 92-99, ISBN 978-954-9526-62-2

[34] Sullivan D., "Previous Query" Refinement Coming To Hit Google Results <http://searchengineland.com/previous-query-refinement-coming-to-hit-google-results-13743>

[35]Tsianos N., Germanakos P., Lekkas Z., Mourlas C and Samaras G (2009).An Assessment of Human Factors in Adaptive Hypermedia Environments Intelligent User Interfaces, 2009,pp 1-35

[36]Zhou, D., Lawless, S., Wade, V.: Improving search via personalized query expansion using social media.Inf. Retr. 1–25 (2012). doi:10.1007/s10791-012-9191-2

Albena Turnina is a PhD student at Department of Information technology at Sofia University, Bulgaria. She has received many grants amongst which are a research grant from Uninova, a grant for participation in Research Programme at London University as well as grants for participation in PhD courses, seminars and conferences. She holds computer certificates CCNA, Network Security, IBM Certified Database Associate. Her professional experience includes lecturing in the field of computer networks and development of web applications. Her research interests encompass semantic web, search personalization and adaptive search.

The design of the data preprocessing using AHP in automatic meter reading system

Mi-Ra Kim[1], Dong-Sub Cho[2]

**[1] Dept. of Computer Science & Engineering, Ewha Womans University
Seoul, Republic of Korea**

**[2] Dept. of Computer Science & Engineering, Ewha Womans University
Seoul, Republic of Korea**

Abstract

The smart grid is an electrical grid that uses information and communications technology to gather and act on information, such as information about the behaviors of suppliers and consumers, in an automated fashion to improve the efficiency, reliability, economics, and sustainability of the production and distribution of electricity. In the smart grid environment, it is a very important factor that would telemeter the factory, home and building to measure the amount of electricity telemetering would measure the amount of electricity using network and IT technology, and transmit to the server. Using the telemetering, it would measure the real time electrical load, and control the electrical demand. There is a big difference between the data in the automatic meter reading system. It is coming to be important that the efficient data treatability between central control server and the remote automatic meter unit. The delay will be able to occur when controlling a numerous termination provision from server. We design and implement a total method that controls the mass data using the support vector machine(SVM) automatic meter reading system. SVM performs classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories. When the server handle the data from automatic meter reading system, we use the automatic meter reading system's priority using SVM algorithms. In this paper, we propose a method in the remote meter reading data, when processed using SVM and Packet Priority Algorithm using AHP data preprocessing. In Remote meter reading data, the data preprocessing is used to obtain the weights using the AHP. It would be increase the server's efficiency and accuracy.

Keywords: *Automatic meter reading system, Analytic hierarchy process, Support vector machine*

1. Introduction

In the electric power industry, the smart grid would be developed the construction of the next generation intelligent grid and related technologies. The smart grid means the power providers and consumers to optimize energy efficiency and real-time information exchange in both directions, by integrating information and communication technologies into the power grid, smart grid of existing infrastructure. In the fields of electric power anytime, anywhere communications equipment can control ubiquitous environment using the latest IT technology can be implemented. The power generation, transmission, distribution, remote meter reading, etc. scattered terminal equipment attached to the control facility for sending and receiving data through a variety of communications environments. These smart grid environments from the central control of the server between the terminal and efficient data processing methods are becoming increasingly important.

In the smart grid environment, it would cause a delay on the server when processing the number of terminal equipment. There is a big difference between smart grid data and web environment and internal business data. Place in accordance with the importance of the data is attached to the terminal equipment to collect information from the remote terminal equipment, and operating and unattended, there is a difference. Communication disorders as a result of if it is not able to retrieve the data terminal equipment, and the failure of the equipment takes time to recover. Existing Web data and otherwise can get important information to be sent because of research on how to handle the data in a smart grid environment, there is a growing need. In this paper, we propose a method in the remote meter reading data, when processed of data preprocessing using AHP in Packet Priority Algorithm. In Remote meter reading data, the data preprocessing is used to obtain the weights using the AHP. In Chapter 2, we would learn about the related works in AHP and SVM, and in chapter 3, we would design about the data preprocessing using AHP. In chapter 4, we would implement and to evaluate the data preprocessing using AHP in automatic meter reading system. In Chapter 5, we would learn about the conclusions and future research.

2. Related works

2.1 Analytic Hierarchy Process

AHP is an intuitive method for formulating and analyzing decisions. AHP has been applied to numerous practical problems in the last few decades(Shim, 1989). Because of its intuitive appeal and flexibility, many corporations and governments routinely use AHP for making major policy decisions(Elkarmi and Mustafa, 1993). Application of AHP to a decision problem involves four steps[1].

Step 1: structuring of the decision problem into a hierarchical model
It includes decomposition of the decision problem into elements according to their common characteristics and the formation of a hierarchical model having different levels. Each level in the hierarchy corresponds to the common characteristic of the elements in that level. The topmost level is the focus of the problem. The intermediate levels correspond to criteria and sub-criteria, while the lowest level contains the 'decision alternatives'.

Step 2: making pair-wise comparisons and obtaining the judgmental matrix
In this step, the elements of a particular level are compared pairwise, with respect to a specific element in the immediate upper level. A judgmental matrix is formed and used for computing the priorities of the corresponding elements. First, criteria are compared pair-wise with respect to the goal. A judgmental matrix, denoted as A, will be formed using the comparisons. Each entry is formed comparing the row element Ai with the column element Aj:

$$A=(a_{ij})(i,j=1,2,\ldots,\text{the number of criteria})$$

For each pairing within each criterion the better option is awarded a score, again, on a scale between 1(equally good) and 9 (absolutely better), whilst the other option in the pairing is assigned a rating equal to the reciprocal of this value. Each score records how well option “x” meets criterion “Y”. Afterwards, the ratings are normalized and averaged. Comparisons of elements in pairs require that they are homogeneous or close with respect to the common attribute; otherwise significant errors may be introduced into the process of measurement (Saaty, 1990).

Step 3: local priorities and consistency of comparisons
Once the judgemental matrix of comparisons of criteria with respect to the goal is available, the local priorities of criteria is obtained and the consistency of the judgements is determined. It has been generally agreed(Saaty, 1980, 2000) that priorities of criteria can be estimated by finding the principal eigenvector w of the matrix A.

Step 4: aggregation of local priorities
Once the local priorities of elements of different levels are available as outlined in the previous step, they are aggregated to obtain final priorities of the alternatives. For aggregation, the following principle of hierarchic composition(Saaty, 2000) is used:

$$\text{Final priority of House } H_1=\sum_{\mathrm{i}}\left(\begin{array}{c}\text{Local priority of } H_1 \text{ with respect}\\ \text{to } C_{\mathrm{i}}\times\text{ Local priority of}\\ C_{\mathrm{i}} \text{ with respect to the goal}\end{array}\right)$$

2.2 Support Vector Machines

First, Vapnik invented support vector machines[2]. In its simplest, linear form, an SVM is a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin. Margin maximization can be expressed as given in [3] as

$$\begin{array}{l}\min_{w,\xi}\frac{1}{2}\|\mathbf{w}\|^2+\frac{C}{n}\sum_i^n\xi_i,\\ \quad\text{s.t. } \forall i,\ \xi_i\geq 0,\quad \forall i,\ y_i(\mathbf{w}\cdot\mathbf{x}_i+b)\geq 1-\xi_i.\end{array}\tag{1}$$

Using a Lagrangian multiplier, this optimization problem can be converted into a dual form which is a QP problem, where the objective function L1 is solely dependent on a set of Lagrangian multipliers α :

$$\max_\alpha L_1(\alpha)=\sum_i^n\alpha_i-\frac{1}{2}\sum_i^n\sum_j^n\alpha_i\alpha_j\mathbf{x}_i\mathbf{x}_j,\tag{2}$$

subject to the inequality constraints,

$$\forall i,\ 0\leq\alpha_i\leq\frac{C}{n},\tag{3}$$

and one linear equality constraint,

$$\sum_i^n y_i\alpha_i=0.\tag{4}$$

There is a one-to-one relationship between each Lagrangian multiplier and each training example. Once the Lagrangian multipliers are determined, the normal vector *w* and the threshold *b* can be derived from the Lagrangian multipliers:

$$\mathbf{w} = \sum_{i}^{n} y_i \alpha_i \mathbf{x}_i, \; b = -\mathbf{w} \cdot \mathbf{x}_k + y_k \quad \text{for some } \alpha_k > 0. \qquad (5)$$

3. The design of the preprocessing using AHP

3.1 The Packet Priority Algorithm Framework

We would propose packet analysis processing technique of using SVM to improve the efficiency of server connection for the configuration in Fig.1[4]. In this process, data preprocessing was applied using AHP to obtain the weights of each of the data.

The remote equipment would connect to the server through the network. If the remote equipment would connect to the server, the remote equipment's details and information is stored in the database. Data is stored in the database through data preprocessing to build SVM model. We would determine the remote equipment's priority using SVM algorithm. The remote equipment are divided into groups based on priority of each. After we would determine remote equipment's priority set forth by the SVM algorithm, then we would analyze the packet information when remote equipment would specifies the access priority. When remote equipment would connect to the server, Packet Priority algorithm by extracting the information of the IP packet is used to analyze the remote equipment's IP. And appointed remote equipment are classified by priority using SVM. We would set the remote equipment access ranking as a result of this classification.

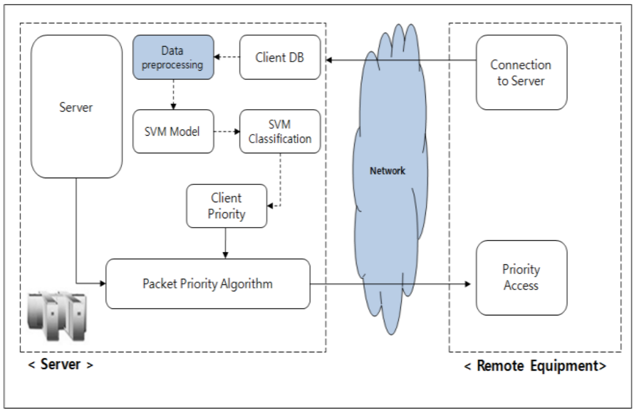


Fig. 1 The Packet Priority Algorithm Framework

There would be divided into two phase that the remote equipment rank analysis and packet processing. The first step is the priority setting phase that would determine using SVM algorithm. It would be grouping the remote equipment data by SVM classification through the resulting model using SVM. Through this process, the priority is determined. The second step is the remote equipment access management through the analysis of IP packet. By analyzing packet when the remote equipment connect to the server, it would be priority classification using the Packet Priority algorithm.

3.2 The data preprocessing using AHP

We would propose the data preprocessing techniques using the AHP in automatic meter reading system. In the database of automatic meter reading system, it would be important field that is usages, area, payment, station field. So we would decide whether to grant the weights in these four field.

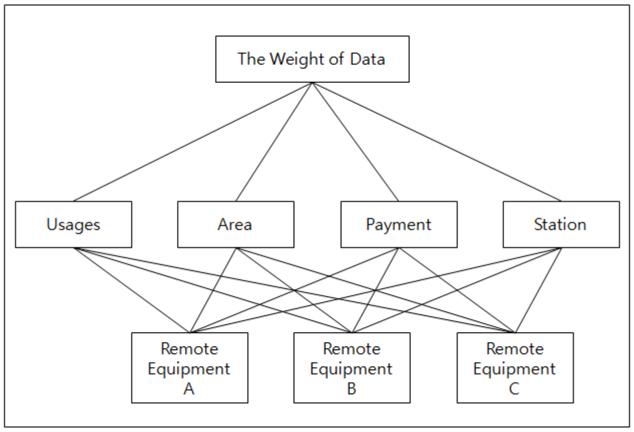


Fig. 2 The weight of data in AHP

We would calculate the comparison matrix about usages, area, payment, station field through the relative comparison. Granted after the comparison value for each field, the relative proportion of the weights were calculated. Table 1 would be the calculation of comparison matrix.

Table 1: The calculation of comparison matrix in AHP

	Usages	Area	Payment	Station
Usages	1.000	3.000	5.000	4.000
Area	0.330	1.000	0.250	0.250
Payment	0.200	5.000	1.000	5.000
Station	0.250	5.000	0.250	1.000
Sum	1.780	14.000	6.500	10.250

The relativity weight of the value of the amount to 0.484, area 0.080, using the amount of the 0.278, Station weighted value of 0.158 was determined. Table 2 is the relativity weight.

Table 2: The relativity weight in AHP

	Usages	Area	Pay	Station	Weight
Usages	0.562	0.214	0.769	0.390	0.484
Area	0.185	0.071	0.038	0.024	0.080
Payment	0.112	0.357	0.154	0.488	0.278
Station	0.140	0.357	0.038	0.098	0.158
Sum	1.000	1.000	1.000	1.000	1.000

4. The implemented and experimentation

In order to apply SVM, we would vectorize the remote meter reading data. In the process of vectorization, we would preprocess granted by the weighting of the data values using the AHP.

Table 3: The database field in automatic meter reading system

Num	Usage	Area	Pay	Stat	Comm	Con	Comm
1	500	Busan	54,000	Factory	On	normal	failure
2	40	Gunsan	3,000	Villa	On	normal	success
3	70	Sokcho	5,000	Villa	On	normal	success
4	200	Jeju	12,000	Apart	On	normal	failure
5	650	Seoul	62,000	Factory	On	normal	success
6	700	Sugi	87,000	Factory	Off	normal	success
7	230	Suwon	17,000	Apart	Off	normal	success
8	90	Daegu	6,000	Villa	On	abnormal	failure
9	170	Inchon	10,000	Apart	On	normal	success
10	540	Muju	60,000	Factory	On	abnormal	success

Table 4: The result of the SVM in the data weight using AHP

Num	1	2	3	4	5	6	7	SVM	Priority
1	261.36	0.72	15,012	1.58	1	1	0	0.99	4
2	19.36	0.24	834	0.158	1	1	1	-1.17	10
3	33.88	0.08	1,390	0.158	1	1	1	-1.08	9
4	96.8	0.08	3,336	0.79	1	1	1	-0.78	6
5	314.6	0.8	17,236	1.58	1	1	1	1.33	2
6	338.8	0.56	24,186	1.58	0	1	1	2.40	1
7	111.32	0.72	4,726	0.79	0	1	1	-0.57	5
8	43.56	0.4	1,668	0.158	1	0	0	-1.04	8
9	82.28	0.64	2,780	0.79	1	1	1	-0.87	7
10	261.36	0.24	16,680	1.58	1	0	1	1.25	3

The relative proportion of the value of the amount to 0.484, area 0.080, using the amount of the 0.278, Station weighted value of 0.158 was determined. Table 3 is the value of a database field in a remote meter reading. Table 4 is the result of the data weight using AHP. Remote metering database in table 3, rough data weighted data preprocessing using AHP values is shown in Table 4. Terminal telemetering so obtained by using the values of Table 4 and apply SVM algorithm to build the SVM model prioritizes. Priority so obtained by the server in a specified order of processing of remote meter reading terminal data processing. Data preprocessing undergone by applying the AHP method, and vectorized data when the value of the experimental results.

5. Conclusions

The automatic meter reading system is growing in importance due to the development of power projects in the smart grid. Through when a surge in demand for remote control through remote meter data management and demand forecasting, demand adjustment is also possible. The automatic meter reading system would command, at the same time, ten thousand down on the automatic meter reading system, data processing delay occurs on the server, or the server could not handle it, because it may fail. To prevent it, the remote meter reading data using the SVM algorithm and Packet Priority Algorithm to prioritize processing framework. In this paper, in order to efficiently handle the remote meter reading data through the framework proposed. Through the comparison, each of the fields of the remote meter reading data and the weights of the field was determined. The weights of the field to affect the priority of remote meter reading equipment.

We would learn about future research by applying different preprocessing techniques that affect the SVM method using AHP and would want to study to compare the SVM algorithm, pre-processing, multi-dimensional research.

References

[1] R. Ramanathan, "A note on the use of the analytic hierarchy process for environmental impact assessment", Journal of Environmental Management, 63, 2001, pp. 27–35.

[2] Changki Lee, Myung-Gil Jang, "Fast Training of Structured SVM using Fixed-Threshold sequential minimal optimization", ETRI Journal, Volume 31, Number 2, April 2009.

[3] V. Vapnik, "Statistical Learning Theory", Wiley, New York, 1998.

[4] Mi-Ra Kim, Dong-sub Cho, "The Streaming Server's Data Processing Technique Using Packet Priority Algorithm", IRACST - International Journal of Computer Science and Information Technology & Security, Vol. 2, No.6, 2012.
[5] Naotoshi Seo, sonots "A Comparison of Multi-class Support Vector Machine Methods for Face Recognition", December 6, 2007.
[6] V. Vapnik, "Statistical learning theory", Wiley, New York, 1998.
[7] S. Knerr, L. Personnaz, and G. Dreyfus, "Nurocosingle-layer learning revisited: A stepwise proce-dure for building and training a neural network", Springer, 1990.
[8] J. Shawe-Taylor J. Platt, N. Cristianini, "Large margin dags for multiclass classification, in Advances in Neural Information Processing Systems", 2000.
[9] J. H. Friedman, "Another approach to polychotomous classification, Technical report", Stanford, Department of Statistics, 1996.
[10] Ji-hye Ok, Dong-sub Cho, "Load Balancing in Distributed System for Packet Mining", ICEE, 2002.
[11] E. Osuna, R. Freund, and F. Girosi, "Training Support Vector Machines: An Application to Face Detection", Proc. CVPR, pp. 130-136, 1997.
[12] J. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines", Microsoft Research Technical Report MSR-TR-98-14, 1998.
[13] I. Tsochantaridis et al., "Support Vector Machine Learning for Interdependent and Structured Output Spaces", Proc. ICML, p. 104, 2004.
[14] B. Scholkopf and A. Smola, "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond", MIT Press, 2001.
[15] Cristianini, N. and Shawe-Taylor, J. "An introduction to support vector machines and other kernel based learning methods", Cambridge University Press, 2000.
[16] A. Bestavros et al. "Distributed Packet Rewriting and its Application to Scalable Web Server Architectures", Proc.6th IEEE Int'l Conf. Network Protocols, IEEE Computer Soc. Press, Los Alamitos, 1988.
[17] Busch, Costas, "A Study on distributed structures", Brown University, 2000.
[18] [Online]. Available: http://svmlight.joachims.org

First Author Mi-Ra Kim received her MS degree in Computer Science and Engineering at Ewha Womans University in 2003. She is currently a PhD candidate in Dept. of Computer Science and Engineering at Ewha Womans University. Her research interests include Automatic Inforamtion Classification, Personalization and Recommendation, Web Mining, Web Programming, Smart-grid.

Second Author Dong-sub Cho received his BS degree in Electrical Engineering, Seoul National University in 1979. He received his MS degree in Electrical Engineering at Seoul National University in 1981. He received his PhD in Computer Engineering at Seoul National University in 1986. He is currently a faculty member of the Department of Computer Science and Engineering, Ewha Womans University. His research interests include Ubiquitous Device and Pervasive Computing, e-Commerce and M-Commerce Design, High Performance Web Server Architecture, and High Available Web Mail Server.

Machine Learning Performance on Face Expression Recognition using Filtered Backprojection in DCT-PCA Domain.

Ongalo Pheobe[1], Huang DongJun[2] and Richard Rimiru[3]

**[1] School of Information Science and Engineering, Central South University
Changsha, Hunan, 410083, PR China**

**[2] School of Information Science and Engineering, Central South University
Changsha, Hunan, 410083, PR China**

**[3] School of Information Science and Engineering, Central South University
Changsha, Hunan, 410083, PR China**

Abstract

An accurate and robust transformed face descriptor that exploits the capabilities of filtered backprojection applied on Discrete Cosine Transform (DCT) and kernel Principal Component Analysis (PCA) methods is proposed. The method is invariant to rotation, variations in facial expression and illumination. Filtered backprojection constructs transform parameters from a set of projections through an image enhancing feature patterns that provide an initialization for subsequent DCT computations. DCT discards high-frequency coefficients that form least significant data to retain a subset of lower frequency coefficients visually significant in the image. The resulting coefficient features are mapped to lower dimensional space using PCA which extracts principal components that form the basis for the neural network classifier. Experiments were carried on JAFEE database and computed results compared with PCA and DCT approach. The results demonstrate significant improvements in results compared to other approaches.

Keywords: *Filtered backprojection, DCT, PCA, Neural Network.*

1. Introduction

Today's world where social media has taken centre stage, face to face communication is shifting towards internet, e-mail, text messaging and telephones. On the other hand faces naturally capture things that are difficult to embrace with spoken words and one such way is by use of face expressions. Face expressions are nonverbal interactive signals through which social information flows to the recipient displaying a person's affective states and intents. Interpreting these affective states gives an insight to a variety of information that can be derived from the face. Face expressions communicate emotions faster and more effective than words. As a result there is need to develop human centred user interfaces that respond readily to naturally occurring, multimodal, human communication with the capacity to receive, initiate course of action and monitor user feedback. Facial recognition accuracy depends heavily on how well the input images are compensated for pose, illumination and facial expression given that variations within a face are much more compared to variations between faces [1].

The existing facial expression recognition literature is broadly divided into three. The holistic approach [3-4], feature-based approach [5-6],[28] and hybrid approach [7]. Holistic approach also referred to as appearance-based approach identifies faces using global representations examples include Principal Component Analysis (PCA), also called Eigenfaces, Linear Discriminant Analysis (LDA), Support Vector Machines (SVM) and Neural Networks (NN). Feature-based approach extracts distinctive facial features such as eyes, mouth, nose as well as other control points from the face. These are processed to reduce input image to a vector of geometric features from which statistical pattern recognition techniques are employed to match faces. The disadvantage of feature approach is difficulty in automatic feature detection and the fact that its left upon the implementer to make arbitrary decisions to which features are relevant. In case the chosen features lack discrimination ability, no amount of subsequent processing can compensate for that intrinsic deficiency [2]. Hybrid approach separately extracts both local and global features and combines them for recognition. As a result it is expected that systems utilizing both local and global features can display more accurate results. Other systems like [8] employ the use machine based learning techniques while [9,10] make use of temporal dynamic information encoded directly in features.

Eigenfaces define feature space that drastically reduce the dimension of the original space, and face identification is carried out in this reduced space. Sirovich and Kirby [11] were the first to utilize Karhunen-Loeve Transform (KLT) to economically represent face images. They demonstrated that any face can efficiently be represented along the eigenface coordinate space, and that a face can be approximately reconstructed using a small collection of eigenfaces. Based on Sirovich and Kirby's findings Turk and Pentland [12,13] noted that projections along eigenfaces could be used as classification features to recognize faces. In [14], authors improved PCA by adding operations to standardize faces with respect to position and size. Also in [13] the authors used PCA on particular features of a face. The features became part of the "feature space," and a distance-to-feature-space (DFFS) metric was used to locate them in an image. This localization served as a pre-processing stage for later normalization, cropping, and classification [1]. Since then, PCA has become a popular method for face recognition. Daw-Tung lin [15] proposed the use of Hierarchical Radial Basis Function Network model to classify facial expressions based on local feature extraction by PCA technique.

Discrete Cosine Transform (DCT) has excellent energy compaction property for highly correlated data this helps in reduction of feature vector dimension. Previously DCT has been used for feature extraction either in a holistic or feature based approach to provide compact subspace representation. In [16] Radon transform was exploited to enhance low frequency components. DCT was used to yield lower dimension feature space while the nearest neighbor classifier was used for classification. Ramasubramanian and Venkatesh [17] combined the use of DCT, PCA and the characteristics of the Human Visual System for face recognition. In [18] Dattatray and Raghunath exploited the use of Radon, DCT and kernel based learning for face recognition using three images per subject they obtained 99.05% on FERET database, 99.32% on ORL database and 99.6% using Yale database. In [19] PCA and LDA were used in DCT domain to derive facial features with reduced dimensionality while in [20] Radon transform based on Particle Swarm Optimization followed by PCA and LDA techniques for face recognition were used. PCA was used for dimension reduction while LDA was used to extract a set of basis vectors which maximize the ratio between class scatter and within class scatter. They achieved a recognition rate of 97.5%.

Approach and Motivation

This paper investigates an alternative holistic approach that can be used for face recognition and compares it to the popular PCA and DCT approach. Facial images are represented as a finite 2-D matrix having local variation in facial intensities resulting from different combinations of abrupt features found in the face. To estimate transform parameters that provide an initialization for subsequent DCT computations, we seek for an accurate mathematical model that can map facial image to the underlying low level dimension space having redundant data and noise removed. Radon transform was used along with the Fourier Slice Theorem (FST) referred to as *filtered backprobagation* in this study to enhance low frequency components in the image. The filtered elemental reconstructions corresponding to each frequency domain pattern are initialized for subsequent DCT computations. DCT transforms the image pattern discarding high-frequency coefficients that separate significant data from the least significant data. This retains a small subset of lower frequency coefficients that are visually significant in an image. PCA is then used to compute eigenvectors in the direction of the largest variance of the training vectors also called eigenfaces. Each eigenface is considered a feature representing a point in a high-dimensional "face". Classification is done using a neural network.

The rest of the paper is organized as follows: Section 2, gives the procedure for data acquisition and reconstruction. Mathematical relationship between Radon transform and Fourier Slice Theorem is also described. Section 3 explains Discrete Cosine Transform process, Section 4 explains PCA while Section 5 describes the applied back propagation neural network process. Section 6, describes the methodology, followed by section 7 which highlights the performance of proposed system based on experimental results. Finally section 8 gives the summary and conclusion.

2. Data Acquisition and Reconstruction

2.1 Image Acquisition and Processing

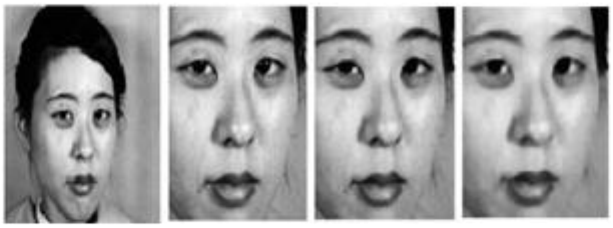
In this study static images were obtained from Japanese Female Facial Expression (JAFEE) database, which contains 213 images of 7 facial expressions posed by 10 Japanese female models [21]. To improve recognition performance against variations in face orientation, noise and illumination, various enhancement procedures were invoked to account for small perturbations in facial geometry and illumination as shown in figure1. Preprocessing measures taken include cropping images to eliminate extrinsic details like hair, neck, ears not central to face expression while eyes, mouth, and nose regions were retained. Image enhancement was done using morphological operators a process of smoothing irregularities and eliminating imperfections such as noise while distorting data of interest as little as possible [22]. Morphological operators opening and closing probe an image with a template called a structuring element. The structuring element is positioned at all possible locations in the image and is compared with other corresponding neighborhood pixels to test whether the element fits within the neighborhood. Morphological opening defined as opening of a set f by a structuring element b is the erosion of *f* by *b* followed by dilation of the result by b as given by Eq.(1).

$$f \circ b = (f \ominus b) \oplus b \tag{1}$$

Erosion filter eliminates isolated facial image details smaller than the structuring element assumed to be noise from the background, followed by dilation which thickens the object. The result of morphological opening is a smoothened object that has noise removed without affecting the shape and size of larger objects in the binary facial object. Morphological opening creates some gaps within an image this are fixed by performing a closing operation on the opening. Morphological closing defined as closing of a set *f* by a structuring element *b* is the dilation of *f* by *b* followed by erosion of the result by *b* as shown by Eq.(2).

$$f \bullet b = (f \oplus b) \ominus b \tag{2}$$

The closing operator smoothens the object border, merges together small features that are closer together and fills up the small gaps in the facial object. The output image Fig.1 (d) is a better quality image for the purpose of interpretation with features clearly visible.



(a) (b) (c) (d)

Fig. 1(a) Original test image (b) a cropped image from a; (c) morphological opening applied on image b; (d) morphological closing applied on image c;

2.2 Radon Transform

Radon transform collects line integrals across an image at different angles capturing directional local features present in the image. The Radon Transform of a 2 D function f(x,y) in (t,θ) plane can be defined as a series of line integrals through f(x,y) at different offsets from the origin [24]. Applying Radon transform on an image f(x,y) for a given set of angles relates to computing the projection of the image along these angles resulting into a profile of intensities. Using the geometry illustrated in Fig.2, the object is represented as a 2-D function f(x, y) and each line integral by (t,θ) parameters.

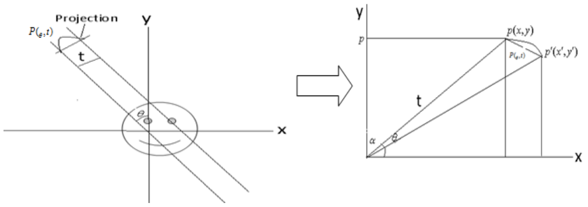


Fig. 2 Illustration of a set of projections through an object at a viewing angle forming the function P.

Radon transform of a 2-D function f(x,y) in (t,θ) is shown in Eq.(3).

$$p(t,\theta) = \int_{\infty}^{\infty}\int_{\infty}^{\infty} f(x,y)\delta(t - x\cos\theta + y\sin\theta)dxdy \tag{3}$$

Where t is the distance of a line from the origin, p(t,θ)is the sinogram, δ(·) is the Dirac delta, $\theta \in [0,\pi]$ is the angle of the line formed by the distance vector and $t \in [-\infty,\infty]$ is the perpendicular offset of the line from the origin. The δ function converts the two dimensional integral to a line integral dl along the line $x\cos\theta + y\sin\theta = t$. The transformed function (t,θ) is the sinogram of f(x,y). A Radon transformed 2D image projected at 125 degrees was computed in Matlab, giving directional lines present in the image see the sinogram Fig. 3. Each projection in the image forms a feature vector from which Radon transformed image intensities were extracted, backprojected and summed up to generate vectors used to approximate the shape of the original object see backprojected image in Fig. 3. To backproject a profile of intensities collected at an angle θ, we replicate the value p(t,θ) at all points along the direction normal to the profile for this angle as formally represented by Eq. 3.

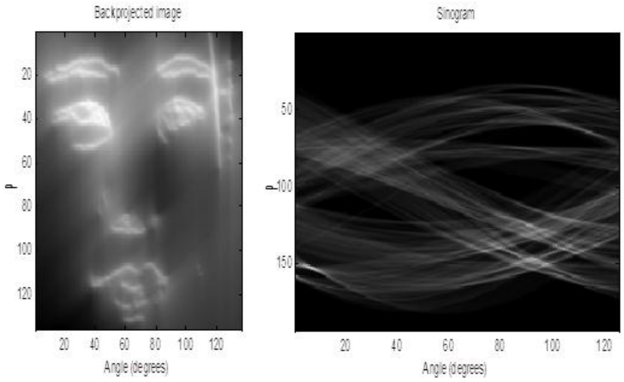


Fig. 3 Backprojected image and its Sinogram computed at 125°

From the results it is visible that backprojecting a Radon transformed 2D image as shown in Fig.3 yields unacceptably blurred image with high density at the center. This results from the overlapping of Fourier-transformed images around the low frequency region. To correct the artifact reformulation of the backprojection approach is inevitable. To do this we use Fourier Slice Theorem.

2.3 Fourier Slice Theorem (FST).

The Fourier Slice Theorem is derived by taking one-dimensional Fourier transform of a parallel projection and noting that it is equivalent to a slice of the two-dimensional Fourier transform of the original object. It follows that a Fourier transform pθ(w) of a projection pθ(t) through an image f(x,y) yields a two dimensional Fourier transform of the image f(u,v) evaluated along the polar lines (wcosθ, wsinθ). Given projection data, we can estimate the object by performing a 2D inverse Fourier transform as follows [24-25].

i. Calculate the inverse of 2D Fourier transforms in the Cartesian co-ordinate.

$$F(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(u,v)e^{-2\pi i(ux+vy)}dudv \tag{4}$$

ii. Then switch to polar coordinate, by letting u=wcosθ, v=wsinθ, which has Jacobin w. Changing differentials using dudv=w dwdθ. The inverse Fourier becomes;

$$f(x,y)=\int_0^{2\pi}\int_0^{\infty}wF(\mathrm{w}\cos\theta,\mathrm{w}\sin\theta)e^{-2\pi iw(x\cos\theta+y\sin\theta)}dwd\theta \tag{5}$$

iii. Manipulation of integration along full rays around a semi-circle limits yields

$$f(x,y)=\int_0^{\pi}\int_0^{\infty}\|w\|F(\mathrm{w}\cos\theta,\mathrm{w}\sin\theta)e^{-2\pi iw(x\cos\theta+y\sin\theta)}dwd\theta \tag{6}$$

From the projection slice theorem

$$P_\theta(w)=F(w\cos\theta,w\sin\theta) \tag{7}$$

Substitute in Eq. (6) to get,

$$f(x,y)=\int_0^{\pi}\int_0^{\infty}\|w\|P_\theta(w)e^{-2\pi iw(x\cos\theta+y\sin\theta)}dwd\theta \tag{8}$$

iv. To ensure that Eq.(8) is indeed a filtered back-projection; filter the projections pθ(t) using a ramp filter ||w|| to generate filtered projections $P_\theta^w(t)$:

$$P_\theta^w(t)=\int_{-\infty}^{\infty}\|w\|P_\theta(w)e^{-2\pi iwt}dw \tag{9}$$

Build up f(x,y) by smearing the filtered projections back across the image.

$$f(x,y)=\int_0^{\pi}P_\theta^w(\mathrm{x}\cos\theta+\mathrm{y}\sin\theta)d\theta \tag{10}$$

Evaluation of the integrand in Eq.(9) requires one dimension interpolation at $t=x\cos\theta+y\sin\theta$. The first part of the filtered back-projection algorithm implementation uses a continuous ramp filter. The second part of reconstruction is performed as per Eq. 10, summing up elemental contributions from each filtered projection to build up the grey scale values for (x,y) pixel. The filtering process is discretized to implement the algorithm. This is accomplished via the Fourier Slice Theorem. The assumption is that the projections are band-limited with band width W, that is $P_\theta(w)$ =0 whenever $\|w\|>W$, where W depends on the sampling size τ. If $\tau\le 1/(2W)$ then sampling theorem requires:

$$P_\theta(t)=\sum_{k=-\infty}^{\infty}p_\theta(k\tau)\frac{\sin 2\pi\mathrm{W}(\mathrm{t}\text{-}k\tau)}{2\pi\mathrm{W}(\mathrm{t}\text{-}k\tau)} \tag{11}$$

Using this representation for a projection the ramp filter in Eq.(9) is rewritten as:

$$P_\theta^w(t)=\sum_{k=-\infty}^{\infty}p_\theta(k\tau)\int_{-W}^{W}\|w\|\left[\int_{-\infty}^{\infty}\frac{\sin 2\pi\mathrm{W}(\mathrm{t}'\text{-}k\tau)}{2\pi\mathrm{W}(\mathrm{t}'\text{-}k\tau)}e^{2\pi iwt'}dt'\right]e^{-2\pi iwt}dw \tag{12}$$

The inner integral in this expression is the Fourier transform of *sinc* function it can be reduced to $\tau e^{2\pi iwk\tau}$. The filtered back projection is expressed as

$$P_\theta^w(t)=\tau\sum_{k=-\infty}^{\infty}p_\theta(k\tau)\int_{-W}^{W}\|w\|\cos\left(2\pi w(k\tau-t)\right)dw \tag{13}$$

Since only the even part survives integration over the interval [-W,W] continuing with discretization process, it is desirable to generate sample points of the filtered projections $P_\theta^w(t)$. Assuming that $\tau=\frac{1}{(2W)}$, we set t=jτ, in the previous equation to get;

$$p_\theta^w(j\tau)=\tau\sum_{k=-\infty}^{\infty}p_\theta(k\tau)\int_{-W}^{W}\|w\|\cos(2\pi w\tau(k-j))dw$$

$$=\tau p_\theta(j\tau)W^2+2\tau\sum_{\substack{k=-\infty\\k\neq j}}^{\infty}p_\theta\int_0^W w\cos(2\pi w\tau)(k-j))dw$$

$$=\frac{1}{\tau}\left[\frac{1}{4}p_\theta(j\tau)-\sum_{\substack{k=-\infty\\(k-j)odd}}^{\infty}\frac{p_\theta(k\tau)}{\pi^2(k-j)^2}\right] \tag{14}$$

From Eq.(14) it is evident that apart from the constant $1/\tau$ the filtered projection data $\{P_\theta^w(j)\}, j=-\infty,\ldots,\infty$ can be obtained by convolving the projection data $\{P_\theta(k)\}, k=-\infty,\ldots,\infty$ with the kernel h(j) given by

$$h(j)=\begin{cases}\frac{1}{4} & \mathrm{j=0}\\ 0 & \mathrm{j=even}\\ \frac{-1}{\mathrm{j}^2\pi^2} & \mathrm{j=odd}\end{cases} \tag{15}$$

The outcome of filtered backprojection is an image with contrasting features (high-frequencies) required for face expression recognition being emphasized as shown in Fig. 4, while blurring (low-frequencies) are minimized.

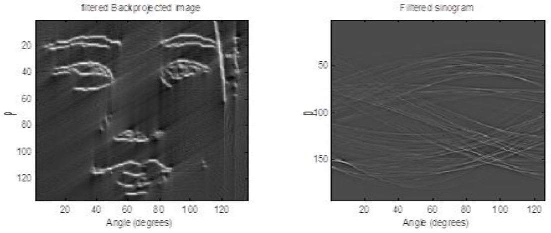


Fig. 4 Filtered backprojected image and its Sinogram computed at 125°

One advantage of filtered backprojection algorithm over frequency domain scheme is that reconstruction procedure begins as soon as the first projection is measured. This speeds up reconstruction procedure reducing the amount of data that must be stored at any given time [24].

3. Discrete Cosine Transform (DCT)

Discrete Cosine Transform is a popular linear projection technique employed in different applications for feature extraction. The superiority of DCT to PCA is that DCT can be realized in a single image or signal, while PCA depends on training samples. DCT is widely used technique in many standards of image coding and compression, like JPEG2000 and MPEG. DCT has the property that, for a typical image, most of the visually significant information about the image is concentrated in a few coefficients. Extracted DCT coefficients can be used as a type of signature that is useful for recognition tasks such as face recognition [1]. Face images have high correlation and redundant information which cause computational burden in terms of processing speed and memory utilization. DCT can be used to transform images from the spatial domain to the frequency domain. Since low frequency components are more visually significant in an image than higher frequencies DCT can been used to discard high-frequency coefficients and quantize the remaining subset of coefficients. This reduces the data volume without sacrificing too much image quality. The 2D-DCT of an M × N matrix A can be defined using Eq. (16).

$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} \cos\left(\frac{\pi(2m+1)p}{2M}\right) \cos\left(\frac{\pi(2n+1)q}{2N}\right), \qquad (16)$$

Where, $0 \leq p \leq M-1$, $0 \leq q \leq N-1$

The values B_{pq} are the DCT coefficients. DCT is an invertible transform, and its two-dimensional inverse discrete cosine transform is given by Eq. (17):

$$A_{mn} = \sum_{p=0}^{M-1} \sum_{q=0}^{N-1} \alpha_p \alpha_q B_{pq} \cos\left(\frac{\pi(2m+1)p}{2M}\right) \cos\left(\frac{\pi(2n+1)q}{2N}\right), \qquad (17)$$

Where, $0 \leq m \leq M-1$, $0 \leq n \leq N-1$.

The values $\propto_p$ and $\propto_q$ are given by Eq. (18)

$$\alpha p = \begin{cases} \sqrt{\frac{1}{M}}, p = 0 \\ \sqrt{\frac{2}{M}}, 1 \leq p \leq M-1 \end{cases}$$

$$\alpha q = \begin{cases} \sqrt{\frac{1}{N}}, q = 0 \\ \sqrt{\frac{2}{N}}, 1 \leq q \leq N-1 \end{cases} \qquad (18)$$

M and *N* are the row and column size of A, respectively. DCT tends to concentrate information, making it useful for image compression applications. Applying DCT to an input sequence decomposes it into weighted sum of basis cosine sequences. To extract DCT coefficients filtered backprojected image was taken as the input image and its DCT computed. The upper left corner of a 2D-DCT matrix contains the most important values, which correspond to low-frequency components within the processed image block. These were used to give a significant reduction in number of feature vectors needed for subsequent processing. For reconstruction DCT inverse Eq. (17) is used. In order to reduce the amount of storage and compute similarity between images for face expression recognition PCA was used. PCA computes a small set of eigenvectors with top Eigenvalues used to build principal components for the feature space.

4. Principle Component Analysis (PCA)

Principal Component Analysis is known for its dimension reduction ability. It uses the least number of dimensions but keeps most of the facial information. Due to its simplicity and robustness, PCA was chosen as the baseline algorithm for face recognition grand challenge (FRGC) evaluation [26]. In this study PCA is used to achieve dimensional reduction by extracting the most representative features of facial data. By keeping low order principal components and ignoring higher ones reduces not only the image size but also the data. To compute PCA, frequency components from DCT were mapped as feature input matrix. First the mean vector of the vector population was computed followed by an approximation of the covariance matrix from a set of feature image matrix. Next eigen vectors and eigenvalues for the covariance transformation were obtained. Eigen vectors are invariant in direction during a transformation as a result they are used to form principal components that represent the dataset. These principal components are called Eigenfaces in Turk and Pentland face detection application and Eigen vehicles in Zhang et al. vehicle detection application [27] they are stored in the database during training for reference. For derivation steps used to compute PCA in this work interested readers are referred to [28]. To extract features that are invariant to illumination and rotations, PCA was used to compute a small set of eigenvectors with top Eigenvalues used to build up image characteristic. In this study, we used Q top eigenvectors where Q represents the number of important features from the eigenspace.

5. Backpropagation Neural Network

A feed forward back propagation Neural Network was used to train the network to recognize face expressions as shown in figure 5. Neural network with biases, sigmoid level, and linear output layer are capable of approximating any function with a finite number of discontinuities. Gradient descent with adaptive learning rate backpropagation was implemented using traingda function in Matlab. The input vector is weighted with appropriate weight matrix. The sum of weighted inputs and bias form input to the transfer function j. Neurons use the differentiable transfer function j to generate the output. As learning progresses across a feed forward back propagation neural network, hidden neurons discover the salient features that characterize the training data from the nonlinear transformed input data that results to a feature space. From this new feature space the classes of interest are separated. The output layer gives rise to facial expressions. Tansig transfer function was used as an activation function for hidden neurons and purelin transfer function was used for output neurons. Each column of matrix P was independent from the other columns resulting in patterns that form feature vectors. To train a NN, the input feature pattern was applied as a stimulus to the first layer of network units, which was propagated through each hidden layer adjusting the weights and bias of the network and testing the training set until an output was generated. The number of epochs were set to 20000, Mean Square Error (MSE) performance function was computed using the Eq. (19)

$$MSE = \frac{1}{N_p} \sum_{i=1}^{N_p} (k_i^a - k_i^d)^2 \tag{19}$$

Where N_p is the number of training patterns in the training set. k_i^a is the actual output of the network for the input pattern i and k_i^d is the desired output of the network for pattern i. The actual network outputs are subtracted from the desired outputs and an error vector is produced. This error vector is the basis for the back propagation step. Errors are passed back through the network by calculating the contribution of each hidden processing layer and deriving the corresponding adjustment needed to produce the correct output. The process is repeated, layer by layer, until each node in the network receives an error signal that describes its relative contribution to the total error. Based on the error signals received, connection weights are readjusted to cause the network to converge towards a state that allows all the training patterns to be encoded.

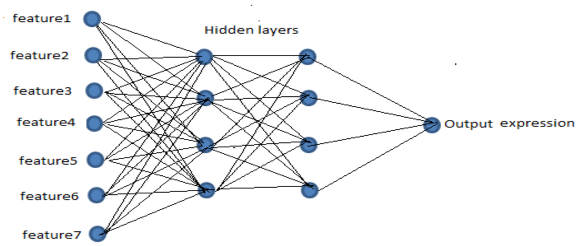


Figure 5. An illustration of neural network training architecture.

6. Methodology

To assess the validity and efficiency of the proposed approach experiments were conducted on Japanese Female Facial Expression (JAFFE) database. The database contains 213 images of 7 facial expressions posed by 10 Japanese female models. Ten individuals posed 3 to 4 expressions of each of the seven facial expressions; happiness, sadness, surprise, anger, disgust, fear and normal expression. JAFFE database presented include these images with minor rotation of camera axis and variations in head poses. Face recognition accuracy depends heavily on how well the input images have been compensated for pose, illumination and facial expression.To improve on direction low-frequency components, Radon transform was applied on preprocessed image data. Being a line intergral Radon transform acts like a low pass filter amplifying low frequency components in the image. Fourier Slice Theorem was then used before image intensities were extracted back-projected and the vectors summed up to generate the approximate shape of the facial object. This output was used to initialize DCT computations. DCT was used to discard high-frequency coefficients and quantize the remaining lower frequency coefficients reducing the data volume without sacrificing too much image quality. To recognize an input face, PCA is used on the subset of DCT coefficients to compute eigenvectors in the direction of the largest variance of the training vectors. During training 136 feature vectors were used while testing phase consisted of 70 feature vectors. Images that were used in the testing set were not included in the training set. A neural network was then trained using the outcome of the PCA process. The output generated was used as a feature map to provide an indication of the presence or absence of face expression feature combinations at the input. For purposes of evaluation and comparison, the following investigations were carried out.

i. Testing the number of Radon projections on face expression recognition
ii. Testing the effect of Morphological opening and closing on filtered backprojection approach for determination of emotional state from facial expression recognition.
iii. Testing the performance of PCA approach used alone for determination of emotional state from facial expression recognition.
iv. Testing the performance of DCT approach used alone for determination of emotional state from facial expression recognition.
v. Testing the effect of white noise robustness on facial data for determination of emotional state from facial expression recognition.

7. Results and Conclusion

7.1 Effect of number of Radon projections

In the proposed approach Radon transform is used to derive directional features with facial image data being projected at 125 degrees. To derive the best projection value, a series of experiments were carried out varying the projection degree between 30 and 180. Table 1 shows results obtained from reconstruction of facial image data for filtered backprojected enhanced data. The results reveal a steady increment in percentage recognition as the number of projections increase until 125 thereafter recognition rate starts to decline with any additional projections. Recognition accuracy was computed using Eq. (20).

$$\text{Recognition Rate} = \frac{\text{Number of Correct Images Classified}}{\text{Total Number of Classifications}} \times 100 \qquad [20]$$

Table1: Effect of number of Radon projections on face expression recognition

No of Radon Projections	***Recognition Rate***
30	94.49
40	93.73
50	94.76
60	95.97
70	96.95
80	96.97
90	97.99
100	98.19
125	**98.99**
150	97.79
180	97.39

7.2 Effect of morphological processing

To evaluate the impact of pre-processing undertaken, we experimented with both morphologically enhanced and non-enhanced images. Non enhanced images gave their highest recognition rate at 97.79% while morphological processed image data gave 98.99% accuracy in determination of emotional state from facial expression recognition as shown in table 2.The confusion matrices presented by table 3 and table 4 present best performances of the experiments carried out in more details. The diagonal entries show the percentage of correct classification for each class and the scores off the diagonal entries show misclassification.

Table 2: Effect of Morphological processing on facial data for expression recognition

No of Radon Projections	**(%) Recognition Rate**	
	Non enhanced data	*Enhanced data*
30	93.69	94.49
40	94.51	93.73
50	95.96	94.76
60	96.36	95.97
70	96.59	96.95
80	96.97	96.97
90	97.39	97.99
100	**97.79**	98.19
125	97.77	**98.99**
150	97.58	97.79
180	*97.15*	*97.39*

Table 3: Confusion Matrix for enhanced filtered back propagation facial data in DCT-PCA domain.

Face expression	***Predicted Emotions***						
	Ang	Dis	Fea	Hap	Nor	Sad	Sur
Ang	100						
Dis		100					
Fea		1.4	97.2				1.4
Hap				98.6	1.4		
Nor					100		
Sad				2.9		97.1	
Surp							100

Recognition rate =98.99%

Table4: Confusion Matrix for non-enhanced filtered backpropagation facial data in DCT-PCA domain.

Face expression	**Predicted Emotions**						
	Ang	Dis	Fea	Hap	Nor	Sad	Sur
Ang	95.8	1.4		1.4		1.4	
Dis		98.6				1.4	
Fea		1.4	95.7				2.9
Hap				97.2	1.4	1.4	
Nor					100		
Sad			1.4	1.4		97.2	
Surp							100

Recognition rate =97.79%

7.3 Performance of PCA and DCT approach

For comparative analysis we implemented PCA and DCT algorithms. In both the algorithms morphologically enhanced facial data were used. The results obtained are illustrated using the confusion matrix given in Table 5 and

Table 6. In all these algorithms the neural network is used as the classifier.

Table 5. Confusion Matrix for morphologically processed filtered backpropagation using PCA approach.

Face expression	***Predicted Emotions***						
	Ang	Dis	Fea	Hap	Nor	Sad	Sur
Ang	100						
Dis		100					
Fea		1.4	97.2				1.4
Hap				98.6	1.4		
Nor					100		
Sad				2.9		97.1	
Surp							100

Recognition rate =98.2%

Table 6: Confusion Matrix for morphologically processed filtered backpropagation using DCT approach.

Face expression	***Predicted Emotions***						
	Ang	Dis	Fea	Hap	Nor	Sad	Sur
Ang	100						
Dis		100					
Fea		1.4	97.2				1.4
Hap			1.4	95.8	1.4	1.4	
Nor					100		
Sad			1.4	1.4		97.2	
Surp							100

Recognition rate =96.6%

7.4 Noise robustness of the proposed approach

The robustness of the proposed approach towards zero mean white noise was tested. We added zero mean white noise with a variance of 0.002 in test images. No noise was added to training images and the results obtained are shown using the confusion matrix in table7.

Table 7: Confusion Matrix showing the effect of zero mean white noise on enhanced filtered back propagation facial data in DCT-PCA domain.

Face expression	***Predicted Emotions***						
	Ang	Dis	Fea	Hap	Nor	Sad	Sur
Ang	100						
Dis		100					
Fea		1.4	95.7				2.9
Hap				97.2	1.4	1.4	
Nor					100		
Sad			1.4	1.4		97.2	
Surp			1.4				98.6

Recognition rate =98.4%

Where facial expressions are depicted by; Ang=Angry, Dis=disgusting, Fea=fear, Hap=Happy, Nor=Nomal, Sur=Surprise and sad faces respectively.

8. Summary and Conclusions

In this study enhanced facial images from JAFEE database were used to accurately determine motional state facial expressions. Morphological operators were used eliminate noise and other irregularities in facial objects. Results demonstrate that preprocessed facial data combined with filtered backprojection using DCT-PCA algorithm can be used to effectively determine emotional state from facial expressions. DCT was used to extract visually significant low frequency components in an image while PCA was used to extract discriminative principle features used by a neural network to accurately determine motional state facial expressions. The feasibility of the proposed approach has been duly tested on JAFEE database. Results show that the proposed approach significantly outperforms other standard methods like DCT and PCA.

Acknowledgments

We thank JAFFE Database for providing us face images for the experiments. This work was supported by the National Natural Science Foundation of China (60873188).

References

[1] Ziad M. Hafed and Martin D. Levine, ”Face Recognition the Discrete Cosine Transform”, International Journal of computer vision 43(3), 167-188,2001.

[2] Rabia Jafri and Hamid R. Arabnia, "A Survey of Face Recognition Techniques", Journal of Information Processing Systems, Vol.5, No.2, June 2009 41

[3] M. Bartlett, G. Littlewort-Ford, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan,”Fully automatic facial action recognition in spontaneous behavior”, In IEEE Int’l Conf. on Automatic Face and Gesture Recognition, pages 223–230, 2006.

[4] B. Jiang, M. Valstar, and M. Pantic, “Action unit detection using sparse appearance descriptors in space-time video volumes”, In Proc. IEEE Int’l Conf. Automatic Face and Gesture Recognition, 2011. In print.

[5] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F. de la Torre, and J. Cohn. “AAM derived face representations for robust facial action recognition in Automatic Face and Gesture Recognition”, 2006. FGR 2006. 7th International Conference on, pages 155 –160, 2006.

[6] M. Valstar and M. Pantic. Combined support vector machines and hidden markov models for modeling facial action temporal dynamics.In ICCV-HCI’07, pages 118–127, 2007.

[7] Y. Tian, T. Kanade, and J. Cohn, “Recognizing action units for facial expression analysis”, IEEE Trans. Pattern Analysis and Machine Intelligence, 23(2), 2001.

[8] G. Zhao and M. Pietikainen," Dynamic texture recognition using local binary pattern with an application to facial expressions", IEEE Trans Pattern Analysis and Machine Intelligence, 2(6), 2007.

[9] Y. Tong, J. Chen, and Q. Ji,"A unified probabilistic framework for spontaneous facial action modeling and understanding", Transactions on Pattern Analysis and Machine Intelligence, pages 1–16, Dec 2010.

[10] T. Simon, M. H. Nguyen, F. D. L. Torre, and J. F. Cohn. "Action unit detection with segment-based svms. Computer Vision and Pattern Recognition", IEEE Computer Society Conference on, 0:2737–2744,2010.

[11] L. Sirovich and M. Kirby, "Low-dimensional Procedure or the Characterization of Human Faces," Journal of the Optical Society of America A: Optics, Image Science, and Vision, Vol.4, pp.519-524, 1987.

[12] M. Turk and A. Pentland, "Eigenfaces For Recognition," Journal Of Cognitive Neuroscience", Vol.3, pp.71-86, 1991.

[13] A. Pentland, B. Moghaddam, and T. Starner, "Viewbased and modular eigenspaces for face recognition," in IEEE Conference on Computer Vision and Pattern Recognition, 1994, pp.84-90.

[14] Akamatsu, S., Sasaki, T., Fukamachi, H. and Suenaga, Y. 1991, "A robust face identification scheme, KL expansion of an invariant feature space", In SPIE Proc. Intell. Robots and Computer Vision X. Algorithms and Techn., 1607:71–84.

[15] Daw-Tung lin, "Facial Expression Classification using PCA and Hierarchical Radial Basis Function Network", Journal of information science and engineering 22, 1033-1046 (2006).

[16] Dattatray V. Jadhav, Raghunath S. Holambe: Radon and discrete cosine transforms based feature extraction and dimensionality reduction approach for face recognition", signal processing 88(10), 2604-2609 (2008)

[17] D. Ramasubramanian, Y.V. Venkatesh, "Encoding and recognition of faces based on the human visual model and DCT", Pattern Recogn. 34 (2000) 2447–2458.

[18] Dattatray V. Jadhav, Raghunath S.Holombe, "Rotation, illumination invariant polynomial kernel fisher discriminant analysis using Radon and discrete cosine transforms based features for face recognition", Pattern Recogn. Lett. 31 (2010) 1002–1009.

[19] W. Chen, M.J. Er, S. Wu, "PCA and LDA in DCT domain", Pattern Recogn. Lett. 26 (2005) 2474–2482.

[20] Hamid, Waleed and Dr. Majed, "Face Recognition Using Improved FFT Based Radon by PSO and PCA Techniques", International Journal of Image Processing (IJIP), Volume (6) Issue (1) 2012.

[21] http://www.kasrl.org/jaffe_download.html

[22] Ongalo P. N. Fedha , Huang Dong Jun , Richard Rimiru, "A Neural Network Based Classifier for a Segmented Facial Expression Recognition System Based on Haar Wavelet Transform", Global Journal of Computer Science and Technology Volume XII Issue VII Version I 7 april 2012.

[23] A. C. Kak and Malcolm Slaney, "Principles of Computerized Tomographic Imaging, Society of Industrial and Applied Mathematics 2001".Also available on Http://www.slaney.org/pct/.

[24] Hugh Murrell, "Computer Aided Tomography",The Mathematica Journal, Vol 6, No. 2, pp.60-65 2001.

[25] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, et al. "Overview of the face recognition grand challenge", IEEE Conf. Computer Vision and pattern recognition, San Diego,CA, 2005.

[26] DS.Z. Li, R.F. Chu, S.C. Liao, L. Zhang, "Illumination invariant face recognition using near-infrared images", IEEE Trans. Pattern Anal. Machine Intell, 29 (4) (2007) 627–639.

[27] Kyungnam Kim,"Face Recognition using Principle Component Analysis", International Conference on Computer Vision and Pattern Recognition, pp.586-591.

[28] Sushil Kumar Paul, Mohammad Shorif Uddin , and Saida Bouakaz ," Extraction of Facial Feature Points Using Cumulative Histogram", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012

First Author: Ongalo Pheobe. Received B.Com. Degree from Bharadhidasan government college for women in 1997 India, and Master of Computer Application(MCA) in 2001 from Pondicherry university India. Since 2003 she has been a lecturer in the department of Computer Science, Egerton University, Kenya. She is currently working towards her Ph.D. degree at the School of Information Science and Engineering, Central South University, Changsha, China. Her research interests include multimedia, neural networks, image processing and pattern recognition.

Second Author :Huang Dongjun, is a professor of Central South University and a doctoral tutor. He currently heads the Department of Computer Engineering, College of Information Science and Engineering. He obtained a master's degree in computer science and technology in May 1996.He completed a PHD degree from Central South University in 2004. In 2007 to 2008 he was at British University of Glasgow graphics and computer vision research group as a visiting scholar. He is committed to teaching and research development, his area of interest includes networking, distributed computing, multimedia systems and applications. Over the last 10 years he has presided over the completion of the National Natural Science Foundation of the State and the school-enterprise cooperation projects 8. He got a Provincial Science and Technology Progress Award (ranked first), made two software copyright, and has published more than 40 academic articles in the "Journal of Software" Electronic Journal and IEEE CVPR well-known publications and conferences, EI, 20. Taught Multimedia Technology and Application "courses for undergraduates and graduate students. Teaching Achievement of Central South University (ranked first), Directive graduated design Outstanding Thesis for school, second prize two, has won a first prize Teaching Quality Excellence Award.

Third Author: Richard Rimiru received his B.Sc. degree from the Department of Mathematics and Computer Science, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya, in 1999; received his M.Sc. degree from the Department of Computer Science, National University of Science and Technology, Bulawayo, Zimbabwe, in 2002. He is currently working towards his Ph.D. degree at the School of Information Science and Engineering, Central South University, Changsha, China. His main research areas are artificial intelligence especially in bio-inspired computing, pattern recognition and image processing.

Interactive Video Platform for E-learning and Remote services

Hsin-Chia Fu[1], Yeong-Yuh Xu[2], Hsiao-Tien Pao[3], Jiabin Wang[4]

[1]**College of Engineering, Huaqiao University,**
QuanZhou, Fujian, China, 362021

[2]**Department of Computer Science and Information Engineering, HungKuang University**
Taichung, Taiwan, R.O.C.

[3]**Department of Management of Science, National Chiao Tung University,**
Hsinchu, Taiwan,

[4]**College of Engineering, Huaqiao University,**
QuanZhou, Fujian, China,

Abstract

This paper proposes a PC based Interactive Video platform for e-learning and remote services. Recently, multimedia technology has been greatly progressed on content bandwidth and picture quality. Problems used to be solved by face to face meetings, now it may be solved over an Internet video meeting in no time delay. Nowadays, people who need help may use a NB or a smart phone to receive all kinds of help or solutions from all over the world. As the gradually mature of cloud computing technology, the generated large amount of audio and video contents by the aforementioned Internet video meetings, can be readily transformed and saved in a searchable video database, such that the proposed platform can provide further more complete, friendly and useful distance services.

Keywords: *Digital Contents, Cloud computing, Interactive video.*

1. Introduction

During the past decades, human excessively wastes of fossil energy, leading to energy shortages and global warming crisis. The insightful people all working actively advocate energy saving and carbon reduction to slow down the advent of the crisis. In recent years, technology new favorites, multinational business leaders, and senior government officials commonly use video conferencing as a tool to reduce travel, to save time, and to increase meeting efficiency. We believe that the next decades will be the best time to develop Internet methods and solutions to help all the people to do day-to-day activities, and to achieve comprehensive energy conservation and carbon reduction.

During the past few years, Internet multimedia technology has outstanding progressive in picture quality and bandwidth reduction. Through the Internet, people at a remote distance can immersive see and hear each other in a live situation. The need now is to integrate multimedia information and network technology, to construct video platforms, so that people who need help around the world can use a NB or a smart phone, to reach experts to get help for solutions of difficult problems.

The Internet breaks through time and space limitations, many daily activities such as meeting, teaching, and a variety of services used to be performed at a specific time and locations, can be easily performed through video connection and communication at anytime and anywhere. Coupled with the increasingly matured cloud computing technology, large amount of audio and video contents were generated by the aforementioned remote video connection, can be converted into Internet video library, so that Internet users can readily check and use these valuable contents.

During the 2008 U.S. presidential election, President Obama campaigned through video conferencing with voters close contact and played streaming video through YouTube, Justin.tv and other audio-visual platform to promote his own policies and vision. The novel way of using Internet multimedia subverts propaganda vehicles, newspapers, posters, television presentations and advertising campaign way. The network will not only be able to quickly release the information, can also provide some insufficient part of the traditional campaign propaganda. If a candidate wants to interact with voters, the past can only be paraded through the streets and running around to shake hands with voters to listen to the voice of the voters. Now use the Internet to publish information of the campaign, twenty-four hours a day at any time to listen to the response of the voters, to greatly expand the level of interaction with voters of various ages. President Obama applied these concepts to interact with voters of different ages, different social status, various living environment to closer distance, eventually get elected overwhelmingly.

We believe that the field of interactive video contact can be used in many types and forms of service applications. Hereby gives a few examples to illustrate.

(1) Agricultural and pastoral assistance: When farming or livestock facing pests and diseases problems, farmers or herders may take on-site photos by a mobile phone and then sent the photos to the interactive video platform to request for help to identify problems and to get solutions. On the other hand, when nearby farmers / pastoralists asked similar questions, it is likely to collect regional timely meteorological or pests disaster reports. District experts may use the received information to submit response strategies, and also to inform other areas for early prevention. This can quickly resolve the difficulties of the farmers and herdsmen, and may also get better understanding of environmental changes in near real-time.

(2) Industry and business services: When equipment or machine failure, in general, a user will notify the vendor to send someone from the field distance away to repair. If on-site personnel may use Internet video devices to take photos or scene video to the vender service departments or stations, then remote maintenance experts can clearly see and hear the situation and condition of the failed equipment over a video screen, and command or instruct on-site people how to fix failed equipment.

(3) Tutoring or teaching assistance: For a long time, a lot of college students in addition to take courses in school, but also work part-time tutoring. If interactive video devices are available at both student and tutor sides, tutor and students may have one-on-one and just like face to face tutoring at anytime and anyplace. Through interactive video connection, learners may quickly and almost immediately achieve tutoring assistance, and tutors may also save lot not necessary trips to reach students.

Interactive video platform, comprehensively applies streaming video and multimedia network communication technologies for distance people without long-distance travel, to query and answer daily life, study and work questions and problems in an energy-saving manners. After difficult problems were solved, the platform will use its video server cluster to edit and to save the problem-solving video contents at Internet AV database for later searching and referencing by many Internet users.

2. Previous works

In this paper, Internet video communication mechanisms are proposed to provide users at distance away, an interactive cooperation method to undertake research and/or assistance to solve problems. As mentioned above, an important function of the proposed platform is the video content transferring method and architecture between users. There are two commonly used architectures for contents transferring: Client-Server architecture, and Peer-to-Peer (P2P) architecture. For example, MSN's messenger chat room uses Client-Server architecture, and Yahoo's Messenger uses P2P way to send information.

The advantage of using client-server architecture is that the information can be centralized and managed by a server to achieve better security control; and its drawback is when the server fails, it is difficult to backtrack and to get lost data back. The advantage of Peer-to-Peer architecture is its capability of dispersing foregoing server, thus the relative reliability of the system will be higher; and its drawback is smoothing audio and video data transmission will be affected when Peer to Peer network bandwidth is lower.

The following are briefings of related literature and commercial available video communication systems: (1) Skype video call, (2) Windows Live Messenger, (3) Co-Life, (4) JoinNet, and (5) GAIA, etc.

(1) Skype Video call [1]**:**

Skype uses P2P technology to provide users real-time high quality audio and video communication. When the dialog parties are connected with smoothly network quality, the Skype sound quality is similar to ordinary telephone. Skype can provide two-party or multi-party video dialogue at the same time. Due to the P2P technology, Skype provides no real-time video recording function. Since storing video contents between the dialogue parties is the major resource that an interactive video platform (IVP) can use to provide Internet users with long-term accumulated useful information from previous communication activities. Some new version of the audio and video server software has been started to provide P2P functionality for more flexibility in audio and video communication services like Skype. In addition, users of Skype video do not allow markings on the video screen to illustrate the designated or important objects.

(2) Windows Live Messenger (MSN) [2]:

MSN is an instant multi-functional messaging (text, speech, and video dialogue) software developed by Microsoft Corporation. MSN uses Client-Server architecture, and handles most of the functions through a server, which is very convenient for task management. By going through an authentication procedure, MSN users may use message server (switchboard server) to relay/receive messages. Currently (January 2013), MSN does not provide audio and video recording/playback, focus markers and multi-language text communications.

(3) Co-Life [3]:

Co-Life is developed by the National Center for High-Performance Computing (NCHC) in Taiwan, Co-Life provides: complete network video conference, and public broadcast speech functions. Co-Life uses Client - Server architecture, to provide text, images and video communication between participant users. Co-Life provides whiteboard to help connected users to emphasize key projects, and to mark important items and/or objects. The Co-Life whiteboard can only be applied to a few specific text and/or image formats for interactive communication.

(4) JoinNet [4]:

JoinNet is a on-line meeting software, developed by HomeMeeting to provide users with video conferencing, whiteboard, synchronized web browsing and video conference recordings. JoinNet is designed according to Client-Server architecture, so the clients are not connected to each other, and the meeting server is responsible for text, images and even video exchange between participants. JoinNet uses the whiteboard as the main tool for participants to discuss the text documents, or marking important points between each other.

(5) GAIA[5]:

GAIA (Global Agriculture Information Alliance) was proposed by the Jigga-Dongxi team of National Chiao-Tung University in Taiwan to participate the 2010 Microsoft Imagine Cup competition, and won the first place on the section of **Looking to the future 2020**. GAIA pointed out that in 2020 the use of the Internet cloud computing and mobile phone mobile devices can help 800 million poor small farmers around the world at any time to obtain valuable atmospheric information from satellite, also to consult experts around the world, for information on agriculture, pest, and financial relief programs. Therefor the value of agricultural production can be greatly improved, so as to increase the personal household income, so as to reduce the disparity between the rich and the poor around the world.

We think that the goal of Global Agricultural Information Alliance may use the proposed IVP to comprehensively promote and achieve the desired objectives.

3. IVP design

Figure (3.1) shows how a video network connection can be used to allow remote experts (mentors) commanding/ teaching field personnel (learner) to operate and/or to maintain machinery or equipment. Learners may use a Webcam to take instant video of field devices, instead of just show the video at his computer screen, and also may use the Internet to pass instant video to experts, for the operation or maintenance suggestions and/or solutions for difficult problems.

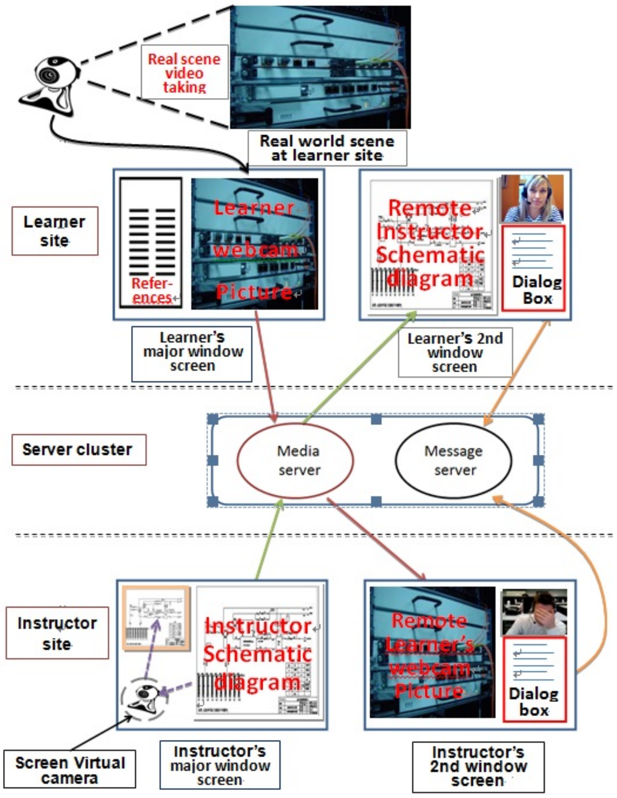

Real scene video taking
Real world scene at learner site
Learner site
Refer-ences
Learner webcam Picture
Remote Instructor Schematic diagram
Dialog Box
Learner's major window screen
Learner's 2nd window screen
Server cluster
Media server
Message server
Instructor site
Instructor Schematic diagram
Remote Learner's webcam Picture
Dialog box
Screen Virtual camera
Instructor's major window screen
Instructor's 2nd window screen

Figure 3.1: The flow diagram of an interactive video platform to be used to allow remote experts (mentors) commanding or teaching field personnel (learner) to operate and/or to maintain machinery or equipment.

In addition, if the instructor wants to show related electronic documents and/or images, or schematic diagrams, he may display these documents at his own screen first, and then use some *virtual screen camera software* [6] to convert these documents into real-time video and to pass to learners' side. Since interaction between teacher and learner has long been recognized as an important mechanism to strengthen communication effects, both instructor or learner may also want to draw simple marks or symbols over the documents displayed at video screens on both site, as shown in Figure (3.2): "⊂⊃" indicates an area or a range to be watched out), "←,↑,→, and↓” indicates a focusing point), “✕” means to be deleted, and “___” marks a text segment to be emphasized.

These simple markings can be easily used to remind or to emphases important text items or image objects, so that users at distance away may have a face-to-face like discussing with each other. Additionally, when the platform users have speaking language barrier problems, the platform is also available online translator to provide multi-language translation dialog box to solve the voice communication difficulties. Thus the platform can greatly expand its scope of services to almost everywhere around the world.

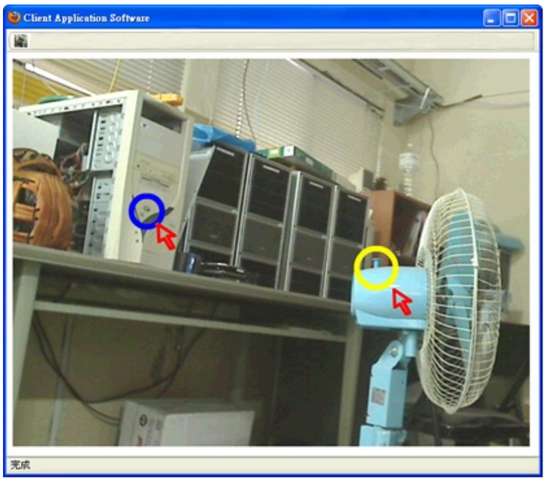


Figure 3.2: An example of using pointing marks to indicate a particular item or an image object on a video display screen. For example, users at distance away may use different color markings to tell each other where the “***fan steering gear***” is and how to turn on a “***computer power switch***” respectively.

As mentioned above, we believe that interactive video platform particularly suitable to be used in developing countries. In general, there are a lot of remote undeveloped areas in developing countries, thus interactive video platform can be immediately used to narrow the distance from the people of the metropolitan areas and the remote undeveloped areas, so that the life, education, and industrial and commercial standards of remote areas can be helped and improved rapidly.

4. Architecture of interactive video platform

As shown in Figure 4.1, the interactive video platform (IVP) uses the Client-Server architecture to achieve its structure functionality needs. A user at client-side may set up available video devices first, and may use a browser to connect the IVP for audio and video communicate with other users at the platform. The server clusters at the platform process the video contents from sending users and dispatch the video contents to designated users. The platform contains three types of different function server clusters: Web Server (WBS), Media Server (MDS), and Message Server (MGS).

When there are more and more users at an IVP, its workload will eventually transcend the service capability of a single server. Thus, the server will not be able to deal with the instant needs of users, and cause the users to wait for a long time or unable to get connected to the platform. Eventually, the IVP's service quality is greatly reduced.

Therefore, how to establish a platform for scalability to meet the ever-increasing load, has become an important issue for the architecture design of the platform.

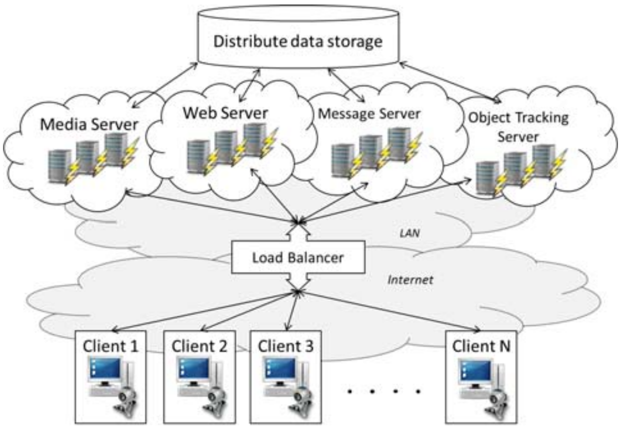


Figure 4.1: The architecture diagram of the proposed interactive video platform (IVP).

Suppose the financial budget to construct and to maintain a server cluster is limited, how to provide high quality services to a large number of users is a topic of concern for cost-effectiveness. Therefore, scalability, availability, and cost-effectiveness are three key points to the design of network based architecture for the proposed IVP. The balanced design considerations of the proposed IVP are summarized as follows.

Web Server (WBS) Design: Web server provides client side users a browser to connect other users and to use various types of server functions available at an IVP. A user, whether to be an expert or not, must be a registered user to use a browser to connect the IVP. In addition to general personal information, e.g., user name, email address, nickname, mobile number, registration at IVP also includes personal interests, skills and expertise level. According to the skills and the expertise level of the online user, WBS connects associated experts for the on-line user.

Media Server (MDS) Design: Media Server provides users of various video associated functions available at IVP, such as video publishing, sharing to editing, storage, and reuse. Adopting Client-Server architecture for the design of IVP is mainly for the convenience of server-side, to centralize editing and management of audio and video contents and database, and to reduce the workload and hardware requirements of the client-side.

Message server (MGS) Design: Among various interactive processes at the IVP, prompt and clear message communication between users is very important. If there is only video interaction available, it may cause semantic ambiguity or emphasis vague situation between users. Therefore, additional interactive text message communication is added and provided by MGS. The benefits of interactive text message communication are at least three fold: be able to achieve careful expressions of semantics, be able to repeatedly view and read messages, be able to provide multinational language translation. A MGS completes each interactive communication service by linking associated users and experts in a ***discussion chamber***, and is responsible for manipulation and management of each discussion chamber like mechanism. Each of the aforementioned streaming video transfer is also controlled by the mechanism of associated discussion chamber.

In addition to the aforementioned three types of servers, when the loads of an interactive video platform reach its upper limit, the load balance server may add extra servers to enhance the service capabilities, and also the toughness of the server clusters. Thus, the major architecture design goals: i.e., *scalability*, *availability* and *cost-effectiveness* for the proposed IVP can be satisfied.

5. Benchmark and evaluation of IVP server cluster

According to the aforementioned design descriptions, a prototype of the interactive video platform has been implemented at http://140.113.216.64. Basically, the design of IVP needs to provide a web interface, video streaming, discussion rooms, and some other associated functions. Suppose these services are provided by a single server, and when the number of users increases, the host server must gradually to withstand the decline in service quality. Below, we will present how to use cheap PC components and Open source software (e.g., dual-core CPU, motherboard, 4G RAM, SATA-300 7200rpm hard drive, Fedora OS, Apache, the WBS, Adobe FMS and MySQL database) to compose the proposed functional servers. Then, the composed servers are tested and evaluated under different using and working environments. The performance testing results will be important references for the design of the proposed IVP.

The followings are the hardware and software component specifications of the servers to be used in the proposed platform:

- **CPU**: Intel Pentium *Dual E2200* @ 2.20GHz;
- **RAM**: 4GB;
- **Hard Disk**: Seagate ST3160815AS, 160GB SATA-300 7200rpm Hard Drive;
- **OS**: Fedora 11;
- **WBS: Apache 2.0**;
- **MDS: Adobe Flash Media Server 3.5**;
- **DATABASE**: MySQL.

Benchmark of Web servers (WBS):

On the WBS testing, we chose to use Apache Bench (AB) [7] to test the system reaction time under different number of users. Since the response time of most well-known websites, is less than 0.5 second. As shown in Table 5.1, when the number of the on-line users is lesser than 200, there are 90% of users (i.e., 180 users) feel the system response time of about 0.5 seconds. And when the number of Internet users increased to 1000 the average response time of the system is increased to about 1.9 seconds. Thus, if the proposed platform wants to provide the same level of quality of service, it is suggested either to increase the number of WBS, or to enhance performance of the existing WBS.

Benchmark of Media servers (MDS):

The MDS benchmark testing and measuring includes CPU load, hard disk load and network load three parts. The system resource monitoring program -- **iostat** [8] in Linux and 700kbps video streams are used to test and to measure the performance of MDS. The Network load associated with MDS video I/O is estimated by linear interpolation.

Figure 5.1(a) shows the relationship between the CPU load vs. the number of played video clips in a MDS video server. For example, when 20 different video clips are playing in one MDS, the associated CPU idle rate is still as high as 99%. It seems that video playing consumes very small amount of CPU load. As proposed in [9], Adobe Flash Media Server may consume up to 50% of CPU load, when 1000 different video clips are playing at the same time.

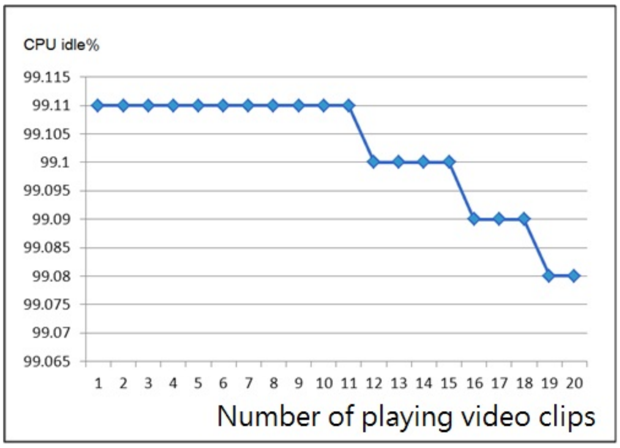
Table 5.1: The Apache benchmark response time (in ms) of the proposed WBS under different number of on-line users.

% of response users / Response time / Number of users	50%	75%	95%	100%	Average response time
1	3	3	3	371	3.450
100	159	177	271	7003	176.715
200	257	368	737	26180	348.594
300	341	444	1099	22792	526.609
400	435	637	1779	43596	709.118
500	418	562	3473	39098	893.391
600	493	794	3578	46094	1070.507
700	402	540	5677	49204	1249.806
800	389	568	5682	49922	1480.069
900	350	712	6608	52564	1633.891
1000	298	448	5543	58450	1929.697

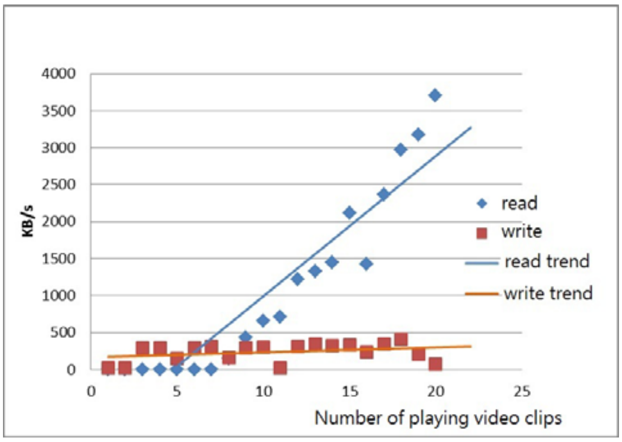
Figure 5.1(b) shows the testing of hard disk I/O workload vs. the number of played video clips. Since the major workload of video playing is disk read, thus the video disk write generates less than 500 KBPS for 20 or less video clips, and this I/O rate may last until all the RAM memory are all used up. On the other hand, the I/O rate of video disk read may be increased in proportion to the number of played video clips. According to the trend of least-squares fitting in Figure 5.1 (b), for each additional video clip playback, the HDD load will be increased by approximately 200 KBPS, thus the following descriptions will only focus on the disk read.

According to the design of Adobe Flash Media Server [10], when broadcasting video streaming, each video clip will be partitioned into 256KB chunks. Therefore, the hard disk I/O capacity is limited by its random access capability of 256KB block. By using the **iometer** [11] to benchmark a hard drive with 256KB chunks, the upper bound of average transfer rate is 35.87MBPS. According to this transfer rate, a hard disk drive could bear 175 playing video clips with 700kbps I/O bandwidth, which generates 130Mbps network load, and consumes10% of CPU load.

In terms of network load, the Adobe flash media server [10] suggests the traffic rate of a single MDS network interface should be controlled to be under 70% of the available network bandwidth, and reserves at least 700kbps bandwidth for each playing video clip to ensure smooth play back video streaming. In the followings, it is suggested that 70% of the total network interface bandwidth is used as the maximum total video streaming flow, which is also used to estimate the number of simultaneous playback of video clips.



(a)



(b)

Figure 5.1: (a) The relational diagram between the numbers of simultaneous playback video clips vs. the CPU load of a MDS; (b) The relational diagram between the read and write bandwidth vs. the number of simultaneous playback video clips of a MDS hard disk drive

Benchmark of Message servers (MGS):

First of all, the relation between multiple discussion chambers vs. associated MGS CPU load is tested. As shown in Figure 5.2, operating multiple discussion chambers occupies only a very small amount of CPU load. The 2nd test is about the relationship between the number of discussion chambers and the number of users in a chamber vs. the MGS network traffic flow rates. The test data is a string of English characters of length 50. Suppose 1000 discussion chambers are initiated at the same time, and each discussion chamber contains an average of 20 people, and one character string is sent out per second per chamber, then the needed network bandwidth is approximately 200 KB/Sec. The testing results are shown in Figure 5.3.

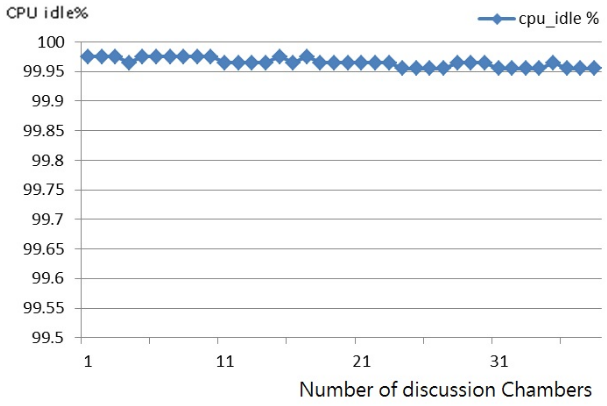


Figure 5.2: The relationship between the MGS CPU load vs. the number of discussion chambers.

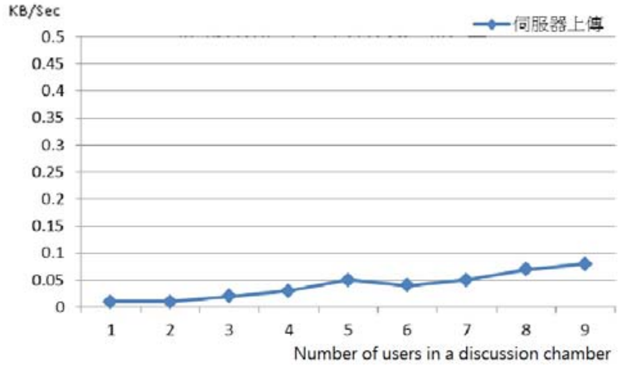


Figure 5.3: The testing results show the relationship between the MGS network traffic flow rates vs. the number of users in a discussion chamber.

6. Conclusions

This paper describes how to use the rapid development of network multimedia technologies [12, 13, 14], to construct an interactive video platform for e-learning and remote services. An Internet user may use a simple PC based video equipment (such as Webcam) to connect the proposed platform to achieve variety of assistance from experts around the world at any time. In the near future, we plan to implement a cloud computing based interactive video platform to test the functional performance, the user friendliness, and the robustness of the network architecture of the proposed video platform. Then, we may build a larger platform to perform real world test and evaluation of e-learning and remote services over Internet.

References

[1] S. A. Baset, H. Schulzrinne, “An Analysis of the Skype Peer-to-Peer Internet Telephony Protocol,” IEEE Infocom’06, Spain, Apr. 2006.

[2] MSN - A collection of Internet sites and services provided by Microsoft, http://www.msn.com

[3] Co-Life: Interactive On-Line Meeting / Speech system, http://meeting.colife.org.tw/index_en.aspx

[4] JoinNet User’s Guide Version 5.2.0, http://download.homemeeting.com/joinnet/doc/JoinNet_User_Guide.pdf.

[5] http://www.imaginecup.com/upload/students/2010/pdf/Awards/2010-Awards-1st- Envisioning.pdf.

[6] http://mosax.sakura.ne.jp/fswiki.cgi?page=SCFH+DSF

[7] Apache Bench -A command line computer program for measuring the performance of HTTP web servers bundled with Apache HTTP Server, http://httpd.apache.org

[8] iostat - A computer system monitor tool used to collect and show operating system storage input and output statistics, http://en.wikipedia.org/wiki/Iostat

[9] Large-scale streaming deployments with Flash Media Interactive Server 3.5, http://www.adobe.com/content/dam/Adobe/en/devnet/flashmediaserver/articles/fmis_largescale_deploy/fmis_largescale_deploy.pdf

[10] FMS White paper - Adobe Flash Media Server 3, http://www.adobe.com/products/flashmediaserver/pdfs/FlashMediaServer3_WhitePaper_ue.pdf.

[11] iometer - An I/O subsystem measurement and characterization tool for single and clustered systems, http://www.iometer.org.

[12] R. Ponce-Medellin, G. Gonzalez-Serna, R. Vargas and L. Ruiz, "Technology Integration around the Geographic Information: A State of the Art", IJCSI, Volume 5, pp17-26, October 2009.

[13] Yeong-Yuh Xu and Hsin-Chia Fu, "*Visual Keyword Based Image Retrieval*", IJCSI, Volume 9, Issue 3, No. 2, pp20-28, May 2012.

[14] Chadi Riman, "A Remote Robotic Laboratory Experiment *Platform* with Error Management," *IJCSI*, Volume 8, Issue 1, January 2011.

Hsin-Chia Fu received the B.S. degree from National Chiao-Tung University in Electrical and Communication engineering in 1972, and the M.S. and Ph.D. degrees from New Mexico State University, both in Electrical and Computer Engineering in 1975 and 1981, respectively. From 1981 to 1983 he was a Member of the Technical Staff at Bell Laboratories. From 1983 to 2012 he has been on the faculty of the Department of Computer science and Information engineering at National Chiao-Tung University, in

Taiwan, ROC. Currently, he is a distinguished professor at the College of Engineering, Huaqiao University, QuanZhou, Fujian, China. From 1987 to 1988, he served as the director of the department of information management at the Research Development and Evaluation Commission, of the Executive Yuan, ROC. From 1988- 1989, he was a visiting scholar of Princeton University. From 1989 to 1991, he served as the chairman of the Department of Computer Science and Information Engineering. From September to December of 1994, he was a visiting scientist at Fraunhofer-Institut for Production Systems and Design Technology (IPK), Berlin Germany. His research interests include digital signal/image processing, VLSI array processors, and neural networks. Dr. Fu was the co-recipient of the 1992 and 1993 Long-Term Best Thesis Award with Koun Tem Sun and Cheng Chin Chiang, and the recipient of the 1996 Xerox OA paper Award. He has served as a founding member, Program co-chair (1993) and General co-chair (1995) of International Symposium on Artificial Neural Networks. He is presently serving on Technical Committee on Neural Networks for Signal Processing of the IEEE Signal Processing Society. He has authored more than 100 technical papers, and two textbooks ``PC/XT BIOS Analysis'', and ``Introduction to neural networks'', by Sun-Kung Book Co., and Third Wave Publishing Co., respectively. Dr. Fu is a member of the IEEE Signal Processing and Computer Societies, Phi Tau Phi, and the Eta Kappa Nu Electrical Engineering Honor Society.

Yeong-Yuh Xu received his B.S. degree in electrical engineering in 1995 from National Sun Yat-Sen University, Kaohsiung, Taiwan. He received his M.S. and Ph.D. degree in computer science and information engineering from National Chiao-Tung University, Hsinchu, Taiwan, in 1997 and 2004, respectively. From 2005 to 2009, he served as a Postdoctoral Fellow in the Department of Computer Science and Information Engineering of National Chiao-Tung University, Hsinchu, Taiwan. Currently, he is an assistant professor in the Department of Computer Science and Information Engineering, Hungkuang University, Taichung, Taiwan. Dr. Xu's research interests include pattern recognition, neural networks, and content-based image/video retrieval.

Pao Hsiao-Tien is a full professor at the Department of Management of Science, National Chiao Tung University, Hsinchu, Taiwan.

Wang Jiabin is an associated professor at the College of Engineering, Huaqiao University, QuanZhou, Fujian, China.

Performance evaluation of apriori with memory mapped files

Anuradha.T[1], Dr.Satya Pasad.R[2]and Dr.Tirumalarao.S.N[3]

**[1]Department of ECM,KL University
Guntur,A.P.,India**

**[2] Department of CS&Engineering, Acharya Nagarjuna University
Guntur,A.P.,India**

**[3] Department of CSE,Narasaraopeta Engineering college
Guntur,A.P.,India**

Abstract

The concept of memory mapped files reduces the I/O data movement by mapping file data directly to the process address space. This is best suitable for the data mining applications which involve accessing large data files. The recent improvement in parallel processor architectures is the multi-core architectures. To get the real benefit from these architectures we have to redesign the existing serial algorithms so that they can be parallelized on multi-core architectures. OpenMP is an API for parallel programming which make a serial program to run in parallel without much redesigning job. Our main concern in this paper is to evaluate the performance of apriori using linux mmap() function compared to fread() function in both the serial and parallel environments. Experiments are conducted with both simulated and standard datasets on multi-core architectures using openMP threads. Our experiments show that mmap() function gives better results than fread() function with both serial as well as parallel implementations of apriori on dual core.

Keywords: *apriori****,*** *fread(), mmap(), multi-core, OpenMP*

1. Introduction

The concept of memory mapping of files was introduced to reduce the overhead of file management. Mmap() function is a unix/linux function which simplifies processing of file data (1).The applications which need huge input/output overhead like network or other applications are using mmap()(2).Many unix/linux functions like grep, fgrep, egrep and the unix pipe facility use memory mapping concept for large data files. Avadis Tevanian describes an approach of file mapping facility under Mach operating system and mentions that useful performance gains can be achieved by using Mach's memory mapping(1). John Heidemann explains that CPU utilization can be reduced by using memory mapped files instead of stdio when sending large files(3).Joseph Jang in his blog has clearly shown the better performance results of mmap() over fread() and iostream(4).

A multi-core processor contains two or more actual processors integrated on the same chip and the performance gains from multi-core processors can be obtained based on the software that can run on multiple cores simultaneously(5).OpenMP is an application program interface for developing shared memory parallel programming(6). It works on the concept of multithreading. Master thread works sequentially and when the parallel region encounters, master thread forks child threads and work along with them(7). In our previous papers we have evaluated the performance of the popular data mining algorithm apriori on dual core with OpenMP threads compared to the serial implementation (8,9).Our present paper mainly concentrates on comparing the benefit of mmap() over fread() in the implementation of the apriori algorithm. The results will compare the performance of mmap() over fread() in serial and parallel implementations of apriori with different datasets at different support counts.

2. Related Work

Apriori is the popular algorithm for achieving the important functionality of data mining known as frequent itemset mining(FIM) or association rule mining(ARM) (10).As data mining deals with large volumes of data, scalability can be achieved by parallelizing the algorithm. (11,12,13)M.J.Zaki has presented a survey paper on parallel and distributed association rule mining.(12) Rakesh Agrawal and John C.Shafer proposed two parallel algorithms known as count distribution and data distribution based on apriori.(11). Zaıane et al has proposed a parallel algorithm for finding frequent item-sets using FP-growth algorithm (13).Pattern mining researchers are also designing parallel algorithms on the recent multi core architectures.(14) Li Liu2, et.al., proposed a cache conscious FP array mechanism for implementing FP-growth algorithm on multi core processors.(15).S.Tatikonda et al., proposed frequent subtree mining from a tree structured data on multi-core architecture.(16) Research is also going on implementing data mining algorithms on multi-core architectures using openMP threads.

Anuradha et al., presented the performance evaluation of parallel apriori on dual core compared to serial execution with different data sets and also by changing the number of threads.(8,9).S.N Tirumalarao et al., studied the performance of k means clustering algorithm on multi-core architectures.(2) S.Mohanavalli et al., implemented parallel k-means algorithm with openMP and distributed algorithm with MPI . They have also compared the performance of these implementations with hybrid model which is the combination of openMP and MPI.(17)Memory mapped files concept was initially used in designing the unix based operating system internals. Avadis Tevanian et al., explains the system call for file mapping in Mach operating system.(1) Direct accessing of user programs to device memory and the files and its advantages in Linux memory management are explained in (18).But only a little research is done in finding the effect of memory mapped files compared to normal file reading operation on data mining algorithms. S.N.Tirumalarao et al.,studied the performance of memory mapped files on k-means clustering algorithm.(2)

3. Theoretical background

3.1 Apriori

Apriori is the popular algorithm for finding frequent itemsets from a transactional database. It was proposed by Agrawal and Srikant(10,19).It consists of two functions :

1. Finding the candidate k-itemsets: Initially, every item in the given database will be in the candidate k-itemset where k=1. For finding the next candidate k-itemsets (k=2 ,3 etc.), we have to join previous frequent k-itemset with itself. For example for finding candidate 2-itemset, we have to join frequent 1-itemset with itself.
2. Finding the frequent k-itemset: For finding frequent k-itemset, we have to find the count of each item in the candidate k-itemset. If the count of the item is more than a pre-specified threshold called minimum support count, it will be placed in the frequent k-itemset.

The major principle of apriori is that all the subsets of a frequent itemset should also be frequent.(10)

3.2 OpenMP

OpenMP is an API for shared memory parallel programming for C,C++ and Fortran. It is very easy to port it on different shared memory architectures.(20) OpenMP has different pragmas which direct the compiler to use the openMP constructs. If the compiler does not support openMP, the program will run sequentially. (6,21,22) To use the openMP constructs in a c program we have to include omp.h header file in our program. The parallelism in openMP is achieved by multiple threads. By using the fork-join model it makes the program to run in serial and parallel modes. Initially master thread runs the program in serial mode and when #pragma omp parallel construct is encountered, the master thread fork the child threads and runs the program in parallel mode along with child threads. Once the parallel part is over, again the child threads join the master thread and the program runs in the serial mode. The number of threads will be decided by the omp-set_num_threads () library function .

3.3 Mmap()

Mmap() function is useful when the process need to access the data from a large file. In fread() function, data must be first copied to the user space buffer before it is being copied to the process address space. Mmap() function avoid this extra copy operation as the file is directly mapped to the address space of a process. (23,2).In the mmap() function, we have to specify the starting address in the process address space from where the file should be mapped and how many bytes of the file we have to map starting from the offset.

4. Mapping Of Apriori On Dualcore

The parallelization of apriori is done based on the count distribution algorithm proposed by agrawal and shafer.(11) Here we follow the partitioning concept and data parallel strategy. The transactional database is partitioned into number of parts equal to the number of threads.

4.1 Algorithm for parallelizing on dual core with 2 threads:

```
Input: Transactional database,TDB  with transactions $TR_1$,$TR_2$,…,$TR_n$
where a transaction $TR_i$ is the random combination of any items from $item_1$ to $item_{10}$
and n is the number of records in the database,
Minimum support count, msc
Output:frequent k-itemsets where k=1,2,3 etc.
Step1 : k=1
candidate k-itemset ={$item_1$, $item_2$, $item_3$, $item_4$, $item_5$, $item_6$, $item_7$, $item_8$, $item_9$, $item_{10}$ }.
Step2: SET_ OMP_NUM_THREADS =2
 partition the given database into three partitions.
#pragma omp parallel
/* begin parallel region
#pragma omp sections
{
 omp section
{
   Find the local count of each item in the candidate k-
   itemset in  partition1
}
omp section
{
    Find the local count of each item in the candidate k-
    itemset in  partition2
 }
}
 /* end of parallel region  */
Global count of each item in candidate k-itemset=sum of the local counts
If the global count of any item is>msc, place the item in frequent k-itemset
Step 3:
K=k+1
Join frequent k-itemset with itself  to find the next candidate k-itemset
Step4:Repeat steps 2 and 3 until any subset of  candidate k-itemset is not frequent.
```

In the above algorithm, The #pragma omp parallel construct directs the compiler to enter into the parallel region. There are two types of work sharing constructs in openMP to divide the work parallel- Loops and Sections. We are using sections construct. Here we are creating two threads and each thread will work on each section. Because we are using multi-core processors each thread will run on separate cores there by making the execution faster compared to serial execution. For running the algorithm with three threads, we divide the database into 3 partitions and set number of threads equal to 3 and for four threads, we divide the database into 4 partitions and set number of threads equal to 4. When the number of threads are more than the number of cores , the threads will share the cores.

5. Experimental Work

The experimentation is carried out on Intel Pentium Dual-core with processor speed 1.6GHz and 3GB RAM . To get openMP compatibility, we have used Fedora 9 Linux (Kernel 2.6.25-14, Red Hat nash version 6.0.52) equipped with GNU C++(*gcc* version 4.3) for our experimentation. Different randomly generated transactional datasets with 2,4,6,8 and 10lakh records are used. Each dataset consists of any random combination of items from $item_1$ to $item_{10}$. Our algorithm is also tested with the standard accident dataset (24) from UCI repository. We have used all 3,40,183 transactions and 1 to 10 items of the accident dataset for testing purpose. The experimentation is done at different support counts. To test the effect of memory mapped files on apriori, the algorithm is run in serial mode by taking different datasets and different support counts separately with fread() and mmap() functions.The real time, user time and system time results of fread() versus mmap() are also compared by running the program parallelly on dual core processor using OpenMP threads by setting number of threads=2,3 and 4. The speed up of parallel apriori is compared with fread() and mmap() functions. The %of mmap benefit of the parallel implementations of apriori are compared by changing the number of threads .

6. Experimental Results

The following notations are used in the tables and graphs in the paper.

nrl- number of records in lakhs
SAF –serial apriori with fread()
SAM-serial apriori with mmap()
pmsc-percentage of minimum support count
PAFD-parallel apriori with fread() on dual core
PAMD-parallel apriori with mmap() on dual core
Prmb-percentage of real time mmap() benefit
Pumb-percentage of user time mmap() benefit
Psmb- percentage of system time mmap() benefit

The following results are observed from the experiments:

1. Percentage of real time mmap() benefit, Prmb values of SAM are compared to PAM by changing pmsc and keeping nrl constant for

different random data sets. Prmb values of PAM are more compared to SAM for all data sets.(Fig.6, Table.1) Prmb values of PAM are also increased compared to SAM for accident dataset at different pmsc values.(Table.3).Prmb values of PAM and SAM are also compared at each support count by changing nrl values. Prmb values are more for PAM compared to SAM at all support counts.(Fig.7,Table.4) Scalability of PAM is more compared to PAF for different datasets at different support counts(Fig.8, Fig.9, Table.2, Tabe.5).

2. Percentage of user time mmap() benefit, Pumb values and system time mmap() benefit psmb values of SAM are also compared to PAM for different data sets at different support counts. PAM gives more benefit compared to SAM in all the cases.(fig.10-13,Table.6-11)

6.1 Observations of serial vs parallel apriori with mmap()

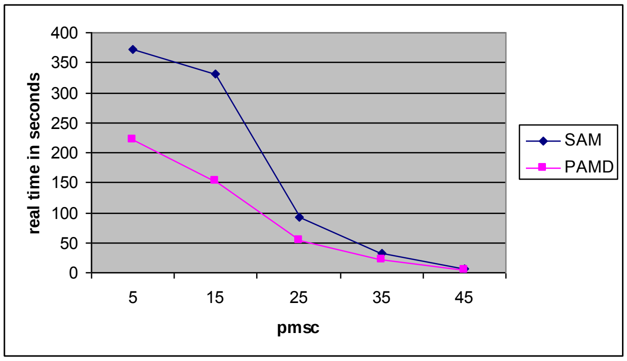


Fig. 1: SAM vs PAMD real time values for 8 lakh data

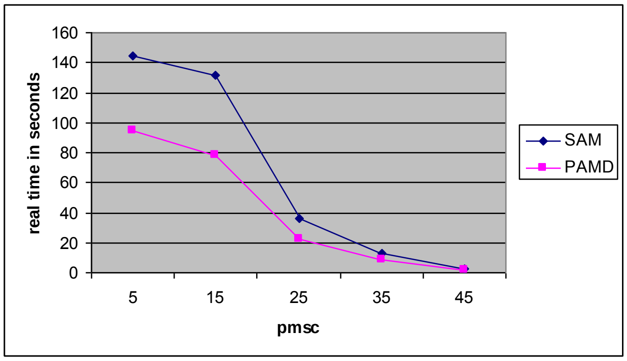


Fig. 2: SAM vs PAMD real time values for accident data

As the percentage of real time mmap() benefit , prmb is more with 3threads compared to 2 and 4 threads in most of the cases, all the PAF and PAM values in graphs and tables indicated in this paper correspond to the values obtained by parallelizing apriori on dual core with three threads.

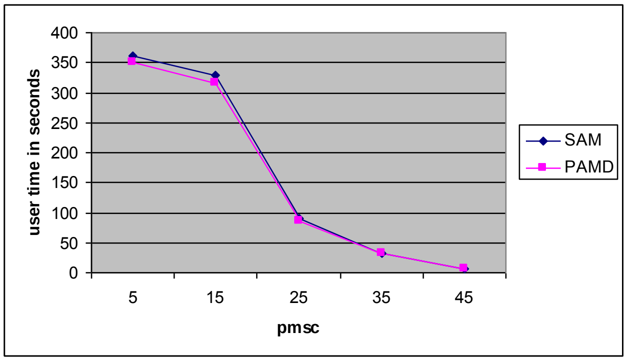


Fig. 3: SAM vs PAMD user time values for 8 lakh data

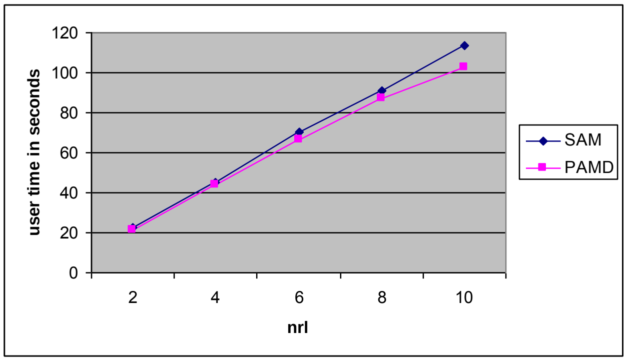


Fig. 4: SAM vs PAMD user time values at pmsc=25

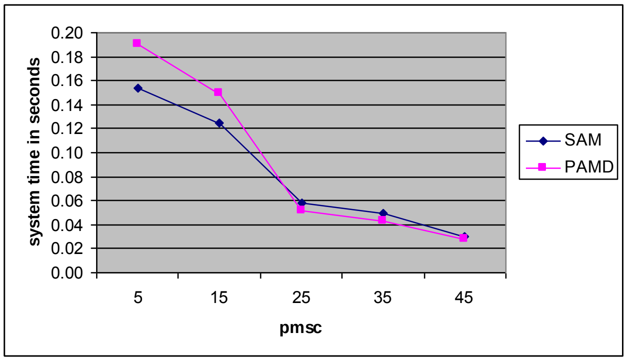


Fig. 5: SAM vs PAMD system time values for accident data

6.2 Observations of real time mmap benefit for serial vs parallel apriori with mmap()

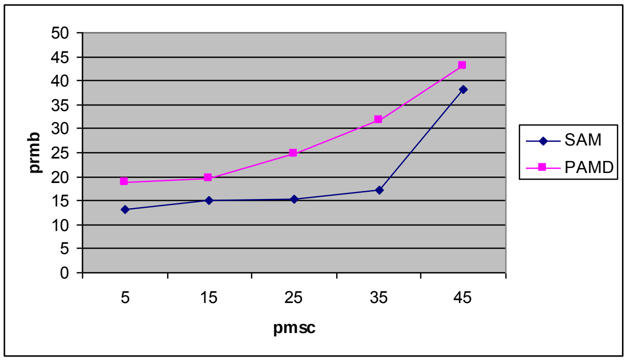


Fig. 6: SAM vs PAMD real time mmap benefit for 8 lakh data

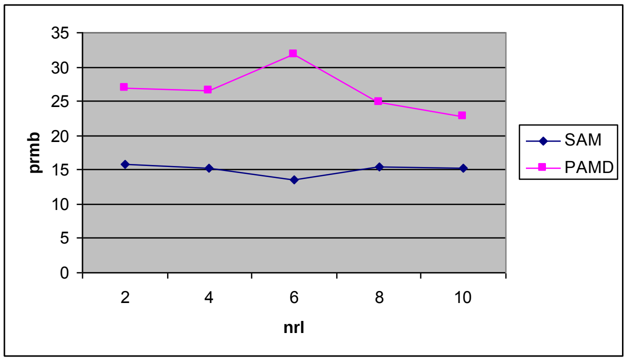


Fig. 7: SAM vs PAMD real time mmap benefit at pmsc=25

6.3 Observations of real time speed up for parallel fread() Vs parallel mmap()

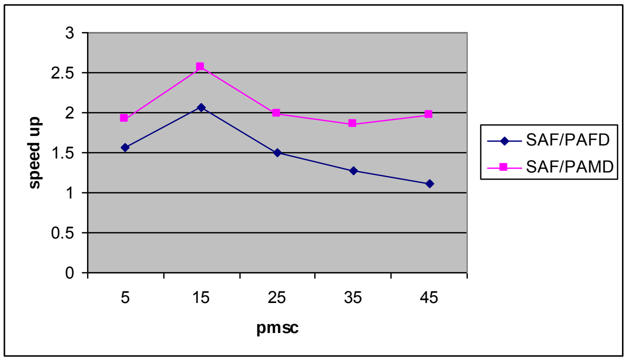


Fig. 8 comparison of speed up of PAFD vs PAMDfor 8 lakh data

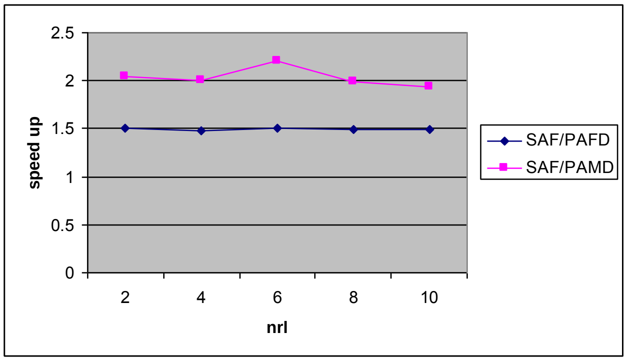


Fig. 9 comparison of speed up of PAFD vs PAMD at pmsc=25%

6.4 Observations of user time mmap benefit for serial vs parallel apriori with mmap()

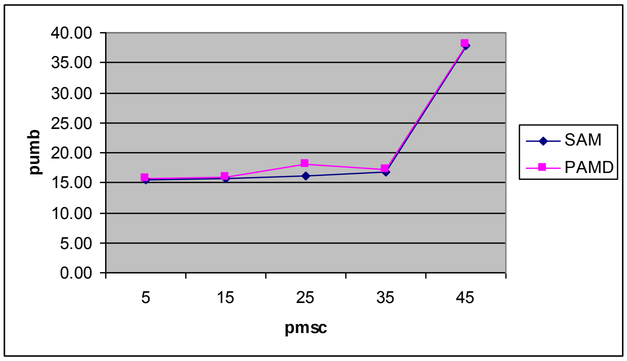


Fig. 10: SAM vs PAMD user time mmap benefit for 8 lakh data

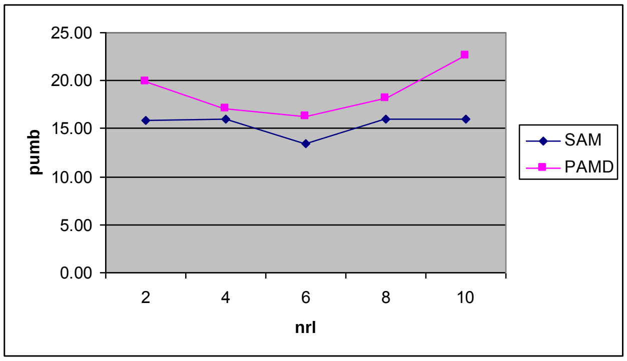


Fig. 11: SAM vs PAMD user time mmap benefit at pmsc=25

6.4 Observations of system time mmap benefit for serial vs parallel apriori with mmap()

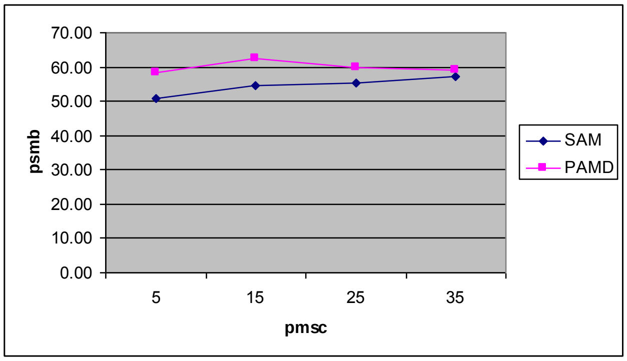


Fig. 12: SAM vs PAMD system time mmap benefit for 8 lakh data

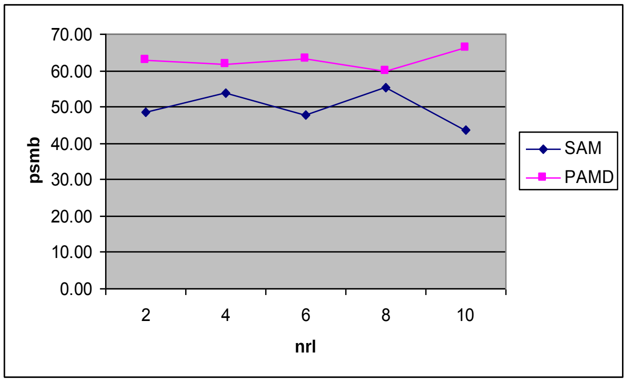


Fig. 13: SAM vs PAMD system time mmap benefit at pmsc=25

Table 1: Real time values for random data with nrl=10

Pmsc	SAF	SAM	pmb	PAFD	PAMD	pmb
5	537.54	455.95	15.18	342.01	263.65	22.91
15	487.97	418.32	14.27	236.5	184.77	21.87
25	135.77	114.98	15.31	90.9	70.13	22.85
35	49.5	40.15	18.89	38.6	26.25	32.00
45	11.4	7.05	38.16	10.2	5.34	47.65

Table 2: Real time speed up val ues for random data with nrl=10

pmsc	SAF	PAFD	PAMD	SAF/PAFD	SAF/PAMD
5	537.54	342.01	263.65	1.57	2.04
15	487.97	236.5	184.77	2.06	2.64
25	135.77	90.9	70.13	1.49	1.94
35	49.5	38.6	26.25	1.28	1.89
45	11.4	10.2	5.34	1.12	2.13

Table 3: Real time values for accident data

Pmsc	SAF	SAM	pmb	PAFD	PAMD	pmb
5	169.12	144.1	14.79	118.32	94.27	20.33
15	153.39	131.33	14.38	98.52	78.52	20.30
25	43.25	36.22	16.25	29.5	22.44	23.94
35	15.87	13.12	17.33	12.2	8.59	29.6
45	3.6	2.2	38.89	3.24	1.80	44.56

Table 4: Real time values for random data with pmsc=15

nrl	SAF	SAM	pmb	PAFD	PAMD	pmb
2	98.1	83.53	14.85	47.4	36.36	23.30
4	194.65	166.49	14.47	95.08	74.46	21.69
6	293.55	250.7	14.60	141.85	110.69	21.97
8	390.65	332.13	14.98	189.69	152.44	19.64
10	487.97	418.32	14.27	236.5	184.77	21.87

Table 5: Real time speed up values for random data with pmsc=15

nrl	SAF	PAFD	PAMD	SAF/PAFD	SAF/PAMD
2	98.10	47.40	36.36	2.07	2.70
4	194.65	95.08	74.46	2.05	2.61
6	293.55	141.85	110.69	2.07	2.65
8	390.65	189.69	152.44	2.06	2.56
10	487.97	236.50	184.77	2.06	2.64

Table 6: user time values for random data with nrl=10

pmsc	SAF	SAM	pmb	PAFD	PAMD	pmb
5	536.58	453.71	15.44	520.52	430.14	17.36
15	486.38	416.16	14.44	469.79	397.01	15.49
25	135.39	113.81	15.94	132.74	102.69	22.64
35	48.98	40.09	18.15	49.02	39.07	20.30
45	11.22	7	37.61	11.20	6.82	39.11

Table 7: user time values for accident data

pmsc	SAF	SAM	pmb	PAFD	PAMD	pmb
5	168.37	143.95	14.50	168.24	143.73	14.57
15	152.94	131.14	14.25	152.78	130.37	14.67
25	43.08	36.21	15.95	43.18	36.13	16.33
35	15.46	12.92	16.43	15.51	12.86	17.09
45	3.53	2.38	32.58	3.58	2.37	33.80

Table 8: user time values for random data with pmsc=15

nrl	SAF	SAM	pmb	PAFD	PAMD	pmb
2	97.66	83.33	14.67	94.04	79.74	15.21
4	193.99	166.12	14.37	187.70	160.64	14.42
6	292.63	249.4	14.77	282.06	237.36	15.85
8	389.55	328.59	15.65	376.21	316.68	15.82
10	486.38	416.16	14.44	469.79	397.01	15.49

Table 9: system time values for random data with nrl=10

pmsc	SAF	SAM	pmb	PAFD	PAMD	pmb
5	0.82	0.38	53.66	1.1	0.41	62.73
15	0.7	0.29	58.57	0.95	0.34	64.21
25	0.32	0.18	43.75	0.58	0.195	66.38
35	0.25	0.11	56.00	0.27	0.113	58.15
45	0.17	0.08	52.94	0.18	0.081	55.00

Table 10: system time values for accident data

pmsc	SAF	SAM	pmb	PAFD	PAMD	pmb
5	0.30	0.15	49.17	0.45	0.19	57.96
15	0.25	0.13	49.60	0.38	0.15	60.53
25	0.14	0.06	58.27	0.14	0.05	62.32
35	0.10	0.05	52.88	0.10	0.04	55.67
45	0.06	0.03	49.15	0.06	0.03	52.54

Table 11: system time values for random data with pmsc=15

nrl	SAF	SAM	pmb	PAFD	PAMD	pmb
2	0.17	0.08	52.94	0.19	0.08	57.89
4	0.28	0.14	49.29	0.43	0.16	62.79
6	0.53	0.22	59.43	0.55	0.2	63.64
8	0.57	0.26	54.74	0.77	0.29	62.34
10	0.7	0.29	58.57	0.95	0.34	64.21

7. Conclusions

The performance of apriori with memory mapped files concept compared to standard fread() function for reading data from the transactional database is identified by using linux mmap() function. The mmap() function shows better performance than fread() in real time, user time and system time.The percentage of mmap benefit is more in parallel apriori compared to serial apriori .

References

[1] Tevanian, Avadis, et al. "A unix interface for shared memory and memory mapped files under mach." Dept. of Computer Science Technical Report, Carnegie Mellon University (1987).

[2] S. N. Tirumala Rao, E. V. Prasad, and N. B. Venkateswrlu, “A Critical Performance Study of Memory Mapping on Multi-core Processors: An Experiment with K-means Algorithm with Large Data Mining Data Sets”, IJCA (0975-8887)2010 Volume1-No. 9.

[3]Optimized performance analysis of Apache-1.0.5 server,www.isi.edu

[4]fread/ifstream, read/mmap performance results www.lastmind.net.

[5] “Multi-core Procesor” From wikipedia ,the free encyclopedia.Available: en.wikipedia.org/wiki/Multi-core processor [Accessed: May 24,2012].

[6] OpenMP Architecture , "OpenMP C and C++ ApplicationProgramInterface", Copyright © 1997-2002 OpenMP Architecture Review Board.http://www.openmp.org/

[7]OpenMP® Programming forTMS320C66x multicore DSPs © 2011 Texas Instruments Incorporated Printed in U.S.A.

[8] Anuradha T ,Satya Prasad R, S.N Tirumalarao "Parallelizing Apriori on Dual Core using OpenMP". International Journal of Computer Applications 43(24):33-39, April 2012.

[9] Anuradha T ,Satya Prasad R,S.N. Tirumalarao "Parallelizing Apriori on Dual Core with multiple threads" International Journal of Computer Applications 50(16):9-16, July 2012.

[10] Agrawal R, Srikant R "Fast algorithms for mining association rules" In: Proceedings of the 1994 international conference on very large data bases (VLDB'94), 1994 Santiago, Chile, pp 487–499

[11] R. Agrawal and J. Shafer "Parallel mining of association rules" IEEE Trans. Knowl. Data Eng., vol. 8, pp. 962–969, Dec. 1996.

[12]M.J. Zaki 1997 "parallel and distributed association mining: A survey" IEEE Concur, vol. 7, pp. 14–25, Dec. 1997.

[13] O. R Zaiane,M. El-Hajj, and P. Lu "Fast parallel association rule mining without candidacy generation" in Proc. ICDM, 2001, [Online].Available: citeseer.ist.psu.edu/474 621.html, pp. 665–668.

[14] Laurent, Anne, et al. "Pgp-mc: Towards a multicore parallel approach for mining gradual patterns." Database Systems for Advanced Applications. Springer Berlin/Heidelberg, 2010.

[15] Liu, Li, et al.2007 "Optimization of frequent itemset mining on multiple-core processor." Proceedings of the 33rd international conference on Very large data bases. VLDB Endowment, 2007.

[16]. Shirish Tatikonda, Srinivasan Parthasarathy 2008 "Mining Tree Structured Data on Multicore Systems", VLDB '08, August 2430, 2008, Auckland, New Zealand

[17] Mohanavalli, S., S. M. Jaisakthi, and C. Aravindan.2011 "Strategies for Parallelizing KMeans Data Clustering Algorithm." Information Technology and Mobile Communication (2011): 427-430.

[18] Memory management in Linux for linux device drivers Third edition eMatter Edition Copyright © 2005 O'Reilly & Associates

[19] Jiawei Han and Micheline Kamber "Data Mining concepts and Techniques", 2nd edition 2006 Morgan Kaufmann Publishers, San Francisco.

[20] Diaz, Javier, Camelia Muñoz-Caro, and Alfonso Niño. "A Survey of Parallel Programming Models and Tools in the Multi and Many-Core Era." Parallel and Distributed Systems, IEEE Transactions on 23.8 (2012): 1369-1386.

[21] Kent Milfeld 2011 "Introduction to Programming with OpenMP" September 12th 2011, TACC

[22] Ruud van der pas "An Overview of OpenMP" NTU Talk January 14 2009

[23]Chapter12 "Shared memory Introduction" www.kohala.com/start/unpv22e/unpv22e.chap12.pdf

[24] K Geurts, G Wets, T. Brijs and K. Vanhoof, "Profiling High Frequency Accident Locations Using Association Rules", Electronic Proceedings of the 82th Annual Meeting of the Transportation Research Board, Washington, January 12-16, USA, 2003,18p.

Discuss the Development of Computer-Aided Industrial Design Technology

Jun YAO[1]

[1] First Author, Corresponding Author **School of Arts and Design，China University of Mining and Technology**
Xuzhou, Jiangsu 221116, China

Abstract

The direction of the development of the industrial design gradually approaches mechanical-electrical integration and informational and electronic products. With an increasing improvement of its technical content, and better social economic conditions, people's consumption concept are getting more and more different. Consumer concerns not just the functionality and quality of the product, more and more people are starting to focus on the appearance of the product, degree of innovation, environmental protection, and so on, which brings a higher degree of difficulty to the industrial technology. It is because the increasing demands of people and the industrial design, many scholars are increasingly concerned about the industrial design in recent years. With the continuous development of computer technology, a wide variety of hardware and software are developed, and a variety of ever-changing technologies are attracting the industrial design talents as well.

Keywords: *Computer-Aided Industrial Design, industrial design technology, CAID*

Computer-aided industrial design (Computer-Aided Industrial Design, CAID) software system is a software system provides automated support for the products of industrial design from the product form, color, decoration, man-machine environment during the product conceptual design stage. It is one of the effective tools to support product innovation design. In recent years, CAID technology has become one of the hot research spots of innovative design and computer-aided technology.

1. Introduction

With the development of information technology and computer network technology, the world economy is undergoing a profound revolution around the "network economy". This revolution has dramatically changed the face of the world economy and the manufacturing environment. The diversification and personalization of consumer's demand have led to the market dynamic variability; manufacturing enterprises are no longer isolated individual resources, but a member of the social system. Facing the trend of increasing competition and product complexity, enterprises should enhance cooperation and participate in dynamic manufacturing system restructuring, in order to change the tactics of manufacturing companies, and establish a digitized, flexible and agile networked manufacturing mode.

2. Development of Computer-Aided Industrial Design Technology

Currently, the domestic and international CAID's research focuses on the application research of computer-aided modeling technology, human-computer interaction in CAID, smart technology, and emerging technology. And it also introduces the design module of some well-known CAD / CAM / CAPP commercial software industry.

2.1 Computer-Aided Modeling Technology

In the area of CAID technology, computer-aided modeling technology is mainly reflected in the shape of the free-form surface design and sketch design. In the free-form surface design, the product appearance freeform surface design study is an important content of the CAID. And the surface feature design is an important development in the design of free-form surfaces. The surface feature design includes three parts, namely basic surface, mobile features and collusion graphics.

2.2 The High-Tech of CAID

Currently, the market began to slip from the emerging technologies in the high-tech of CAID, such as virtual reality, genetic algorithms and so on. But how to use these technologies well in CAID field, this would be carried out with some of the traditional technologies effectively, thereby to approach some CAID related research. The collaborative, parallel design is now one of the main development directions of this technology. As for the industrial design, the sense of product design is very important, a detailed study of the product functions, principles, shape must be carried out. Many scholars observe on a variety of contents in some parallel environment in different angles of the technology and deep into it to understand and explore. But the starting point of other scholars is from the CSCW perspective, they have carried out a detailed analysis of the collaborative design and studied the use of this design, explored the model of engineering and industrial design work.

2.3 Intelligent Technology of CAID

Currently, the Intelligent CAD has witnessed a considerable degree of development. Integrated Intelligent Design System (I2CAD) provides an integral computer support to the design during the whole process. As it relates to the creative thinking as well as the frequent human-computer interaction, industrial design, in particular, needs auxiliary artificial intelligence technology. In the industrial design process of creative design thinking, the translation of the designer’s conception fast into sketch is a fairly complex process behavior, and this process is known as the stage of concept function. Many scholars make researches on the stage of concept function from the designer's creative design thinking, design knowledge representation, and put forward their own views with a combination of sketch design. Design grammar, it is the formal description method by refining and abstracting the elements of shape, color, and shape of the object and its generation principle from the angle of design methodology. It is one of the foundations of intelligent design system. At present, the design grammar of the industrial design community scholars mainly includes the pose grammar, color grammar, shape grammar and modeling conversion grammar.

2.4 Key Technologies to Be Solved

(1)Research of modern design methodology. Based on the development direction of the modern industrial design, makes research on the qualitative design process and the design method with the similar accurate method to lay the theoretical foundation of CAID from the design object itself.

(2) Research of innovative design technology. Follow the principles and norms obeyed by the design process of research design thinking process and computer support; explore a wide range of innovative techniques, study innovative design principles, methods and techniques in-depth.

3. Influence of the Computer to Industrial Design

3.1 Influence of the Computer to The Concept And Method Of Industrial Design

Due to the development of computer software, the product has made a great progress in the design of the degrees of freedom. In the traditional design, the expression of hyperboloid and the free-form surface is very troublesome, and it often needs to produce a solid model to express clearly. It is also a difficult thing to change model again into engineering drawing. Therefore, in the design, the designer always avoids the use of free-form surfaces, which makes the design conservative. Today, with the use of computer to generate data model, all these difficulties are gone, and the relationship between design and manufacturing is closer. The use of computers makes us change the design criteria. Traditional design puts high demands on the effect of expression, it often takes that whether the drawing production is sophisticated, the line is light, and the color is uniform as an important criterion of evaluation. However, this criterion loses its meaning due to the computer-precise data and sophisticated output. Meanwhile, we put the evaluation criteria on the evaluation of the merits of the design. Moreover, the Computer-aided design has shortened the product development cycle. On one hand, it increases the efficiency of the work; on the other hand, it eliminates many steps of the traditional design performance. Especially on the program modifications and adjustments, it is very convenient to modify because the computer retains the whole process of design.

3.2 Product Modeling CAID

Traditional mechanical design and manufacturing is cumbersome and difficult to modify because of the use of artificial mapping, it generally only draws view. It is very difficult to draw a perspective view for complex models, designers can only base on its plan to imagine the finished model after the three-dimensional model, which is very difficult to design and production. However, computer-assisted cartography changes all these. For example, a common base, the computer maps out its three-view, the shaft side maps, and cross-sectional view only in5 min, and computer automatically marks all sizes. when you want to change a size, all views are automatically amended accordingly axis the angle of the side of the map, cross-sectional view of the cutting position can be adjusted and can be displayed to direct three-dimensional effect. Simulating products work environment, rendering module, assigning different materials, designing products appearance, drawing idea sketches and design effect diagram, mimic motion effects, analyzing movement interference, with this end, we can watch last made to improve efficiency and to avoid losses caused by ill-considered design at the design stage.

3.3 Conceptual Effect Drawing Stage of Modeling Design

In the early stages of the modeling creative design, industrial design promotes hand-painted, for hand-painted is the most natural way for designers to capture the inspiration, as long as there is a piece of paper, a pen, they can record their inspirations at anytime, anywhere. Hand drawing even should be a means of inspiration record because inspiration cannot be controlled. Only accumulating in daily life, will you be able to come in handy in a lot of materials. When you combine these materials with your personal style as well as product design orientation, you can go back to work in front of the computer.

4. Characteristics and Application

4.1 The Main Features

CAID technology has unparalleled advantages than the traditional industrial design, industrial designers can free to express a creative idea to display their talents if master it. It can enhance the quality of the overall product design, strengthen the competitiveness of the product market, and has the following characteristics: high-quality three-dimensional space software system set a three-dimensional solid modeling, static coloring, complex lighting model, and the multimedia animation in one, vivid image. It ensures the high quality of the design through advanced design tools. Flexibility with such high-tech tools for creative design directly on the system, using 3D solid modeling techniques for geometric modeling of objects, such as a color design, material editing, form, texture depicting real-time rotation transformation, rapid real image generate output, many different styles, program evaluation and testing. It can easily be modified until satisfied. The system is to optimize the design.

4.2 The Application of High-Tech Research in CAID

Currently, these emerging technologies of virtual reality, neural networks, genetic algorithms and parallel design, collaborative design method are the hot spots of the majority of scholars. The introduction of these technologies into CAID field, combination of traditional optimization design, fuzzy technology, intelligent technology to CAID study also gradually win the attention of scholars.

Parallel design, collaborative design is one of the trends of modern design. In the field of industrial design, especially product design, it is necessary to study the parallel and collaborative design mechanism of product features, principle, and layout. From Concurrent engineering point of view, some scholars have explored deeply into the parallel design process and design environment for concurrent engineering design technology.

4.3 Positioning and Its Implementation

Analysis through in-depth investigation of a number of factories, refrigerators, machine tools, and other typical product development process, various design techniques can be seen in the application of the entire product development process and the role of various designers. Based on accurate position, they can develop CAID system development strategy and implementation method. Computer-aided industrial design (CAIDS) is developed in accordance with the CAID technical principles and methods of computer-aided design software system. It is shown in Fig.1.

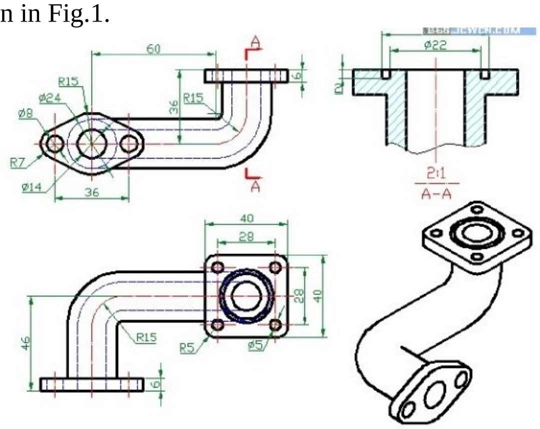


Fig. 1 Computer-aided design software system

It should be in accordance with the thinking of system engineering, completely take industrial design theory and methods as guidance for the smart innovative product development and design system. it first does form design and human-computer design in the machine; then imports the product model into design platform, including design of color, decoration, material, etc.; at last, it comes to the conclusion of the product modeling program modeling program expression including renderings of product modeling, design, evaluation, and engineering geometric model. The product designed through the system has the features of good shape and beautiful color, pleasant, high quality, efficient, animation and others.

4.4 Application Examples

It can effectively form a product family or a different design for the evaluation and selection. Currently it mainly uses algorithms include adaptive neural network and morphological differences in residual algorithm to control the generation of the new design, and the method of adaptive neural network is applied boarder. Figure 2 is its calculation model.

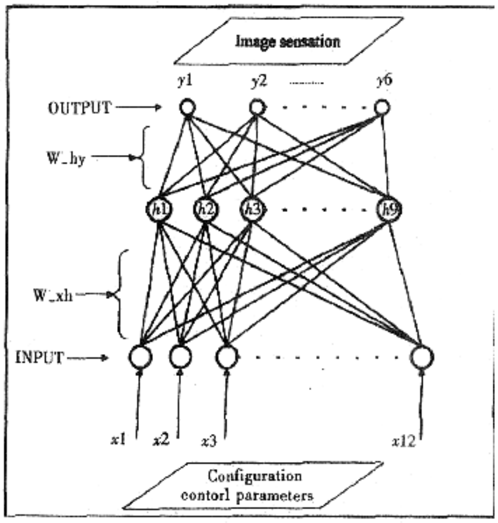


Fig. 2 Calculation mode

5. Developing trends of CAID

From perspective of industrial design, with the further development of CAD, artificial intelligence, multimedia, virtual reality technology, there must be a deeper understanding of the design process and a new level of design thinking of design mind. CAID will make industrial design develop in the direction of diversification, optimization, integration, make human-computer interaction more natural, innovative design more advanced and effective. From the entire product design and manufacturing trends, parallel design, collaborative design, intelligent design, virtual design, agile design, full life cycle design, all these design approaches represent the development direction of modern product design patterns. With the further development of the technology, on the basis of the information, product design patterns will inevitably develop in the direction of digital, integrated, networked, and intelligent computer-aided. Industrial design trends must be consistent with the above trends, and eventually establish a unified design support model. The industrial designers and engineering designers should gradually converge toward unification.

5.1 Functional Components and Data Modeling Platform Interactive Mode

When the design system uses a COM component technology to the development of various functional modules, the data between the functional components and modeling platform interactive mode select and transplanted directly determines the design of the system performance.

The mode C is the tool interface independent, be a separate component. The purpose of this development model to facilitate the expansion of the tool interface updates. Because of close interaction interface with the user, in some cases, the possibility of change of the interface is relatively large. When the tool interface needs to be changed, according to the mode C development tools components only need to change the code to interface components. C / S mode software are more suitable for with mode C development components. Of course, these three development modes have their pros and cons. Mode C enhanced maintainability, due to excessive interface calls and reduce the operating efficiency of the tool.

5.2 Concepts and the Way Influence of Computer To Industrial Design

The impact of the development of computer software, computer industrial design concepts and have made great progress on the way to make the product design in degrees of freedom. The traditional design of hyperboloid, the expression of the free-form surface is very troublesome, and often need to produce a solid model to express clearly. Model again into engineering drawing, is a difficult thing. Therefore, in the design, designers always avoid the use of free-form surfaces, which makes the design has become conservative. Today, the use of computer generated data model, all these difficulties are gone, and the closer relationship between design and manufacturing. The use of computers make us changed the design criteria. Traditional design with high demands on the effect of expression, often drawing production whether sophisticated, the lines whether light quite, and the color is uniform as an important criterion of evaluation. Computer-precise data, sophisticated output effect; this criterion loses its meaning. While the evaluation criteria on the evaluation of the merits of the design. Computer-aided design to shorten the product development cycle. On one hand, promote the efficiency of the work; on other hand, eliminating the need for many of the traditional design performance steps. Program modifications and adjustments, because the computer retains the whole process of design, modify it very convenient. In this way, than the traditional design in the development of a new product should be shortened from one-third to one-half of the cycle, and some can even shorten the cycle multiplied

5.3 Internet to Influence the Industrial Design

Today the momentums Internet (Internet) connectivity worlds also expand the presence of art and design space toward the broader digitization of the art design. Not connected to the Internet, the art design is still subject to geographical and time constraints, and with the Internet, information can travel at the speed of light, art and design works taken around the world. People either own creation of electronic files in the network, the existing work of art can also be converted to digital images on the network around the world who are interested to browse in the world of digital-based can also communicate with others their own point of view or order their favorite works.

5.4 Features and Advantages of Virtual Building Design Method

Based "virtual building" intelligent building components of the multi-angle, providing a fully 3D environment. Virtual Building smart objects, all the building components are parameterized contains special properties of building components, such as dimensions, materials, performance, cost and other comprehensive information intelligent three-dimensional objects.

5.5 Design Functional Requirements of The System

Focused thesis considered from the perspective of the application of industrial design, form design method to analyze the design system, and does not involve other aspects. The papers from the form layout, proportion, linear, detail four aspects to consider the form of the computer-aided design system functional requirements of industrial design:

Provide layout design support, in particular the application of surface segmentation technology, auxiliary designer morphology facade.

Provide morphology proportion design support, provides the proportion of information in a timely manner for the designer, assisted designers tune. Whole, the design has the priority ratio between the forms of graphics, auxiliary designer overall ratio proportional relation.

To provide a form of linear design support, especially to generate characteristic curve, the organization of free curve function, such as the number of free-form curve adjusted mutual the congeners curve relationship or proportion curve relationship.

6. Design of Computer-Aided Morphological System

Industrial design form of the computer-aided design system (referred to as the design of the system, the same below) should include the five aspects of domain knowledge, design goals, the designers, the design process and the resources available. On the basis of analysis of the morphological design requirements of computer-aided industrial design, selection modeling platform and systems development techniques, and modeling platform to build a computer-aided morphological design system. It is shown in Fig.3.

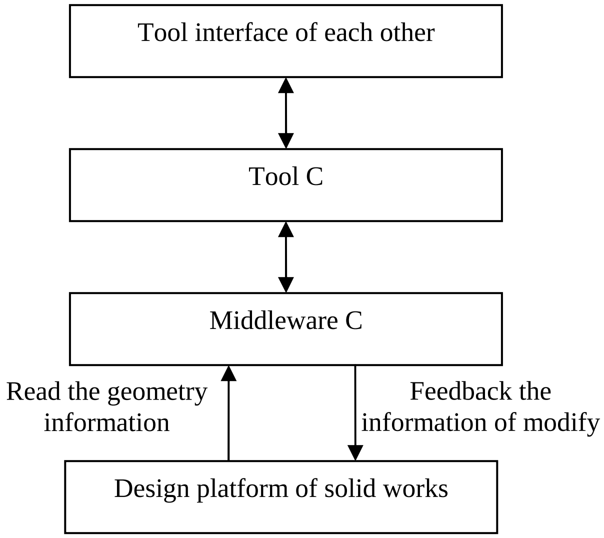

Tool interface of each other
Tool C
Middleware C
Read the geometry information
Feedback the information of modify
Design platform of solid works

Fig.3 Mode C

7. Conclusions

Collaborative industrial design system, based on the design of the product, we have to borrow the morphology typical electromechanical products shape our self-developed auxiliary design system design, color design and evaluation module, they are integrated into a collaborative environment, thus saving the development cycle of the system, and to avoid unnecessary duplication of development. Has been analyzed in the previous section of this paper characteristics of VRML This graphic file formats, pointed out that the hierarchical tree structure is particularly suitable for the organization and storage of industrial design products, so choose VRML to build the cooperative industrial design process product model, coupled with interactive features on Java3D and call the relationship between them, so Here take a look at how they are down collaborative environment industrial design activities.

Product designs in collaborative environment include three part of designed by the client, concurrency control, server-side storage. Concurrency control is responsible for conflict resolution, which is at the same time a product model can only be a designer calls. Because the server side is the use of a relational database for program storage, so the database record locking can solve the problem of access to the conflict.

References

[1] Rajesh Kumar Goutam, Sanjay Kumar Dwivedi, "Search Engines Comparison on the Basis of Session Duration and Click Hits", International Journal of Computer Science Issues, Vol. 8, No. 2, 2011, pp:179-183.

[2]WANG Hai-bo, "Computer Aided Industrial Design", Journal of Anhui University of Technology, No.2, 2005, pp:23-26.

[3]Johannes Behrisch, Mariano Ramirez, Damien Giurco, "Representation of Ecodesign Practice: International Comparison of Industrial Design Consultancies", Sustainability, Vol.3, No.10, 2011, pp: 1778-1791.

[4] G.L. Hu, X. Zhu, “Comprehensive evaluation of population, resources, environment and economic system of Xinjiang: Based on the principal component analysis”, Ecological Economy, No. 6, 2009, pp. 67-69.

[5] S. Li, W. Qiu, and Q.L. Zhao, “Quantitative relationship between environmental quality and economic development of Heilongjiang province”, Journal of Harbin Institute of Technology, Vol.38, No. 11, 2006, pp.1986-1988.

[6] Elmira Moghaddami Khalilzad, Sanam Hosseini, "Recovery of Faulty Cluster Head Sensors by Using Genetic Algorithm (RFGA) ", International Journal of Computer Science Issues, Vol.9, No. 4, 2012, pp: 141-145.

[7] Gert Pasman, Ingrid Mulder, "Bringing the Everyday Life into Engineering Education", International Journal of Advanced Corporate Learning, Vol. 4, No.1, 2011, pp: 25-31.

Jun Yao Male, Han nationality, born in September 1979, Jiangsu people, Art and Design Institute of China University of Mining and Technology, associate professor, Postgraduate, master degree; research direction for industrial design. University Road 1 Art and Design College of China University of Mining and Technology, Xuzhou City, Jiangsu province.

Research on Zero-Knowledge Proof Protocol

Wang Huqing[1,2], Sun Zhixin[3,4]

[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2]College of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing,China
[3]College of Internet of Things , Nanjing University of Posts and Telecommunications, Nanjing, China
[4]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

Abstract

Zero-knowledge proof protocol is a critical component of cryptography, which in recent years has raised increasing concern of many scholars.Its application field is very extensive,and it has made breakthrough progress in many aspects,including mathematics and network safety and so on. This article launches the elaboration from the concept, nature, mathematics theory, general proof process of the zero-knowledge proof , focusing on the application research of polynomial function root, graph isomorphism, cloud storage service, RFID, proxy digital signature and identity authentication etc.Finally, the direction for further research is summed up.The systematic introduction to zero-knowledge proof protocol has important theoretical guidance and practical significance on attracting more scholars involved in the research as well as expanding application fields.

Keywords:zero-knowledge proof; identity authentication; digital signature; cloud storage; polynomial function root

1. Introduction

Zero-knowledge proof ,a very interesting and important applicative topic[1]，which has attracted a lot attention of peer scholars, having achieved abundant achievements in many security related fields , has became one of the hot issues in the research fields of cryptography[2][3].

1.1 Main ideas

The main idea of zero-knowledge proof is as follows: P (the prover) had some secret information. P wanted to prove to V (the verifier) by taking other proof process without revealing anything other than the fact that it knows in order to prevent the confidential information from leaking to anyone (including V or any other third party). We call this technology which can achieve the purpose of proof without revealing anything " zero-knowledge proof(ZKP) "[4].

1.2 Example analysis

Nineteen eighties, Goldwasser et al first put forward the concept of zero-knowledge proof [5]. We can understand the zero- knowledge proof from the case in life:

(1) Supposing a room can only be opened by a key with nothing else available,The prover P wanted to prove himself owning the room key without revealing the key to the verifier V to prevent leakage. Therefore, if V determined that the room had a certain object, P can take out the object to prove himself having the key. This proof process is zero-knowledge proof, and " knowledge " is the key [6].

(2) zero-knowledge proof of color blind with red and green ball[7]

Surpposing the verifier V is color blind.There are now two balls,one is red, another is green.The two balls are exactly the same besides their color.It is required to prove to V that the two balls are truly different because they seem to be identical to him. We adopt the following method: let V hold a ball in each hand, standing opposite P. P can also see this two ball without telling V which is red and which is green. Then, let him take hands behind his back, he can decide randomly whether to swap the balls or not. The probability of exchanging or not each accounted for 1/2. Finally, he takes balls from back and asks P to answer whether V has exchanged two balls. According to the color of the ball, P can simply judge. Repeatedly, if P can answer correctly everytime, then V will believe that these two balls color are different to a large enough probability . The process also belongs to zero-knowledge proof, and " knowledge " is the color of the ball.

(3) The most typical example of zero-knowledge proof is the Cave model which was put forward by Jean-Jacques Quisquater and Louis Guillou[7]. As shown in Figure 1:

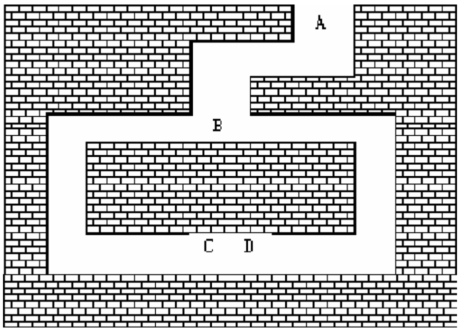


Fig.1 : the cave model.

1.3 the general process of zero-knowledge proof protocol

The general process of using zero-knowledge proof protocol is shown in following figure:

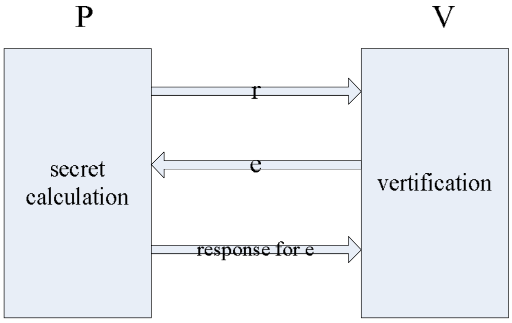


Fig 2:the general process of zero-knowledge proof protocol

(1)The prover P sents promise random number r to the verifier V.
(2)V sents random challenge value e to P.
(3)P calculats secretly and sents the result to V as the challenge-response for second step.
(4)V verifies the response. If the verification fails, the process of proof will end.Otherwise, the above steps will be repeated for N times. If every verification can be successful , V will receive P's proof in great probability.

1.4 Nature of zero-knowledge proof protocol

Zero-knowledge proof protocol has the following three important properties.In other words, using zero-knowledge proof protocol to prove a problem, the proving system must meet the following requirements:
(1) Completeness
If the prover P and the verifier V comply with the general process of zero- knowledge proof protocol strictly, the proof is considered to be successful and P is credible.
(2) Rationality
If it failes once in N times of verification,the proof is regarded as failed and P is a fake prover who is unreliable.
(3) Zero-knowledge
During the verification,V can't obtaion any privacy or important information,let alone anything about the knowledge,except to believe that P dose have it.Even though V verifies repeatly,he can not prove the existing fact to others anymore.

2.Key mathematical knowledge applied to zero-knowledge proof protocol

The commonly used algorithm theory in zero-knowledge proof , which is similar to public-key cryptosystems in cryptography ,is mainly based on the following key mathmetic problem:
(1)the square root problem of modulo n
Given a positive integer n , a∈Zn, if there exists x∈Zn that makes x2=a(mod n), then x is called as a square root of modulo n.
(2) the calculation problem of the discrete logarithm
Given a prime number p,and a,which is one of the primitive element on finite field Zp .To b on Zp,looking for one and only integer c that makes a^c≡b(mod p). In general, the problem is difficult if you are looking forward p, and there is still no algorithm to calculate polynomial of discrete logarithm. The method based on the elliptic curve discrete logarithm is commonly used.
(3) Large integer factorization problem
The factorization of a large integer M which is N digit, is usually impossible to be done in O (N) , but rather to up to O (exp (N)) level.

3. Typical application and Implementation of zero-knowledge proof protocol

3.1 Zero-knowledge proof of polynomial function root

Assuming that P has got a solution x0 to an integral coefficient high-order polynomial function f (x), he wants to prove himself without revealing x0 or any information about x0 to V.This is a zero-knowledge proof problem of polynomial function root . In document [8], the author introduces multiple discrete logarithm problem and puts forward zero-knowledge proof algorithm of polynomial function root , on the basis of solving problem of discrete logarithm.
Assuming integral coefficient high-order polynomial function f (x) = $\sum_{i=0}^{n} a_i x^i$, the process is shown as follow:

(1) prover P and verifier V choose p and generating element α of Z_p^*

(2) P calculates $\beta_i = \alpha^{x^i} \bmod p, i = 1,2,\cdots n.$ and sends the result to V

(3) P proves himself having the solution x_0 to establish $\beta_i \equiv \beta_{i-1}^{x_0} (\bmod p), i = 1,2,\cdots n$ to V by using zero-knowledge proof method of multiple discrete logarithm.

(4) V verifies $\prod_{i=0}^{n} (\beta_i)^{\alpha_i} \equiv 1(\bmod p)$

The above process is repeated many times.

3.2the Isomorphism problems

Assuming there are 2 undirected graphs G1=<V1，E1> and G2=<V2,E2> , the number of their vertices are equal and the number of their sides are equal .If there exists a replacement, φ ,which makes any （u，w）∈E1 meet φ （u，w)∈E2 , the two graphs are isomorphic. If P wants to prove that it knows the replacement φ ,which satisfies the conditions while P does not want to disclose any relevant information ,it is a zero-knowledge proof problem.

The proof is given in the literature [9]:

(1)P randomly selects a replacement γ , in the role of γ , converts Figure G1 into Figure H and sends it to V .

(2)V randomly selects e∈{0,1} and sends it to P .

(3)P calculates based on the value of e ,if e=0，sends $\pi = \gamma$; if e=1,sends $\pi = \gamma\ \varphi$ -1

(4)V Verifies whether H= π Ge is established .

The above process is repeated many times.

3.3 Cloud storage services security management

The security, reliability, and availability of cloud storage services are the important factors which result in the widely uses of the cloud storage business .In order to obtain the user's trust, the cloud storage service provider must provide the users with recoverability proofs(POR，a proof of retrievability)[10],and prove the integrity and security of the data.Considering the security of POR protocol , there are mainly two points which are listed as following :

(1)How to prevent prover from deceiving the verifiers by tampering data in the process of verification ?

(2)For publicly verifiable POR protocol , how to prevent malicious verifier to store data by analyzing the response of public challenges?

Based on the above two security issues ,in the POR protocol ,we can consider using the zero-knowledge proof .The completeness and rationality of the zero-knowledge proof method can ensure the security of the first point and the “zero-knowledge” characteristic of the zero-knowledge proof method can ensure the security of the second point .

In the literature [11], the author suggests the zero-knowledge data recoverability proof protocol model, which can prevent the provers` deception and the leak of validation data . The certification process is in strict accordance with the process of Figure 2 mentioned above.The algorithm mainly uses the bilinear map group system S= （p,G,GT,e) where G and GT of two large prime numbers p-group of order and will need to verify that the data file F into a partition F= $\{m_{i,j}\} \in Z_p^{n\times s}$, the secret data of each sub-block $\{m_{i,j}\}$ is protected by the random number $\lambda_j \in Z_p$ and tag $\{\sigma_i\}$ is protected by the random number $\gamma \in Z_p$ .Moreover to avoid malicious verifier from obtaining $\{\lambda_j\}$ and γ , the author takes advantage of a committed method to protect by using H_1^{λ} and $e(\prod_{i=1}^{s} u_i^{\lambda_i}, H_2)$ where $H_1 = h^{\alpha}$, $H_2 = h^{\beta}$, α and β are random numbers of $\in_R Z_p$; h is an anti-collision Hash function, $u_i = g^{\tau_i} \in G$, τ_i is a random number which $\in Z_p$ and its range: （i=1,⋯s), E is the mapping of the bilinear map group system .

3.4 Applications in the RFID

RFID(Radio Frequency Identification) is mainly composed of the readers, tags and databases.Since there are enough energy and computing resources between the database and the reader ,traditional cryptographic algorithms can be considered in terms of safety , it is commonly believed that the passage between the two is secure . But the communication media between the reader and the tag is a wireless media ,which is completely exposed to the attacks and other unauthorized readers . The transfer of information lacks confidentiality and there is a possibility of malicious attacks.For example, an attacker can replay legitimate reader or tag signal to counterfeit in order to get the trust of the label or reader . Thus the attacker obtains secret information .To be against such attacks, to verify the identity of a legitimate reader or tag and not to disclose any private information is especially important . To achieve the purpose of proving legal identity without disclosing any private information ,scholars have thought of using zero-knowledge proof protocol to deal with the problems.In the literature [13] and [14] ,the author shows the mutual authentication protocol based on zero-knowledge proof protocol RFID which are between the reader and tags . Here, we can give out the zero-knowledge authentication method based on the reader of elliptic curve . Fig.3:

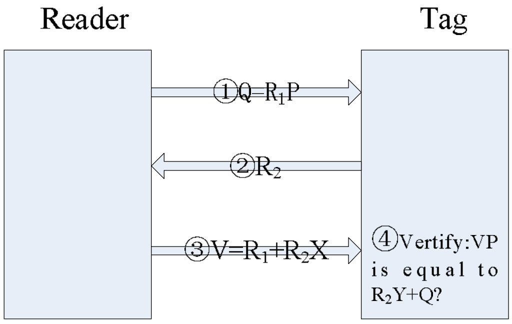


Fig.3: zero-knowledge proof protocol applied in RFID

The algorithm principle of this authentication is to use the elliptic curve to deal with discrete logarithm problems .Here is an introduction of the parameters related to what mentioned above:

① Finite field selection GF(2m)；parameters a and b，define the elliptic curve

E: $Y^2 = X^3 + aX + b(p > 3)$ or

$Y^2 + XY = X^3 + aX + b(p = 2)$

② P denotes the basis point

③X is the private key，Belong to the random intervals in the range of [1.n-1]；Y is the public key，Y=XP

④ R1 and R2 are both random integers，and R1∈[1,n-1]，R2≥1.

3.5 proxy digital signature

Currently，the two most popular public-key digital signature methods ： one is the RSA digital signature method which is based on the decomposition of large factor difficult ， and the other is the ELGamal type digital signature method which is based on the finite field GF (p) looking for discrete logarithm difficult . The proxy signature is needed when the real signer is absent .It is neither feasible nor safe to obtain the real signer's private key in computing.So,the common solution is to produce proxy signature ,but not the proxy signer .Proxy signature has the following properties :

① The dissimilarity between Proxy signature and normal signature，

② The Unforgeability. Only the original signer and designated agent signer can produce effective proxy signature，

③ Verifiability .Judging from the proxy signature ,the validators can believe the original signer identity the signature news，

④ Identifiability. The original signer can recognize the identity of the proxy signers from the proxy signature.

⑤ Non-repudiation. The proxy signer cannot deny the proxy signature which was erected and approved by himself.

Based on the requirements above and the characteristics of zero-knowledge proof, we can consider using zero-knowledge proof agent agreement principles to realize the Proxy Digital Signature. Literature [15][16][17] all put forward specific ways which use zero-knowledge proof agent agreement to realize the Proxy Digital Signature.The literature [15] is based on the problems of computing discrete logarithm.By using zero-knowledge digital signatures ,it puts forward not only a proxy signature scheme ,but also a multiple proxy signature scheme .These two scheme can effectively prevent the original signer from faking the proxy signer and forging proxy signature key for proxy signature .The literature [16] puts forward a zero-knowledge digital signature scheme based on RSA，but the zero-knowledge proof agreement in literature [16]continues using the first Fiat - Shamir identity authentication protocol process.In literature [17],it indicates that by using the zero-knowledge proof thoughts ,we can prevent others from faking the infomation owner to ask the proxy signers for the signature.In the following fig.4 ,it shows the algorithm flowchart of the realization of proxy signature by using the zero-knowledge proof.

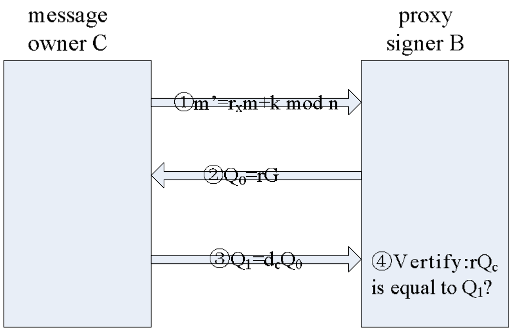


Fig.4: Zero-knowledge proof protocol applied in proxy digital signature

The principle of the algorithm is also calculated based on the problems of the elliptic curve discrete logarithm. The

parameters related to what mentioned above are as follows :
①Let Fq be a finite field with Q elements, E is the elliptic curve ;
② Set G as a base, n is the order of G;
③ k is a random integer in the range of [1, n - 1], R = kG = (rx, ry), m '= RxM + k mod n, m R is the information of the signature ;
④ r is a random integer in the range of [1, n - 1] , Q0=rG ;
⑤dc is the private key of the message owner C， Qc is the public key of C , Qc=dcG.

3.6 Identity Authentication

If “zero-knowledge” means identity information, then zero-knowledge proof protocol can be applied to identity authentication.The common authentication requires transmission of password or personal secret information ,which will give attackers loopholes to attack more or less[18][19]. With zero-knowledge proof of identity , one can prove himself legal to the system without transfering above information .There are large amounts of documents show that zero-knowledge proof protocol has great superiority and importance in the field of authentication presently[20][21][22][23]. Fiat-Shamir authentication is the first that had been proposed, while it is the most basic zero-knowledge authentication scheme. The process of Fiat-Shamir authentication is shown in fig.5:

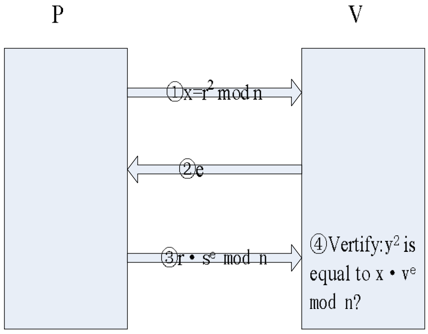


Fig.5: Fiat-Shamir authentication process

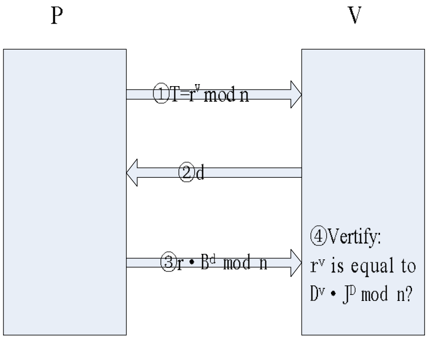


Fig.6: Guillou-Quisquater authentication process

The introduction of the parameters about Fiat-Shamir authentication process :
① n= $p \times q$,n is a random modulus , p and q are two large primes.
② s is the private key of the prover P . v is the public key of P . s and n are coprime. $v = s^2 \bmod n$
③ r, a committed random number,is a random integer,which∈[1,n-1].
④ e∈{0，1},a challenging bit.
In addition, we give the Guillou-Quisquater authentication, a classic authentication process. The process of the Guillou-Quisquater authentication is shown in figure 6:
The introduction of related parameters :
①J，v，n are public keys , n= $p \times q$, p and q are two large primes.
②B is the private key ,which meets $J \times B^v = 1 \bmod n$.
③ r, a committed random number,is a random integer,which∈[1,n-1].
④ d∈[0,v-1], a challenging bit .

4 Conclusions and prospects

zero-knowledge proof protocol has become a very important component in cryptographic algorithms and security protocols.In this paper,the main idea,nature, general proof process, mathematical theory and specific applications of zero-knowledge proof protocol are introduced.Zero-knowledge proof protocol has the advantage of zero leakage proof, so it can be applied to prove many key issues,like many classic mathematical problem: the polynomial function roots, the graph isomorphism,as well as other NP problem[24], such as the Sudoku games [25]. The prover can prove that he has the method to solve some problem and he does not worry about the method revealed. Zero-knowledge proof protocol are very useful in the field of network and information security too,like authentication, digital signatures, etc.It is very important that proving to each other the identify of the user without revealing the user information in authentication and digital signatures.In order to effectively prevent unauthorized users impersonating legitimate users,we can use zero-knowledge proof protocol to authenticate.

Research on a large number of documents, we can see that during the process of zero-knowledge proving,it must meet its three nature including completeness,rationality and zero-knowledge and the procedure must follow the steps of Fig. 2.

In our future study,we will focus on using zero-knowledge proof theory to solve more problems and exploring its broader applications.In this paper,for many applications,some simple proof algorithm and process is given, the calculation of mathematical formula and the details of the proving procedure are not given explanation in-depth,the complexity of communication and computation is not considered yet.All those will be the content of our next task.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under Grant(No. 60973140, 61170276), the National High Technology Research and Development Program of China (863 Program) (No. 2009AA01Z212),the Natural Science Foundation of Jiangsu Province under Grant(No.BK2009425),the construction project of Jiangsu Province advantage discipline “ Information and Communicaiton Engineering”,the research project of Nanjing University of Posts and Telecommunications(No. NY210034).

References

[1]Lindell,Y;Zarosim,H.Adaptive Zero-knowledge Proofs and Adaptively Secure Oblivious Transfer.[J].Journal of Cryptology.24(4),pp.761-799,2011

[2]Garg,Sanjam;Jain,Abhishek;Sahai,Amit.Leakage-resilient zero knowledge.31st Annual International Cryptology Conference,CRYPTO 2011.pp:297-315.

[3]Lin,Huijia;Pass,Rafael;Tseng,Wei-Lung Dustin;Venkitasubramaniam,Muthruamakrishnan.Concurrent non-malleable zero knowledge proofs. 30th Annual International Cryptology Conference,CRYPTO 2010.pp:429-446.

[4]Bayer,Stephanie;Groth,Jens.Efficient zero-knowledge argument for correctness of a shuffle.31st annual International Conference on the Theory and Applications of Cryptographic Techniques,EUROCRYPT 2012.pp:263-280.

[5] S.Goldwasser,S.Micali,C.Rackoff.The Knowledge Complexity of interactive Proof Systems.Proceddings of the 17th ACM Symposium on Theory of Computing 1985,291-304.

[6] http://baike.baidu.com/view/1228083.htm.

[7]Zhang Yinbin. Research on Zero-Knowledge Proof and Its Applications[D].China:Huaibei Normal University,2011

[8]Li Xi,Wang Daoshun.Zero-knowledge proof protocol of the roots of polynomial functions[J].Journal of Tsinghua University(Science and Technology),2009,Vol.49,No.7,pp:999-1002.

[9] Goldreich O,Micali S,Wigderson A.Proofs that yield nothing but their validity and a methodology of cryptographic protocol design [J].FOCS,1986.174-187

[10Juels A,Kaliski-Jr B S.Pors:Proofs of retrievability for large files.In:Proceedings of the 2007 ACM Conference on Computer and Commnunications Security,CCS 2007.Alexandria:ACM,2007.584-597.

[11] Zhu y;Wang HX;Hu ZX;Ahn,GJ;Hu,HX.Zero-knowledge proofs of retrievability.Science China-Information Sciences.54(8),pp.1608-1617,2011.

[12]Huang,YJ;Lin,WC;Li,HL.Efficient Implementation of RFID Mutual Authentication Protocol.IEEE Transactions on Industrial Electronics.59(12),pp.4784-4791.2012.

[13] Liu,H;Ning,HS.Zero-Knowledge Authentication Protocol Based on Alternatiove Mode in RFID Systems.IEEE Sensors Journal.11(12),pp.3235-3245,2011

[14]Wang Xiao-mei,Zhang Qiu-jian.Mutual Authenticantion for RFID based on elliptic curve and zero knowledge[J].Computer Engineering and Applications.pp:1-4,2012.

[15]Tan Zuo-wen,Liu Zhuo-jun.Proxy Signature Schemes Based on Signature of Zero-Knowledge[J].Computer Science.Vol.31,No.11,pp:70-72,2004.

[16] Qi,CM;Cui,SM.A Zero-Knowledge Proof of the RSA Digital Signature Scheme.1st International Symposium on Computer Network and Multimedia Technology.Vol(1and 2),pp.1037-1040,2009

[17]Zhang Jian-zhong,Ma Wei-fang.Blind Proxy Blind Signature Scheme on Elliptic Curve[J].Vol.36,No.11,pp:126-127,2010.

[18]Upadhyay,Saurabh;Singh,Sanjay Kumar.Video Authentication:Issues and Chanllenges[J].International Journal of Computer Science Issues.Vol.9,No.1-3,pp:409-418,2012.

[19]Malempati,Sreelatha,Mogalla,Shashi.Enhanced authentication schemes for instruction prevention using native language passwords[J]. International Journal of Computer Science Issues.Vol.8,No.4-1,pp:356-362,2011.

[20] Jaafar,Abdullah M;Samsudin,Azman.Visual zero-knowledge proof of identity scheme:A new approach.2nd International Conference on Computer Research and Development.pp.205-212,2010.

[21] Naranjo,JAM;Antequera,N;Casado,LG;Lopez-Ramos,JA.A suite of algorithms for key distribution and authentication in centralized secure multicast environments.[J].Journal of Computaional and Applied Mathematics,236(12),pp.3042-3051,2012

[22] Camenisch,J;Gro,T.Efficient Attributtes for Anonymous Credentials.ACM Transactions on Information and System Security.15(1),SI(4),2012.

[23] Chang Qing,Zhao Fang.Research of Authentication System Based on Zero-knowledge Proof[J].Value Engineering,pp:167,2010.

[24] Minh-Huyen Nguyen;PSalil Vadhan.Zero knowledge with efficient provers.Poceedigns of the thirty-eighth annual ACM symposium on Theory of computing.pp,287-295,2006

[25] Chien,Yu-Feng; Hon,Wing-Kai.Cryptographic and Physical Zero-Knowledge Proof:From Sudoku to Nonogram.5th International Conference on Fun with Algorithms.Vol(6099), pp:102-112,2010.

Wang Huqing received master degree from Nanjing University of Posts and Telecommunications in 2004.She has been worked in her home university since 2004. Now,she is pursuing the PHD degree in the field of network security under the guidance of Dr.Sun Zhixin in Nanjing University of Aeronautics and Astronautics

Sun Zhixin has around twenty years both teaching and research experience.He is the leader of College of Internet of Things in Nanjing University of Posts and Telecommunications and he is the editor of many journal like computer communication,Journal on Communication ,Journal of Software and so on.

Representation of Knowledge patterns for Semantic Web

Rostislav Miarka[1]

[1] **Department of Informatics and Computers, Faculty of Science, University of Ostrava**
30. dubna 22, 701 03 Ostrava, Czech Republic

Abstract

This article presents representation of knowledge patterns in RDF language. The approach to knowledge representation can be used in Semantic Web as a tool of finding some additional RDF assertions in the frame of linked data. The article introduces the term "knowledge pattern" and dividing the knowledge patterns into two groups – to Top-level knowledge patterns and to Domain knowledge patterns. Examples of top-level knowledge patterns for sentences in the English language are also a part of this article.

Keywords: *Knowledge patterns, RDF graph language, Semantic web, English sentence.*

1. Introduction

The term "Semantic web" refers to the vision of Web of linked data. Nowadays, the Semantic web is also called Web of Data, which is enabled by Semantic web technologies. Among the Semantic web technologies are RDF, OWL, SKOS and SPARQL. The main goal of Linked data is to allow users to share structured data in the web environment. The term "Linked data" was introduced by Tim Berners-Lee in his talk named Linked Data Web Architecture. This term represents the way of publishing and linking of structured data on the web.

Linked data uses RDF (Resource Description Framework) in two ways. The first way is the use of RDF data model to publish structured data on the web. The second way is the use of RDF links to link data sources.

2. Knowledge patterns

Knowledge patterns were first introduced in [1] and they are closely related to the creation of ontologies or knowledge bases. The term knowledge pattern represents a general structure (pattern) of knowledge. The pattern itself is not a part of the target knowledge base or ontology. While using a knowledge pattern, the general terms from pattern are renamed to special terms from the modeled domain. Renaming of terms is called morphism. The main purpose of knowledge patterns is a reuse of knowledge.

We propose to divide knowledge patterns to two groups – Top-level knowledge patterns and Domain knowledge patterns.

2.1 Top-level knowledge patterns

The original idea of knowledge patterns was based on the fact that the pattern itself is not a part of the target knowledge base and it is integrated to the knowledge base by renaming its symbols. Knowledge patterns which correspond to this description are called Top-level knowledge patterns and they form the first group of knowledge patterns. Such knowledge patterns can be used in any knowledge base or ontology.

2.2 Domain knowledge patterns

The second group of knowledge patterns is called Domain knowledge patterns. Nowadays, many ontologies, vocabularies or knowledge bases are created by web ontological languages. Languages RDF+RDFS and OWL are the most important web ontological languages. In these languages, the ontology contains definition of classes and relation among them. Relations are also called properties. The OWL language enables using two types of properties – object properties and datatype properties. Object property defines relation between two classes; datatype property defines attributes of a class. The attribute is represented by a data type, such as string, integer or date.

The particular structure of classes and properties creates knowledge pattern on lower level of abstraction. While using this type of knowledge patterns, instances of classes (individuals) and instances of properties are created. These instances represent concrete objects from the modeled domain. This type of knowledge pattern is called Domain knowledge pattern.

We propose to represent knowledge patterns in RDF+RDFS[1] languages.

[1] In the following, RDF(S) will be used as an abbreviation for RDF+RDFS.

3. Representation of knowledge patterns in RDF(S)

The basic building block of RDF(S) is an RDF triple, which represents one statement. The statement is in form

subject – predicate – object.

The subject represents a term which is described in the statement. The predicate represents property assigned to the subject. The object represents the value of this property. It can be other term or a literal value (string, integer, date). Terms and properties in RDF(S) are identified by URI (Uniform Resource Identifier) references. RDF(S) contains only binary relations, more complex relations must be decomposed to a set of binary relations. The RDF(S) language can be represented in two forms – as a graph and in a text form.

The RDF graph language is used to create graph representation of RDF. Ontology or knowledge base in the RDF graph language is represented by a directed graph. Nodes in this graph represent subjects and objects of statements, arcs represent relations (predicates). The basic building block of an RDF graph is an RDF triple, which represent one statement. The statement is in form

subject – predicate – object.

Its graph representation is shown in next figure.

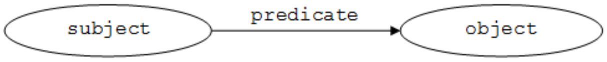


Figure 1. RDF triple.

Text form of RDF is called RDF serialization. It can have more forms. Among these forms is RDF/XML, N3 notation, N-triples, RDFa. Serialization called RDF/XML is the mostly used type of serialization. It is based on the XML language.
For representation of knowledge patterns, we propose to use the RDF(S) language [6], [7]. In the case of graph representation, it is used the RDF graph language which is extended by quantifiers and enables to state negation [4]. A classical RDF graph uses only solid lines. To distinguish the knowledge patterns from the classical RDF statements, we use dashed lines. The example of a classical RDF triple and a triple which represents knowledge pattern is shown in Figure 2.

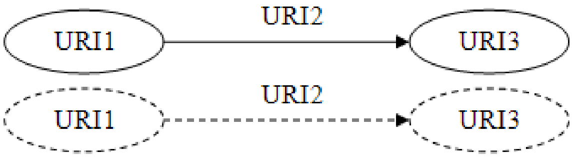


Figure 2. Classical RDF triple (above) and RDF triple representing knowledge pattern (below).

The important part of using knowledge patterns is morphism. By the help of morphism, the general terms from the pattern are mapped to special terms from the problem domain. Morphism is represented by the relation of specialization – the general term is renamed to a special term from the domain of interest. Mapping of one term will be represented by the one classical RDF triple (with a solid line). The subject of this triple will be a special term from the domain of interest, the predicate will be "isa" and the object will be the general term from the pattern. The example of morphism of one term is shown in Figure 3.

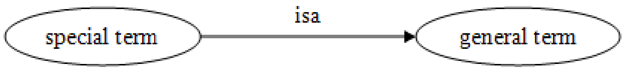


Figure 3. Morphism.

Representation of knowledge patterns in a text form will be described by the RDF/XML serialization. Vocabulary for description of knowledge patterns was defined. The schema of this vocabulary is shown in next figure.

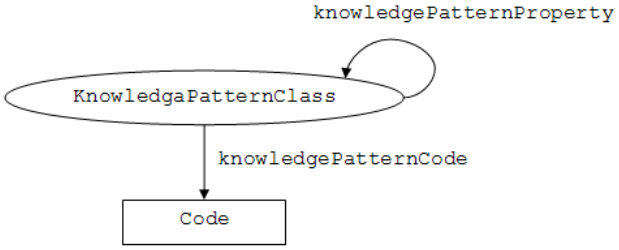


Figure 4. Schema of vocabulary for knowledge patterns description.

The class called *KnowledgePatternClass* represents one class (i.e. one term) from knowledge pattern. It is an abstract term which represents any class. All classes which belong to knowledge pattern are subclasses of *KnowledgePatternClass*. Classes in the knowledge pattern are connected by properties. An abstract property which connects two classes in the pattern is called *knowledgePatternProperty*. All properties which belong to knowledge pattern are subproperties of this property. The property *knowledgePatternCode* assigns the code of the knowledge pattern (literal value) to all classes form the pattern. The code of patterns is its identifier. While using concrete knowledge patters, morphism of terms is defined by the relation *isa*. This relation defines mapping of one special term to the general term.

The vocabulary for description knowledge patterns is available online: http://www.r-miarka.net/kp.rdf.

4. Examples of knowledge patterns

In the rest of the article, there will be presented examples of knowledge patterns, concretely of top-level knowledge patterns. There will be examples of knowledge patterns for conversion of sentences from the English language to RDF(S).

Words of a natural language form vocabulary of this language. Words are divided to word classes, such as noun, pronoun, verb etc. Grammar of a language determines the way of constructing a sentence. Each language has its own grammar. Grammar of a language does not work with work classes; it works with parts of sentences, which are also called constituents of sentences. There are two basic parts of sentence – the subject and the predicate. These two parts are to be found in almost each sentence. For example, imperative sentences (“Run!”, “Stop!” etc.) belong to exceptions of this rule. Apart from subject and predicate, there are additional parts, which extend the sentence – object, attribute and adverbial complement (adverb).

For the process of construction of sentence, it is important, which part the particular word represents. The term word order is closely related to this process. Word order of language determines the order of parts in a sentence. There exist two basic types of word order – fixed word order and free word order. Fixed word order has relatively strict rules for ordering parts in sentence. This type of word order is used by Germanic languages, e.g. English. Free word order allows changing the order of parts in sentence, according to the actual context. This type of word order is typical for languages which enable declension and inflexion. Among these languages is Czech language.

The essential parts of each sentence are subject and predicate. To mark the basic types of word orders, we use three parts – subject, predicate and object. All three parts are marked with a letter. Subject is marked by the letter S, predicate is marked by the V (verb) and object is marked by O. Combination of these three basic parts determines the type of word order [2], [3]. There are six basic types of word order: SVO (subject verb object), SOV (subject object verb), VOS (verb object subject), VSO (verb subject object), OSV (object subject verb) and OVS (object verb subject). Languages using the SVO word order include English, Romance languages, Bulgarian and Chinese. Languages using the SOV word order include Japanese, Mongolian, Turkish and Korean. Word order marked as VSO is used by the following languages: Classical Arabic, Insular Celtic languages and Hawaiian. Word order VOS is used in the Fijian language (Fiji) and the Malagasy language (Madagascar). The OSV word order is used by languages Xavante (Brazil) and Warao (Venezuela). The OVS word order is used by the Hixkaryana language (Brazil).

4.1 Structure of English sentence

The English language uses a fixed word order marked as SVO (subject verb object). The letter O in this abbreviation marks a direct object. The basic word order can be a little different because in the sentence can appear an auxiliary verb (will, can, would etc.) or indirect object. Sentences can express different things. It can state a fact – declarative sentences (positive or negative). It can ask about some things – questions. It can give commands – imperative sentences. Because this article is focused on knowledge patterns, which are closely related to ontologies or knowledge bases, the only type of a sentence which is important is declarative sentence. Questions or imperative sentences are not a part of any knowledge base or ontology.

Declarative sentences can state positive facts or negative facts. The structure of both positive and negative sentence is quite similar in the English language. A sentence in a negative form contains auxiliary words “do not” in addition.

The simplest type of positive declarative sentence is a bare sentence, which contains only a subject and predicate. Examples of this type of a sentence are “It snows.”, “I hope.” etc.

A more complex type of sentence is that which contains an object in addition. This type of a sentence is in form SVO (subject verb object). Examples of this type of a sentence are “I know Michael.”, “I like oranges.”, etc. Another type of a sentence contains an adverbial complement in addition. It can be an adverbial complement of manner, place or time. A sentence can contain one of these complements or a combination of them. A sentence containing all three types of adverbial complements has its word order marked as SVOMPT (subject verb object manner place time). Examples of this type of a sentence are “I play the piano daily.”, “I speak English very well.” etc. Another type of a sentence contains an attribute in addition. The attribute extends the subject or the object in the sentence. Examples of attributes are sizes, colors, i.e. “small”, “big”, “red”, “blue”, etc.

Table 1 Shortcuts and full URIs

Shortcuts	URI
ISA	http://en.wikipedia.org/wiki/Is-a
subject	http://en.wikipedia.org/wiki/Subject_(grammar)
predicate	http://en.wikipedia.org/wiki/Predicate_(grammar)
object	http://en.wikipedia.org/wiki/Object_(grammar)
attribute	http://en.wikipedia.org/wiki/Is-a
Prince Charles	http://en.wikipedia.org/wiki/Charles,_Prince_of_Wales
Elizabeth II	http://en.wikipedia.org/wiki/Elizabeth_II
foaf:knows	http://xmlns.com/foaf/0.1/knows
is	http://en.wikipedia.org/wiki/Is
young	http://en.wiktionary.org/wiki/young
Prince William	http://en.wikipedia.org/wiki/Prince_William
likes	http://en.wikipedia.org/wiki//Like
football	http://en.wikipedia.org/wiki/Football_(soccer)

Types of a negative declarative sentence are the same as for a positive declarative sentence. Negation in a sentence is related to a verb and it changes the verb to its opposite meaning. Examples of this type of a sentence are “I do not know John.”, “Michael does not like ice-cream.”, etc.

Identification of nodes and edges in an RDF graph is realized by the help of URIs, which can be quite long. RDF graph with full URIs would be confusing to the users. It is possible to use shortcuts for full URIs. In the following RDF graphs will be used shortcuts which are shown in Table 1.

4.2 Knowledge pattern for sentence in basic form

The first knowledge pattern presented here is knowledge pattern for conversion of a sentence in the basic form to RDF. This type of a sentence contains three parts – subject, predicate and object. In English, the order of parts is as follows – subject – verb – object. This sentence forms one RDF triple. The figure 5 shows this triple as a knowledge pattern. In the following, this pattern will be marked as KPS01.

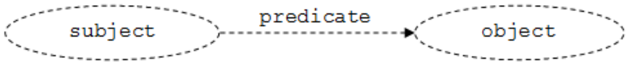


Figure 5. KPS01.

Let us consider that the root element of all following descriptions in RDF/XML will be the same. Its form will be as follows:

```
<rdf:RDF
   xmlns="http://www.r-miarka.net/kp.rdf#"
   xml:base="http://www.r-miarka.net/kp.rdf"
   xmlns:foaf="http://xmlns.com/foaf/0.1/"
   xmlns:wikien="http://en.wikipedia.org/
      wiki/"
   xmlns:rdfs="http://www.w3.org/2000/01/
      rdf-schema#"
   xmlns:rdf="http://www.w3.org/1999/02/
      22-rdf-syntax-ns#">
...
</rdf:RDF>
```

The representation of KPS01 in RDF/XML, which uses the vocabulary kp.rdf, is shown below.

```
<rdfs:Class rdf:about=
      "http://en.wikipedia.org/wiki/
      Subject_(grammar)">
   <rdfs:label xml:lang="en">Subject
      </rdfs:label>
   <rdfs:comment xml:lang="en">Subject in
      a sentence</rdfs:comment>
   <rdfs:subClassOf rdf:resource=
      "http://www.r-miarka.net/kp.rdf#
      KnowledgePatternClass"/>
   <kp:knowledgePatternCode>KPS01
      </kp:knowledgePatternCode>
   <wikien:Predicate_(grammar) rdf:resource=
      "http://en.wikipedia.org/wiki/
      Object_(grammar)"/>
</rdfs:Class>

<rdfs:Class rdf:about=
      "http://en.wikipedia.org/wiki/
      Object_(grammar)">
   <rdfs:label xml:lang="en">Object
      </rdfs:label>
   <rdfs:comment xml:lang="en">Object in
      a sentence</rdfs:comment>
   <rdfs:subClassOf rdf:resource=
      "http://www.r-miarka.net/kp.rdf
      #KnowledgePatternClass"/>
   <kp:knowledgePatternCode>KPS01
      </kp:knowledgePatternCode>
</rdfs:Class>

<rdf:Property rdf:about=
      "http://en.wikipedia.org/wiki/
      Predicate_(grammar)">
   <rdfs:label xml:lang="en">Predicate
      </rdfs:label>
   <rdfs:comment xml:lang="en">Predicate in
      a sentence</rdfs:comment>
   <rdfs:subPropertyOf rdf:resource=
      "http://www.r-miarka.net/kp.rdf#
      knowledgePatternProperty"/>
   <rdfs:domain rdf:resource=
      "http://en.wikipedia.org/wiki/
      Subject_(grammar)" />
   <rdfs:range rdf:resource=
      "http://en.wikipedia.org/wiki/
      Object_(grammar)" />
</rdf:Property>
```

The example of using knowledge pattern KPS01 is the sentence:

Prince Charles knows Elizabeth II.

Prince Charles is the subject of the sentence, Elizabeth II is the object and the predicate (relation) is knows. To represent of the predicate, property “knows” from vocabulary FOAF will be used. Morphism for terms from this sentence in a graph form is shown in next figure.

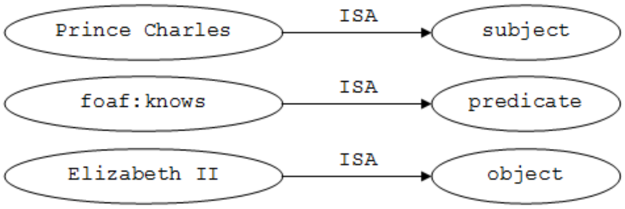


Figure 6. KPS01 - morphism.

After renaming the terms, the resulting RDF graph will look like that in next figure. This graph contains only one RDF triple.

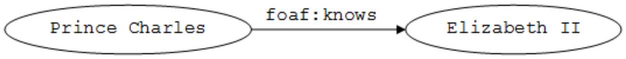


Figure 7. KPS01 - result.

Representation of pattern KPS02 in RDF/XML is shown below.

```
<rdf:Description rdf:about=
      "http://en.wikipedia.org/wiki/
      Prince_Charles">
   <kp:isa rdf:resource=
      "http://en.wikipedia.org/wiki/
      Subject_(grammar)" />
</rdf:Description>

<rdf:Description rdf:about=
      "http://xmlns.com/foaf/0.1/knows">
   <kp:isa rdf:resource=
      "http://en.wikipedia.org/wiki/
      Predicate_(grammar)" />
</rdf:Description>

<rdf:Description rdf:about=
      "http://en.wikipedia.org/wiki/
      Queen_Elizabeth_II">
   <kp:isa rdf:resource=
      "http://en.wikipedia.org/wiki/
      Object_(grammar)" />
</rdf:Description>
```

The result in RDF/XML contains one RDF triple:

```
<rdf:Description rdf:about=
      "http://en.wikipedia.org/wiki/
      Prince_Charles">
   <foaf:knows rdf:resource=
      "http://en.wikipedia.org/wiki/
      Queen_Elizabeth_II" />
</rdf:Description>
```

4.3 Knowledge pattern for sentence with attribute by subject

The second example of knowledge pattern is one for a sentence in the basic form which contains an attribute in addition. An attribute can extend the subject or the object in a sentence. An attribute can be represented by color (red, blue etc.) or size (small, big etc.). In this case, it will represent a sentence with an attribute by the subject. A graph representation of this knowledge pattern is shown in next figure.

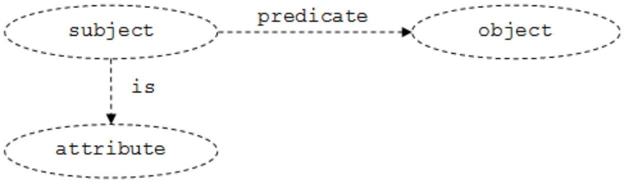


Figure 8. KPS02.

The RDF graph in the figure contains two RDF triples. The first triple represents a sentence in the basic form (see KPS01), the second triple represents an extension of the subject by an attribute. The predicate in this triple is general relation “is”. According to the type of attribute, it can be different. For color it could be “hasColor”, for size “hasSize”, etc. The node which represents the subject connects these two triples to the resulting RDF graph. For representation of nodes in the RDF graph are used URI identifiers. The node with a label “subject” has the same URI, so it can be displayed only once in the resulting RDF graph.

Representation of pattern KPS02 in RDF/XML is shown below.

```
<rdfs:Class rdf:about=
     "http://en.wikipedia.org/wiki/
     Subject_(grammar)">
  <rdfs:label xml:lang="en">Subject
     </rdfs:label>
  <rdfs:comment xml:lang="en">Subject in
     a sentence</rdfs:comment>
  <rdfs:subClassOf rdf:resource=
     "http://www.r-miarka.net/kp.rdf#
     KnowledgePatternClass"/>
  <kp:knowledgePatternCode>KPS02
     </kp:knowledgePatternCode>
  <wikien:Predicate_(grammar) rdf:resource=
     "http://en.wikipedia.org/
     wiki/Object_(grammar)"/>
  <wikien:Is rdf:resource=
     "http://en.wikipedia.org/wiki/
     Attribute"/>
</rdfs:Class>

<rdfs:Class rdf:about=
     "http://en.wikipedia.org/wiki/
     Object_(grammar)">
  <rdfs:label xml:lang="en">Object
     </rdfs:label>
  <rdfs:comment xml:lang="en">Object in
     a sentence</rdfs:comment>
  <rdfs:subClassOf rdf:resource=
     "http://www.r-miarka.net/kp.rdf#
     KnowledgePatternClass"/>
  <kp:knowledgePatternCode>KPS02
     </kp:knowledgePatternCode>
</rdfs:Class>

<rdf:Property rdf:about=
     "http://en.wikipedia.org/wiki/
     Predicate_(grammar)">
  <rdfs:label xml:lang="en">Predicate
     </rdfs:label>
  <rdfs:comment xml:lang="en">Predicate in
     a sentence</rdfs:comment>
  <rdfs:subPropertyOf rdf:resource=
     "http://www.r-miarka.net/kp.rdf#
     knowledgePatternProperty"/>
  <rdfs:domain rdf:resource=
     "http://en.wikipedia.org/wiki/
     Subject_(grammar)" />
  <rdfs:range rdf:resource=
     "http://en.wikipedia.org/wiki/
     Object_(grammar)" />
</rdf:Property>

<rdfs:Class rdf:about=
     "http://en.wikipedia.org/wiki/
     Attribute">
  <rdfs:label xml:lang="en">Attribute
     </rdfs:label>
  <rdfs:comment xml:lang="en">Attribute by
     a subject</rdfs:comment>
  <rdfs:subClassOf rdf:resource=
     "http://www.r-miarka.net/kp.rdf#
     KnowledgePatternClass"/>
  <kp:knowledgePatternCode>KPS02
     </kp:knowledgePatternCode>
</rdfs:Class>

<rdf:Property rdf:about=
     "http://en.wikipedia.org/wiki/Is">
  <rdfs:label xml:lang="en">Is</rdfs:label>
  <rdfs:comment xml:lang="en">
     Verb is</rdfs:comment>
  <rdfs:subPropertyOf rdf:resource=
     "http://www.r-miarka.net/kp.rdf#
     knowledgePatternProperty"/>
  <rdfs:domain rdf:resource=
     "http://en.wikipedia.org/wiki/
     Subject_(grammar)" />
  <rdfs:range rdf:resource=
     "http://en.wikipedia.org/wiki/
     Attribute" />
</rdf:Property>
```

An example of use of this knowledge pattern is sentence

Young Prince William likes football.

An attribute of this sentence is “young”. Morphism for this sentence in a graph form is shown in next figure.

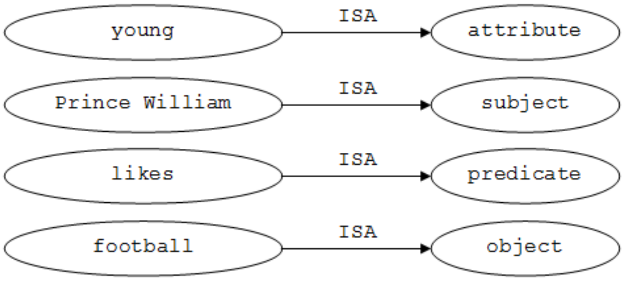


Figure 9. KPS02 - morphism.

Result after renaming the terms is shown in figure 10.

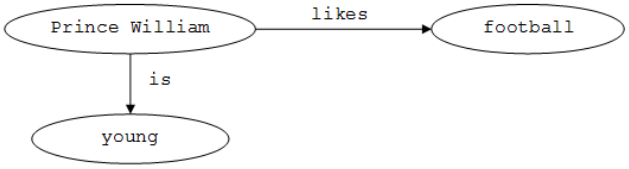


Figure 10. KPS02 - result.

The morphism for the given sentence in RDF/XML will look as follows:

```
<rdf:Description rdf:about=
     "http://en.wiktionary.org/wiki/young">
  <kp:isa rdf:resource=
     "http://en.wikipedia.org/wiki/
     Attribute" />
</rdf:Description>
```

```
<rdf:Description rdf:about=
     "http://en.wikipedia.org/wiki/
     Prince_William">
  <kp:isa rdf:resource=
     "http://en.wikipedia.org/wiki/
     Subject_(grammar)" />
</rdf:Description>

<rdf:Description rdf:about=
     "http://en.wikipedia.org/wiki/Like">
  <kp:isa rdf:resource=
     "http://en.wikipedia.org/wiki/
     Predicate_(grammar)" />
</rdf:Description>

<rdf:Description rdf:about=
     "http://en.wikipedia.org/wiki/
     Football_(soccer)">
  <kp:isa rdf:resource=
     "http://en.wikipedia.org/wiki/
     Object_(grammar)" />
</rdf:Description>
```

Result after using morphism is shown further. Both triples are connected together.

```
<rdf:Description rdf:about=
     "http://en.wikipedia.org/wiki/
     Prince_William">
  <wikien:Like rdf:resource=
     "http://en.wikipedia.org/wiki/
     Football_(soccer)"/>
  <wikien:Is rdf:resource=

  "http://en.wiktionary.org/wiki/young"/>
</rdf:Description>
```

5. Conclusions

Knowledge patterns can help with reusing knowledge. Knowledge patterns can be divided to two groups – Top-level knowledge patterns and Domain knowledge patterns. This article contains the proposal of representation of knowledge patterns in RDF(S) – in a graph form (RDF graph) and in a text form (RDF/XML). The usage of representation in RDF(S) enables to use knowledge patterns in the Semantic web (also called Web of data or Data web). Two examples of knowledge patterns and their use are a part of this article. The future work will be aimed at finding other Top-level knowledge patterns and at finding Domain knowledge patterns.

References

[1] P. Clark, J. Thompson and B. Porter, “Knowledge patterns,” In *Handbook on Ontologies*, S. Staab and R. Studer, Eds. Berlin: Springer-Verlag, 2004, ISBN 3-540-40834-7, pp. 191–207.

[2] D. Crystal, “The Cambridge Encyclopedia of Language (2nd edition ed.)”, Cambridge: Cambridge University Press, 1997, ISBN 0-521-55967-7.

[3] M. S. Dryer, “Order of Subject, Object, and Verb“, http://linguistics.buffalo.edu/people/faculty/dryer/dryer/DryerWalsSOVNoMap.pdf

[4] A. Lukasová, M. Vajgl and M. Žáček, “Reasoning in RDF graphic formal system with quantifier,” *Proceedings of the International Multiconference on Computer Science and Information Technology*, 2010, ISBN 978-83-60810-22-4, pp. 67–72.

[5] S. Staab, M. Erdmann, A. Maedche and S. Decker “Ontologies in RDF(S),” [online] http://www.ida.liu.se/ext/etai/received/semaweb/010/paper.pdf.

[6] W3C: Resource Description Framework (RDF): Concepts and Abstract Syntax, http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

[7] W3C: RDF Primer, http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

Rostislav Miarka graduated fromt the University of Ostrava (Czech Republic). The focus of his dissertation is a formal representation of knowledge patterns in ontological languages. The topic of the PhD thesis corresponds to the content of the article. He is an author more than 10 scientific publications.

Noise analysis and column FPN suppression technology

Xiao-fen JIA, Bai-ting Zhao

School of Electrical and Information Engineering, Anhui University of Science & Technology
Huainan 232001, China

Abstract

Noise is an important indicator that affects the image quality. Among the different processing strategies implemented in image sensors, the effect on the noise is complicated. The two types of noise that influences the imaging quality of the digital camera are analyzed firstly. The optical signal and reset signal of the typical correlated double sampling (CDS) are read out through two column amplifiers (CA), respectively. The performance of the two column amplifiers are not exactly the same, resulting in itself will bring the column FPN. To suppress the column FPN effectively, a single amplifier CDS scheme is proposed, which have only one CA, the power consumption and area of the proposed CDS circuit is reduced by half.

Keywords: *Column FPN, Correlated double sampling (CDS), Column amplifier (CA), Denoising.*

1 Introduction

In recent years, rapid research and development have helped the digital imaging device’s application more and more widely. The high-quality image acquisition referred to as the focus of attention. Among the different processing strategies implemented in image sensors such as motion detection, feature extraction and adaptive dynamic range enhancement, the effect on the noise are complicated. However, in modern imaging systems, acquired images pass through many stages of digital processing, which can introduce different noise. Acquired images form image sensors are disturbed by kinds of noise interference [1]. Image sensors are characterized as active pixel sensors and passive pixel sensors depending on the readout circuit [2], they incorporate an amplifier in each pixel and amplifying the collected charge outside the array, respectively. However, a variation caused by mismatch in the cell circuit or in the column readout circuits mean that the sensors suffers from fixed pattern noise that degrades the quality of the resulting image. Therefore denoising is particularly important.

In section 2, different noise caused by many stages, such as processing strategies implemented in image sensors and different stages of image processing is discussed. The correlated double sampling circuit with single column amplifier is proposed in section 3. Finally, conclusion is given in section 4.

2 Noise analysis

The imaging process of digital camera image includes two steps, that is, image sampling and image processing. The optical signal collected from optical lens is converted into electric signal through the image sensor in the image sampling process. Through the analog front-ends: the correlated double sampling (CDS), the programmable gain amplifier (PGA) and the digital-to-analog (ADC), the electric signal can be converted into digital signal, that is, mosaic image. The acquired mosaic image is then processed by image processing, which are demosaicing, denoising, enhancing, color correction and so on. The algorithms used in the two stages are complex, several processes are nonlinearity and the effect of different process on the noise is complicated. However, the noise that affecting the image imaging quality consist of two parts, which is caused in the process of the image sampling and the image processing, respectively. There are denoising 1 and denoising 2 of Fig 1.

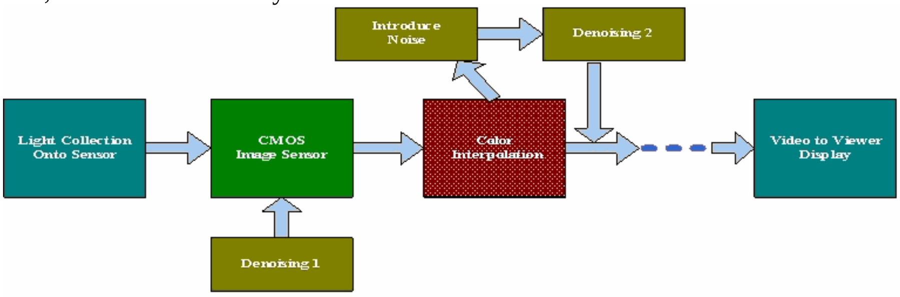


Fig.1 Imaging flow chart

Denoising 1 is processed the noise caused by the image sensor. The image sensor noise can be divided into two categories, the pattern noise and the random noise. The random noise is obvious in the case of low light, which includes thermal noise, KTC noise, dark current shot noise and 1/f noise. The pattern noises include two types, that is, fixed pattern noise (FPN) and photo response non uniformity (PRNU). The latter is related to the illumination pattern noise component, which is independent of time, associated with the signal. Fixed pattern noise is one of the primary limitations to image quality in image sensors, caused by the mismatch between individual pixels or columns of the image sensor, and had much greater effect than the random noise. It is spatial in nature and ideally does not change with time for a particular imager. Two types of FPN have been reported, namely, pixel FPN, which is caused by a mismatch in the cell circuit, and column FPN, caused by a mismatch in the column readout circuits. The main problem of column FPN in an image is not its actual magnitude, but its perceptual effect observed by the human visual system, column FPN can be observed as vertical stripes in an image that are visible even if the magnitude of the column FPN is much lower than the pixel FPN, that is, the column FPN is much more visible than pixel FPN [3, 4]. Although it is hard to quantify the perceptual difference between pixel and column FPN, it has been proposed [5] that random column FPN is five times more harmful to the perceived image quality than pixel FPN. So the column FPN degrades the quality of resulting image, we should reduce it endeavor. Snoeij et al [6] proposed a dynamic column FPN reduction technique, which make the initial column FPN of 0.69% of full scale is made nearly invisible in the measured images.

The target of denoising 2 is the noise caused by the step of image processing, which include the change of the noise structure caused by the demosaicing process and the noise caused by the order of demosaicing and denoising. The influence of demosaicing algorithms on the effects of denoising has been discussed [7]. In the case of noise exist, the demosaicking process will sharpening zoom high-frequency noise, causing noise pollution, changing the structure of the input noise, and forming pseudo-edge, caused the noise analysis becomes very complicated.

3 Correlated Double Sampling Circuit

Nixon et al [7] proposed a typical double sampling circuit to reduce the fixed pattern noise. The pixel FPN is reduced by reading out the pixel and reset signals to the column amplifier (CA). The column FPN is eliminated by reading out the output signal of the column amplifier and the compensation signal to the output column amplifier. The optical signal and reset signal of the typical CDS go through two CA, respectively. The performance of the two column amplifiers are not exactly, resulting in itself will bring the column FPN. To suppress the column FPN, a switch was added in the two paths of the optical signals and reset signals [9], sampled once when the effective optical signals were read-out. The method can reduce the column FPN effectively, with the cost of add once sampling and increase the power consumption. The column FPN can be observed as vertical stripes in an image that seriously affecting image quality. In order to effectively reduce FPN column, the paper introduces a design scheme of single amplifier CDS, as shown in Fig 2.

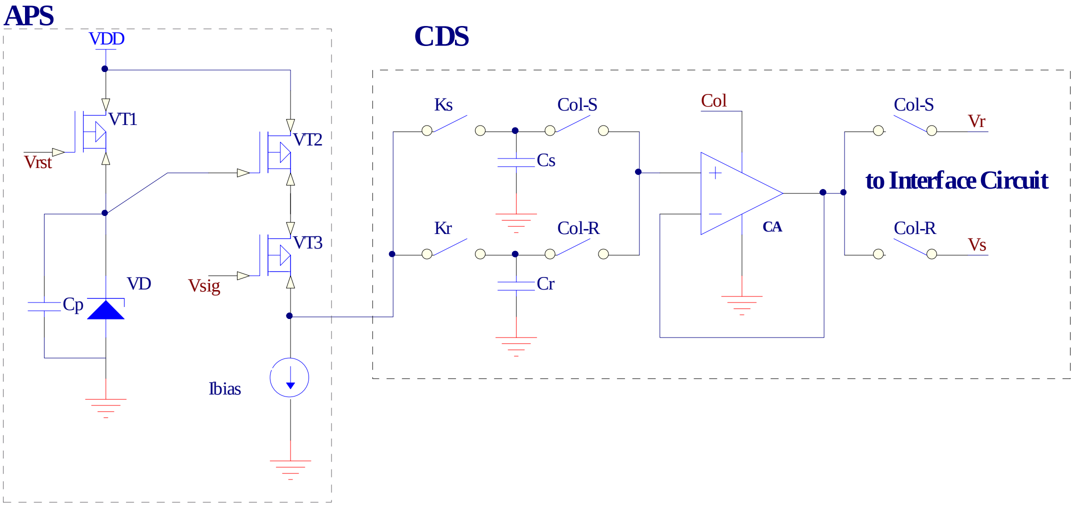


Fig 2 The proposed single amplifier CDS

There is only one single column amplifier in Fig 2, so the readout signal from APS should be sampling twice. That is, sampling the optical signal V_{sig} to C_s using K_s, then, sampling the reset signal V_{rst} to C_r using K_r. Finally, the effectively optical signal V_{opt} can be obtained with the subtraction of V_{sig} and V_{rst}. Since these two values are read out through the same path, the column FPN is greatly reduced in one difference operation, so that it can be ignored.

The workflow of the proposed single amplifier CDS circuit can be described in two steps:

(1) Sample. Sampling the optical signal V_{sig}, in this process, *Col_S* is effectively, the node voltage value of V_s can be gotten with the following equation.

$$V_S(n) = V_{stg} + \Delta \qquad (1)$$

(2) Hold. Sampling the reset signal V_{rst}, in this process, *Col_R* is effectively, the node voltage value of V_r can be gotten with the following equation.

$$V_r(n+1) = V_{rst} + \Delta \qquad (2)$$

The effectively optical signal V_{opt} can be obtained from equation (1) and equation (2), that is, the difference operation of them, the result as follows.

$$V_{opt}(n) = V_r(n+1) - V_S(n) = V_{rst} - V_{stg} \qquad (3)$$

In which, Δ is the column FPN caused by the column amplifier.

Compared with the traditional CDS circuit, there is only one CA in each column, so the power consumption and area of the proposed CDS circuit is reduced by half. The optical signal V_{sig} and the reset signal V_{rst} are read out through the same path, the column FPN is greatly reduced, so that it can be ignored. The results of the proposed scheme are compared with the algorithms of [8] and [9], which can be seen in Table 1. Obviously, the column FPN of the proposed scheme greatly reduced, both power consumption and area decreased as well.

Table 1 the compare results

	[8]	[9]	the proposed
area	small	small	small
power consumption	big	moderate	small
column FPN	moderate	big	small

3.1 Single Column Amplifier

The most simple column amplifier is source follower, which has great influence by the manufacture process, and has poor output linearity. So, the op-amp is generally used to take the place of column amplifier, and has been widespread in many fields.

Symmetrical operational transconductance amplifier (OTA) is widely used for the wide swing, big gain bandwidth product, the input and output can be shorted at the same time. Fig 3 is the traditional symmetrical OTA [10]. In it, *Col_sel* is column choose signal, p_{bisa} is offset voltage, V_i is input, V_o is output. We use it as the single column amplifier in the proposed correlated double sampling circuit.

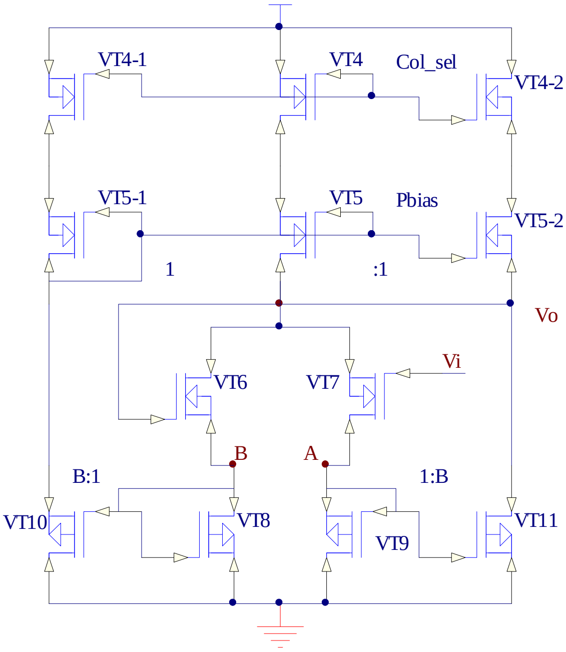


Fig.3 The traditional symmetrical OTA

3.2 Interface Circuit

The proposed CDS circuit can not output the optical signals and the reset signals simultaneously, can not connected with PGA directly, because there is only one CA. So it is necessary to design an interface circuit, as shown in Fig 4. The OTA in it is symmetrical operational transconductance amplifier, which has been introduced in the last part. The node voltage V_s and V_r are the input of the interface circuit, which have the same meaning with Fig 2. V_o is the output of the interface circuit, which connected with PGA.

The working principle of the interface circuit is presented below.

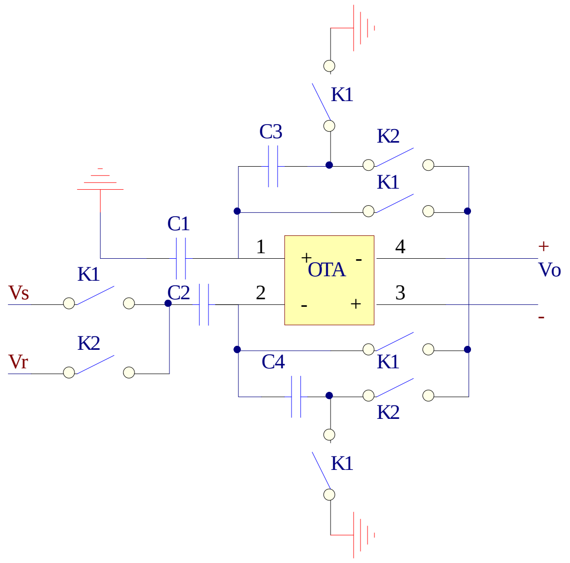


Fig 4 The interface circuit between CDA and PGA

Firstly, clock S_1 effective, sampling the optical signal V_{sig} and preserve it to C_2.

Then, disconnect S_1, clock S_2 effective, sampling the reset signal V_{rst}, and subtract the optical signal V_{sig} at the same time.

Thus, the output signal V_{rst} -V_{sig} can be obtained, which is the output of interface circuit and connected with PGA, the result can be gotten with the following equation.

$$V_o == V_{rst} - V_{stg} \tag{4}$$

5 Conclusions

Denoising and demosaicing are two important aspects of the image imaging process. The noises affecting image imaging quality consist of two parts, which are image sensor noise caused in the process of the image sampling and the noise caused by demosaicing. The proposed CDS technology used one CA to read out optical signal and reset signal sequential, can greatly suppress the column FPN of the image sensor noise. The power consumption and area of the proposed CDA circuit is reduced by half simultaneously.

Acknowledgments

The research is supported by the Youth Foundation of Anhui University of science & technology of China under Grant No.12257, No.2012QNZ06, the Doctor Foundation of Anhui University of science & technology of China under Grant No.11223, and the Guidance Science and Technology Plan Projects of Huainan under Grant No.2011B31.

References

[1] X. LI, Zh. J. Song, J. WANG, M. M. LI, “Image De-noising Based on Total Least Squares”, Computer Engineering, vol. 36, no. 24, 2010, pp. 206–210.

[2] A. Elouardi, S. Bouaziz, A. Dupret, L. Lacassagne, J.O. Klein, R. Reynaud, “Time Comparison in Image Processing: APS Sensors Versus an Artificial Retina Based Vision System”, Meas. Sci. Technol. vol. 18, 2007, pp. 2817–2826.

[3] Hui Tian. “Noise Analysis in CMOS Image Sensors”, PhD Dissertation, USA: Stanford University, 2000.

[4] G. Fikos, L. Nalpantidis, S. Siskos, “A Compact APS with FPN Reduction and Focusing Criterion Using FGMOS Photocell”, Sensors and Actuators A: Physical, vol. 147, no. 2, 2008, pp. 419–424.

[5] D. Sacket, “CMOS Pixel Device Physics”, 2005 IEEE ISSCC Circuit Design Forum: Characterization of Solid-State Image Sensors, San Francisco, CA. 2005.

[6] M. F. Snoeij, A. J. P. Theuwissen, K. A. A. Makinwa, J. H. Huijsing, “A CMOS Imager With Column-Level ADC Using Dynamic Column Fixed-Pattern Noise Reduction”, IEEE Journal of Solid-State Circuits, vol. 41, no. 2, 2006, pp. 3007–3015.

[7] LI Xuan, HOU Zhengxin, XU Hongyu,et al, “Denoise Method Study on CMOS Image Sensor”, Computer Engineering and Applications, vol. 47, no. 8, 2011, pp. 167-169.

[8] R. H. Nixon, S. E. Kemeny, B. Pain, C. O. Staller, E. R. Fossum, “256x256 CMOS Active Pixel Sensor Camera on a Chip”, IEEE Journal of Solid-State Circuits, vol. 31, no. 12, 1996, pp. 2046-2050.

[9] S. Decker, R. D. MeGrath, K. Brehmer, G. G. Sodini, “A 256x256 CMOS Imaging Array with Wide Dynamic Range Pixels and Column Parallel Digital Output”, IEEE Journal of Solid-State Circuits, vol. 33, no. 12, 1998, pp. 2081-2091.

[10] Y. Degerli, F. Lavernhe, P. Magnan, J. Farre. “Non-stationary Noise Response of Some Fully Differential on-chip Readout Circuits Suitable for CMOS Image Sensors”, IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, vol. 46, no. 12, 1999, pp. 1461-1474.

First Author Mrs. Jia received the Master degree in control science and engineering, from the Harbin Institute of Technology. Currently, she is a lectorate at Anhui University of Science & Technology, Electrical and Information Engineering College. Her research interests include Image processing and Rough sets.

Second Author Dr. Zhao received the Master degree in control theory and control engineering from the Qingdao University of Science & Technology, in 2005. He received the Ph.D. degree in control science and engineering, from the Harbin Institute of Technology. Currently, he is a lectorate at Anhui University of Science & Technology, Electrical and Information Engineering College. His research interests include Image processing, intelligent control and Rough sets.

Demosaicing Algorithm for Color Filter Arrays Based on SVMs

Xiao-fen JIA, Bai-ting Zhao

**School of Electrical and Information Engineering, Anhui University of Science & Technology
Huainan 232001, China**

Abstract

One color filter array (CFA) used in a digital camera allows only one of the red-green-blue primary color components to be sensed at each pixel, and interpolating the other missing two components by methods known as demosaicing. A novel support vector machines (SVMs) based demosicing algorithm is proposed to reduce edge artifacts and false color artifacts effectively. The proposed algorithm is a four-step method. Firstly, construct middle plane K_r or K_b on the mosaic image. Secondly, train SVMs with the trained samples constructed on the middle plane. Thirdly, interpolate the unknown value of the middle plane K_r or K_b. Finally, calculate the missing pixel value. Experimental results showed that the proposed approach produced visually pleasing full-color result images and obtained better PSNR values than other demosaicing algorithms

Keywords: *Demosaicing, Color filter array (CFA), Image interpolation, Support vector machines (SVMs).*

1. Introduction

In recent years, rapid research and development have helped make digital imagers more and more widespread in daily life. People's requirements to the image quality are more rigorous. The different processing strategies implemented in image sensors, and the different stages of image processing are more important. Demosaicing is one of the significant stages of image processing.

To capture a color image, three image sensors are needed to simultaneously sense the three-primary colors: red (R), green (G) and blue (B). However, to minimize the size, cost and complexity, designers employ a single image sensor overlaid with a color filter array (CFA) to acquire the color image. With this scheme, only one pixel value of the three-primary colors is sensed. To restore a full-color image, the two missing color values at each pixel need to be estimated from the adjacent pixels. This process is commonly known as CFA interpolation or demosaicking.

Bilinear interpolation is the simplest method for CFA interpolation, in which the missing color value is filled with the average of its neighboring CFA samples in the same color. It introduces errors in the edge region with blurred result images and produces color artifacts. To obtain more accurate and visually pleasing results, many sophisticated CFA interpolation methods have been proposed. In [1] an effective color interpolation algorithm (ECI) using signal correlation to get better image quality is provided. The frequency response of this approach is better than the conventional methods especially in high frequency. Another enhanced ECI interpolation approach (EECI) which effectively used both the spatial and the spectral correlations is proposed in [2], and it provided effective scheme to enhance two existing state-of-the-art interpolation methods. In [3] a universal demosaicking algorithm (UD) is provided employing an edge-sensing mechanism and a post-processor to unify existing interpolation solutions. Tsai and Song [4] exploited high-frequency information of the green channel to reduce the aliasing error in red and blue channels. In [5], Lian et al designed an efficient filter for estimating the luminance at green pixels and presented an adaptive filtering approach to estimating the luminance at red and blue pixels. Hos et al designed several new CFA patterns based on the ideal of minimizing the demosaicing error [6], and used the adaptive weighting method to get full color image. A SVMs based error correction scheme is provided in [7] to improve interpolation accuracy of result images. Recently, a novel SVMs based image interpolation method for gray images employed the local spatial property information is proposed in [8], and experimental data showed that SVMs based interpolation can provide high quality interpolation result images. In this paper, SVMs based interpolation is used for demosaicing.

The remainder of this paper is organized as follows. In section 2, SVMs is briefly introduced. In section 3, the details of the proposed demosaicing approach is described. Section 4 is the experimental results of the methods under comparison. Finally, conclusion is given in section 5.

2. SVMs

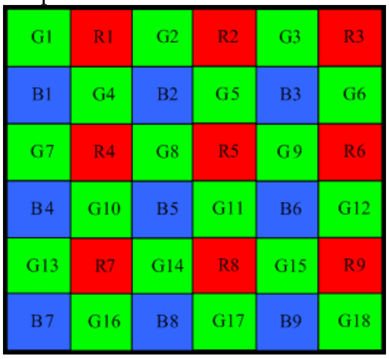
SVM is built on the basis of statistical learning theory with optimal ways to solve the problem of machine learning. Which have been used successfully for many supervised classification tasks, regression tasks and novelty detection tasks [9-12]. Support vector regression (SVR) is a function approximation approach applied with SVM. A wide range of image processing problems have also been solved with

SVMs. The basic idea of SVR is mapping the data in the current space with linear non-separable case to a high dimensional feature space in which the data point is separable. A training data set $T=\{(x_i,y_i)\}_{i=1}^{m}$ consists of m points $\{x_i,y_i\}$, $i=1,2,...,m$, $x_i\in R^d$, $y_i\in R^d$, where, x_i is the i-th input pattern and y_i is the i-th output pattern. The aim of SVR is to find a function $f(x)=<\omega,\phi(x)>+b$ to obtain eventual targets y corresponding x.

The kernel function $k(x_i,x)=<\phi(x_i),\phi(x)>$ is used to implement the nonlinear mapping, which can be selected as linear kernel, polynomial kernel, radial basis function (RBF) kernel, or two layer neural kernel.

3. Proposed Algorithm

The most popular CFA filter pattern is Bayer pattern in which the color components are placed in an orderly fashion as showed in Fig 1 [1-3]. Although other patterns can also be processed with our proposed algorithm, Bayer pattern is regarded as the default CFA pattern in our algorithm description.



G1	R1	G2	R2	G3	R3
B1	G4	B2	G5	B3	G6
G7	R4	G8	R5	G9	R6
B4	G10	B5	G11	B6	G12
G13	R7	G14	R8	G15	R9
B7	G16	B8	G17	B9	G18

Fig.1 Bayer pattern of CFA

Image interpolate rely heavily on color correlations, which include spatial and spectral correlations. The image spectral correlation between the R, G, B channels can be represented as K_r plane and K_b plane, where $K_r=G-R$ and $K_b=G-B$ [1]. For real-world images, the contrasts of K_r and K_b are quite flat over small regions, and this property is suitable for interpolation.

The SVM-based interpolation is performed to G channel, B channel and R channel respectively. Four steps are needed when interpolating an unknown pixel value no matter in which channel. We summarize the procedure as follows.

(1) Construct middle plane K_r or K_b on the mosaic image.

(2) Train SVM with trained samples constructed by the known values on the K_r plane or K_b plan.

(3) Interpolate the unknown values of the K_r plane or K_b plane using the trained SVM.

(4) Calculate the unknown pixel value using the interpolated K_r or K_b values.

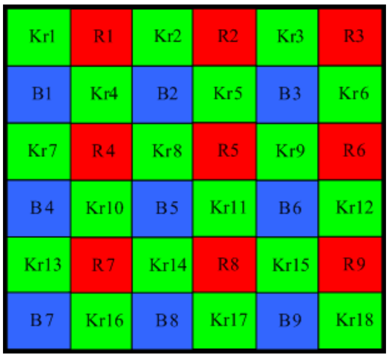
When using SVMs, the samples are constructed by selecting the neighbor pixels. The principle of selecting neighbor pixels region is the trained mode similar with the forecast mode. The forecast mode is determined by the position of the same color pixels around the neighbor regions.

Firstly, interpolate G channel.

Step1: Interpolate the G color value with known R.

(1) The plane of K_r is constructed for SVMs training.

We can calculate K_r value for the pixels with known G color value employing the two adjacent known R color values. Fig 1 shown, pixel G$_3$ is in the place of odd row, the corresponding K_r value can be calculated with $K_{r3}=G_3-(R_2+R_3)/2$. Pixel G$_5$ is in the even row, the corresponding K_r value can be calculated wit $K_{r5}=G_5-(R_2+R_5)/2$. For the special brim column or row pixels, for instance, G$_{13}$ and G$_{16}$, we can obtain the corresponding K_r value with $K_{r13}=G_{13}-R_7$ and $K_{r16}=G_{16}-R_7$, respectively. After the K_r plane for all the pixels with known G color value is estimated, as shown in Fig 2, this K_r plane can be used for SVMs training.



Kr1	R1	Kr2	R2	Kr3	R3
B1	Kr4	B2	Kr5	B3	Kr6
Kr7	R4	Kr8	R5	Kr9	R6
B4	Kr10	B5	Kr11	B6	Kr12
Kr13	R7	Kr14	R8	Kr15	R9
B7	Kr16	B8	Kr17	B9	Kr18

Fig.2 Kr plane for G channel interpolate

(2) Interpolate the K_r values of the pixels with known R in the K_r plane using SVMs.

Every pixel with known K_r value in the above K_r plane is selected as center pixel to construct three samples for SVMs training. Output patterns of these samples are the K_r values of the center pixel. The input pattern is the four-dimensional vector constituted by the K_r values of four neighbor pixels around the center pixel. For example, K_{r8} is selected as center pixel, one input pattern can be comprised of K_{r2}, K_{r7}, K_{r14} and K_{r9}. Another input pattern constituted by K_{r4}, K_{r10}, K_{r11} and K_{r5}. The third pattern is made up of K_{r1}, K_{r13}, K_{r15} and K_{r3}. All these samples are used for SVMs training. The trained SVMs can be employed to estimate K_r value of the pixel with known R color value. For example, when the input pattern constituted by K_{r5}, K_{r8}, K_{r11} and K_{r9} is used, K_{r5}corresponding R_5 can be obtained with the trained SVMs.

(3) For the pixel i with known R color value the G color value is estimated as $G_i = K_{ri} + R_i$

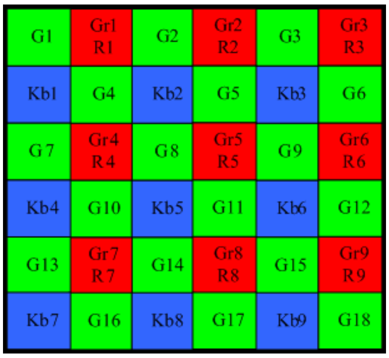
Step2: Interpolate the G color value with known B.

Likewise, the plane of K_b can be constructed, and all the K_b values of the pixels with known B color value can be estimated with SVMs. Then, the G color value of the pixel i with known B color value can be estimated with $G_i = B_i + K_{bi}$. Now, we can obtain all G color value of the image, which can be considered as the known pixels in the second pass.

Secondly, interpolate B channel.

Step1: Interpolate the B color value with known R.

Similarly with the work in G channel, the plane of K_b can be constructed for SVMs training. The K_b value of the pixel i with known B color value can be calculated as $K_{bi} = G_i - B_i$, where G_i has been estimated in the first pass. And we get the K_b plane showed in Fig 3. In this plane, SVMs are trained with the samples constructed from pixels with known K_b value. K_b value of the center pixel is the output pattern for the samples. Two input patterns of the center pixel can be used to construct samples for SVMs training. For example, when K_{b5} is selected as the center pixel, one of the two input patterns is constitutive of K_{b1}, K_{b7}, K_{b9} and K_{b3}, Another one is comprised of K_{b2}, K_{b4}, K_{b8} and K_{b6}. After all the examples are used for SVMs training, the trained SVMs can be used to estimate K_b value of the pixel with known R color value. For example, K_{b5} corresponding G_{r5}/R_5 can be estimated with trained SVMs employing the input pattern constituted by K_{b2}, K_{b5}, K_{b6} and K_{b3}. Thus, all the K_b values of the pixels with known R color value can be estimated.



G1	Gr1 R1	G2	Gr2 R2	G3	Gr3 R3
Kb1	G4	Kb2	G5	Kb3	G6
G7	Gr4 R4	G8	Gr5 R5	G9	Gr6 R6
Kb4	G10	Kb5	G11	Kb6	G12
G13	Gr7 R7	G14	Gr8 R8	G15	Gr9 R9
Kb7	G16	Kb8	G17	Kb9	G18

Fig.3 Kb plane for B channel interpolate

Step2: Interpolating the B color value with known G.

So far, all the rest pixels with unknown K_b values in K_b plane are the pixels with known G color values. These unknown K_b value can also be estimated using the trained SVMs. For examples, K_{b9} corresponding G_9 can be estimated with the input pattern constructed from K_{b3}, K_{b5} (corresponding G_{r5}/R_5), K_{b6} and K_{b6}(corresponding G_{r6}/R_6). Then the B color value of the pixel i could be calculated with $B_i = G_i - K_{bi}$. Now, we get the B color channel of the image.

Thirdly, interpolate R channel just like the interpolation to B channel.

4. Experiments

The experiments are performed in Matlab 2G memory, 3.0GHz single-core CPU and the SVM tools for Matlab [12] are used. In order to verify the effect of the proposed algorithm, some standard test images that have been widely used in other literatures and a wide range of real images are used in our experiments. Some of these test images are showed in Fig 4. Bilinear interpolation, ECI interpolation [1], EECI interpolation [2], UD interpolation [3], Hos et al. [6] (CFA4b Adaptive), and our proposed approach are used in our experiments. In these experiments, the γ -SVR with radial basis function kernel is employed for the SVMs based interpolation, and all parameters in the SVMs tool are set to default. Peak signal to noise ratio (PSNR) value between the source image and the result image is employed to compare different demosaicing algorithms. PSNR is calculated for all the images showed in Fig 4 and listed in Table 1. It is obvious that the proposed approach gets the highest average PSNR value. Hos's algorithm [6] obtained high PSNR of the image Sails, Mountain, and Sky. The common characteristic of the three images are with fewer edges.

Experimental result images of image Sailboat employing different demosaicing approaches are zoomed and illustrated in Fig 5. It can be observed that the ECI interpolation blur the image edges with visible artifacts appeared in the edge regions, such as sail edge. Color artifacts are also appeared obviously in the people region and sailboat mark word region in the result images of EECI, UD and Hos's algorithm. Our proposed SVMs based approach obtains the best visual result with less edge artifacts and less color artifacts. These observation results are consistent with PSNR value listed in Table 1. Experimental result images of real image Family are illustrated in Fig 6. We can also find edge artifacts and color artifacts appeared in the result images of ECI, UD, EECI and Hos's algorithm, especially in the house edge region. The proposed approach produces less edge artifacts and less color artifacts. These observations indicate that our proposed approach keeps the edge details effectively and produces less color artifacts.

Table.1: PSNR of different demosaicing approachs

Image	ECI	EECI	UD	[6]	Proposed
Wall	26.15	25.33	25.83	29.13	**29.57**
House	26.04	28.57	28.86	31.06	**31.35**
Building	20.92	22.25	22.91	24.39	**24.51**
Face	18.13	19.13	16.89	19.99	**20.30**
Sails	24.02	26.10	26.02	**26.97**	26.81
Girl	25.40	27.40	27.43	28.44	**28.57**
Lighthouse	23.67	25.02	24.12	26.24	**26.60**
Sailboat	22.54	22.92	22.97	25.90	**26.02**
Plane	20.12	27.10	26.15	27.32	**27.56**
Mountain	22.90	24.46	24.98	**27.99**	27.86
Tree	20.89	21.88	20.21	22.19	**22.67**
Bridge	20.94	23.39	23.36	26.03	**26.15**
Sky	26.64	28.04	31.23	**33.42**	33.03
Family	21.92	23.63	25.47	27.34	**27.53**
Average	22.88	24.66	24.75	26.89	**27.04**

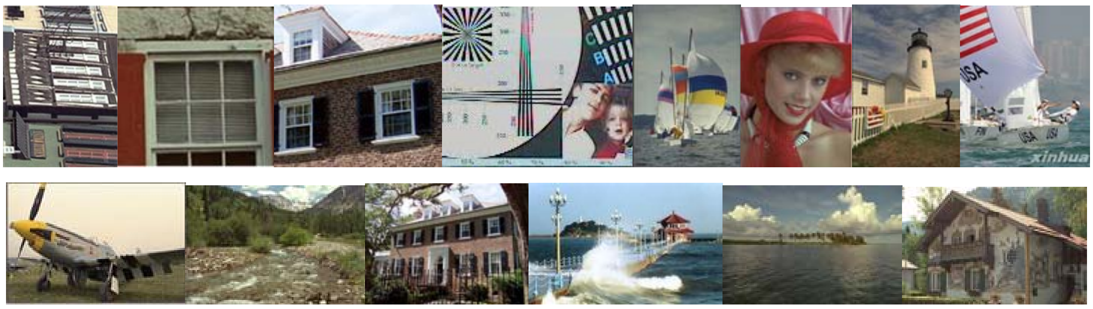
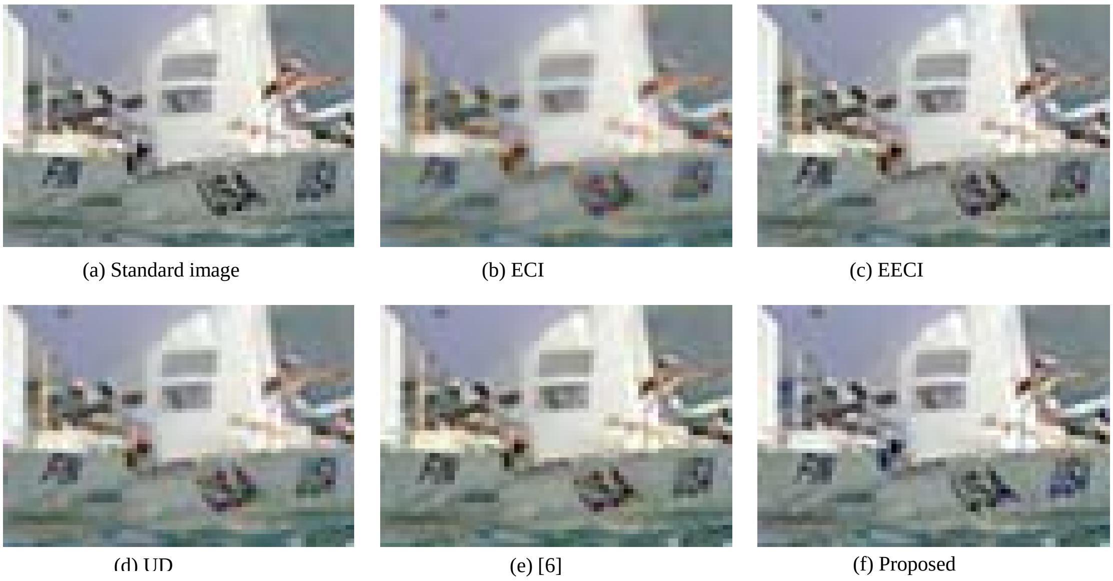


Fig.4 Some test images



(a) Standard image

(b) ECI

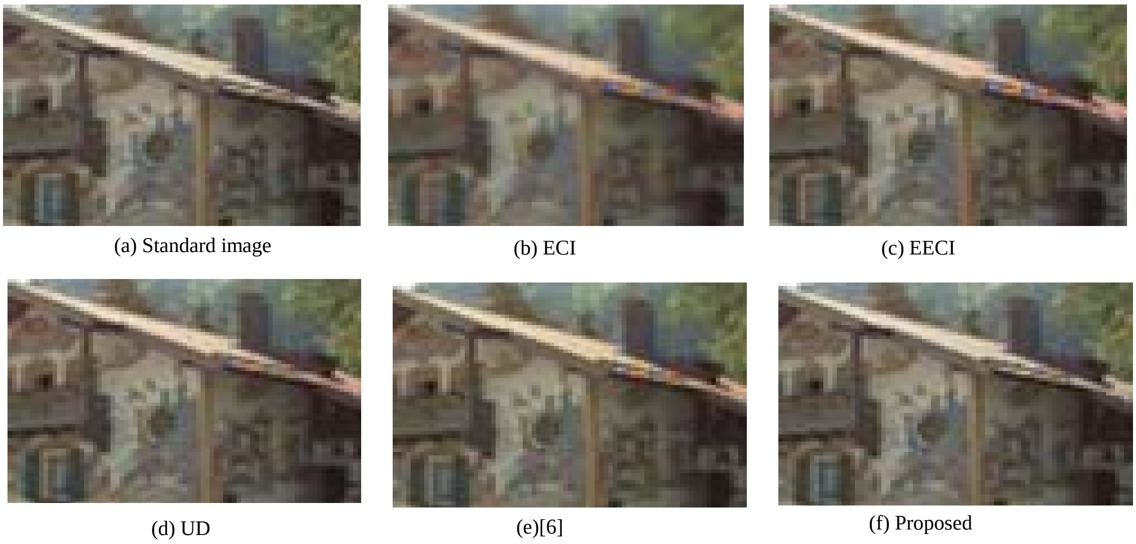
(c) EECI

(d) UD

(e) [6]

(f) Proposed

Fig.5 Zoomed region of the demosaiced image Sailboat



(a) Standard image (b) ECI (c) EECI

(d) UD (e)[6] (f) Proposed

Fig.6 Zoomed region of the demosaiced image Family

5. Conclusions

Based on the insights gained from our study, SVMs can ensure the accuracy of the interpolation results by its properties of global optimal and generalization ability, the mosaic image can be interpolated effectively with the combination of image correlation and SVMs. The proposed demosicing algorithm can reduce edge artifacts and false color artifacts effectively, have excellent effect to the image with more edge. The experimental results show that the proposed algorithm obtains higher PSNR value and produces visually pleasing full-color images.

Acknowledgments

The research is supported by the Youth Foundation of Anhui University of science & technology of China under Grant No.12257, No.2012QNZ06, the Doctor Foundation of Anhui University of science & technology of China under Grant No.11223, and the Guidance Science and Technology Plan Projects of Huainan under Grant No.2011B31.

References

[1] S. C. Pei, I. K. Tam, “Effective Color Interpolation in CCD Color Filter Array Using Signal Correlation”, IEEE Trans.on Circuits and Systems for Video Technology, vol. 13, no. 6, 2003, pp. 503-513.

[2] L. L. Chang, Y. P. Tan, “Effective Use of Spatial and Spectral Correlations for Color Filter Array Demosaicking”, IEEE Trans. on Consumer Electronics, vol. 50, no. 1, 2004, pp 355-365.

[3] R. Lukac, K. N. Plataniotis, “Universal Demosaicking for Imaging Pipelines with an RGB Color Filter Array”, Pattern Recognition, vol. 38, no. 11, 2005, pp. 2208-2212.

[4] C. Y. Tsai, K. T. Song, “A New Edge-adaptive Demosaicing Algorithm for Color Filter Arrays”, Image and Vision Computing, vol. 25, no. 9, 2007, pp. 1495-1508.

[5] N. X. Lian, L. Chang, Y. P. Tan, V. Zagorodnov, “Adaptive Filtering for Color Filter Array Demosaicking”, IEEE Trans. on Image Processing, vol. 16, no. 10, 2007, pp. 2515-2525.

[6] P. W. Hao, Y. Li, Z. C. Lin, E. Dubois, “A Geometric Method for Optimal Design of Color Filter Arrays”, IEEE Trans. on Image Processing, vol. 20, no. 3, 2011, pp 709-722.

[7]J. Wang, L. Ji, “Image Interpolation and Error Concealment Scheme Based on Support Vector Machine”, Journal of Image and Graphics, vol. 7(A), no. 6, 2002, pp. 558-564.

[8] L. Y. Ma, Y. Shen, and J. C, Ma. “Local Spatial Properties Based Image Interpolation Scheme Using SVMs”, Journal of Systems Engineering and Electronics, vol. 19, no. 3, 2008, pp. 618-623.

[9] N. Y. Deng, Y. J. Tian, “A Novel Data Mining Method: SVM”, Science Press, Beijing, 2004.

[10] S. Zheng, J. W. Tian, and J. Liu, “Research of SVM-based Image Interpolation Algorithm Optimization”, Journal of Image and Graphics, vol. 10, no. 3, 2005, pp. 338-343.

[11] H. Z. Wang, R. Zhang, F. K. Liu etc, “Improved Kriging Interpolation Based on Support Vector Machine and Its Application in Oceanic Missing Data Recovery”, Proc. of the 2008 International Conference on Computer Science and Software Engineering, vol.4, 2008, pp. 726-729.

[12] K. S. Ni, T. Q. Nguyen. “Image Super-resolution Using Support Vector Regression”. IEEE Trans. on Image Processing, vol. 16, no. 6, 2007, pp. 1596-1610

[13] C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~ cjlin /libsvm.

First Author Mrs. Jia received the Master degree in control science and engineering, from the Harbin Institute of Technology. Currently, she is a lectorate at Anhui University of Science & Technology, Electrical and Information Engineering College. Her research interests include Image processing and Rough sets.

Second Author Dr. Zhao received the Master degree in control theory and control engineering from the Qingdao University of Science & Technology, in 2005. He received the Ph.D. degree in control science and engineering, from the Harbin Institute of Technology. Currently, he is a lectorate at Anhui University of Science & Technology, Electrical and Information Engineering College. His research interests include Image processing, intelligent control and Rough sets.

Application of Volterra LMS Adaptive Filter Algorithm Based on Gaussian Distribution

Xinling Wen[1], Dongfang Luo[2]

[1] Zhengzhou Institute of Aeronautical Industry Management, Zhengzhou, 450015, China

[2] Henan College of Finance & Taxation, Zhengzhou, 451464, China

Abstract

This paper mainly studied the LMS adaptive filter algorithm to the Volterra system model. Through the construction of the second order Volterra system model, the application of respectively selecting the first order and second order variable step length de-correlation Volterra LMS algorithm in gaussian noise environment, when the input signals in different correlation coefficient, the iteration times are not more than 2000 times and all items can realize the convergence, which prove the accuracy of the algorithm paper presented. The Volterra LMS adaptive filter algorithm can be effectively applied into the mechanical vibration damping and noise elimination, which has a broad application prospect.

Keywords: *Volterra series, adaptive filter algorithm, LMS, system identification, gaussian distribution.*

1. Introduction

Adaptive filter research began in the 1950's. Widrow and Hoff, etc first puts forward the least mean square (LMS) algorithm. [1] LMS algorithm has the advantages of simple structure, small amount of calculation, and easy to realize real-time processing, so, in the field of the low noise elimination, spectrum enhancement, and system identification, etc, which has been widely used. [2] Among them, the traditional adaptive linear filtering theory based on the gaussian noise model has more mature, and has a wide range of applications in many engineering fields. However, with the expansion of signal processing field, nonlinear filter adaptive filter algorithm is gradually becoming the research hot spot in the world. Volterra adaptive filter algorithm have been successful applied in the military industry and other signal processing or modeling, which reflect its accuracy.

In this paper, we mainly research the Volterra LMS adaptive filtering algorithm [3]based on the gauss noise environment, and through the simulation of the fault vibration model of the nonlinear system identification, which proved the performance of the Volterra LMS adaptive filter algorithm.

2. Volterra series nonlinear system

A discrete causal nonlinear Volterra system relationship between the input signal $x(n)$ and its output $y(n)$ can be expressed as the Volterra series formula. [4]

$$\begin{aligned} \boldsymbol{y}(n) &= h_0 + \sum_{m_1=0}^{\infty} h_1(m_1)x(n-m_1) \\ &+ \sum_{m_1=0}^{\infty}\sum_{m_2=0}^{\infty} h_2(m_1,m_2)x(n-m_1)x(n-m_2)+... \\ &+ \sum_{m_1=0}^{\infty}\sum_{m_2=0}^{\infty}...\sum_{m_p=0}^{\infty} h_p(m_1,m_2,...,m_p)x(n-m_1)x(n-m_2)...x(n-m_p)+... \end{aligned} \tag{1}$$

Among the formula (1), $h_p(m_1,m_2,...,m_p)$ is called as p orders Volterra kernel coefficient. It is called linear kernel when p=1. Volterra series can be seen as the Taylor series expansion with memory circumstance, which can approach to any continuous nonlinear system model. Formula (1) expresses there are infinity numbers Volterra kernel to the nonlinear system. In the fact application, we should carry out truncation process in the practical application.

Truncation process contains two aspects of the order number p and memory depth N. How to truncate is relevant to the specific nonlinear system type and the performance of the requirements. Usually, only considering the second order truncation model, that is p=2, and hypothesis h_0 = 0, memory depth is N. System can be simplified:

$$\begin{aligned} \boldsymbol{y}(n) &= \sum_{m_1=0}^{\infty} h_1(m_1)x(n-m_1) \\ &+ \sum_{m_1=0}^{N-1}\sum_{m_2=m_1}^{\infty} h_2(m_1,m_2)x(n-m_1)x(n-m_2) \end{aligned} \tag{2}$$

Among the formula (2), we suggest the kernel of Volterra series is symmetrical. To any of $p!$ numbers $m_1,m_2,...,m_p$ transposition, $h_p(m_1,m_2,...,m_p)$ is equation.

Thus, formula (2) has $N(N+3)/2$ numbers Volterra kernel. Considering the symmetrical character to the Volterra series. We can define system kernel quantity in n times.

$$\begin{aligned}\boldsymbol{H}(n)=[&h_1(0;n),h_1(1;n),\cdots,h_1(N-1;n)\\ &,h_2(0,0;n),h_2(0,1;n),\cdots,\\ &h_2(0,N-1;n),h_2(1,1;n),..,h_2(N-1,N-1;n)]^T\end{aligned} \quad (3)$$

The same we can define system input vector in n time.

$$\begin{aligned}\boldsymbol{X}(n)=[&x(n),x(n-1),\cdots,x(n-N+1),x^2(n)\\ &,x(n)x(n-1),\cdots,x(n)x(n-N-1),\\ &x^2(n-1),..,x^2(n-N+1)]^T\end{aligned} \quad (4)$$

Thus the output can be expressed as formula (5) in n times.

$$y(n)=\boldsymbol{H}^T(n)\boldsymbol{X}(n) \quad (5)$$

Formula (2) and (5) state that a nonlinear system can be state extended to express as linear combination of the input vector $X(n)$ each component, which is the advantages of the nonlinear system Volterra series expressed.

3. Adaptive volterra filter

If the known system has the form style as the formula (5), but its kernel vector $\boldsymbol{H}(n)$ is unknown, so, we can use Figure 1 to identify the system kernel vector $\boldsymbol{H}(n)$ similar to the linear style. In the Figure 1, $\boldsymbol{W}(n)$ is Volterra filter coefficient vector with the length of $M=N(N+3)/2$.

If we define the Volterra filter coefficient vector $\boldsymbol{W}(n)$ is: $w(n)=[w_0(n),w_1(n),...,w_{M-1}(n)]^T$, then the output of Volterra filter $c(n)$ is : $c(n)=\boldsymbol{W}^T(n)\boldsymbol{X}(n)$, among the formula, $\boldsymbol{X}(n)$ is the input vector of the formula (3).

The purpose of the system identification is changing filter coefficient vector $\boldsymbol{W}(n)$ through a adaptive algorithm, which to make error information $e(n)$ into minimum in a sense. That is to say, it will make some a cost function $J(n)$ of $e(n)$ into minimum. When the cost function $J(n)$ approaching to minimum, we can think $H(n)\approx\boldsymbol{W}(n)$.

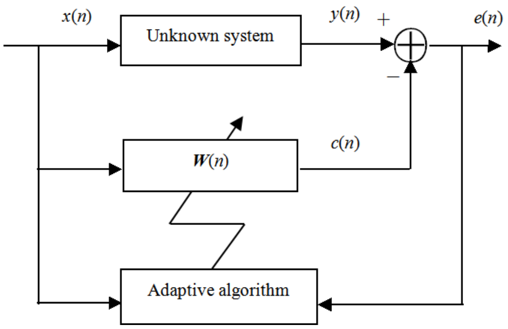


Fig. 1 Identification method based on adaptive filter.

If we define the cost function $J(n)$ as formula (6).

$$J(n)=e^2(n)=[y(n)-c(n)]^2=[y(n)-\boldsymbol{W}^T(n)\boldsymbol{X}(n)]^2 \quad (6)$$

We can calculate the $\boldsymbol{W}(n)$ differential coefficient to $J(n)$. And make the changing direction is opposite to the differential coefficient direction of $\boldsymbol{W}(n)$. Then we can get the optimal $\boldsymbol{W}(n)$ recursion algorithm. That is LMS algorithm. [5] LMS algorithm process can be concluded as formula (7). [6]

$$\begin{aligned}e(n)&=y(n)-\boldsymbol{W}^T(n)\boldsymbol{X}(n)\\ W(n+1)&=W(n)+u\boldsymbol{X}(n)e(n)\end{aligned} \quad (7)$$

Among the formula (7), the initial value of $\boldsymbol{W}(n)$ can be determined according to the prior knowledge, or simply select $\boldsymbol{W}(n)=[0,0,...,0]^T$. Step length u decides convergence speed, tracing character, and the stability of the LMS algorithm. In order to assure the convergence speed, and stability system, we can adopt standardization method to determine the step length u.

$$u_n=u\,\|\,\boldsymbol{X}(n)\,\|^2 \quad (8)$$

Among the formula (7), $u_n(0<u_n<2)$, which called standardization step length. LMS algorithm has the advantage of little calculation amount, but because system has nonlinear character, it makes the correlation matrix eigenvalue of input signal extend to big, and lead the convergence speed to slow. In order to speed up the convergence speed, in this paper, we adopt different step length to the system linear part and the nonlinear part. [7] The first order and second order terms of Volterra series use different convergence factor, the weight vectors iterative formula of LMS algorithm [8] is shown as bellow.

$$\boldsymbol{W}(n+1)=\boldsymbol{W}(n)+\begin{bmatrix}\mu_1 & \cdots & 0 & 0 & \cdots & 0\\ 0 & \ddots & 0 & 0 & \ddots & 0\\ 0 & \cdots & \mu_1 & 0 & \cdots & 0\\ 0 & \cdots & 0 & \mu_2 & \cdots & 0\\ 0 & \ddots & 0 & 0 & \ddots & 0\\ 0 & \cdots & 0 & 0 \cdots & & \mu_2\end{bmatrix}\boldsymbol{X}(n)e(n) \quad (9)$$

If we adopt scalar style. It is shown as formula (10).

$$w(1,m_1;n+1)=w(1,m_1;n)+\mu_1 e(n)x(n-m_1)$$
$$w(2,m_1,m_2;n+1)=w(2,m_1,m_2;n)+\mu_2 e(n)x(n-m_1)x(n-m_2) \quad (10)$$

Among the formula (10), $m_1=0,1,\cdots,N-1$; $m_2=0,1,\cdots,N-1$. And in order to assure algorithm convergence in statistical sense, Volterra LMS algorithm convergence factor must choose in the following range shown as formula (11).

$$0<\mu<\frac{1}{tr[\boldsymbol{R}]}<\frac{1}{\lambda_{\max}} \quad (11)$$

Among the formula (11), $\lambda_{\max}$ is the eigenvalue of maximum of the input vector autocorrelation matrix $\boldsymbol{R}=E[\boldsymbol{X}(n)\boldsymbol{X}^T(n)]$. Form the style we can see, the condition of convergence of Volterra series is same as the traditional LMS algorithm, but because the definition of the input vector $\boldsymbol{X}(n)$ is different, which making Volterra LMS method input vector autocorrelation matrix $\boldsymbol{R}$ contains the input signal of high order statistics. So, even in white noise condition, it will lead matrix characteristic value expansion. When the correlation of the input signal become stronger, the convergence speed will become slow, even cannot realize convergence. [9]

4. Algorithm simulation and character analysis

This algorithm is used for a nonlinear system model [10] damping and de-noising, it was assumed that identify a nonlinear time-invariant system input and output relationship for:

$$\begin{aligned}c(n)=&0.7512x(n)+0.3467x(n-1)+0.1231x(n-2)\\&+0.6892x^2(n)+0.2154x^2(n-1)\\&-1.4893x^2(n-2)+0.5625x(n)x(n-1)\\&-1.5903x(n-1)x(n-2)+2.3467x(n)x(n-2)\\&+noise(n)\end{aligned}$$

The first order kernel coefficient respectively is 0.7512, 0.3467, and 0.1231; the second kernel coefficient respectively is 0.6892, 0.2154, -1.4893, 0.5625, -1.5903, and 2.3467. The input signal is $x(n)=ax(n-1)+noise(n)$, among it , $noise(n)$ is gaussian white noise with mean value is 0 and variance is 1. We set the input signal signal to noise ratio is 20dB. We adopt weight coefficient error to analyze algorithm performance. And take the second order Volterra filter with the memory length N of 3. The simulation result is obtained through 50 times independent simulation after taking average. When the input signal is for weak correlation signal, namely a=0.3, the kernel coefficient convergence curve of Volterra series by the Volterra LMS algorithm calculating is shown as Figure 2.

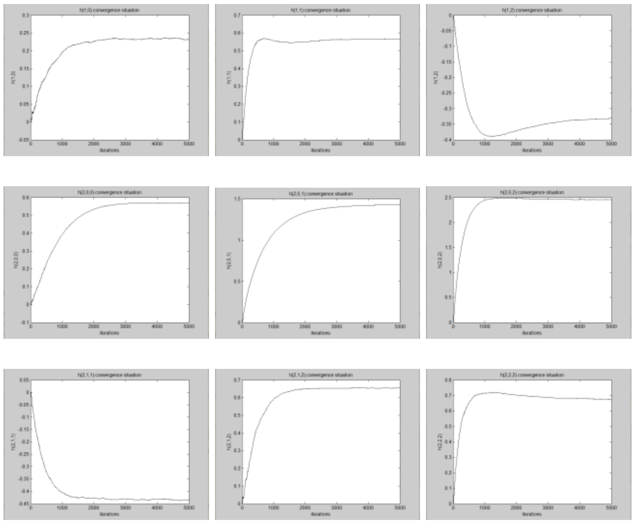
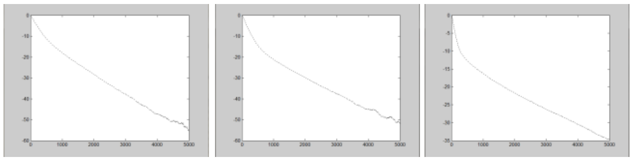


Fig. 2 Each weight value convergence situation comparison under the weak correlation.

From the Fig. 2 we can see, in the weak correlation cases, Volterra LMS algorithm can realize fast convergence, steady state disorder quantity is relatively low. However, in the medium correlation and strong correlation cases, the input signal weight value still can realize convergence, but the convergence speed is slow. Figure 3 is given in three related intensity input signal under the condition of weight coefficient error norm curve of the Volterra LMS algorithm.



a=0.3 a=0.6 a=0.9

Fig. 3 Weight coefficient error bound norm.

From the Fig. 3 we can see, with the increasing of the input signal correlation strength (From 0.3, 0.6 until 0.9), the weight coefficient error norm curve all decrease, that is error is smaller by smaller, but with the increasing of correlation strength, the mismatch error can not achieve minimum in a relatively short period of time.

5. Conclusion

This paper studies the model of Volterra LMS adaptive filter algorithm, through using the advantage of variable step length and decorrelation, respectively adopts different convergence factor to the first order and second order terms, which improving the performance of Volterra LMS algorithm. Simulation shows that this algorithm in different input signal correlation situation has better convergence performance and steady state performance. This Volterra LMS adaptive filter algorithm can better able to used in mechanical model damping and noise elimination, which has a broad application prospect.

Acknowledgments

This paper is supported by Young Scientific Research Fund Project of Zhengzhou Institute of Aeronautical Industry Management. (No. 2012133002)

References

[1] Widrow B, Stearns S.D, “Adaptive Signal Processing”, Engle Wood Cliffs, NJ: Prenticer Hall, 1985.

[2] Lv Zhensu, Xiong Jingsong, “A Novel Improved Variable Step-size LMS Algorithm”, Signal Processing, Vol. 24, No. 1, 2008, pp.144-146.

[3] Xiaoyan Cheng, Guoqing Dang, “Research on the Step-Variable Self-Adaptive Filtering Technology based on Improved Least Mean Square Algorithm”, JCIT, Vol. 7, No. 11, 2012, pp. 202-208.

[4] Wang Guangsen, Wang Cheng, “Nonlinear Systems Identification Based on Adaptive Volterra Filter”, Electronics Opitcs & Control, Vol.12, No.2, 2005, pp. 42-44.

[5] Koh T, Powers E J, “Second-order Volterra Filtering and Its Application to Nonlinear System Identification”, IEEE Transactions on Acoustic, Speech, Signal Processing, ASSP-33, Part 6, 1985.

[6] Li Aihong, Xiao Shanzhu, Zhang Eryang, “A Method of Adaptive Volterra Predistortion Based On the Discrete Wavelet Transform”, Signal Processing, Vol. 25, No.1, 2009, pp. 40-43.

[7] GRIFFITH D W, ARCE G R, “A Partially Decoupled RLS Algorithm for Volterra Filters”, IEEE T-SP, Vol. 47, No. 2, 1999, pp. 579-582.

[8] Suma S.A., Dr. K.S.Gurumurthy, "New Improved echo canceller based on Normalized LMS Adaptive filter for Single talk and Double talk Detection, Subband echo cancellation, Acoustic Echo cancellation", JNIT, Vol. 1, No. 2, 2010, pp. 61-74.

[9] Luo Yongjian, Wu Yinsheng, Sun Jun., “A Kind of Improved Adaptive Volterra Filter”, Journal of Data Acquisition & Processing, Vol 24, No.5, 2009, pp.676-679.

[10] Tong Bing Zhao Chunhui, “A new adaptive method for nonlinear system modeling”, Applied Science and Technology, Vol. 31, No. 10, 2004, pp. 4-6.

Xinling Wen received the Master degree in data collection and signal processing from North China University of Water Resources and Electric Power, in 2009. Currently, she is a Lecturer at Zhengzhou Institute of Aeronautical Industry Management. Her interests are in data collection and signal processing and nonlinear system modeling.

Research and Application of BSS Algorithm on The Gearbox Fault Diagnosis Based on The MMI Criterion

Yu Chen[1], Haitao Jiang[2]

[1] Zhengzhou Institute of Aeronautical Industry Management
Zhengzhou, 450015, China

[2] Jiaozuo Teachers College
Jiaozuo, 454001, China

Abstract

This paper presented a kind of blind source separation (BSS) technology and applied it into the gearbox fault diagnosis through the blind mixing signal separation. The algorithm based on the natural gradient fixed step-length was used to calculate the statistical independent source signal estimate value, and successfully extracted the fault information according to the separation signal power spectrum based on the minimum mutual information (MMI) criterion. The gearbox fault condition can be diagnosed effectively through the experiment proved, which provided a new method to the mechanical equipment fault diagnosis and running state monitor.

Keywords: *BSS, MMI, natural gradient, mechanical equipment, fault diagnosis.*

1. Introduction

Through the monitoring and analysis to the mechanical equipment vibration noise, according to the noise sound level and the frequency changing to judge the fault position and reason has become one of the important means and methods, which has been widely applied. Among them, blind signal separation (BSS) [1] can separate each source signal estimation value under the assumption that each source signal is statistical independence each other from the mixed signal sample. At present, the several typical blind signal separation algorithms basically have Sejnowski and Bell's Infomax algorithm [2], natural gradient algorithm, Cardoso's EASI algorithm [3], inverse iteration algorithm, JADE algorithm [4] and Hyvarinen's Fast ICA algorithm [5], etc.

Blind signal separation algorithm is to establish cost function based on the information theory, higher order statistics[6], etc, and to optimize the objective function by using the optimization algorithm. Among them, the gradient algorithm is a kind of classic unconstrained optimization algorithm with the simple principle, and easy to realize with the equal variation characteristics, which can realizes the online calculation. So, the gradient algorithm [7] are widely used in the blind signal separation filed.This paper uses the technology to separate the aliasing vibration noise into irrelevant signals, thus provide the basis for the diagnosis of gearbox fault monitoring.

2. Blind source separation principle

Blind source separation can be expressed as the following formula. [8]

$$X(t) = AS(t) + N(t) \tag{1}$$

Among the formula (1), $S(t) = [S_1(t), S_2(t)...S_n(t)]^T$ is n d source vector, $X(t) = [x_1(t), x_2(t)...x_m(t)]^T$ is m d observation signal vector, $\boldsymbol{A}$ is m×n d mixing matrix, the elements express the mixed signal, $N(t) = [n_1(t), n_2(t)...n_m(t)]^T$ is m d noise. Blind source separation's goal is to estimate a separation matrix only according to the observation signal $X(t)$ without any prior knowledge.

$$Y(t) = WX(t) \tag{2}$$

Among the formula (2), $Y(t) = [y_1(t), y_2(t)...y_n(t)]^T$ is n d separation signal vector, if *WA*=*P*, *P* is a replacement array, which to get the aim to recover source signal. [9]

3. BSS algorithm based on MMI

The basic method of minimum mutual information [10] is to select suitable neural network weight matrix $\boldsymbol{W}$ to make output $Y(t)$ each component has minimum dependence. In an ideal situation until tends to zero, which to achieve the purpose of separation. Thus, we can use entropy to express the dependency among the signals [11]. And we can adopt the formula (3) to get.

$$I(W) = -H(X) - E\{\log|\det(W)|\} + \sum_{i=1}^{n} H(y_i;W) \qquad (3)$$

We can make the mutual information amount $I(\boldsymbol{W})$ into minimum when selecting $\boldsymbol{W}$, which is MMI criterion. Because input signal entropy $H(Y;W)$ has nothing with the selecting $\boldsymbol{W}$. So, we can make it simply and build the cost function shown as formula (4).

$$\begin{aligned} L(\boldsymbol{W}) &= -E\{\log|\det(W)|\} + \sum_{i=1}^{n} H(y_i;W) \\ &= -E\{\log|\det(W)|\} - E\left\{\sum_{i=1}^{n} \log p(y_i)\right\} \qquad (4) \end{aligned}$$

If we use different nonlinear transformation $g_i(y_i)$ to each component $y_i = \sum{}^{'} wi_j x_j$, and make $Z = (g_1(y_1),...,g_n(y_n))$ is output after transformation. Then the combination entropy $H(Z;W)$ of each component of Z can be expressed as formula (5).

$$H(Z;W) = H(X) + E\{\ln|\det(W)|\} + \sum_{i=1}^{n} E\{\ln g_i^{'}(y_i)\} \qquad (5)$$

We can get natural gradient of combination entropy $H(Z;W)$ relative to separation matrix $\boldsymbol{W}$.

$$\begin{aligned} \frac{dW}{dt} &= \eta(t)\frac{\partial H(Z,W)}{\partial W} W^{-T}W \\ &= \eta(t)(W^{-T} - E\{\phi(Y)X^T\})W^{-T}W \qquad (6) \end{aligned}$$

Among the formula (6), $W^{-T} = (W^{-1})^T$, $\phi(Y) = (-\frac{g_1^{''}(y_1)}{g_1^{'}(y_1)}),...,-\frac{g_n^{''}(y_n)}{g_n^{'}(y_n)})$, $\eta(t)$ is learning step length, expectation item $E\{\phi(Y)X^T\}$ can be instead of instantaneous value $\phi(Y)X^T$. Then the adaptive iteration formula to $\boldsymbol{W}$ is shown as formula (7).

$$W(t+1) = W(t) + \eta(t)(1 - \phi(Y)Y^T)W(t) \qquad (7)$$

Formula (7) is iterative formula of natural gradient algorithm, it is thus clear that natural gradient algorithm can avoid $\boldsymbol{W}$ inversion, which making the calculation amount decreasing.

4. Gearbox fault diagnosis

Based on the minimum mutual information criterion of natural gradient separation algorithm, we carry out simulation to a gearbox fault diagnosis. [12] From the gearbox structure, it is comprised of the shaft, bearings, gears and spare parts, etc. So, we can acquire the vibration signal through installing sensors upon the gearbox, the collection signal is consisted by all kinds of vibration source signal and other noise interference, this kind of compound may be additive and multiplicative or other some more complex form. For simplicity, we can assume that the signal is linear through the transmission of the gearbox, the gearbox vibration signal can be said for the linear superposition signal of the gear meshing signal, bearing signal and noise signal, etc. According to the theoretical analysis, the gear meshing signal $s_1(t)$ can be represented as:

$$s_1(t) = \sum_{k=1}^{K} A_k(t)\cos(2\pi k f_m t + \varphi_k(t)) \qquad (8)$$

Among the formula (8), f_m is meshing frequency, $\varphi_k(t)$ is phase, k is harmonic order times, and bearing fault vibration signal can be expressed as formula (9).

$$s_2(t) = \sin(2\pi k f_f t)(1 + \beta\sin(2\pi f_c)) \qquad (9)$$

Among the formula (9), f_f is the fault characteristic frequency related with fault, f_c is shaft speed frequency. Meshing frequency f_m is integer times of the shaft speed frequency, the two frequences are related. Fault characteristic frequency f_f is rolling element bearing ring rolling in the drive frequency or its harmonic component, it is not relevant with the meshing frequency and shaft speed frequency. But the fault vibration signal $s_2(t)$ is fault characteristic signal or its harmonic and bearing vibration signal superposition, such as the formula (9) shows, so the gear meshing vibration signal $s_1(t)$ and fault vibration signal $s_2(t)$ is not related, they can be regarded as two independent component in blind separation model. [13]

For the meshing frequency f_m 1168 Hz, the gear shaft speed for 2920r/min, the rolling body spalling fault frequency f_f 256 Hz, [14] and we sample the signal in 10KHz in 0.4s, the waveform is shown as Fig. 1.

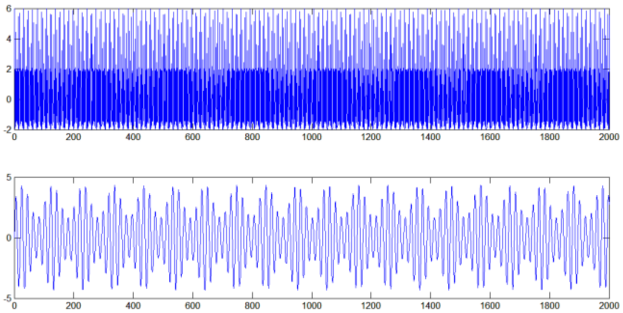


Fig. 1 Source signal.

The gear meshing vibration signal $s_1(t)$ and fault vibration signal $s_2(t)$ are blind mixed into the observation signal, which is respectively shown as Fig. 2.

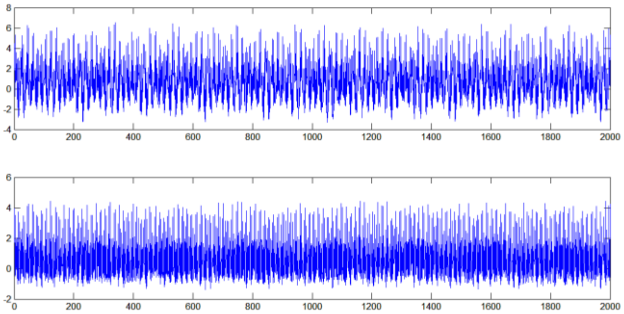


Fig. 2 Observation signal.

The observation signal power spectrum is shown as Fig. 3. The traditional vibration signal processing method is directly obtained and diagnosed through the observation signal. Because the characteristics of vibration source signal in the sensor is obtained during the mutual mixed or various nonlinear distortion, the noise jamming is big, transmission channel is complex, which often can not obtain the very good separation effect. [15]

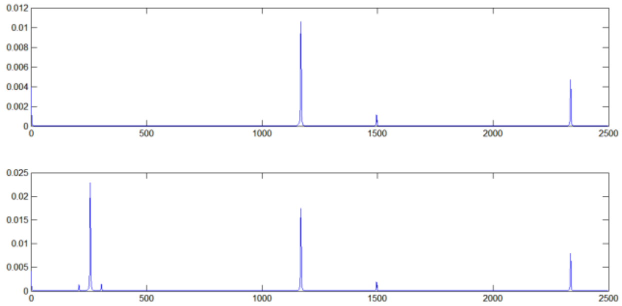


Fig. 3 Observation signal power spectrum.

We can separate the mixed signal by using the natural gradient algorithm, the separation signal is shown as Fig. 4.

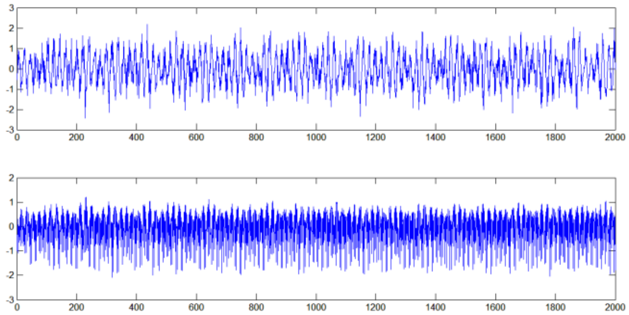


Fig. 4 Separation signal waveform.

When we carry out the power spectrum to the separation signal, the power spectrum of the separation signal is respectively shown as Fig. 5.

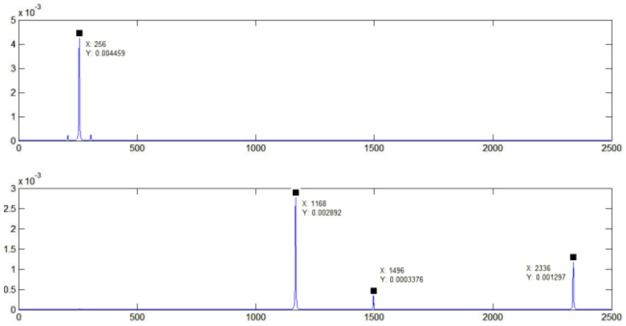


Fig. 5 Separation signal power spectrum.

From the Fig. 5 we can see, the power spectrum of the separation signal is approximation to the power spectrum of the source signal. The bearing fault frequency of 256Hz and the gear meshing frequency of 1168 Hz, its 2 times harmonic frequency can be reflected in the separated signal power spectrum. In addition, 1496 Hz signal is the composition result of the fault frequency and the spindle rotation frequency. The comparison result between the separation signal frequency and source signal is shown as Table 1.

Table 1: Comparison between the separation signal and the source signal

Data Type	f_f **(Hz)**	f_m **(Hz)**	**2** f_m **(Hz)**	≈ $f_f + f_m$ **(Hz)**
Source signal	256	1168	2336	1496
Separation signal	256	1168	2336	1496

From the Table 1 we can see, the frequency separation result by MMI algorithm this paper presented is same as the source signal, which proved the algorithm has high separation precision. And we can judge the fault reason by using the algorithm.

5. Conclusion

In view of the gradient blind separation algorithm is a kind of online algorithm, which can realize the vibration signal online separation, and through the experimental analysis, the natural gradient algorithm can better separate the vibration signal, and we can adjust the value of the separation matrix according to the system's model characteristic, and get better separation results. Because the natural gradient algorithm has the equal changing characters, the amount of calculation, convergence speed and steady state performance is contradictory, step length in selecting also need comprehensive consider.

We used the BSS algorithm this paper presented to carry out the fault signal separation and fault diagnosis, the results show that based on the minimum mutual information criterion, the natural gradient blind source separation method not only can well separate blind mixed signal, but also can effectively realize the gearbox fault type diagnosis, which express the method has broad application prospects in the rotating mechanical equipment fault diagnosis.

Acknowledgments

This paper is supported by Youth Scientific Research Fund Project of Zhengzhou Institute of Aeronautical Industry Management. (No. 2012133001)

References

[1] Zhang Xianda, Bao Zheng, “Blind Signal Separation”, Acta Electronica Sinica, Vol. 29, No.12A, 2001, pp. 1766-1771.

[2] Li Xiaojun, Li Xiaolong, Zhang Xianda, Blind source separation: classificaiton and frontiers, Journal of Xidian University, 2004, 31(3) : 399- 404.

[3] Bell A J, Sejnowski T J. An information-maximization approach to blind separation and blind deconvolution[J]. Neural Computation, 1995, 7(6): 1129-1159.

[4] Cardoso J.F, Lalleld B. H, Equivariant adaptice source separation, Signal Processing, IEEE Transactions, Vol. 44, No. 12,1996: 3017-3030.

[5] Hyvarinen A, A fast fixed-point algorithm for Independent component analysis, Neural Computation, 1997, 9 (7) :1483 – 1492.

[6] Aapo Hyvarinen n, Independent component Aanlysis: Algorithm and Applications, Neural networks, 2000(13): 411-430.

[7] Hyvarinen A, Fast and robust fixed-point algorithms for independent component analysis, IEEE Trans, Neural Networks, 1999,10(3): 26-34.

[8] Sun Shouyu, Zheng Junli, Wu Dewei, Research on Blind Source Separation Based on Natur al Gradient Algorithm, Journal of Air Force Engineering University(Natural Science Edition), No. 3, Vol. 4, 2003: 50-54.

[9] Jiao Fangfang, Feng Zhihong, Yang Guiqin, Research on Blind Source Separation and Blind Signal Extraction, Radio Engineering,2011, Vol. 41, No. 9, pp. 11-14.

[10] Wu Zuolun, Yang Shixi, Study on mechanical noise separation based on technique of blind source, Journal of Zhejiang University of Science and Technology, Vol. 15, No. 4, 2003, pp. 219-223.

[11] Fu Weihong, Yang Xiaoniu, Novel algorithm for step size adaptive blind source separation based on natural gradient, J. Huazhong Univ. of Sci. & Tech. (Nature Science Edition), No. 10, Vol. 35, 2007, pp.18-20.

[12] Chen Zhongsheng, Yang Yongmin, Shen Guoji, Application of Independent Component Analysis to Early Diagnosis of Helicopter Gearboxes, Mechanical Science and Technology,2004, Vol. 23, No. 4: 481-483.

[13] Shi Qingbin, Ma Jiancang, Study on Blind Source Separation of Mechanical Vibration Signal, Measurement & Control Technology, 2008, Vol. 27, No. 5, pp.78-80.

[14] Zhang Yongxiang, Ming Yantao, Wang Honglei, Application of blind source separation based on minimum mutual information to gearbox fault diagnosis, Machinery Design & Manufacture, No. 6,2006 :87-89.

[15] Ma Jiancang, Shi Qingbin, Zhao Shuyuan, Zhang Qunfang, Blind Source Separation for Nonlinearly Mixed M echanical Vibration Signals,Noise and Vibration Control, No.6, 2008, pp. 5-8.

Yu Chen received the Master degree in data collection and signal processing from Northwestern Polytechnical University, in 2009. Currently, he is an Associate Professor at Zhengzhou Institute of Aeronautical Industry Management. His interests are in data collection and signal processing and nonlinear system modeling.

HAITAO JIANG received the Master degree in Circuit and System professional from Northwestern Polytechnical University, in 2009. Currently, he is an Lecturer at Department of Physics and Electronics engineering, Jiaozuo Teachers College. His interests are in intelligent sensor data collection and signal processing.

The Mathematical Statistics Theory Application on the Price Fluctuation Analysis

Jintao Meng[1], Jing Li[2]

[1] **Zhengzhou Institute of Aeronautical Industry Management**
Zhengzhou, 450015, China

[2] **Zhengzhou Railway Vocational & Technical College**
Zhengzhou, 450052, China

Abstract

Grain price and output fluctuation are the normal state of market economy. It is one of the most important economic researches to understand grain price and output fluctuation law, which provides theory basis for the macroeconomic regulation and control. According to the cobweb model theory, the relationship between citrus production and price is accord with the divergence type of cobweb model .This means that simply relying on market regulation can make fluctuation between production and price bigger, go against citrus production and cultivation, thus, affecting the interests of farmers. It is well-known most farmers are concerned about the future price trend and the probability of price fluctuation. This paper uses mathematical statistics theory to study the citrus price changes, and the corresponding change trend, providing a theoretical basis for majority of farmers to better estimate citrus price change trend.

Keywords: *cobweb theory, citrus price fluctuation, mathematical statistics, confidence interval.*

1. Introduction

Currently, more and more attention to the development of the agricultural economy. In such a good market background, it is the theme that the farmers expand production. Each region drastically increases corresponding capital investment. But the market is ruthless. Like roller coaster, citrus fruit price is elusive, and many large investments only get more debt, which becomes the burden of some farmers.

Combining with China Statistical Yearbook 2009, related data of citrus price and production released by the China Citrus Web in 2009, and the cobweb theory[2,3,4], we derive the relationship between citrus production and market price , see [5, 6]:

$$\begin{cases} D_t = 2850.02 - 527.96P_t \\ S_t = 2935.9 - 546.72P_{t-1} \end{cases}$$

Where D_t : the demand function, S_t : the supply function, P_t : the price at t time, P_{t-1} : the price at $t-1$ time.

According to the cobweb model theory, this conforms to the second case of the cobweb model. That is, when $|\mu|>|\beta|$, $\lim_{t\to+\infty} p_t$ does not exist, and is tend to be infinite. This shows that with the passage of time, the change in price range is bigger, and the actual price will be getting further away from the equilibrium one, see [1]. Only relying on market to regulate the citrus production would lead to its price and yield far from the equilibrium point, see [7]. The citrus price fluctuation every year may make a phenomenon of “low price hurting farmers " happen repeatedly.

The citrus farmers are most concerned about citrus price prediction and the corresponding possibility. Combining and using the mathematical statistics, the paper will analyze previous price data, and predict the citrus fruit prices and the corresponding probability to provide more scientific price prediction.

2. Mathematical Statistics Preliminaries

As a discipline born in the turn of 20th century, mathematical statistics is widely used as a branch of mathematics. Based on probability theory, mathematical statistics uses experimental and observational data to study random phenomenon, so as to make reasonable estimation and judgment on the objective law of the research object.

Because a large number of random phenomena will necessarily show its regularity, theoretically, observation of the random phenomena enough times will make the regularity of the research object clearly present. But in reality, people are often unable to observe all the research objects(or called overall), only observe or test some (or called the sample) to obtain limited data.

In mathematical statistics certain common objects are collectively referred to as the overall, and its size and scope is decided by the specific research and investigation purpose. Each member constituting the overall is called individual. The number of individuals in the overall is referred to as the overall capacity. The overall distribution is generally unknown. Sometimes the distribution type (such as common normal distribution, binomial distribution, etc.) can be known, but the specific parameters of the distribution (such as expectation, variance, etc.) are unknown. And the task of mathematical statistics is to statistically deduce the unknown distribution of the overall according to some individual data. In order to judge what distribution the overall obeys or estimate what value the unknown parameters should take, we can extract several individuals from the overall to observe, get some data to study the overall, and judge the distribution of the overall and make reasonable estimate on unknown parameters through statistically analyzing the data. The general method is according to certain principles to extract several individuals from the overall to observe. This process is called sampling.

Obviously, the observation result of each individual is random, which can be regarded as a random variable value. The $i-th$ individual indicators extracted from overall X is regarded as $X_i(i=1,2,\ldots,n)$, then X_i is a random variable; and $x_i(i=1,2,\ldots,n)$ can be used as specific observation value of individual index X_i. The X_1, X_2,... , X_n are sample values, individual number in the sample is called sample capacity (or sample size).

Normally, It is assumed that the sample is the one independent and identically distributed.
The overall and sample are two basic concepts in mathematical statistics. On the one hand, the sample is from the overall, naturally with information of the overall, and based on that, some of the characteristics of the overall (distribution or the parameters of distribution) can be studied. On the other hand, using sample to study the overall can be time-saving (especially for destructive sampling experiment). The problem of deducing distribution of the overall X by a sample X_1, X_2,... , X_n is known as the problem of statistical inference.

In order to deduce the overall from the sample, some appropriate statistic is needed to be constructed, by which unknown overall is got. Such sample statistics is sample function. To construct an excluding overall is a function of unknown parameters for the sample statistics. A function of unknown parameters is constructed as the sample statistics for the sampled sample. Commonly used statistics have:

Sample mean： $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}X_i$.

Sample variance： $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$

The point of origin of Sample (k order)：

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^{\,k}, k=1,2,\cdots$$

The sample’s original dot pitch (k order)：

$$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k, k=2,3,\cdots$$

The commonly used statistical distribution: standard normal distribution, t - distribution, χ^2 - distribution, F - distribution. These distributions have formal tables to look for quantile, or to find the probability value by quantile, so they are commonly used.

There are two ways to estimate the unknown parameter in the overall distribution: one is point estimation, another is called interval estimation: (1) point estimate: if the overall distribution is known, but some parameter in it is unknown, the method of giving reasonable estimate of the unknown parameters from the extracted sample is point estimation, The commonly used point estimation method is distance estimation and maximum likelihood estimation. Distance estimation method is to make the sample distance of the corresponding order number to approximate the overall distance according to the number of unknown parameters of the overall, and solving simultaneous equation can get unknown parameter approximation value. Maximum likelihood estimation method is based on the maximum likelihood function of independently and identically distributed property and structure from the sample, then finds out extreme value point, and lets the extreme value approximate the unknown parameters. (2) the interval estimation - confidence interval: suppose θ is the unknown parameter of the overall distribution , X_1 , X_2,... , X_n are samples taken from the overall X, for a given number $1-\alpha,(0<\alpha<1)$, if the existence of statistics value $\varsigma=\varsigma(X_1,X_2,\cdots,X_n)$, $\xi=\xi(X_1,X_2,\cdots,X_n)$ makes $P\{\varsigma<\theta<\xi\}=1-\alpha$, random interval (ς,ξ) is called $1-\alpha$ bilateral confidence interval of θ, $1-\alpha$ is the confidence degree, and then ς and ξ are respectively the bilateral lower confidence limit and bilateral upper confidence limit of θ.

The meaning of confidence degree $1-\alpha$: If repeatedly sampled many times in the process of random sample, multiple sample values X_1, X_2,... , X_n can be got. Each corresponding sample value has aconfidence interval (ς,ξ). Each interval contains either a truth value of θ or not. According to Bernoulli large numbers law, When the sampling frequency k is sufficiently large, the frequency of true value θ in the interval is close to the confidence degree $1-\alpha$. That i-s, the number of the intervals containing the true value of θ are about $k(1-\alpha)$, and the number of the intervals not containing the true value θ are about $k\alpha$, e.g. Suppose $1-\alpha = 0.95$, repeatedly sample-d 100 times, about 95 intervals contain the true value of θ, and about five interval does not contain the true value of θ.

Confidence degree and estimation precision are a pair of contradiction, the greater the confidence degree, the greater the probability of the true value θ in the confidence interval (ς,ξ), the longer t-he interval (ς,ξ), and the lower the estimate accuracy of unknown parameter θ is. Conversely, the higher the estimate accuracy of the parameter θ , the smaller the length of the confidence interval (ς,ξ), the lower the probability of the true value in (ς,ξ), and the smaller the confidence degree $1-\alpha$ is. The general rule is: in the guarantee of confidence degree, the estimation accuracy should be improved.

The common method of seeking confidence interval is: on the basis of point estimate, appropriate function U containing samples and parameters to be estimated should be constructed. And confidence interval can be derived from the known function U and the given confidence degree.

The general steps: (1) Select Some better estimator $\hat{\theta}$ from unknown parameter θ. (2) Center on $\hat{\theta}$ and construct a function $U = (X_1, X_2, \cdots, X_n, \theta)$ that depends on the sample and parameter θ; And if the distribution of the function is known (independent of θ), the random variable with such nature is called the pivot function. (3) For a given confidence degree $1-\alpha$, determine λ_1 and λ_2, and make $P\{\lambda_1 \le U \le \lambda_2\} = 1-\alpha$. When λ_1 and λ_2 in $P\{U \le \lambda_1\} = P\{\lambda_2 \le U\} = \frac{\alpha}{2}$ are selected, generally, the quantile table can be used. (4) After identical deformation of the inequality $\lambda_1 \le U \le \lambda_2$, it turns into $P\{\varsigma < \theta < \xi\} = 1-\alpha$, and then (ς,ξ) is the bilateral confidence interval of the confidence degree $1-\alpha$ of θ. In some practical problems, $P\{U \le \lambda_1\} = \alpha$ or λ_1 and λ_2 in $P\{\lambda_2 \le U\} = \alpha$ are only needed to meet, after identical deformation of the inequality, it becomes $P\{\varsigma < \theta\} = 1-\alpha$ or $P\{\theta < \xi\} = 1-\alpha$, so confidence interval similar to $(\varsigma, +\infty)$ or $(-\infty, \xi)$ can be got.

3. Data Analysis

Based on the cobweb model theory, price change of citrus is accord with divergent situation of the cobweb model. Further analysis is needed to get the price change trends on the basis of the original price information, and the possible ranges of values, providing a theoretical analysis for the next planting plans.

To analyze the pricing trend of citrus, we have a statistics of the wholesale prices in some area (Unit: 50 Yuan per 50kg). Due to the bigger price change in individual time periods, it is obviously not appropriate to analyze the daily price. We have 30 different prices for statistical analysis, and divide the prices into eight price intervals. Although some errors mean values of grouped sample and variance of the approximate calculation are used in mathematical statistics, after the analysis of the practical problems, the errors are in the acceptable range.

Table 1: the statistics data

Price change interval	Statistical number n_i	Price mean value x_i	$n_i x_i$
[120,130)	1	125	125
[130,140)	3	135	405
[140,150)	6	145	870
[150,160)	14	155	2170
[160,170)	4	165	660
[170,180)	1	175	175
[180,190)	0	185	0
[190, 200)	1	195	195
total	30		4600

price mean of the drawn-out sample can be easily calculated:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{4600}{30} \approx 153.33$$

Similarly, the variance approximation value of the sample can be calculated:

$$S^2 \approx \frac{1}{n-1}[\sum_{i=1}^{k} n_i x_i^2 - \frac{1}{n}(\sum_{i=1}^{k} n_i x_i)^2]$$
$$\approx 172.99.$$

Suppose price fluctuation of citrus accords with most widely normal distribution in nature, if $X \sim N(\mu, \sigma^2)$, among which μ and σ^2 are unknown, the above sample sampling is made to construct pivot function:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

If the confidence degree $1-\alpha = 0.95$, then $T = \frac{153.33-\mu}{13.15/\sqrt{30}} \sim t(29)$, we can obtain the confidence interval [148.41, 158.25] on the confidence degree 0.95.

If the confidence degree is $1-\alpha = 0.99$, then $T = \frac{153.33-\mu}{13.15/\sqrt{30}} \sim t(29)$, we can obtain the confidence interval [146.71, 159.95] on the confidence degree 0.99.

If the confidence degree is $1-\alpha = 0.8$, then $T = \frac{153.33-\mu}{13.15/\sqrt{30}} \sim t(29)$, we can obtain the confidence interval [150.19, 156.47] on the confidence degree 0.8.

According to different requirements, a different confidence degree can be chosen, so as to find the desired confidence interval.

4. Conclusion

Combining with the price information of citrus in some area, the paper used mathematical statistics to predict the price trend. It also can predict more agricultural prices and the corresponding probability, make farmers of economic crops predict prices of agricultural products more purposefully and scientifically and adopt active planting ways to reduce risk and get more income.

References

[1] Richard H Dai. et al. Chaos economics, Shanghai Translation Publishing House, 1996.

[2] Li Zhongmin and Zhang Shiying, "Nonlinear cobweb model of dynamic analysis". Quantity Economic Research, No. 2, 1997, pp. 45-51.

[3] Huang Zelin, "Dynamic Analysis on nonlinear and Disequilibrium Cobweb Model ". Mathematics in Practice and Theory, Vol. 34, No. 3, 2004, pp. 40-45.

[4] Wang Jun and Yang Fuchun, " Some Sufficient and Ecessary Conditions of Convergent Cobweb Model ". Mathematics In Economics, Vol. 23, No. 4, 2006, pp. 364-369.

[5] He Jin and Qi Chunjie, "Empirical Study on Price Formation and Profit Distribution of Citrus in China". Journal of Northwest Agriculture and Forestry University (Social Sciences Edition), Vol. 9, No. 6, 2009, pp. 36-43.

[6] Lu Xiaoxu and Zhang Jie, "Based the cobweb model theory citrus production and price fluctuations analysis". The rural economy, No. 8, 2010, pp. 60-62.

[7] Shao Lu and Sheng Yajun, "Based on the the Cobweb model China's grain price determination mechanism analysis", Agricultural Economics, Vol. 22, No. 9, 2011, pp. 110-111.

Jintao Meng received the Master degree from Zhengzhou University in 2008, Operations research and control theory professional. Lecturer of Zhengzhou Institute of Aeronautical Industry Management.

Jing Li graduated from Zhengzhou University in 2004, Operations research and control theory professional. Lecturer of Zhengzhou Railway Vocational & Technical College.

Fault Tolerant Circuit Design Using Evolutionary Algorithms

Hui-Cong Wu
School of Information Science and Engineering, Hebei University of Science and Technology
Shijiazhuang 050018, China

Abstract

With the rapid development of semiconductor technology and the increasing proliferation of emission sources, digital circuits are frequently used in harsh electromagnetic environments. Electrostatic Discharge (ESD) interferences are gradually gaining prominence, resulting in performance degradations, malfunctions and disturbances in component or system level applications. Conventional solutions to such problem are shielding, filtering and grounding. This paper presents an evolvable hardware platform for the automated design and adaptation of a motor control circuit. The platform uses EHW to automate the configuration of FPGA dedicated to the implementation of the motor control circuit. The ability of the platform to adapt to certain number of faults was investigated through introducing single logic unit fault and multi-logic unit faults. Results show that the functionality of circuit can be recovered through evolution. It also shows that the placement of faulty affect the ability of GA to evolve correct circuit, and the evolutionary recovery ability of the circuit descends with the number of fault units increasing.

Keywords: *Evolvable Hardware, Fault Tolerant, Motor Control Circuits*

1. Introduction

Brushless motor are frequently employed in the speed regulation of many driving systems. The performance and sustained reliability of the motor control circuit are of great importance. Usually the control circuits designed in SCM or DSP are easy to damage in extreme environmental conditions, such as electromagnetism interference and high-energy radiation.

Recently, fault tolerant systems are widely used in space applications where hardware deteriorates due to damages caused by aging, temperature drifts and high-energy radiation. In this case, human intervention is difficult or impossible; the systems must therefore maintain functionality themselves. Conventional fault tolerant systems employ techniques such as redundancy, checking-pointing and concurrent error detection. These techniques all rely on the presence of additional redundant and add considerable cost and design complexity. In most cases, it can't satisfy the application requirements [1].

As a newly emerging but promising research field, evolvable hardware (EHW) [2-4] may provide alternative approaches and new mechanisms for the design of fault tolerant systems. EHW is based on the idea of combining reconfigurable hardware devices with GA to perform reconfiguration autonomously. Which refers to the characteristics of self-organization, self-adaptation and self-recovery. With the use of evolutionary computation, evolvable hardware has the capability of autonomously changing its hardware architectures and functions. It can maintain existing function in the context of degradations or faults in conditions where hardware is subject to faults, temperature drifts, high-energy radiation, or aging.

As to logic or digital circuits, gate-level evolution usually takes logic gates as the basic units or building-blocks. Many researchers in this field prefer extrinsic evolution at gate-level because it is generally applicable to various circuits and its outcomes are comparatively formal and consequently analyzable. Many encouraging results for gate-level evolution of logic circuits have been demonstrated [5]. Nanjing University of Aeronautics and Astronautics has had the online fault tolerant evolution of digital circuits and analogy circuits on FPGA and FPTA respectively [6-8]. The Jet Propulsion Laboratory (JPL) performs research in fault tolerant, long life, and space survivable electronics for the National Aeronautics and Space Administration (NASA). JPL has had experiments to illustrate evolutionary hardware recovery from degradation due to extreme temperatures and radiation hardware environments. Their experiment results demonstrate that the original functions of some evolved circuits, such as low-pass filters and the 4-bit DAC, could be recovered by reusing the evolutionary algorithm that altered the circuit topologies [9-11].

This paper presents an evolvable hardware platform for the automated design and adaptation of brushless control circuits. The platform employs a genetic algorithm to autonomously configure the FPGA dedicated to the implementation of the motor control circuit. The ability of the platform to adapt to a certain number of faults was investigated through introducing single logic unit fault and multi-logic unit faults.

The paper is organised as follows: section 2 presents the architecture of the fault tolerant platform. Section 3 describes the evolutionary design process, such as the chromosome representation, the design of fitness function, and the adaptation strategy for GA parameters. Section 4 illustrates experiments on fault tolerance through evolution of the brushless motor control circuit. Concluding and future research are given in section 5.

2. Fault Tolerant Platform

The motor control circuit is selected as an initial study experiment objects. The motor achieves the phases changing operation with electronic circuit. The control system structure is illustrated in Fig. 1. It includes three parts, the motor control circuit, the drive module and the brushless motor itself. The brushless motor checks the position of the rotors by using 3 location sensors. It produces three position feedback signal *S0, S1* and *S2*. When the Rotor rotates 360 degrees along the same direction, the position signal *S0, S1* and *S3* have a total of six states combination as shown in Table 1. The motor control circuit triggers each switch (*M0, M1, M2, M3, M4, M5*) in the drive module in accordance with the certain order.

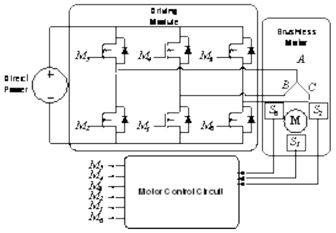


Fig. 1 The motor control system.

The motor control circuit fault tolerant evolution environment is shown in Fig. 2.The platform comprises of FPGA, GA evolution module, VHDL coding conversion module and FPGA development tool software environment.

Alter EP1K50 FPGA, which is capable of partial dynamic reconfiguration, was adopted as the experiment hardware. It provides a Joint Test Action Group (JTAG) system interface connected to the computer parallel port, through which the circuit configuration bits can download to FPGA to validate its functionality.

The evolution module is the core part of the system. Circuit structure is represent by chromosome. The simulated evolution is used to evolve a good set of architecture bits that determine the functions and interconnections of the logic units in FPGA.

The VHDL coding conversion module together with Quartus II integrated software can realize the conversion from chromosome representation to circuit structure. The internal configuration structure of FPGA chips is unknown for normal users. After the best chromosome is derived from generations of genetic operation in evolution module, it is expressed by VHDL; then Quartus II (which is an integrated software environment developed by the FPGA providers) compile and translate the VHDL to FPGA configuration bits. In QUARTUS II environment, TCL script language can be used together with QUARTUS command to accomplish the whole process from formulation of VHDL program to download of FPGA configuration bits.

3. Evolutionary Circuit Design

Evolutionary algorithms are used for circuit design. Circuit representation, fitness evaluation, and parameters choice are crucial ingredients of effective evolutionary circuit design.

3.1 Chromosome Representation

A correct circuit representation is the base for effective design. There are many approaches to circuit description, such as binary code, min-term code and functional-level code. The direct approach to EHW encodes circuit's architecture bits as chromosomes, which specify the connectivity and functions of different hardware components (of the gate level) of the circuit.

According to the motor control circuits, there are three bits inputs *S0, S1* and *S3* which represent the feedback signals of the rotors' position, and six bits outputs (*M0, M1, M2, M3, M4, M5*) which control the six switches of the driving module to ensure that the rotor can change to next position correctly.

Fig.2 shows the computational model for gate-level evolution of the brushless motor control circuit. The evolution area is an array of 8*5. Because the first column works as inputs and the last as outputs, the two columns won't participate in the evolution. The actual evolutionary area is the form of a rectangular array that consists of logic units in 8 rows by 3 columns. Each logic unit comprised has 2 inputs, one output and perform 4 functions AND, OR, NAND, NOR.
H0 is the input column, there are 3 logic units which accept the primary inputs *S0, S1* and *S2* respectively，*H1,H2,H3* are implication evolution columns，there are 8 logic units in each column, the total 24 logic units are the redundancy resources to be evolved. *H4* is the output column, logic units in this column, which act as the 6 interfaces, connect to the outputs of the circuit *M0, M1, M2, M3, M4, M5*.

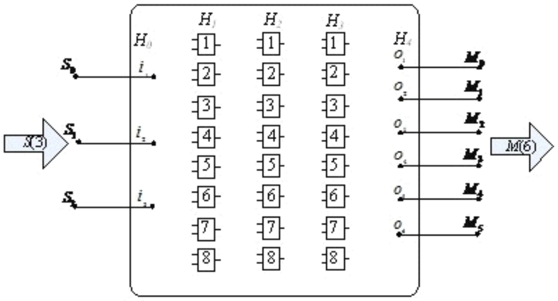


Fig. 2 The computational model of motor control system.

The configuration array which represents interconnections and functions of the logic units is shown as following:

$$C_{0,1}=\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,l} \\ b_{1,1} & b_{1,2} & \cdots & b_{1,l} \\ w_{1,1} & w_{1,2} & \cdots & w_{1,l} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,l} \\ \vdots & \vdots & \vdots & \vdots \\ w_{m,1} & w_{m,2} & \cdots & w_{m,l} \end{bmatrix} \qquad (1)$$

$$C_{k-1,k}=\begin{bmatrix} a_{k,1} & a_{k,2} & \cdots & a_{k,l} \\ b_{k,1} & b_{k,2} & \cdots & b_{k,l} \\ w_{1,1} & w_{1,2} & \cdots & w_{1,l} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,l} \\ \vdots & \vdots & \vdots & \vdots \\ w_{l,1} & w_{l,2} & \cdots & w_{l,l} \end{bmatrix}(2\leq k\leq K-1) \qquad (2)$$

$$C_{K\text{-}1,K}=\begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ w_{l,1} & w_{l,2} & \cdots & w_{l,n} \end{bmatrix} \qquad (3)$$

The configuration array includes two parts, the functional array and the connectional array, the functional array expressed as aij in the first two rows represents the functions of logic units in each column. The connectional array expressed as wij in the rest rows represents the interconnections of each logic units in current column with the logic units on its next left column.

Each logic unit comprised 4 functions can be encoded in column vector format $(a_{k,i},b_{k,i})^T$, suppose there are L logic units in each column, the function column vector of L logic units compose the functional array, which can be expressed as following:

$$F_{k-1,k}=\begin{bmatrix} a_{k,1}\ a_{k,2}\cdots a_{k,l} \\ b_{k,1}\ b_{k,2}\cdots b_{k,l} \end{bmatrix}(1\leq k\leq K-1) \qquad (4)$$

The column vector $(a_{k,i},b_{k,i})^T$ express the function of the logic unit in column k and row i , $1\leq k\leq K-1$,and $1\leq i\leq l$ 。As outputs interfaces, there is no according function array in column k.

In formulation (1), C0,1 represents the configuration array between column H0 and H1. Here the value of m is 3 according to 3 inputs in column H0; the value of L is 8 according to 8 logic units in column H1. Wi,j represents the connection relationship of logic unit in column H1 with each logic unit in previous column H0. the value of Wi,j is '1' or '0'.

In formulation (2), the configuration array Ck-1,k represents the configuration array between column Hk-1 and Hk. The value of K is 8 according to the total column number. In combinational circuits, feedback is prohibited; and the logic units in column n only receive inputs from the next left column n-1 and not allowed to receive inputs from others. In the configuration array Ck-1,k, wi,j='1' represents the logic unit in column k-1and row i is connected to the logic unit in column k and row j. On the contrary,' wi,j='0' represents these two logic units aren't connected. The value of l is 8 according to 8 logic units in each column.

In formulation (3), Ck represents the connectional array between column Hk-1 and Hk. Act as outputs interfaces, there is no according function array in column K, the value of n is 6 according to the 6 outputs.

To ensure that each pair of configuration array corresponds with one correct logic circuit, some limitation rules should be observed:

①In the connection array $C_{0,1}$, it is prohibited that all of the elements in each row are '0'in order to ensure that the primary circuit inputs *S0*, *S1*, *S2* can be all connected to the first column *H1*.
②There should be at least two '1' in each column of the connection array, because each logic unit input is routed to only two units from the next left column.
③Since the output layer only achieves the signal output function, the connected matrix in each column must have only one element that is '1', in order to ensure that all components of the system output can be connected with the logic unit.

If the configuration array violates above limitation rules, it corresponds to a invalidate circuit. In such cases we set the fitness to '0' in the evolution process.

3.2 Fitness Evaluation

For problems of gate-level evolution, design objectives mainly include expected functions, efficiency of resource usage (in terms of gate count) and operating speed of circuits (estimated with Maximal Propagation-Delay (MPD)). Although a functionally correct circuit with fewer logic gates and fewer number of gates contained in the longest signal chain of the circuit is usually preferable, the main purpose in this paper is to investigate the capacity of fault recovery using EHW in case of faults. Therefore, the design objective only concerns with the expected functions or behaviors, which have been specified by truth table in section 2. Thus, the functional fitness value of the evolved circuit is calculated as

$$F=\sum_{i=1}^{n}\sum_{j=1}^{m}C_{i,j} \qquad C_{i,j}=\begin{cases}1 & outdata = epdata\\ 0 & outdata \neq epdata\end{cases} \qquad (5)$$

Where *outdata* is output value of current evaluated circuit; *epdata* is output value of expected circuit; the value of the formulation is the numbers of bits of circuit outputs resulted from a specific combination of inputs, which scores 'correct' if its measured value equals that specified. For the brushless motor control circuit which include 3 inputs (6 input combinations) and 36 output-bits, in order to get a smoother landscape that is consequently easier to search, each of the output-bits is counted independently instead of treating them as a whole when computing fitness. The biggest fitness value is 36.

3.3 Population initialization and population size

In general, populations of configuration-strings were randomly generated. In this paper, population seeding and population recall approaches that were proposed in [1] were applied. The method of population seeding involves taking the fittest solution stored from the previous evolutionary run as the seed, and placing it into a population of 29 randomly generated configuration-strings. Population recall simply involves re-introducing the most recent population of configuration-strings evolved, and using this as the initial start point. In the evolution design process when errors were introduced, a randomly selected final population, evolved for motor control circuit with no faults in the FPGA, was used as the initial population for the recall approach. Both population seeding and recall enable the GA to adapt to the faulty FPGA architecture and produce correct circuits with target fitness considerably faster than when a population of configuration-strings is randomly generated.

A relative large GA population size is desirable for effective searching because diversity of chromosomes can be easily preserved in the population. However, in this experiment, considering seeding population and recall population is applied, seed chromosome in a small population has much greater probability of being optimized by the GA search compared to a chromosome in a large population. To let the seed chromosome increasing operating efficiency, we made a compromise and set the population size to 30.

3.4 Adaptation strategy for GA parameters

Some GA parameters, especially probabilities of crossover and mutation, P_c and P_m, have large effects on GA's performances; and their optimal values are usually impossible to be predefined to suit various problems and states of GA [13]. In our approach, P_c and P_m are varied with the individuals' distribution and GA's genetic processes so as to maintain diversity in the population and sustain the convergence capacity of the GA. Diversity in the population, which measures how diversely the individuals are distributed in the phenotype space, is estimated .

Evolutionary processes of the GA are simply identified with the diversity or distribution of the individuals, which is represent by δ_t. Pc and Pm are designed to adapt themselves in the following ways

$$P_c = \begin{cases} P_{C0} & 0<\delta_t<k1 \\ P_{C1} + \dfrac{(P_{c0}-P_{c1})(1-\delta_t)}{1\text{-}k1} & k1<\delta_t<1 \end{cases} \qquad (7)$$

$$P_m = \begin{cases} P_{m0} & 0<\delta_t<k2 \\ P_{m1} + \dfrac{(P_{m1}-P_{mo})(\delta_t-k2)}{1\text{-}k2} & k2<\delta_t<1 \end{cases} \qquad (8)$$

where, P_{c0} and P_{m0} are initial values of P_c and P_m respectively, it is usually feasible to let $P_{c0} = 0.8$ and $P_{m0} = 0.1$ due to the above adaptation strategy; According to the above equations, P_c and P_m will decrease as a whole during a GA run; meanwhile they will respond to changes of individuals' diversity reflected by δ_t. In this way, a higher P_c and a higher P_m to speed up the genetic search at the first evolution stage, a lower P_c and a higher P_m to improve the quality of elitist solutions at the final stage, and a lower P_c but a higher P_m, resulted from an increasing δ_t, to prevent the GA from premature convergence at all stages.

4. Experiment

The objective of the experiments was to recover the functionality of the motor control circuit implemented on FPGA. When one or more logic units can't be used, evolution was applied to obtain a new circuit solution that recovered the circuit's functionality. In this experiment, faults were introduced by setting all connections with the corresponding fault logic unit to '0'. Different numbers of faults were introduced for experiments. Firstly, we evolved a motor control circuit in case all 24 logic units available; Secondly, single faults in different position of the circuit and multi-faults in column H1 were introduced respectively; and then evolutionary process was carried out to recover the circuit topology with the same functionalities. In order to investigate the fault-tolerance ability, three technical indexes are defined here: the convergence rate, the average fitness, and the average evolution generations. The convergence rate is defined of the proportion of functionality recovery times in every 10 times evolution. The average fitness describes the average correct bits of the actual response corresponding to objective response in every 10 times evolution. The average evolutionary generations that reflects self-recovery speed to every corresponding type faults denotes the average generations to implement self-recovery in every 10 times evolution.

4.1 Single Logic Unit Fault

The aim is to test that the platform has good fault recovery ability for single logic unit fault. When fault is introduced to logic unit in the H1 column, the convergence rate is 100%; that is to say, the motor control circuit evolved can recover from single logic unit fault completely. But when fault is introduced to H0 and H2 column, the correct circuit can't be evolved correctly all the time. That is because the fault unit is near the inputs and outputs position; the placement of fault logic unit has crucial impact on fault tolerant ability. It will greatly affect the ability of the GA to evolve high quality circuit. Faults close to the inputs or outputs will have a more detrimental effect than those distributed in the centre column. Table 2 illustrates the experimental results with single fault introduced.

Table 2:Experimental results with single fault introduced

Position of Faults	***Average Evolution Generation***	***Average Fitness***	***Evolution Time***	***Average convergence rate***	***Number of Logic Units***
H2	496	30.54	96	100%	17
H1	623	27.4	153	90%	16
H3	637	28.1	151	70%	18

4.3 Multi-Logic Unit Fault

Here increasing number of logic unit faults are introduced to illustrate fault-tolerance ability of the motor control circuit respectively. The experiment results are shown in table 3.

Table 3 indicates that the fault tolerant ability of FPGA descends obviously with the number of fault logic units increasing. Especially when four logic units fault occur, the average convergence rate is no more than 30%; the average fitness diminishes obviously and the average evolutionary generations increase rapidly. The average number of logic units used to implement the circuit reduces as the number of faults increases.

Table 3: Experimental results with increasing numbers of faults introduced

Number of Faults	***Average Evolution Generations***	***Average Fitness***	***Evolution Time***	***Average convergence rate***	***Number of Logic Units***
2	637	28.1	151	81%	18
3	858	28.85	221	62%	16
4	1575	27.12	315	29%	15
5	3000	20.4	732	0%	-

From the experimental results above, we can know that the number of fault logic units is closely related to the fault tolerant ability; that is to say, with the number of fault logic units increasing, evolutionary recovery of the same circuit needs more evolutionary generations, and the average fitness and convergence rate descend evidently. The reason consists in the increasing number of fault logic units makes the signal paths which are used to accurately transfer signals become less; consequently to evolve the objective circuit topologies become more difficult; so the fault tolerant ability is affected obviously. We also find that if 4 logic units cause faults, the correct functional circuit can't be evolved; that is to say, the most permissive faults are 4 logic units.

An example configuration of the motor control circuit evolved with 4 logic unit faults is illustrated in Fig. 3. the objective of this work was not explicitly to design more efficient circuits but to show that it is possible to evolve an alternative circuit in case of fault occur in the original circuit, thus the functionality can be recovered.

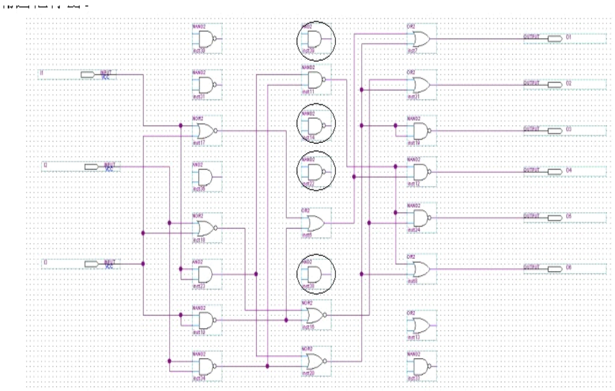


Fig. 3 The evolved motor control circuit with four logic unit faults introduced.

5. Conclusions

A fault tolerant hardware platform for the automated design of brushless motor control circuit has been presented. The platform uses the principle of EHW to automate the configuration of FPGA dedicated to the implementation of the motor control circuit. Our experiments show that it is possible to recover the function of motor control circuit through evolution when faults are introduced, thus the fault tolerant capability has been approved. Furthermore, the ability of the platform to adapt to increasing numbers of faults was investigated by introducing faulty to different locations of the topology structure. Results show that the functional circuit can be derived from single logic unit fault and multi-logic unit faults; the most permissive faults are four logic units. Of course the placement of faulty logic units will influence the ability of GA to evolve high quality circuit, fault directly on logic units which are connected to the inputs and outputs will have a more detrimental effect than those distributed in the centre of the topology structure. It also shows that the evolutionary recovery ability of the motor control circuit descends obviously with the number of fault logic units increasing.

The real attractiveness and power of EHW comes from its potential as an adaptive hardware while operating in a real physical environment. Further work will focus on On-line evolution in electromagnetism interference environment, which poses a great challenge.

Acknowledgments

The work presented in this paper has been funded by Personnel Administration Department of Heibei Province Research Project for People Return after Study Abroad, Research Plan of the Science and Technology Department of Hebei Province (12210212), and Hebei University of Science & Technology Research Foundation for the Doctoral Program.

References

[1] Arslan, B. I., Thomsom, T. R.: Evolutionary Design and Adaptation of High Performance Digital Filters with an Embedded Reconfigurable Fault Tolerant Hardware Platform. Software Computing, vol. 8, 2004,pp. 307-317. Springer, Berlin.

[2] Higuchi, T., Niwa, T., Tanaka, T., H. de Garis, and Furuya, T.: Evolvable Hardware with Genetic Learning: A first step toward building a Darwin machine. In: Proc. of Second International Conference On the Simulation Adaptive Behavior (SAB'92), Cambridge, MA.1992, pp. 417–424.

[3] Thompson, A.: Hardware Evolution: Automatic Design of Electronic Circuits in Reconfigurable Hardware by Artificial Evolution. University of Sussex, Doctoral Thesis, 1996.

[4] Yao, X., Higuichi, T.: Promises and Challenges of Evolvable Hardware, IEEE Trans. On Systems Man and Cybernetics-Part C: Applications and Reviews, 1999,vol. 29, pp. 87–97.

[5] Zhao, S. G., Jiao, L. C.: Multi-objective Evolutionary Design and Knowledge Discovery of Logic Circuits Based on an Adaptive Genetic Algorithm. Genetic Programming and Evolvable Machines, vol. 8, Springer, Berlin,2006,pp.195-210.

[6] Gao, G. J., Wang, Y. R., Cui, J., Yao, R.: Research on Multi-objective On-line Evolution Technology of Digital Circuit Based on FPGA Model. In: Proc. of 7th International Conference of Evolvable System: From Biology to Hardware, Wuhan, China, 2007,pp. 67-76.

[7] Ji, Q. J., Wang, Y. R., Xie, M., Cui, J.: Research on Fault-Tolerance of Analog Circuit Based on Evolvable Hardware. In: Proc. of 7th International Conference of Evolvable System: From Biology to Hardware, Wuhan, China, 2007, pp. 100-108.

[8] Yao, R., Wang, Y. R., Yu, S. L., Gao, G. J.: Research on the Online Evolution Approach for the Digital Evolvable Hardware. In: Proc. of 7th International Conference of Evolvable System: From Biology to Hardware, Wuhan, China, 2007,pp. 57-66.

[9] Stoica, A., Keymeulen, D., Arslan, T., Duong, V., Zebulum, R.S., Ferguson, I., Guo, X.: Circuit Self-Recovery Experiments in Extreme Environments. In: Proceedings of the 2004 NASA/DoD Conference on Evolution Hardware, Seattle, WA, USA, 2004, pp. 142–145.

[10] Stoica, A., Keymeulen, D., Zebulum, R.S., Thakoor, A., Daud, T., Klimeck, G., Jin, Y., Tawel, R., Duong, V.: Evolution of Analog Circuits on Field Programmable Transistor Arrays. IEEE Computer Society Press, Los Alamitos In: Proc. of the Second NASA/DOD Workshop on Evolvable Hardware, 2000,pp. 99–108.

[11] Ricardo, S., Zebulum, R.S., Keymeulen, D., Duong, V., Guo, X., Ferguson, M.I., Stoica, A.: Experimental Results in Evolutionary Fault-Recovery for Field Programmable Analog Devices. In: Proceedings of The 2003 NASA/Dod Conference on Evolvable Hardware, Chicago, IL, USA, 2003,pp. 182–186 .

[12] Zhao, S. G.: Study of the Evolutionary Design Methods of Electronic Circuits. PhD. Dissertation (in Chinese), Xidian University, Xi_an, China, 2003.

[13]Chu, J.: Study of the Fault Tolerant Bionic circuit Model. PhD. Dissertation (in Chinese), Ordnance Engineering College, Shijiazhuang, China, 2009.

First Author

Hui-cong Wu get her doctor's degree from Shijiazhunag Mechanical Engineering College in 2007, She is an associate professor in Hebei University of Science & Technology. She visited The University of Birmingham, UK from 2009 to 2010. Her research interests include evolutionary computation, software engineering, data mining.

A Cooperative Spectrum Sensing Scheme Based on Trust and Fuzzy Logic for Cognitive Radio Sensor Networks

Yonghua Wang[1,2], Yuehong Li [1] , Fei Yuan [3] and Jian Yang [1]

**[1] School of Automation, Guangdong University of Technology
Guangzhou, 51006, China**

**[2] Shenzhen Key Laboratory of High Performance Data Mining
Shenzhen, 518055, China**

**[3] School of Information Science and Technology,Sun Yat-sen University
Guangzhou, 51006, China**

Abstract

This paper proposes a cooperative spectrum sensing scheme based on trust and fuzzy logic for Cognitive Radio Sensor Networks (CRSN). The CRSN nodes use the T-S fuzzy logic to make local decisions on the presence or absence of the primary user's (PU) signal, and then use a censoring method to only allow the relatively reliable decisions sent to the fusion center. Utilizing a trust evaluation scheme based on the factors such as local sensing difference, sensing location factors, and sensing channel conditions for each node. Combing the majority rule and the trust values of the nodes, the fusion center makes the final decision. Simulation results show that the proposed scheme could improve the detect probability effectively.

Keywords: *Cognitive Radio Sensor Networks, trust, fuzzy logic, Cooperative Spectrum Sensing.*

1. Introduction

Most Wireless Sensor Networks (WSNs) operate in the unlicensed ISM (Industrial, Scientific and Medical) frequency band, for example, the 2.4 GHz band. While such bands are also used by other wireless applications such as WiFi, Bluetooth, cordless phones, RFID, microwave ovens, etc. And in a large-scale wireless sensor network, due to the event-driven nature, when an event occurs many WSN nodes need to transmit the event signal data simultaneously. Therefore, the interference or collision probability increases. The Cognitive Radio (CR)[1] that enables higher spectrum efficiency by dynamic spectrum access [2][3]can be exploited by WSN to solve this problem. The WSN comprised of sensor nodes equipped with CR is defined Cognitive Radio Sensor Networks (CRSN)[4].

To mitigate the problem of uncertainty in spectrum sensing in a cognitive radio network, cooperative spectrum sensing can be used. Different techniques were proposed for cooperative spectrum sensing. The simplest method is to use an OR or AND operation among the received sensing results [5]. Combing techniques based on maximal ratio combining (MRC) and equal gain combining (EGC) were investigated in [6], an optimal linear cooperation scheme base on a likelihood ratio test (LRT) has been proposed in [7]. In [8], the censor-based cooperative spectrum sensing has been proposed to save energy. And a censor-based cooperative spectrum sensing scheme using Takagi and Sugeno's (T-S) fuzzy logic for cognitive radio sensor networks was proposed in [9]. But in these schemes, the CRSN nodes are often assumed to be trustworthy. In practice, there are malicious CRSN nodes sending false reporting values to the Fusion Center (FC) [10], which will induce the FC to make wrong decisions. Based on the previous work in [9] and [11], we proposed a cooperative spectrum sensing scheme based on trust and fuzzy logic for CRSN. The nodes use the T-S fuzzy logic to make local decisions on the presence or absence of the primary user's (PU) signal, and then use a censoring method to only allow the relatively credible decisions sent to FC. And a trust scheme on the basis of the factors such as local sensing difference, sensing location factors, and sensing channel conditions for each node is implemented. The FC makes the final decision according to the majority rule and the trust weights of the nodes. This method can improve the sensing performance while saves the node's energy.

2. Cooperative Spectrum Sensing Scheme

2.1 Local Spectrum Sensing and Decision

Suppose each CRSN node using the energy detection. The hypotheses if the primary user is present (H_1) or not (H_0) are as follows:

$$x_i(t)=\begin{cases} n_i(t): & H_0 \\ h_i(t)s(t)+n_i(t): & H_1 \end{cases} \tag{1}$$

Where $x_i(t)$ is the received signal by the i -th node, $n_i(t)$ is the thermal noise, $h_i(t)$ is the channel gain from the PU to the i -th node, and $s(t)$ is the transmit signal from the PU.

The local test static of the i -th node using energy detection is:

$$x_{Ei}=\sum_{k=0}^{N-1}\left|x_i(k)\right|^2, i=1,2,\cdots,M \tag{2}$$

Where $x_i(k)$ is the k -th sample of received signal at the i -th node, M is the number of nodes, and N is the number of samples, $N=2TW$, where T and W are detection time and signal bandwidth, respectively[12].

Let the two fuzzy sets Low and High depict the i -th node detected energy x_{Ei} , and their membership functions are[12]:

$$\mu_{Low}(x_{Ei})=\begin{cases} 1 & ,if\ x_{Ei}\le\mu_{0i} \\ e^{\frac{(x_{Ei}-\mu_{0i})^2}{2\sigma_{0i}^2}} & ,otherwise \end{cases} \tag{3}$$

$$\mu_{High}(x_{Ei})=\begin{cases} 1 & ,if\ x_{Ei}\ge\mu_{1i} \\ e^{\frac{(x_{Ei}-\mu_{1i})^2}{2\sigma_{1i}^2}} & ,otherwise \end{cases} \tag{4}$$

And the Fig. 1 shows the shapes of the above functions [9] [12].

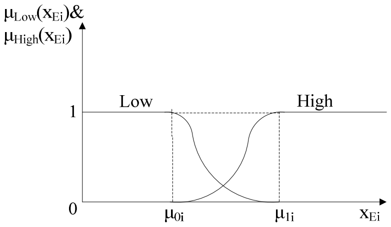


Fig.1 The membership functions of the two fuzzy sets

Based on the T-S fuzzy method, the local decision of the i -th node [9] is:

$$Ld_i=\frac{-\mu_{Low}(x_{Ei})+\mu_{High}(x_{Ei})}{\mu_{Low}(x_{Ei})+\mu_{High}(x_{Ei})} \tag{5}$$

And the fuzzy rule sets are [9] [12]:

Rule 1: IF (x_{Ei} is *Low*) THEN ($Ld_i=Ld_{\min}$)

Rule 2: IF (x_{Ei} is *High*) THEN ($Ld_i=Ld_{\max}$)

Where Ld_i is the local decision of the i -th node, $Ld_{\max}$ and $Ld_{\min}$ are the maximum and minimum value of Ld_i respectively. $Ld_{\max}=1$ means the PU is present and $Ld_{\min}=-1$ means the PU is absent.

Given the censoring threshold value $C\in(0,1)$,the local soft decision Ld_i is transmitted to the FC if and only if $Ld_i<-C$ or $Ld_i>C$.

2.2 The Trust Value of CRSN Nodes

Combine the factors such as the difference of local sensing; sensing location, sensing channel condition [11], the FC establishes trust value for each node.

The difference of local sensing is the difference of the sensing value of a single node with the average value of all nodes. The smaller the difference, the more reliable of the node. The difference of local sensing of the l -th node is:

$$D_l=\left|Ld_l-\frac{1}{Ns}\sum_{l=1}^{Ns}Ld_l\right| \tag{6}$$

Where Ns is the number of nodes that sent the local decision to the FC.

The sensing location factor includes two aspects: the distance of the l -th node to PU, the distance of the l -th node to the FC, and denoted by LO_a and LO_b respectively. The sensing location factor is:

$$LO_l=LO_a^l\times LO_b^l \quad l=(1,2,\cdots,Ns) \tag{7}$$

Because the sensing channel is non-ideal, so the information transmission is prone to error. Based on the channel SNR, define the sensing channel condition factor of the l -th node:

$$CC_l=SNR_l \quad l=(1,2,\cdots,Ns) \tag{8}$$

Carry out the standardizing to the three factors, and get the following expressions [11]:

$$D_l^{'} = \frac{D_{\max} - D_l}{D_{\max} - D_{\min}} \tag{9}$$

$$LO_l^{'} = \frac{LO_{\max} - LO_l}{LO_{\max} - LO_{\min}} \tag{10}$$

$$SC_l^{'} = \frac{SC_{\max} - SC_l}{SC_{\max} - SC_{\min}} \tag{11}$$

Where $D_{\max}$ and $D_{\min}$ is the maximum and minimum value of the sensing location factor respectively, and the others are similar.

The trust value of the l -the node is:

$$T_l = \alpha D_l^{'} + \beta LO_l^{'} + \gamma SC_l^{'} \tag{12}$$

Where $\alpha + \beta + \gamma = 1$.

2.3 Data Fusion at the FC

The FC receives Ns local decisions from CRSN nodes. According to the majority rule[9], the final decision fusion rule is:

$$H = \begin{cases} H_1 : \sum_{j=1}^{Ns} Ld_j T_j > 0 \\ H_0 : otherwise \end{cases} \tag{13}$$

3. Simulation Results and Analysis

To evaluate the performance of the proposed cooperative spectrum sensing algorithm for CRSN and compare the performance with some exiting methods, the Monte-Carlo simulations are carried out with 100,000 samples under the following conditions: The number of CRSN nodes M is 7, the PU signal is likely-equally BPSK signal, The noises at the sensing and control channels are Gaussian with zero mean and unit variance, The number of samples N is 300.

The performance of the scheme is compared with the performance of the OR/AND fusion rules, the equal gain combination (EGC) based scheme [6], and the fuzzy logic scheme [9].

Firstly, the performance of the factors that local sensing difference, sensing location factors, and sensing channel condition is evaluated. Let the SNRs for the nodes are -20,-18,-16,-14,-12, -10,-8 dB, respectively, and $C = 0.2$. From Fig.2, it can be seen that the three factors all have the effect on the detection performance, and the local sensing difference has the minimum effect while the sensing channel condition factor has the greatest effect. The scheme has the best sensing performance when the $\alpha = \beta = \gamma = 1/3$.

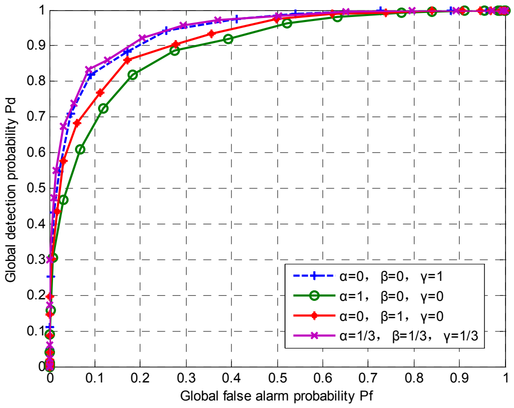


Fig.2. The effect of the α, β, γ

Secondly, the impact of the SNR on the performance is evaluated. The simulation was carried out under conditions that the SNRs of the first six nodes are -15,-14,-13,-12,-11,-10 dB, respectively, and the SNR at the 7th node is -14,-12,-10,-8,-6 dB respectively, $\alpha = \beta = \gamma = 1/3$,and $C = 0.2$. From Fig.3, it can be seen that P_M increases while the SNR at the 7th node decreases.

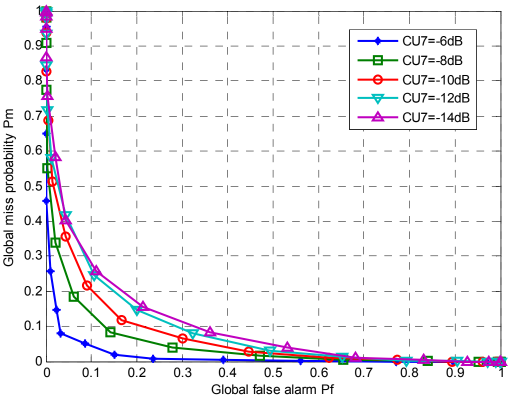


Fig.3. The impact of the SNR

Finally, the simulations are conducted under the $\alpha = \beta = \gamma = 1/3$, the SNRs for the nodes are -20,-18,-16,-14,-12,-10,-8 dB, respectively, and $C = 0.2$. The ROC curves of our scheme and compared schemes are depicted in Fig. 4. It can be seen that the proposed scheme has the highest detection probability than the OR/AND fusion rules based scheme, the EGC based scheme, and the fuzzy logic scheme. It means that the method of node trust evaluation improves the FC decision accuracy.

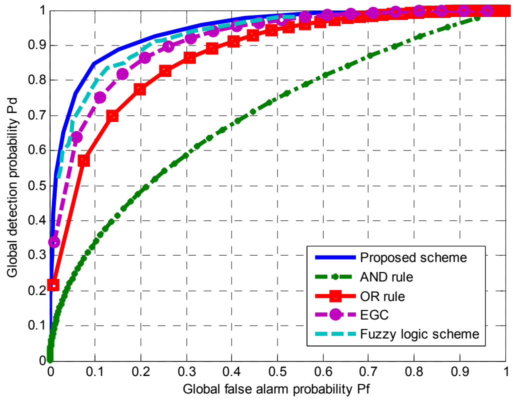


Fig.4 ROC curves of proposed scheme vs. comparison schemes

4. Conclusions

By introducing the trust scheme into the fuzzy logic scheme to represent the reliability of the nodes, a cooperative spectrum sensing scheme based on trust and fuzzy logic for CRSN was proposed in this paper. Analysis and simulations show that while keep the merits the fuzzy logic scheme that reduce the number of nodes reporting its local decision to the FC and save the energy of the CRSN nodes, the proposed scheme can improve the detection probability effectively compared with the OR/AND fusion rules based scheme, the EGC based scheme, and the fuzzy logic scheme. It is useful for the CRSN because the hardware, low energy restriction and the performance are favorable.

Acknowledgments

This research was supported by The National Natural Science Foundation of China. (No. 61102034),Shenzhen New Industry Development Fund under grant (No.CXB201005250021A), the Developing Fund for Innovative Research Team of Guangdong University of Technology(No. GDUT2011-10), and the Ph.D. Programs Foundation of Guangdong University of Technology (No. 103042).

References

[1] J. Mitola, G. Q. Maguire, "Cognitive Radios:Making Software Radios More Personal",IEEE Personal Communications,Vol. 6,No. 4, 1999,pp. 13-18,.

[2] Chaoub, Abdelaali; Ibn-Elhaj, Elhassane, "Comparison between Poissonian and Markovian Primary Traffics in Cognitive Radio Networks ",International Journal of Computer Science Issues, Vol.9, No .2, 2012, pp. 63-73.

[3] Yadav, Shrikrishan;Roy, Krishna Chandra, "Tradeoff Analysis of Bit-Error-Rate (BER) in Cognitive Radio Based on Genetic Algorithm",International Journal of Computer Science Issues, Vol. 9, No. 1, 2012 , pp.390-394.

[4] O.B. Akan,O.B.Karli,O.Ergul, "Cognitive Radio Sensor Networks",IEEE Networks,Vol.23, No.4, 2009,pp.34-40.

[5] Tevfik Yucek, Huseyin Arslan, "A Survey of Spectrum Sensing Algorithms for Cognitive Radio Applications",IEEE Communications Surveys & Tutorials, Vol. 11, No.1, 2009, pp. 116-130.

[6] J.Ma,G.Zhao,Y.Li, "Soft Combination and Detection For Cooperative Spectrum Sensing in Cognitive Radio Networks", IEEE Transactions on Wireless Communications, Vol.7, No.11, 2008 , pp:4502-4507.

[7] Z. Quan, S. Cui, and A. H. Sayed, "Optimal Linear Cooperation for Spectrum Sensing in Cognitive Radio Networks", IEEE Journal of Selected Topics in Signal Processing, Vol. 2, No.1, 2008, pp. 28-45.

[8] S.Maleki, A.Pandharipande, G.Leus,"Energy-Efficient Distributed Spectrum Sensing for Cognitive Sensor Networks", IEEE Sensors Journal, Vol.11, No.3,2011, pp.565-573.

[9] T.K.XUAN,I.KOO , "A Censor-Based Cooperative Spectrum Sensing Scheme Using Fuzzy Logic for Cognitive Radio Sensor Networks", IEICE Transctions on communications, Vol. E93-B, No.12, 2010, pp.3497-3500.

[10] W.Lang, Y.Zhu, H.Li,"Trust Level Based Byzantine Attack-Exclusion Cooperative Spectrum Sensing Scheme for Wireless Cognitive Sensor Networks in Smart Grids",in 2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2012, pp:21–25.

[11]Zhang Guanghua, Ding Renshuang, Huang Lijing, "Using Trust to Establish Cooperative Spectrum Sensing Framework",Procedia Engineering, Vol.15, 2011,pp.1361–1365.

[12]T.K.XUAN,I.KOO , "Cooperative Spectrum Sensing using Kalman Filter based Adaptive Fuzzy System for Cognitive Radio Networks", KSII Transctions on Internet and Information System,Vol. 6, No.1, 2012, pp.287-304.

Yonghua Wang has received his B.S. degree in Electrical Engineering and Automation from Hebei University of Technology in 2001, the M.S. degree in Control Theory and Control Engineering from Guangdong University of Technology in 2006, and Ph. D degree in Communication and Information System from Sun Yat-sen University in 2009.Currently he is a lecturer in school of automation in Guangdong University of Technology since 2009. His current research interests include cognitive networks, RFID, etc.

Yuehong Li has received his B.S. degree in Electrical Engineering and Automation from Hangzhou Dianzi University in 2010.He is now working towards his M.S. degree in Control Theory and Engineering at Guangdong University of Technology.

Fei Yuan has received his B.S.degree in Communication Engineering and M.S. degree in Control Theory and Control Engineering from Guangdong University of Technology in 2007,2010 respectively. He is now working towards his Ph.D. degree in Communication and Information system at Sun Yat-sen University. His research interests include wireless networks,data mining, etc.

Jian Yang has received his B.S. and M.S. degree in Biomedical Engineering from Northeastern University in 2003,2006 respectively. He has received Ph. D degree in Communication and Information system from Sun Yat-sen University in 2009. Currently he is a lecturer in school of automation in Guangdong University of Technology since 2009. His research interests include RFID,wireless communications etc.

A high precision 3D dynamic measurement prototype system based on color-coded fringe and phase shifting

Dahui Qin[1], Jianjun Liu[1,2] , Fuzhen Liu[1] and Xiaojun Dai[1]

[1] School of Civil Engineering and Architecture, Southwest Petroleum University, Chengdu, China;

[2] State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation (Southwest Petroleum University), Chengdu, China;

Abstract

Accurate dynamic measurement of the 3-D shape of moving objects is a rapidly expanding field, with applications in entertainment, design, and manufacturing. In this work, the two CCD/CMOS cameras are calibrated with a flexible and accurate camera calibration method. Then, a new approach based on color-coded fringe is proposed for high precision 3D dynamic measurement. The method is based on phase shifting and stereo vision. Combination of color coding and phase shifting can make automatic phase unwrapping, stereo vision can accomplish stereo matching and 3D points reconstruction automatically according to phase map after unwrapping. An experimental result is presented to demonstrate the performance of the method.

Keywords: *3D dynamic measurement; color-coded fringe; phase shifting; camera calibration.*

1. Introduction

Accurate dynamic measurement of the 3-D shape of moving objects is a rapidly expanding field, with applications in entertainment, design, and manufacturing. Among the existing 3-D shape measurement techniques, the techniques based on stereo vision using digital projection of structured light and recording by CCD/CMOS cameras are increasingly used due to their fast speed and non-contact nature. In this work, we developed a 3D dynamic measurement system consists of two CCD/CMOS cameras and a DLP projector. There are two key techniques in this system: camera calibration and absolute phase unwrapping using an image

Calibration of camera is a prerequisite for the extraction of precise 3D information from some images. Much work about camera calibration has been done in the photogrammetry community[1] , and also in computer vision[2-4]. In this work, we use the flexible camera calibration method proposed by Zhang[3] . It only requires the camera to shoot a planar pattern shown at a few (at least three) different orientations. Either the camera or the planar pattern can be freely moved. The motion need not be known. Radial and tangential lens distortion is modeled.

After the two cameras are calibrated, our system becomes able to perform a 3D dynamic measurement. In this work, we proposed a new method based on color coding , phase shifting and stereo vision for high precision 3D dynamic measurement. Combination of color coding and phase shifting can make automatic phase unwrapping, stereo vision can accomplish stereo matching and 3D points reconstruction automatically according to absolute phase map after unwrapping.

2. Camera calibration

2.1 Pinhole model

In order to perform 3D measurement with our system, the two cameras must be calibrated. The camera is modeled by the pinhole model. A 3D point $M=[X,Y,Z]^T$ in the world coordinate system and its 2D image projection point $m=[u,v]^T$ in a camera image coordinate system are related by:

$$s\tilde{m}=A[R \quad t]\tilde{M} \qquad (1)$$

Where tilde means homogeneous coordinates, $\tilde{M}=[X,Y,Z,1]^T$ and $\tilde{m}=[u,v,1]^T$, s is an arbitrary scale factor, A is the intrinsic matrix consists of the intrinsic parameters, [R t] is the extrinsic matrix. R and t are the extrinsic parameters, and they are the rotation matrix and the translation vector which relate the world coordinate system to the camera coordinate system. The intrinsic matrix A consists of the intrinsic parameters and is given by:

$$A=\begin{bmatrix}\alpha & \gamma & u_0\\ 0 & \beta & v_0\\ 0 & 0 & 1\end{bmatrix} \qquad (2)$$

Where α and β are scale factors consist of a focal length of a lens and a size of a cell of the image sensor. γ

is a parameter which represents the skewness of the two axes in the image coordinate system. (u_0, v_0) is the coordinate of the principal point.

2.2 Radial distortion

The image taken with a camera is distorted by the lens, and that image is different from the image taken with the ideal pinhole camera. Therefore, the lens distortion must be considered to deal with the model of the imaging system which is close to the actual camera[3]. The most commonly used correction is for the radial lens distortion that causes the actual image point to be displaced radially in the image plane. The radial distortion can be approximated using the following expression:

$$\left.\begin{aligned}\delta_{xr} &= x\left(k_1 r^2 + k_2 r^4 + k_3 r^6 + \cdots\right)\\ \delta_{yr} &= y\left(k_1 r^2 + k_2 r^4 + k_3 r^6 + \cdots\right)\end{aligned}\right\} \quad (3)$$

Where δ_{xr} and δ_{yr} are the radial distortion values, (x, y) is the ideal normalized image coordinates, k_1 , k_2 , k_3 , $\cdots$, are the parameters describing the radial distortion, and $r = \sqrt{x^2 + y^2}$.

2.3 Tangential distortion

For Accurate dynamic measurement of the 3-D shape of moving objects, the tangential distortion of the camera lenses must be taken into account. The expression for the tangential distortion is often written in the following form:

$$\begin{aligned}\delta_{xt} &= 2p_1 xy + p_2\left(r^2 + 2x^2\right)\\ \delta_{yt} &= 2p_2 xy + p_1\left(r^2 + 2y^2\right)\end{aligned} \quad (4)$$

Where δ_{xt} and δ_{yt} are the tangential distortion values, (x, y) is the ideal normalized image coordinates, p_1 and p_2 are the parameters describing the tangential distortion, and $r = \sqrt{x^2 + y^2}$.

2.4 Camera calibration

In this work, we use the flexible camera calibration method proposed by Zhang[3] . The method can obtain the intrinsic and extrinsic parameters, and the distortion parameters of the lens according to 3D coordinates on the model plane and their correspondences of image coordinates.

Zhang's calibration procedure is as follows:

1. Shoot a few images of the model plane under different orientations by moving either the plane or the camera.

2. Detect the image feature points.

3. Preliminary estimate the five intrinsic parameters and all the extrinsic parameters using the closed-form solution.

4. Refine all parameters, including lens distortion parameters, by maximum likelihood estimation.

3 3D dynamic measurement

Phase-shifting method has been used extensively in optical metrology to measure 3-D shapes of objects at various scales. In this paper, a phase-shifted sinusoidal fringe patterns are recorded in a color image, from which the phase information at every pixel is obtained. This phase information is then converted to xyz coordinates of the object surface after the system is calibrated. In this work, we use a DLP projector projects a computer generated color-coded image onto the object. The fringe patterns included by a color-coded image, which are deformed by the object surface, are captured by two cameras synchronously. Then a phase-wrapping and phase-unwrapping algorithm and a phase-to-coordinates conversion algorithm are used to reconstruct the 3D geometry. DLP projector can project color image, and at same time eliminate phase shift error.

3.1 color coding method

For 3D dynamic measurement, the key is that 3D reconstruction must be acquired through one image. So three-step phase shifting information must be in a image. Color image include three channels (red, green and blue channels), and a channel include a phase shifting 2/3 π fringe. The cameras of the 3D dynamic capture color-coded fringe which is illustrated in Fig.1.

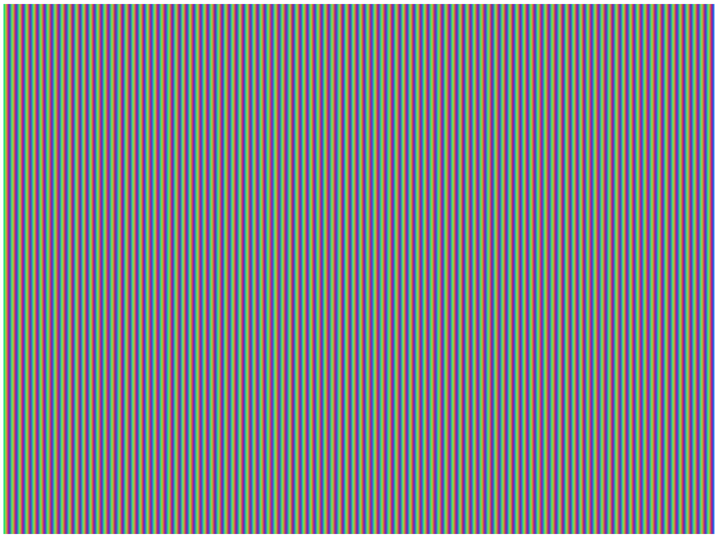


Figure 1. Color-coded fringe image

A color-coded frnge image include three different phase information. The 24 bits true color image is projected by DLP projector, in which each pixel values are divided into R, G, B component of three primary colors. So each channel image is obtained.

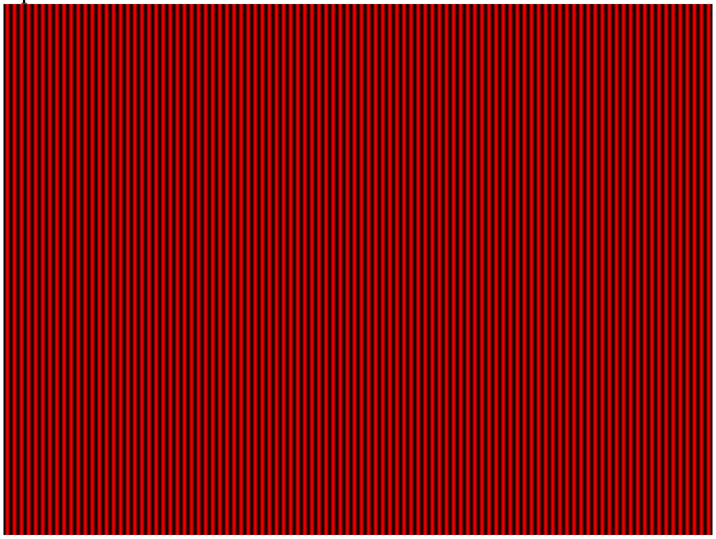


Figure 2. Red channel fringe image

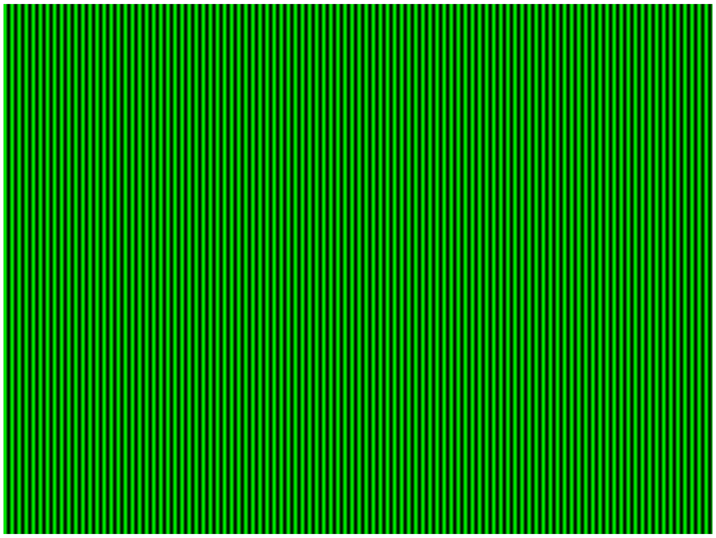


Figure3. Green channel fringe image

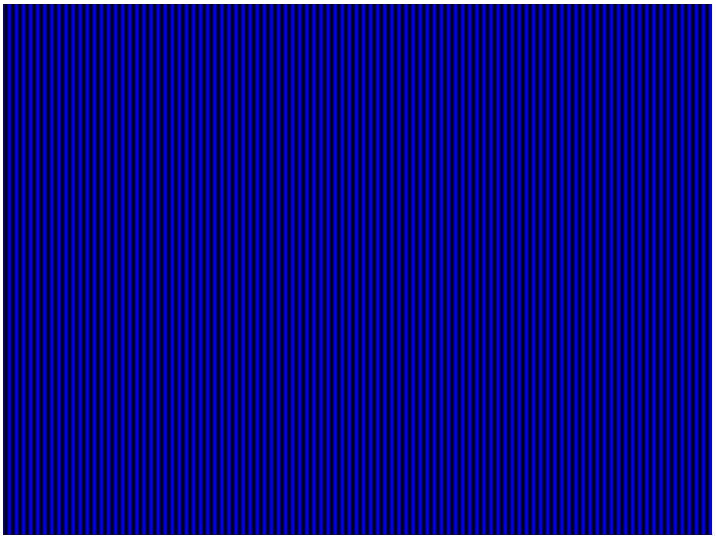


Figure4. Blue channel fringe image

3.2 Phase shifting

Many different phase-shifting algorithms have been developed[5]. A well known method for determining these parameters is the so called three-step phase shifting . Let us assume that a total of three phase-shifted fringe images are captured, each with a phase shift of $\delta_i (i=1,2,3,4)$. Then the intensity of the i-th image can be represented as:

$$I_i(x,y) = I^{'}(x,y) + I^{''}(x,y)\cos[\phi(x,y)+\delta_i] \quad (5)$$

Where $I^{'}(x,y)$ is the average intensity, $I^{''}(x,y)$ is the intensity modulation, and $\phi(x,y)$ is the phase to be determined. By solving equation (5) simultaneously, we can obtain the phase:

$$\varphi(x,y) = tg^{-1}\frac{I_3 - I_2}{I_1 - I_2} \quad （6）$$

Phase $\phi(x,y)$ in algorithm (6) is the so-called modulo 2π phase at each pixel whose value ranges from 0 to 2π . If the R, G and B fringe patterns are obtained, phase unwrapping is necessary to remove the sawtooth-like discontinuities and obtain a continuous phase map[5].

3.3 Stereo vision

In order to compute the 3D coordinate, we need to obtain correspondences of the two images. In this work, we divide the color-coded fringe image into three channels fringe images, and then a series of contour lines and obtain pixel point sets in these lines and a series of phase grey curves through three-step phase shifting algorithm described in section 3.1. Then we implement the matching point sets in the left-right images according to the each absolute value of phase grey lines. Finally we get the corresponding points by epipolar constraint within the matching point sets. Assuming the corresponding points to be (u_r, v_r) and (u_l, v_l), the world coordinates of the point to be (x_w, y_w, z_w), we have the following equation that transform the world coordinates to the camera image coordinates:

$$s\{u_r, v_r, 1\}^T = P_r\{x_w, y_w, z_w\}^T \qquad (7)$$
$$s\{u_l, v_l, 1\}^T = P_l\{x_w, y_w, z_w\}^T$$

Where P_r and P_l are the calibrated matrix for the two cameras,

$$P_r = A_r[R_r \quad T_r] \qquad (8)$$
$$P_l = A_l[R_l \quad T_l]$$

From Equations 7 and 8 we can obtain four linear equations:

$$\begin{aligned} &f_1(x_w, y_w, z_w, u_r) = 0, \\ &f_2(x_w, y_w, z_w, v_r) = 0, \\ &f_3(x_w, y_w, z_w, u_l) = 0, \\ &f_4(x_w, y_w, z_w, v_l) = 0. \end{aligned} \qquad (9)$$

Where u_r, v_r and u_l, v_l are known. Therefore the world coordinates (x_w, y_w, z_w) of the point can be solved by least square algorithm.

3.5 3D dynamic measurement system

In this paper, we developed a 3D measurement system consists of two CCD/CMOS cameras and a DLP projector. The system overview is shown in figure 4.The camera image resolution is set 1024×820 and the projector image also is 1280 × 1024. The procedure of 3D dynamic measurement with this system is as follows. Firstly, the two cameras are calibrated with a flexible and accurate camera calibration method. As soon as the cameras calibration is finished 3D dynamic measurement becomes to be able to perform. Secondly, the DLP projector projects a color-coded fringe image which designed in section 3.1 to the object surface and the two CCD/CMOS cameras capture the one modulated image simultaneously respectively. Then, we compute the actual phase of the two cameras using three-step phase shifting and get two phase maps. Finally, we accomplish stereo matching and points reconstruction automatically according to the two phase maps after unwrapping.

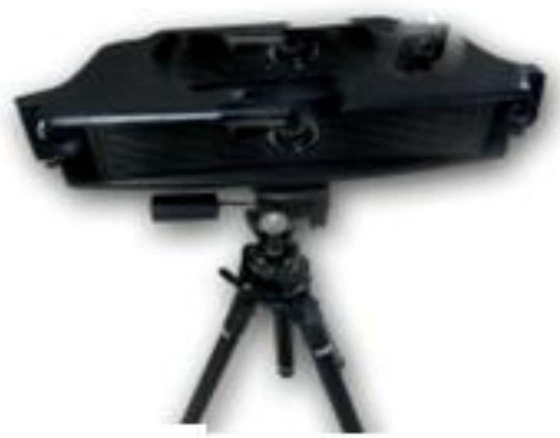


Fig.5. 3D measurement system

4 Experiments

In this work, a 400mm×300mm high precise calibration board with 99 circle marks is used to obtain calibration data sets for the camera calibration. In order to ensure the precision of the camera calibration, twelve groups of calibration images are captured from different orientation, which is shown in Fig.6. Then the coordinates of center of circles were extracted automatically. We calibrate the two cameras using the extracted coordinate of central of the circle. The cameras calibration results are shown in Table 1.

Table 1. Cameras calibration results

		Left camera	Right camera
intrinsic parameters	Focal Length	422.27651861 5421.97530654	3102.24690491 3029.02660688
	Skrew:	0	0
	Principal Point	562.45214993 340.19617783	627.16132910 287.89371002
	p_1	0.00109245	0.01331403
	p_2	0.00219015	-0.00099325
	k_1	0.22341062	-0.04855767
	k_2	6.16265442	0.71134609
extrinsic parameters	R	0.95804156 -0.00892189 0.28649044 0.00091801 -0.99941480 -0.03419366 0.28662786 0.03302195 -0.95747273 0.99086	0.99572594 -0.09228766 -0.00358354 -0.09191234 -0.98637979 -0.13640759 0.00905401 0.13615395 -0.99064632
	T	-100.25315756 101.35010168 1396.12329433	77.63567398 188.32795872 1306.35646977

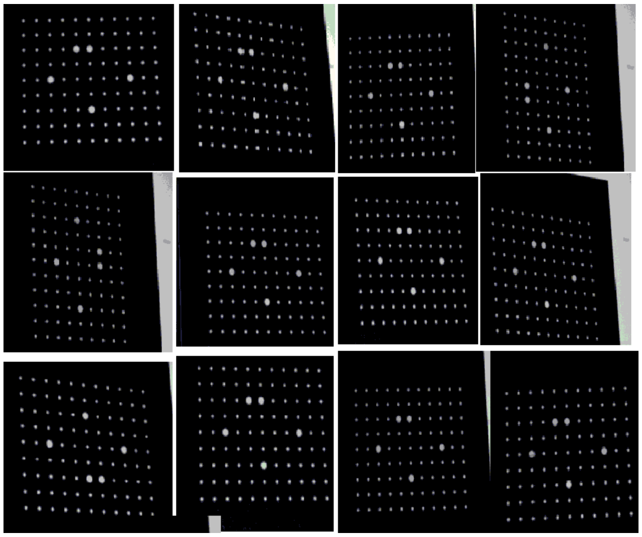
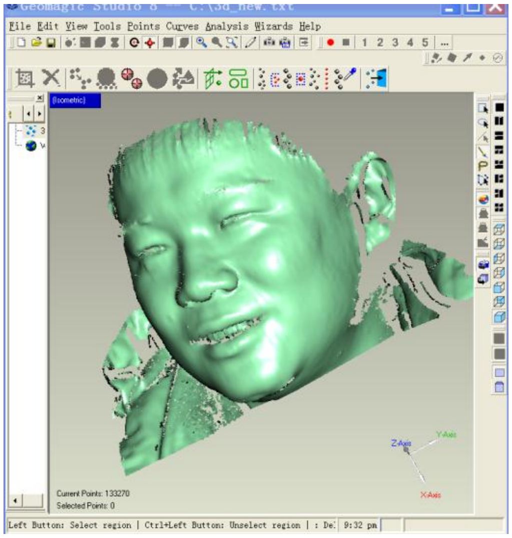


Fig.6 Calibration images

To verify the performance of the method proposed in this work, a face model is measured. Pair of color fringe images is captured during the measurement process. The measurement result is shown in Fig.7 .


Geomagic Studio 8 -- C:\3d_new.txt
File Edit View Tools Points Curves Analysis Wizards Help
1 2 3 4 5 ...
(Isometric)
Z-Axis
Y-Axis
X-Axis

Current Points: 133270
Selected Points: 0

Left Button: Select region | Ctrl+Left Button: Unselect region | : De| 9:32 pm

Fig.7 3D measurement result of the face model

5. Conclusion

In this paper, we have developed a 3D dynamic measurement system consists of two CCD/CMOS cameras and a DLP projector. By using a flexible calibration method, the system can be constructed easily. After the cameras calibration, we have proposed a method based on color-coded fringe, phase shifting and stereo vision for high precision 3D dynamic measurement. A face model has been measured and the reconstructed 3D model is very smooth

Acknowledgement

This paper is financially supported by Natural Science Foundation of China (Grant No.51174170) and National Science and Technology Major Project of China under Grant No. 2011ZX05013-006.

References

[1] B. Ergun and I. Baz. Design of an expert measurement system for close-range photogrammetric applications. Optical Engineering, 45.5 (5), 2006, 053604-5.

[2] R. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses." IEEE J. Robotics and Automation, 3 (4), 1987, 323-344.

[3] Z. Zhang. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (11), 2000, 1330-1334.

[4] Y. Zhang and Z. Zhang. Camera calibration technique with planar scenes. Machine Vision Applications in Industrial Inspection XI, Santa Clara, CA, USA, SPIE, 2003.

[5] D. C. Ghiglia and M. D. Pritt. Two-Dimensional Phase Unwrapping: Theory, Algorithms, and Software, John Wiley and Sons,1998.

A new hybrid artificial bee colony algorithm for global optimization

Xiangyu Kong[1], Sanyang Liu[2], Zhen Wang[3]

[1.] Department of Applied Mathematics, Xidian University, Xi'an 710071, China

[2.] Department of Applied Mathematics, Xidian University, Xi'an 710071, China

[2.] Institute of Information and System Computation Science, Beifang University of Nationalities, Yinchuan 750021, China

Abstract
To further improve the performance of artificial bee colony algorithm (ABC), a new hybrid ABC (HABC) for global optimization is proposed via exploring six initialization methods. Furthermore, to balance the exploration and exploitation abilities, a new search mechanism is also developed. The algorithms are applied to 27 benchmark functions with various dimensions to verify its performance. Numerical results demonstrate that the proposed algorithms outperforms the ABC in global optimization problems, especially the HABC algorithm with random initialization and HABCO algorithm with orthogonal initialization.

Keywords: *Artificial bee colony algorithm, Initialization methods, Search mechanism, Differential evolution.*

1. Introduction

Global optimization problems arise in almost every field of science, engineering and business. By now, learning from life system, many optimization methods have been developed to solve global optimization problems, such as genetic algorithms (GAs) [1,2], ant colony optimization (ACO) [3], differential evolution (DE) [4] and particle swarm optimization (PSO)[5]. These kinds of algorithms can be named as artificial-life computation. Recently, Karaboga [6] proposed a new kind of optimization technique called artificial bee colony (ABC) algorithm for global numerical function optimization, which simulates the foraging behavior of honey bee swarm. A set of comparison experimental results show that ABC algorithm is competitive to some conventional bio-inspired algorithms with an advantage of employing fewer control parameters [7]. Due to its simplicity, ABC algorithm has been applied to solve many kind of real-world problems, for instance, leaf-constrained minimum spanning tree problem [8], flow shop scheduling problem [9], inverse analysis problem [10], radial distribution system network reconfiguration problem [11], clustering problem [13], TSP problems [14], and so on.

According to the applications showed above, ABC algorithm seems to be a well-performed algorithm. However, similar to other population-based algorithms, there still are insufficiencies in ABC algorithm, such as slower convergence speed for some unimodal problems and easily get trapped in local optima for some complex multimodal problems [7]. It is well known that for the population-based algorithms the exploration and the exploitation abilities are both necessary facts. The exploration ability refers to the ability to investigate the various unknown regions to discover the global optimum in solution space, while the exploitation ability refers to the ability to apply the knowledge of the previous good solutions to find better solutions. The exploration ability and the exploitation ability contradict to each other, so that the two abilities should be well balanced to achieve good performance on optimization problems. So far as we know that the search equation of ABC algorithm is good at exploration but poor in exploitation.

Therefore, accelerating convergence speed and avoiding local optima have become two most important goals in ABC algorithm modification. To overcome the issues in ABC algorithm and achieve the two goals above, inspired by PSO and DE, a new search mechanism is proposed in the new hybrid artificial bee colony (HABC) algorithm. In order to balance the exploration ability and the exploitation ability, the random search equation is used in the employed bee stage and the best-guided search equation is used in the onlooker bee stage. In addition, to enhance the convergence speed, six initialization methods are employed and compared, including random initialization, chaotic initialization, opposition-based initialization, inter-cell initialization, chaotic opposition-based initialization, and orthogonal initialization [15]. Experimental results and comparisons denote the effectiveness and efficiency of the proposed HABC algorithms.

The rest of the paper is organized as follows. In Section 2, ABC algorithm is summarized briefly. In Section 3, the proposed hybrid artificial bee colony algorithm is described. In Section 4, experiments are presented and the results are discussed. Finally, a conclusion is provided in Section 5.

2. Overview of artificial bee colony algorithm

By simulating the foraging behavior of bee colonies, artificial bee colony (ABC) algorithm, which is a swarm intelligence-based optimization algorithm, was proposed by Karaboga in 2005 for numerical function optimization [6]. The main steps of ABC algorithm can be described as follows.

```
Initialization
Repeat
    Employed bee stage: Place the employed bees on the food sources in the memory.
    Onlooker bee stage: Place the onlooker bees on the food sources in the memory.
    Scout bee stage: Send the scout bees to the search area for discovering new food sources.
Until (conditions are satisfied)
```

In ABC algorithm, the colony consists of three kinds of bees: employed bees, onlooker bees and scout bees. Half of the colony is employed bees, and the other half is onlooker bees. The employed bees explore the food source and send the information of the food source to the onlooker bees. The onlooker bees choose a food source to exploit based on the information shared by the employed bees. The scout bee, which is one of the employed bees whose food source are abandoned, finds a new food source randomly. The position of a food source is a possible solution to the optimization problem. Denote the food source number as SN, the position of the i th food source as $x_i \quad (i = 1, \cdots, SN)$, which is a D -dimensional vector.

In ABC algorithm, the i th fitness value fit_i for a minimization problem is defined as:

$$fit_i = \begin{cases} \dfrac{1}{1+f_i}, & f_i \geq 0, \\ 1+abs(f_i), & f_i < 0, \end{cases} \tag{2.1}$$

where f_i is the cost value of the i th solution.

The probability of a food source being selected by an onlooker bee is given by:

$$p_i = \frac{fit_i}{\sum_{i=1}^{SN} fit_i}. \tag{2.2}$$

A candidate solution from the old one can be generated as:

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}), \tag{2.3}$$

where $k \in \{1, 2, \cdots, SN\}$, $k \neq i$ and $j \in \{1, 2, \cdots, D\}$ are randomly selected indices, $\phi_{ij} \in [-1, 1]$ is a uniformly distributed random number. The candidate solution is compared with the old one, and the better one should be remained.

If the abandoned food source is x_i, the scout bee exploits a new food source according to:

$$x_{ij} = x_j^{\min} + rand(0,1)\left(x_j^{\max} - x_j^{\min}\right), \tag{2.4}$$

where $x_j^{\max}$ and $x_j^{\min}$ are the upper and lower bounds of the j th dimension of the problem's search space.

3. New Hybrid artificial bee colony algorithm

3.1 Initialization

Population initialization is a crucial step in swarm intelligence algorithms, because it can affect the quality of the solutions and the convergence speed. There are several kinds of initialization methods to generate the initial population. Here, the following initialization methods will be considered in the new hybrid artificial bee colony algorithm, and the efficiency of these methods will be compared.

3.1.1 Random initialization

The random initialization is the most commonly used method to generate initial population. In the original ABC algorithm, the random initialization was employed. The initial population is generated randomly within the range of the boundaries of the parameters, which is:

$$x_{ij} = x_j^{\min} + rand(0,1)\left(x_j^{\max} - x_j^{\min}\right), \tag{3.1}$$

where $i \in \{1, 2, \cdots, SN\}$, $j \in \{1, 2, \cdots, D\}$.

3.1.2 Chaotic initialization

Chaos is found in non-linear dynamical systems, which is a deterministic random-like process. Chaos is apparently random and unpredictable but it also presents regularity, and it has a very sensitive dependence upon the initial condition and parameters. Mathematically, chaotic maps may be considered as source of randomness. Because of these randomness and sensitivity dependence on the initial conditions of chaotic maps, it has been considered as an initialization method for the heuristic algorithms to improve the global convergence by escaping from local optima. [16]

A chaotic map is a discrete time dynamic system running in chaotic state, which is: $x_{k+1} = f(x_k)$, $0 < x_k < 1$, $k = 1, 2, \cdots$.

According to the definition of $f(\cdot)$, the chaotic maps include several types, such as logistic map, circle map, gauss map， sinusoidal iterator, and so on. Here, the sinusoidal iterator is employed as a chaotic map in colony initialization. The initial equation is defined as follows:

$$x_{ij} = x_j^{\min} + ch_{k,j}\left(x_j^{\max} - x_j^{\min}\right), \tag{3.2}$$

where $ch_{k,j} = \sin(\pi \cdot ch_{k-1,j})$, $ch_{k-1,j} \in (0,1)$, $k = 1, 2, \cdots, K$, is the chaotic sequence.

3.1.3 Opposition-based initialization

According to [17], random initial solutions are further from the solution than their opposite random initial solutions, which can accelerate convergence. The initial population is generated as follows:

$$ox_{ij} = x_j^{\min} + x_j^{\max} - x_{ij}, \tag{3.3}$$

where $x_{ij} = x_j^{\min} + rand(0,1)\left(x_j^{\max} - x_j^{\min}\right)$.

3.1.4 Chaotic opposition-based initialization

Based on the properties of chaotic maps and opposition-based learning methods, a new initialization approach employs both chaotic maps and opposition-based learning methods [18]. The new initialization approach is given as follows:

$$ox_{ij} = x_j^{\min} + x_j^{\max} - x_{ij}, \tag{3.4}$$

where $x_{ij} = x_j^{\min} + ch_{k,j}\left(x_j^{\max} - x_j^{\min}\right)$, $ch_{k,j} = \sin(\pi \cdot ch_{k-1,j})$, and $ch_{k-1,j} \in (0,1)$, $k = 1, 2, \cdots, K$, is the chaotic sequence.

3.1.5 Inter-cell initialization

In the population initialization step, it is desired that the initial population can be scattered uniformly over the feasible solution space, so that the algorithm can search the whole solution space evenly. For this, the inter-cell initialization can be employed: the feasible solution space is divided into *SN* subspace and a solution is randomly selected in each one of these subspaces. Here, *SN* is the population size.

3.1.6 Orthogonal initialization

An orthogonal array specifies a small number of combinations that are scattered uniformly over the space of all possible combinations. Thus, the initial population generated by the orthogonal design can be scatted uniformly over the feasible solution space, so that the algorithm can explore the solution space evenly. Here, we employ the orthogonal initialization method based on orthogonal array and quantization technique, which is described in [19] [20].

3.2 New search mechanism

It is well known that both the exploration and exploitation abilities are necessary for the population based algorithms. How to balance these two abilities to achieve good optimization performance is very important.

In section 2, in ABC algorithm, the employed bees explore the new food source and send the information to the onlooker bees; while the onlooker bees exploit the food sources which are explored by the employed bees. It means that in the ABC algorithm the employed bee stage represents the exploration ability of the algorithm, and the onlooker bee stage represents the exploitation ability of the algorithm. The search equation proposed in ABC algorithm is good at exploration but poor at exploitation, so that it will affect the convergence speed of the algorithm. Inspired of PSO [5], in order to improve the exploitation ability of ABC algorithm, take the advantages of the search equation in PSO, the global best solution will be considered in the new search equation in the onlooker bee stage. The modified search equation in onlooker bee stage is described as follows:

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) + \varphi_{ij}(y_j - x_{ij}), \tag{3.5}$$

where $k \in \{1, 2, ..., SN\}$ is a random selected index which is different from i , $j \in \{1, 2, ..., D\}$ is a random selected index, y_j is the j th element of the global best solution, $\phi_{ij} \in [-1,1]$ and $\varphi_{ij} \in [0,1.5]$ are both uniformly distributed random numbers.

Differential evolution (DE) [4] is a population based algorithm to function optimization, whose main strategy is to generate a new position for an individual by calculating vector differences between other randomly selected members in the population. “DE/current-to-rand/1” is a variant DE mutation strategy, which can effectively maintain population diversity according to randomness of the search equation. Motivated by “DE/current-to-rand/1” mutation strategy and based on the property of ABC algorithm, a new search equation in employed bee stage is proposed as follows:

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) + \varphi_{ij}(x_{ij} - x_{lj}), \tag{3.6}$$

where $k, l \in \{1, 2, ..., SN\}$ are random selected indexes which are different from i ; $j \in \{1, 2, ..., D\}$ is a random selected index; $\phi_{ij} \in [-1,1]$ and $\varphi_{ij} \in [-1,1]$ are uniformly distributed random number; ϕ_{ij} and φ_{ij} are both negative or both positive, which can keep the search direction the same.

In general, inspired by DE and PSO, the new search equation and search mechanism are proposed to balance the exploration ability and exploitation ability in ABC algorithm. In the employed bee stage, search equation (3.6) is used to keep the exploration ability of ABC algorithm; while, in the onlooker bee stage, search equation (3.5) is employed to increase the exploitation ability of the algorithm.

3.3 The hybrid artificial bee colony algorithm

Based on the above analysis, the main steps of the new hybrid artificial bee colony are as follows.

Algorithm: Hybrid artificial bee colony algorithm

```
Initialize the food sources by using one of the initialization methods proposed in subsection 3.1, and evaluate the population, $trail_i = 0$, $(i = 1, 2, \cdots, SN)$. $Cycle = 1$.
Repeat
   Step 1: Search the new food source for employed bee according to (3.6) and evaluate its quality.
   Step 2: Apply a greedy selection process and select the better solution between the new food source and the old one.
   Step 3: If solution does not improve $trail_i = trail_i + 1$, otherwise $trail_i = 0$.
   Step 4: Calculate the probability according to (2.2) and apply roulette wheel selection scheme to choose a food source for onlooker bees.
   Step 5: Search the new food source for onlooker bees according to (3.5) and evaluate its quality.
   Step 6: Apply a greedy selection process and select the better solution between the new food source and the old one.
   Step 7: If solution does not improve $trail_i = trail_i + 1$, otherwise $trail_i = 0$.
   Step 8: If $\max(trail_i) > \text{limit}$, replace this food source with a new food source produced by the initialization methods proposed in subsection 3.1.
   Memorize the best solution achieved so far.
   $Cycle = Cycle + 1$
Until($Cycle =$ Maximum Cycle Number)
```

4. Numerical experiments

4.1 Test functions and parameter settings

In this section, the HABC algorithm with six different initialization methods is applied to minimize 27 benchmark functions, as shown in Table 1, 2 and 3. In Table 1 and 2, the dimensions of the benchmark functions are given in the third column. The benchmark functions presented in Table 3 are tested of dimension $D = 30$ and dimension $D = 100$. All the benchmark functions, presented in Table 1, 2 and 3, include unimodal, multimodal, regular, irregular, separable, non-separable and multidimensional. Initial range, characteristics and formulation of these functions are listed in Table 1-3.

In all these benchmark functions, Colville and Rosenbrock have an arrow curving valley. If the direction changes cannot be kept up with and the search space cannot explored properly, these two problems will be hard to optimize. The functions of Schwefel and Zakharov have a high eccentric ellipse. The main difficulty is that their gradients are not oriented along the axes, which needs to balance the orientation and high eccentricity of the ellipse makes it a significant challenge for some algorithms. Ackley, Griewank, Rastrigin and Schwefel are complex multimodal functions with a large number of local optima. To obtain good results for these functions, the search strategy must efficiently balance the exploration and exploitation abilities.

All experiments were repeated 25 times independently for each function. The population size was 100 and the maximum number of generation was 3000 for both ABC algorithm and six HABC algorithms in the experiments. Therefore, all experiments were run for 300,000 function evaluations.

4.2 Experimental results

In this subsection, a set of experiments tested on 27 benchmark functions were performed, which compared the performance of six HABC algorithms with ABC algorithm. The results are shown in Table 4-7 in terms of best, worst, mean, standard deviation and mean time. In these tables, ABC represents the original ABC algorithm; HABCOB represents the HABC algorithm with the Opposition-based initialization; HABCCH represents the HABC algorithm with the Chaotic initialization; HABCIC represents the HABC algorithm with the Inter-cell initialization; HABCCOB represents the HABC algorithm with the Chaotic opposition-based initialization; HABCO represents the HABC algorithm with the Orthogonal initialization and HABC represents the HABC algorithm with the random initialization. The best results are highlighted in boldface.

Compared with the results, the mean values of these six HABC algorithms are equal or close to the optimal ones and the standard deviations are relatively small. All seven algorithms, including ABC and other six HABC algorithms can find optimal solutions on fuctions f_2, f_4, f_6, f_7-f_9, $f_{14}-f_{16}$, f_{21}, and f_{22}, f_{24} with $D=30$. In particular, the HABCCOB algorithm has the smallest standard deviation in function f_7 and f_8, which means

Table 1. Benchmark functions f_1 - f_8 used in experiments. D: Dimension, C: Characteristic, U: Unimodal, M: Multimodal, S: Separable, N: Non-Separable.

No	Range	D	C	Function	Formulation
1	[-4.5,4.5]	2	UN	Beale	$f(x)=(1.5-x_1+x_1x_2)^2+(2.25-x_1$ $+x_1x_2^2)^2+(2.625-x_1+x_1x_2^3)^2$
2	[-100,100]	2	MS	Bohachevsky	$f(x)=x_1^2+2x_2^2-0.3\cos(3\pi x_1)$ $-0.4\cos(4\pi x_2)+0.7$
3	[-10,10]	2	MS	Booth	$f(x)=(x_1+2x_2-7)^2+(2x_1+x_2-5)^2$
4	[-5,10]× [0,15]	2	MS	Branin	$f(x)=\left(x_2-\frac{5.1}{4\pi^2}x_1^2+\frac{5}{\pi}x_1-6\right)^2+$ $10\left(1-\frac{1}{8\pi}\right)\cos x_1+10$
5	[-10,10]	4	UN	Colville	$f(x)=100(x_1^2-x_2)^2+(x_1-1)^2$ $+(x_3-1)^2+90(x_3^2-x_4)^2$ $+10.1((x_2-1)^2+(x_4-1)^2)$
6	[-100,100]	2	UN	Easom	$f(x)=-\cos x_1\cos x_2\exp(-(x_1-\pi)^2$ $-(x_2-\pi)^2)$
7	[-2,2]	2	MN	GoldStein- Price	$f(x)=\begin{bmatrix}1+(x_1+x_2+1)^2\\(19-14x_1+3x_1^2-14x_2+6x_1x_2+3x_2^2)\end{bmatrix}$ $\begin{bmatrix}30+(2x_1-3x_2)^2\\(18-32x_1+12x_1^2+48x_2-36x_1x_2+27x_2^2)\end{bmatrix}$
8	[0,1]	3	MN	Hartman3	$f(x)=-\sum_{i=1}^{4}c_i\exp\left[-\sum_{j=1}^{3}a_{ij}(x_j-p_{ij})^2\right]$ $c=[1.0,1.2,3.0,3.2]$; $a=\begin{bmatrix}3.0&10&30\\0.1&10&35\\3.0&10&30\\0.1&10&35\end{bmatrix}^T$ $p=\begin{bmatrix}0.3689&0.1170&0.2673\\0.4699&0.4387&0.7470\\0.1091&0.8732&0.5547\\0.03815&0.5743&0.8828\end{bmatrix}^T$

Table 2. Benchmark functions f_9 - f_{16} used in experiments. D: Dimension, C: Characteristic, U: Unimodal, M: Multimodal, S: Separable, N: Non-Separable, $n = D$.

No	Range	D	C	Function	Formulation
9	[-5,5]	2	MN	Six Hump Camel Back	$f(x)=4x_1^2-2.1x_1^4+\frac{1}{3}x_1^6$ $+x_1x_2-4x_2^2+4x_2^4$
10	[-10,10]	2	UN	Matyas	$f(x)=0.26(x_1^2+x_2^2)-0.48x_1x_2$
11	[- D , D]	2	MN	Perm	$f(x)=\sum_{k=1}^{n}\left[\sum_{i=1}^{2}(i^k+0.5)((x_i/i)^k-1)\right]^2$
12	[-4,5]	4	UN	Powell	$f(x)=\sum_{i=1}^{n/k}(x_{4i-3}+10x_{4i-2})^2+5(x_{4i-1}+10x_{4i})^2$ $+(x_{4i-2}+10x_{4i-1})^4+10(x_{4i-3}+10x_{4i})^4$
13	[0, D]	24	MN	PowerSum	$f(x)=\sum_{k=1}^{n}\left[\sum_{i=1}^{4}(x_i^k)-b_k\right]^2$ $b=[8,18,44,114]$
14	[0,10]	4	MN	Shekel	$f(x)=-\sum_{j=1}^{m}\left[\sum_{i=1}^{4}(x_i-a_{ij})^2+c_i\right]^{-1}$ $c=\frac{1}{10}[1,2,2,4,4,6,3,7,5,5]^T$ $a=\begin{bmatrix}4.0&1.0&8.0&6.0&3.0&2.0&5.0&8.0&6.0&7.0\\4.0&1.0&8.0&6.0&7.0&9.0&5.0&1.0&2.0&3.6\\4.0&1.0&8.0&6.0&3.0&2.0&3.0&8.0&6.0&7.0\\4.0&1.0&8.0&6.0&7.0&9.0&3.0&1.0&2.0&3.6\end{bmatrix}$
15	[-10,10]	2	MN	Shubert	$f(x)=\left(\sum_{i=1}^{5}i\cos((i+1)x_1+i)\right)$ $\cdot\left(\sum_{i=1}^{5}i\cos((i+1)x_2+i)\right)$
16	[-36,36]	6	UN	Trid6	$f(x)=\sum_{i=1}^{n}(x_i-1)^2-\sum_{i=2}^{n}x_ix_{i-1}$

the results obtained by HABCCOB algorithm on these functions are the most stable. And for function f_{14}, the results obtained by HABC algorithm has the smallest standard deviation; for function $f_{15}-f_{21}$, the results obtained by HABCCO have the smallest standard deviation. All these results indicate that on the above 12 functions, the six HABC algorithms obtain the better solutions than the original ABC algorithm. Furthermore, except the function f_1, f_{18} with $D=30$ and $D=100$, and f_{23} with $D=100$, the six HABC algorithms perform better than the original ABC algorithm, while on these three functions the superiority of ABC algorithm to HABC algorithms is not very obvious in terms of the mean and standard deviation of the solutions. These results indicate that HABC algorithms can balance between exploration and exploitation well.

Compared the results obtained by these six HABC algorithm, we can find that the HABC algorithm with the random initialization performs the best, the HABC algorithm with the Orthogonal initialization. Secondly, the HABC algorithm with the Inter-cell initialization is the worst. It indicates that the population initialization methods will affect the solution obtained by the algorithms.

In order to show the performance of the six HABC algorithms more clearly, Figure 1, Figure 2 and Figure 3 show the mean best function value of some functions. It is clear that for most functions the HABC algorithms have the better performance than the ABC algorithm. Particularly, the HABC algorithm performs the best, which can convergence to the optimum fast and stable, and HABCO the second, and HABCOB, HABCCH and HABCCOB perform similar the third, the HABCIC the worst.

Figure 4, Figure 5 and Figure 6 show the statistical results of the function values for the test functions shown in Figure 1 - 3. Here, box plots are used to illustrate the distribution of these samples obtained from 25 independent runs. The upper and lower ends of the box are the upper and lower quartiles. The line within the box represents the median, and thin appendages summarize the spread a shape of the distribution. Symbol “ + ” indicate for outlier and the notches denote a robust estimation of the uncertainty about the medians for box-to-box comparison. From Figure 4-6, we can see that HABC algorithms can obtain the better and more stable solutions than ABC algorithm does, especially the HABC algorithm, and further verifies the discussion obtained in Table 4-10 and Figure 1-3.

5. Conclusion

In this paper, a new hybrid artificial bee colony algorithm is developed for global optimization problems with six kind of initial methods and new search mechanism. The experimental results tested on 27 benchmark functions show that HABC algorithms are competitive with ABC algorithm, especially the HABC algorithm. The improvement can mainly be attributed to the following reasons. First, the new search mechanism can balance the exploration and exploitation abilities very well, which can both maintain the diversity and improve the convergent speed. Secondly, the initialization methods can affect the quality of the solutions and the convergence speed. Therefore, the HABC algorithms are accuracy and effective for global optimization problems.

It is desirable to further apply HABC algorithms to solving those more complex real-world optimization problems and it will be our further work.

Acknowledgements

This work is supported by National Nature Science Foundation of China (No. 60974082) and Foundation of State Key Lab. of Integrated Services Networks of China.

References

[1] K. S. Tang, K. F. Man, S. Kwong, Q. He. Genetic algorithms and their applications. IEEE Signal Processing Magazine, 1996, 13(6):22-37.

[2] X. Xue, Y. Gu. Global Optimization Based on Hybrid Clonal Selection Genetic Algorithm for Task Scheduling. Journal of Computational Information Systems. 2010, 6(1):253-261.

Table 3. Benchmark functions f_{17} - f_{27} used in experiments. C: Characteristic, U: Unimodal, M: Multimodal, S: Separable, N: Non-Separable, $n = D$.

No	Range	C	Function	Formulation
17	[-32,32]	MN	Ackley	$f(x) = -20\exp\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}\right) - \exp\left(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)\right) + 20 + e$
18	[-10,10]	UN	Dixon-Price	$f(x) = (x_1 - 1)^2 + \sum_{i=2}^{n} i(2x_i^2 - x_{i-1})^2$
19	[-600,600]	MN	Griewank	$f(x) = \frac{1}{4000}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n}\cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$
20	[-10,10]	MN	Levy	$f(x) = \sin^2(\pi y_1) + \sum_{i=1}^{n-1}\left[(y_i - 1)^2(1 + 10 \sin^2(\pi y_i + 1))\right] + (y_n - 1)^2(1 + 10\sin^2(2\pi y_n))$ $y_i = 1 + \frac{x_i - 1}{4}, \quad i = 1, \cdots, n.$
21	[0, π]	MS	Michalewicz	$f(x) = -\sum_{i=1}^{n}\sin(x_i)\left(\sin(ix_i^2/\pi)\right)^{2m}$, $m = 10$
22	[-5.12,5.12]	MS	Rastrigin	$f(x) = \sum_{i=1}^{n}\left[x_i^2 - 10\cos(2\pi x_i) + 10\right]$
23	[-30,30]	UN	Rosenbrock	$f(x) = \sum_{i=1}^{n-1}[100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2]$
24	[-500,500]	MS	Schwefel	$f(x) = \sum_{i=1}^{n} -x_i \sin\left(\sqrt{\lvert x_i \rvert}\right)$
25	[-100,100]	US	Sphere	$f(x) = \sum_{i=1}^{n} x_i^2$
26	[-10,10]	US	SumSquares	$f(x) = \sum_{i=1}^{n} ix_i^2$
27	[-5,10]	UN	Zakharov	$f(x) = \sum_{i=1}^{n} x_i^2 + \left(\sum_{i=1}^{n} 0.5ix_i\right)^2 + \left(\sum_{i=1}^{n} 0.5ix_i\right)^4$

Table 4. Best, worst, mean, standard deviation and mean time value obtained by ABC and other six HABC through 25 independent runs on benchmark function f_1 - f_8 .

		Best	Mean	Worst	Std	Mean-Time
f_1	ABC	**9.77e-17**	**1.98e-15**	**1.64e-14**	**3.06e-15**	**16.33508**
	HABCOB	3.50e-09	3.41e-07	1.76e-06	4.16e-07	16.81662
	HABCCH	1.56e-08	5.71e-07	2.10e-06	6.71e-07	16.71493
	HABCIC	2.37e-09	5.20e-07	5.25e-06	1.05e-06	17.12752
	HABCCOB	3.63e-10	3.22e-07	9.12e-07	2.58e-07	16.92151
	HABCO	1.62e-08	5.35e-07	4.44e-06	8.95e-07	17.23723
	HABC	3.62e-09	5.42e-07	2.92e-06	6.76e-07	17.28006
f_2	ABC	**0**	**0**	**0**	**0**	**16.75061**
	HABCOB	**0**	**0**	**0**	**0**	17.01885
	HABCCH	**0**	**0**	**0**	**0**	16.87816
	HABCIC	**0**	**0**	**0**	**0**	17.38567
	HABCCOB	**0**	**0**	**0**	**0**	17.09213
	HABCO	**0**	**0**	**0**	**0**	17.42189
	HABC	**0**	**0**	**0**	**0**	17.4868
f_3	ABC	4.43e-20	1.79e-18	8.26e-18	1.77e-18	**15.83831**
	HABCOB	4.44e-21	4.41e-19	1.76e-18	3.94e-19	16.22402
	HABCCH	4.11e-20	**3.70e-19**	1.48e-18	**3.86e-19**	16.09317
	HABCIC	**4.06e-21**	5.70e-19	2.86e-18	6.89e-19	16.60484
	HABCCOB	2.00e-20	7.34e-19	3.24e-18	8.12e-19	16.34202
	HABCO	7.78e-20	4.92e-19	**1.35e-18**	4.08e-19	16.68255
	HABC	7.62e-21	4.16e-19	1.54e-18	4.15e-19	16.59208
f_4	ABC	**3.98e-01**	**3.98e-01**	**3.98e-01**	**0**	**16.20202**
	HABCOB	**3.98e-01**	**3.98e-01**	**3.98e-01**	**0**	16.59222
	HABCCH	**3.98e-01**	**3.98e-01**	**3.98e-01**	**0**	16.47659
	HABCIC	**3.98e-01**	**3.98e-01**	**3.98e-01**	**0**	16.91512
	HABCCOB	**3.98e-01**	**3.98e-01**	**3.98e-01**	**0**	17.45144
	HABCO	**3.98e-01**	**3.98e-01**	**3.98e-01**	**0**	17.83527
	HABC	**3.98e-01**	**3.98e-01**	**3.98e-01**	**0**	17.04465
f_5	ABC	5.82e-03	5.79e-02	1.04e-01	2.83e-02	**17.44701**
	HABCOB	1.98e-03	2.72e-02	9.37e-02	2.58e-02	18.1546
	HABCCH	3.54e-03	3.25e-02	1.16 e-01	2.69e-02	18.11104
	HABCIC	**1.21e-03**	**2.40e-02**	**7.26e-02**	**1.99e-02**	18.53727
	HABCCOB	1.77e-03	3.39e-02	1.22e-01	2.96e-02	18.19926
	HABCO	1.46e-03	2.38e-02	9.17e-02	2.29e-02	17.94445
	HABC	1.39e-03	2.79e-02	1.17e-01	2.82e-02	17.62544
f_6	ABC	**-1**	**-1**	**-1**	**0**	**16.53898**
	HABCOB	**-1**	**-1**	**-1**	**0**	16.86348
	HABCCH	**-1**	**-1**	**-1**	**0**	16.70109
	HABCIC	**-1**	**-1**	**-1**	**0**	17.37636
	HABCCOB	**-1**	**-1**	**-1**	**0**	16.91044
	HABCO	**-1**	**-1**	**-1**	**0**	17.21757
	HABC	**-1**	**-1**	**-1**	**0**	17.15267
f_7	ABC	**3**	**3**	**3**	1.46e-15	**16.6893**
	HABCOB	**3**	**3**	**3**	1.21e-15	16.96251
	HABCCH	**3**	**3**	**3**	1.31e-15	16.8877
	HABCIC	**3**	**3**	**3**	1.15e-15	17.20357
	HABCCOB	**3**	**3**	**3**	**5.73e-16**	16.95133
	HABCO	**3**	**3**	**3**	6.72e-16	17.3129
	HABC	**3**	**3**	**3**	6.15e-16	17.33915
f_8	ABC	**-3.86**	**-3.86**	**-3.86**	1.85e-13	27.04722
	HABCOB	**-3.86**	**-3.86**	**-3.86**	2.13e-15	**27.00847**
	HABCCH	**-3.86**	**-3.86**	**-3.86**	2.17e-15	26.86845
	HABCIC	**-3.86**	**-3.86**	**-3.86**	2.10e-15	27.34851
	HABCCOB	**-3.86**	**-3.86**	**-3.86**	**2.06e-15**	27.05978
	HABCO	**-3.86**	**-3.86**	**-3.86**	2.11e-15	27.53382
	HABC	**-3.86**	**-3.86**	**-3.86**	2.11e-15	27.39167

Table 5. Best, worst, mean, standard deviation and mean time value obtained by ABC and other six HABC through 25 independent runs on benchmark function f_9 - f_{16} .

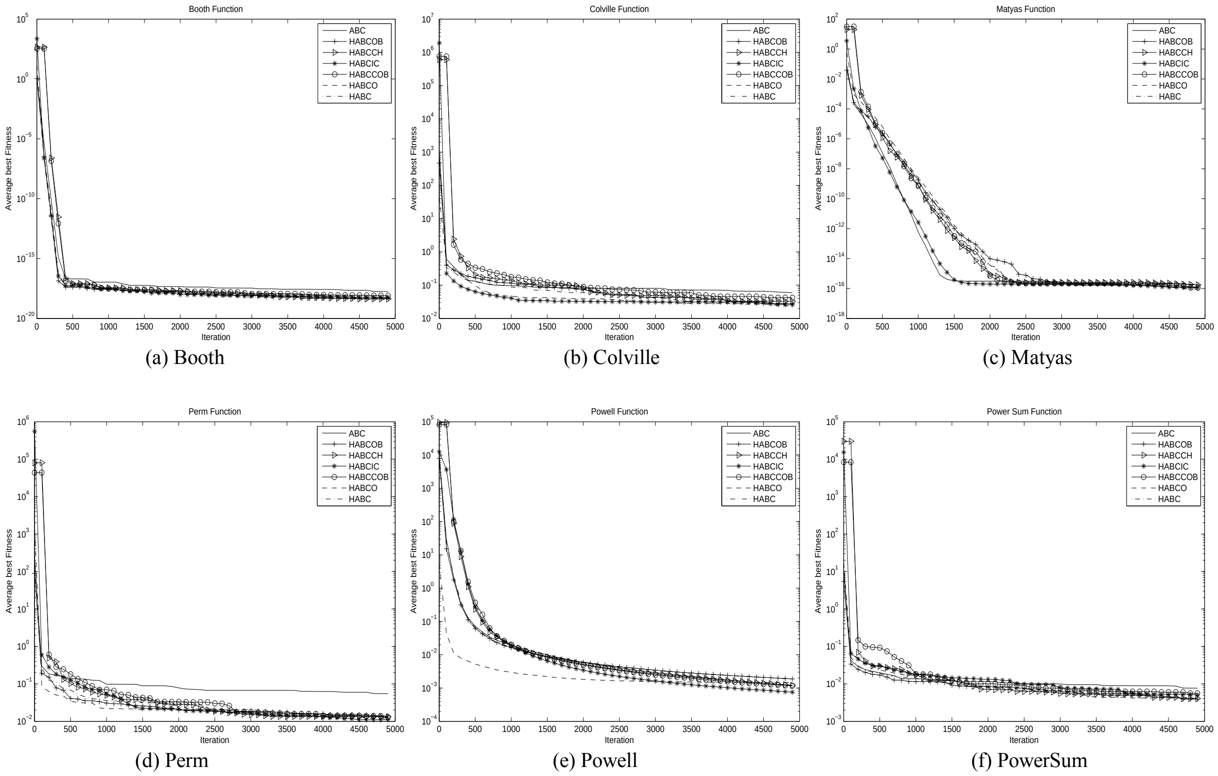
		Best	Mean	Worst	Std	Mean-Time
f_9	ABC	**4.65e-08**	**4.65e-08**	**4.65e-08**	**0**	**16.87600**
	HABCOB	**4.65e-08**	**4.65e-08**	**4.65e-08**	4.44e-17	17.20769
	HABCCH	**4.65e-08**	**4.65e-08**	**4.65e-08**	6.15e-17	17.04306
	HABCIC	**4.65e-08**	**4.65e-08**	**4.65e-08**	**0**	17.47723
	HABCCOB	**4.65e-08**	**4.65e-08**	**4.65e-08**	**0**	17.23549
	HABCO	**4.65e-08**	**4.65e-08**	**4.65e-08**	4.44e-17	17.51557
	HABC	**4.65e-08**	**4.65e-08**	**4.65e-08**	**0**	17.70736
f_{10}	ABC	8.27e-18	1.26e-16	4.59e-16	1.14e-16	**15.75365**
	HABCOB	1.32e-18	1.71e-16	2.81e-15	5.53e-16	16.18157
	HABCCH	1.00e-17	1.48e-16	4.87e-16	1.36e-16	15.91371
	HABCIC	5.29e-18	1.14e-16	3.51e-16	1.01e-16	16.42786
	HABCCOB	**6.36e-19**	1.12e-16	4.64e-16	1.21e-16	16.12124
	HABCO	1.42e-18	1.20e-16	5.32e-16	1.37e-16	16.62848
	HABC	1.85e-17	**8.90e-17**	**3.35e-16**	**9.44e-17**	16.51861
f_{11}	ABC	7.10e-03	5.20e-02	1.65e-01	4.50e-02	**22.32331**
	HABCOB	1.00e-03	1.39e-02	4.17e-02	1.15e-02	22.69221
	HABCCH	2.06e-03	1.32e-02	5.52e-02	1.09e-02	22.46052
	HABCIC	1.77e-03	**1.10e-02**	3.53e-02	8.59e-03	23.07073
	HABCCOB	2.16e-03	1.2e-02	3.92e-02	9.44e-03	22.81637
	HABCO	2.17e-03	1.13e-02	2.78e-02	**6.22e-03**	23.38827
	HABC	**4.54e-04**	1.18e-02	**3.40e-02**	9.32e-03	23.27893
f_{12}	ABC	8.98e-04	1.16e-03	1.38e-03	1.44e-04	**30.42611**
	HABCOB	1.30e-03	1.83e-03	2.82e-03	3.34e-04	31.00383
	HABCCH	6.60e-04	1.16e-03	2.49e-03	3.73e-04	30.83248
	HABCIC	**6.23e-04**	**7.28e-04**	**8.15e-04**	**4.80e-05**	31.27124
	HABCCOB	7.43e-04	1.14e-03	1.95e-03	2.51e-04	31.19106
	HABCO	7.69e-04	1.47e-03	2.39e-03	4. 00e-04	31.76304
	HABC	9.44e-04	1.83e-03	2.76e-03	3.81e-04	31.34683
f_{13}	ABC	**2.22e-04**	7.692e-03	1.83e-02	4.66e-03	**21.53706**
	HABCOB	4.57e-04	**3.906e-03**	1.66e-02	3.95e-03	21.93324
	HABCCH	3.30e-04	4.042e-03	1.84e-02	3.91e-03	21.94299
	HABCIC	4.12e-04	4.663e-03	1.34e-02	4.06e-03	22.28781
	HABCCOB	3.98e-04	5.589e-03	1.25e-02	3.75e-03	22.13560
	HABCO	4.59e-04	4.184e-03	**1.20e-02**	3.12e-03	22.40422
	HABC	2.78e-04	4.219e-03	9.70e-03	**3.01e-03**	22.26531
f_{14}	ABC	**-10.5364**	**-10.5364**	**-10.5364**	1.99e-15	**30.54940**
	HABCOB	**-10.5364**	**-10.5364**	**-10.5364**	2.76e-15	32.45616
	HABCCH	**-10.5364**	**-10.5364**	**-10.5364**	2.61e-15	31.69657
	HABCIC	**-10.5364**	**-10.5364**	**-10.5364**	2.76e-15	31.90898
	HABCCOB	**-10.5364**	**-10.5364**	**-10.5364**	3.03e-15	31.79735
	HABCO	**-10.5364**	**-10.5364**	**-10.5364**	2.90e-15	31.69738
	HABC	**-10.5364**	**-10.5364**	**-10.5364**	**1.36e-15**	31.69646
f_{15}	ABC	**-186.731**	**-186.731**	**-186.731**	1.00e-11	**21.36644**
	HABCOB	**-186.731**	**-186.731**	**-186.731**	2.84e-14	22.21063
	HABCCH	**-186.731**	**-186.731**	**-186.731**	3.01e-14	22.27073
	HABCIC	**-186.731**	**-186.731**	**-186.731**	2.17e-14	22.22425
	HABCCOB	**-186.731**	**-186.731**	**-186.731**	2.84e-14	22.28537
	HABCO	**-186.731**	**-186.731**	**-186.731**	**2.09e-14**	22.09370
	HABC	**-186.731**	**-186.731**	**-186.731**	3.62e-14	22.25839
f_{16}	ABC	**-50**	**-50**	**-50**	1.46e-13	**21.17621**
	HABCOB	**-50**	**-50**	**-50**	1.15e-13	22.05822
	HABCCH	**-50**	**-50**	**-50**	1.21e-13	22.15278
	HABCIC	**-50**	**-50**	**-50**	1.35e-13	22.16214
	HABCCOB	**-50**	**-50**	**-50**	1.57e-13	22.20180
	HABCO	**-50**	**-50**	**-50**	**1.08e-13**	22.03147
	HABC	**-50**	**-50**	**-50**	1.28e-13	22.19336

Table 6. Best, worst, mean, standard deviation and mean time value obtained by ABC and other six HABC through 25 independent runs on benchmark function f_{17} - f_{27} with dimension $D = 30$.

		Best	Mean	Worst	Std	Mean-Time
f_{17}	ABC	2.93e-14	3.81e-14	4.35e-14	3.85e-15	**21.70356**
	HABCOB	**2.58e-14**	3.16e-14	4.00e-14	2.69e-15	22.77789
	HABCCH	2.93e-14	3.27e-14	4.00e-14	2.40e-15	22.17633
	HABCIC	**2.58e-14**	3.16e-14	**3.29e-14**	2.02e-15	22.65846
	HABCCOB	2.93e-14	3.22e-14	4.00e-14	2.71e-15	22.41689
	HABCO	2.93e-14	**3.04e-14**	**3.29e-14**	**1.69e-15**	22.79281
	HABC	2.93e-14	3.06e-14	**3.29e-14**	1.74e-15	22.53482
f_{18}	ABC	1.63e-11	**1.43e-09**	**1.48e-08**	**3.29e-09**	**17.58065**
	HABCOB	1.89e-10	8.14e-06	1.85e-04	3.69e-05	18.01535
	HABCCH	1.16e-10	3.21e-08	2.65e-07	8.06e-08	18.21370
	HABCIC	1.13e-10	3.58e-06	2.30e-05	6.23e-06	18.48829
	HABCCOB	2.80e-11	5.27e-07	1.00e-05	2.00e-06	18.17515
	HABCO	4.21e-09	5.20e-08	4.13e-07	8.45e-08	18.71279
	HABC	**1.45e-11**	1.68e-07	1.38e-06	3.69e-07	18.40993
f_{19}	ABC	**0**	1.60e-16	4.44e-16	1.47e-16	**22.03209**
	HABCOB	**0**	3.55e-17	**1.11e-16**	5.29e-17	22.59067
	HABCCH	**0**	3.11e-17	**1.11e-16**	5.09e-17	22.88428
	HABCIC	**0**	3.11e-17	**1.11e-16**	5.09e-17	23.11554
	HABCCOB	**0**	3.11e-17	**1.11e-16**	5.09e-17	22.86108
	HABCO	**0**	4.88e-17	4.44e-16	9.66e-17	23.47623
	HABC	**0**	**2.66e-17**	**1.11e-16**	**4.84e-17**	23.05462
f_{20}	ABC	3.32e-16	5.02e-16	6.59e-16	7.45e-17	**32.90053**
	HABCOB	2.98e-16	4.16e-16	5.11e-16	6.35e-17	33.64184
	HABCCH	2.45e-16	4.10e-16	5.17e-16	8.13e-17	33.90082
	HABCIC	**1.99e-16**	**3.46e-16**	5.41e-16	7.81e-17	34.08160
	HABCCOB	2.86e-16	4.21e-16	5.07e-16	6.57e-17	33.89444
	HABCO	2.83e-16	4.13e-16	**4.93e-16**	**5.56e-17**	34.66688
	HABC	2.61e-16	4.03e-16	4.97e-16	6.34e-17	34.09351
f_{21}	ABC	-29.6275	-29.6176	-29.6059	6.20e-03	**25.27370**
	HABCOB	**-29.6309**	-29.6288	-29.6206	2.96e-03	25.95365
	HABCCH	**-29.6309**	-29.6271	-29.5882	8.69e-03	26.10902
	HABCIC	**-29.6309**	-29.6266	-29.5881	1.15e-02	26.29809
	HABCCOB	**-29.6309**	-29.6290	-29.6230	2.17e-03	26.17266
	HABCO	**-29.6309**	**-29.6302**	**-29.6273**	**8.40e-04**	26.57591
	HABC	**-29.6309**	-29.6298	-29.6254	1.62e-03	26.33754
f_{22}	ABC	**0**	6.82e-15	5.68e-14	1.89e-14	**18.69893**
	HABCOB	**0**	**0**	**0**	**0**	19.18100
	HABCCH	**0**	**0**	**0**	**0**	19.46644
	HABCIC	**0**	6.82e-15	5.68e-14	1.89e-14	19.58300
	HABCCOB	**0**	**0**	**0**	**0**	19.40121
	HABCO	**0**	**0**	**0**	**0**	19.84200
	HABC	**0**	**0**	**0**	**0**	19.53271
f_{23}	ABC	1.86e-04	1.34e-02	8.19e-02	2.38e-02	**18.19598**
	HABCOB	1.07e-04	1.41e-02	1.10e-01	2.48e-02	18.81008
	HABCCH	9.59e-05	1.07e-01	1.97e-00	3.93e-01	18.93122
	HABCIC	9.73e-04	9.59e-02	1.25e-00	2.51e-01	19.22760
	HABCCOB	6.02e-04	2.33e-01	3.67e-00	7.52e-01	18.93006
	HABCO	6.45e-03	1.38e-02	**2.13e-02**	**4.04e-03**	19.22689
	HABC	**8.18e-05**	**1.13e-02**	6.78e-02	1.78e-02	19.17352
f_{24}	ABC	**3.82e-04**	**3.82e-04**	**3.82e-04**	6.81e-13	**24.67135**
	HABCOB	**3.82e-04**	**3.82e-04**	**3.82e-04**	7.43e-13	25.42231
	HABCCH	**3.82e-04**	**3.82e-04**	**3.82e-04**	7.93e-13	25.00704
	HABCIC	**3.82e-04**	**3.82e-04**	**3.82e-04**	**6.03e-13**	25.22714
	HABCCOB	**3.82e-04**	**3.82e-04**	**3.82e-04**	1.34e-12	25.14297
	HABCO	**3.82e-04**	**3.82e-04**	**3.82e-04**	6.81e-13	25.49795
	HABC	**3.82e-04**	**3.82e-04**	**3.82e-04**	8.91e-13	25.46099
f_{25}	ABC	4.38e-16	5.05e-16	5.48e-16	3.35e-17	**22.23210**
	HABCOB	2.99e-16	4.22e-16	4.97e-16	6.54e-17	23.61736
	HABCCH	2.81e-16	4.19e-16	5.35e-16	6.70e-17	23.68333
	HABCIC	2.99e-16	4.10e-16	5.00e-16	6.85e-17	23.44918
	HABCCOB	2.41e-16	4.14e-16	5.03e-16	8.09e-17	23.86629
	HABCO	**1.21e-19**	**1.95e-16**	**4.84e-16**	**1.98e-16**	23.57562
	HABC	2.92e-16	4.08e-16	4.92e-16	6.91e-17	23.39717
f_{26}	ABC	3.06e-16	4.90e-16	5.51e-16	6.12e-17	**22.68787**
	HABCOB	2.87e-16	4.20e-16	5.21e-16	7.44e-17	23.93296
	HABCCH	3.03e-16	**3.81e-16**	**4.88e-16**	**5.84e-17**	24.04358
	HABCIC	2.61e-16	3.98e-16	5.00e-16	7.75e-17	23.84461
	HABCCOB	**2.50e-16**	3.82e-16	5.15e-16	7.86e-17	24.24357
	HABCO	2.92e-16	3.87e-16	5.05e-16	6.98e-17	24.04727
	HABC	3.05e-16	4.24e-16	5.15e-16	6.41e-17	23.89610
f_{27}	ABC	9.12e+01	1.68 e+02	2.35e+02	3.13e+01	**22.89231**
	HABCOB	1.16e+02	1.76e+02	2.41e+02	3.17e+01	24.36207
	HABCCH	1.83e+02	2.58e+02	3.52e+02	5.15e+01	24.08142
	HABCIC	1.68e+02	2.68e+02	3.30e+02	4.91e+01	24.03064
	HABCCOB	1.50e+02	2.52e+02	3.89e+02	6.66e+01	24.19654
	HABCO	**1.59e-00**	**3.38e-00**	**8.77e-00**	**1.78e-00**	25.14578
	HABC	1.01e+02	1.93e+02	2.52e+02	3.94e+01	24.23884

Table 7. Best, worst, mean, standard deviation and mean time value obtained by ABC and other six HABC through 25 independent runs on benchmark function f_{17} - f_{27} with dimension $D = 100$.

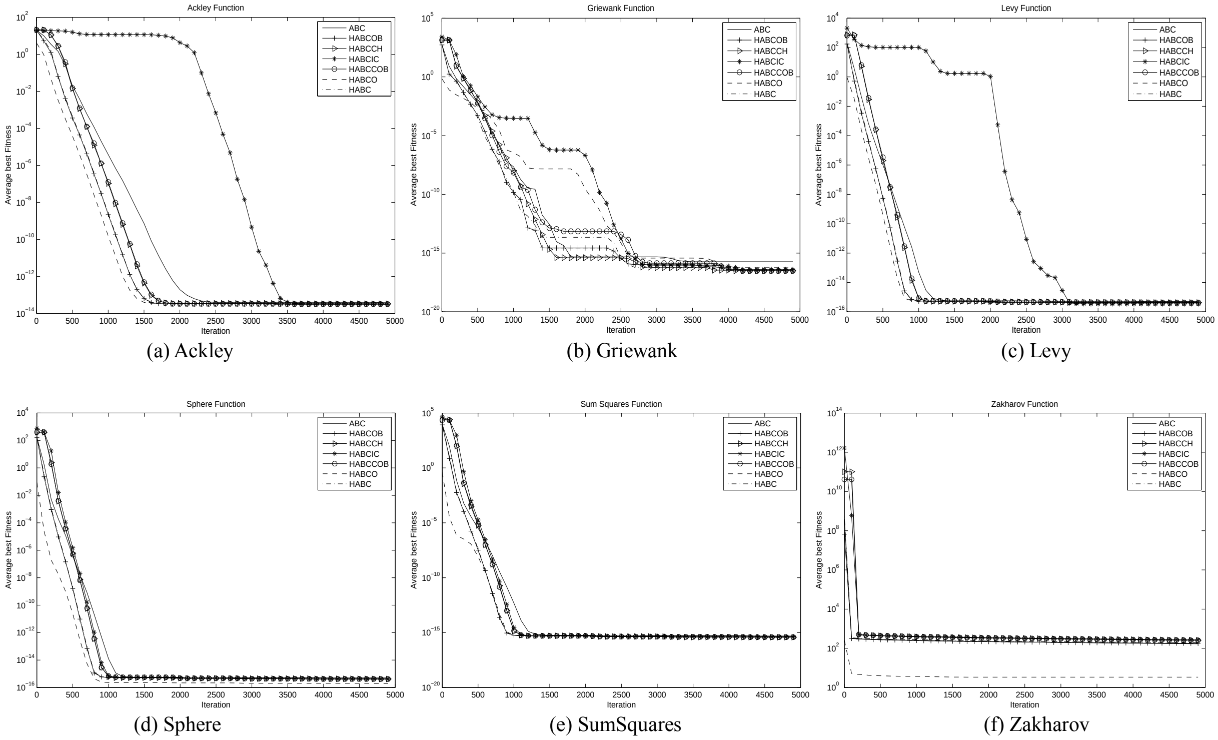
		Best	Mean	Worst	Std	Mean-Time
f_{17}	ABC	8.70e-10	2.16e-09	4.23e-09	9.00e-10	**27.42205**
	HABCOB	2.71e-13	3.14e-13	3.53e-13	2.51e-14	28.17576
	HABCCH	4.70e-13	7.05e-13	1.34e-12	1.75e-13	28.39553
	HABCIC	9.01e-00	1.17e+01	1.35e+01	1.20e-00	29.17666
	HABCCOB	5.84e-13	6.68e-13	7.61e-13	5.31e-14	28.36998
	HABCO	**1.47e-13**	**1.65e-13**	**1.75e-13**	**8.47e-15**	40.89752
	HABC	2.60e-13	3.19e-13	3.77e-13	3.30e-14	27.96116
f_{18}	ABC	3.91e-04	**1.69e-03**	**4.65e-03**	1.10e-03	**26.87399**
	HABCOB	2.58e-04	6.20e-01	7.14e-00	1.57e-00	27.88947
	HABCCH	4.59e-04	7.44e-01	1.01e+01	2.25e-00	27.87438
	HABCIC	1.29e-02	4.50e-00	1.73e+01	5.08e-00	27.26161
	HABCCOB	5.48e-04	5.45e-02	1.15e-00	2.28e-01	28.08510
	HABCO	6.67e-01	6.67e-01	6.68e-01	**2.33e-04**	39.36317
	HABC	**1.04e-04**	5.56e-01	4.90e-00	1.31e-00	27.74415
f_{19}	ABC	9.99e-16	5.21e-15	4.81e-14	9.27e-15	**29.54152**
	HABCOB	9.99e-16	1.20e-13	2.66e-12	5.30e-13	31.11756
	HABCCH	5.55e-16	9.18e-11	2.30e-09	4.59e-10	31.43440
	HABCIC	**4.44e-16**	6.56e-13	1.48e-11	2.96e-12	31.06522
	HABCCOB	8.88e-16	2.06e-11	5.10e-10	1.02e-10	31.43196
	HABCO	4.11e-15	1.20e-07	2.89e-06	5.77e-07	48.57922
	HABC	9.99e-16	**4.20e-15**	**1.89e-14**	**4.98e-15**	30.98284
f_{20}	ABC	2.53e-15	3.14e-15	3.66e-15	2.95e-16	**62.18824**
	HABCOB	2.29e-15	2.65e-15	3.17e-15	2.30e-16	63.43297
	HABCCH	2.30e-15	2.67e-15	3.13e-15	2.02e-16	63.48975
	HABCIC	4.26e-06	1.95e+01	1.02e+02	3.18e+01	62.75135
	HABCCOB	**2.10e-15**	**2.62e-15**	**2.98e-15**	1.86e-16	63.74450
	HABCO	2.53e-15	2.78e-15	3.12e-15	1.89e-16	84.86471
	HABC	2.30e-15	2.80e-15	3.20e-15	**1.72e-16**	63.15849
f_{21}	ABC	-98.3566	-97.9593	-97.4569	2.52e-01	40.77573
	HABCOB	-98.9313	-98.5219	-98.249	1.67e-01	41.93359
	HABCCH	-98.5658	-98.1583	-97.7409	1.84e-01	41.98199
	HABCIC	-97.8113	-92.8133	-88.6395	2.29e-00	**40.44041**
	HABCCOB	-98.5910	-98.1587	-97.4723	2.31e-01	42.11596
	HABCO	-98.9559	-98.7545	-98.5844	9.32e-02	58.16782
	HABC	-98.7921	-98.4558	-98.1394	1.55e-01	41.84345
f_{22}	ABC	3.41e-13	5.34e-09	9.09e-08	1.85e-08	**27.92113**
	HABCOB	**2.27e-13**	5.64e-13	1.48e-12	2.96e-13	29.00650
	HABCCH	3.41e-13	4.17e-04	6. 81e-03	1.52e-03	28.94996
	HABCIC	4.98e+01	1.55e+02	2.51e+02	5.73e+01	28.52347
	HABCCOB	**2.27e-13**	1.50e-03	3.71e-02	7.42e-03	28.95015
	HABCO	**2.27e-13**	5.14e-13	**7.96e-13**	**1.47e-13**	42.21861
	HABC	**2.27e-13**	**4.91e-13**	9.09e-13	1.82e-13	28.88114
f_{23}	ABC	**7.78e-03**	**6.16e-02**	**1.78e-01**	**4.09e-02**	27.95234
	HABCOB	3.25e-02	6.98e-01	3.18e-00	8.34e-01	29.33051
	HABCCH	1.07e-02	1.09e+01	8.30e+01	2.47e+01	28.84404
	HABCIC	7.75e+01	1.26e+02	1.99e+02	3.26e+01	**27.48526**
	HABCCOB	5.56e-03	3.36e-00	3.87e+01	7.76e-00	29.10317
	HABCO	3.70e-02	1.23e-01	4.05e-01	1.19e-01	41.59810
	HABC	1.76e-02	5.71e-01	4.72e-00	9.86e-01	29.22457
f_{24}	ABC	2.02e-03	3.87e+02	8.29e+02	2.09e+02	**25.57455**
	HABCOB	**1.27e-03**	**1.27e-03**	1.28e-03	2.20e-06	27.92126
	HABCCH	3.20e+02	8.77e+02	1.42e+03	2.93e+02	26.63203
	HABCIC	**1.27e-03**	9.53e-00	1.18e+02	3.28e+01	27.17422
	HABCCOB	2.37e+02	9.42e+02	1.66e+03	3.49e+02	26.62176
	HABCO	**1.27e-03**	**1.27e-03**	**1.27e-03**	**3.38e-11**	43.24939
	HABC	**1.27e-03**	3.61e-03	5.97e-03	1.17e-02	27.79104
f_{25}	ABC	2.76e-15	3.24e-15	3.63e-15	2.23e-16	**24.52196**
	HABCOB	2.52e-15	2.80e-15	3.12e-15	**1.68e-16**	25.71166
	HABCCH	**1.88e-15**	2.69e-15	3.20e-15	2.85e-16	25.88816
	HABCIC	1.98e-15	**2.65e-15**	**2.98e-15**	2.59e-16	25.58481
	HABCCOB	2.32e-15	2.72e-15	2.99e-15	1.79e-16	25.91885
	HABCO	2.19e-15	2.76e-15	3.17e-15	1.98e-16	37.45111
	HABC	2.26e-15	2.66e-15	3.20e-15	2.51e-16	25.60685
f_{26}	ABC	2.50e-15	3.19e-15	3.77e-15	3.78e-16	**24.92382**
	HABCOB	**2.02e-15**	2.68e-15	3.20e-15	2.34e-16	26.53891
	HABCCH	2.06e-15	2.72e-15	3.17e-15	2.55e-16	26.69612
	HABCIC	2.25e-15	2.70e-15	2.99e-15	2.11e-16	26.46492
	HABCCOB	2.10e-15	**2.63e-15**	2.98e-15	2.17e-16	26.83566
	HABCO	2.28e-15	2.64e-15	3.18e-15	2.04e-16	38.26213
	HABC	2.07e-15	2.69e-15	**2.97e-15**	**2.02e-16**	26.41392
f_{27}	ABC	1158.267	1283.235	1372.793	**4.38e+01**	**25.23831**
	HABCOB	1212.111	1342.483	1470.248	5.96e+01	27.00475
	HABCCH	1696.155	1913.789	2113.579	1.17e+02	26.95500
	HABCIC	1189.182	1361.141	1476.671	7.33e+01	26.82238
	HABCCOB	1542.838	1873.756	2065.895	1.39e+02	26.88027
	HABCO	1130.100	1367.226	1476.271	7.73e+01	38.83916
	HABC	1220.647	1323.562	1466.346	6.51e+01	26.65520



(a) Booth (b) Colville (c) Matyas

(d) Perm (e) Powell (f) PowerSum

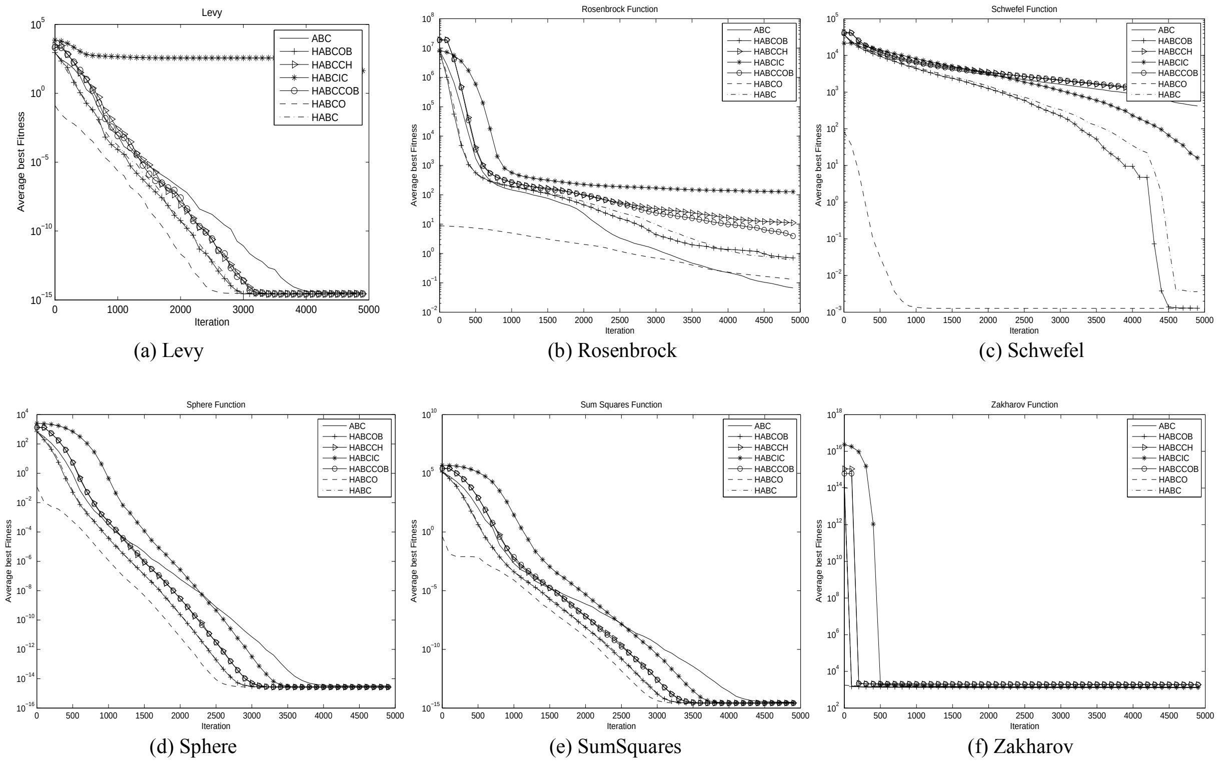
Figure 1. Mean of best function values for test functions.



(a) Ackley (b) Griewank (c) Levy

(d) Sphere (e) SumSquares (f) Zakharov

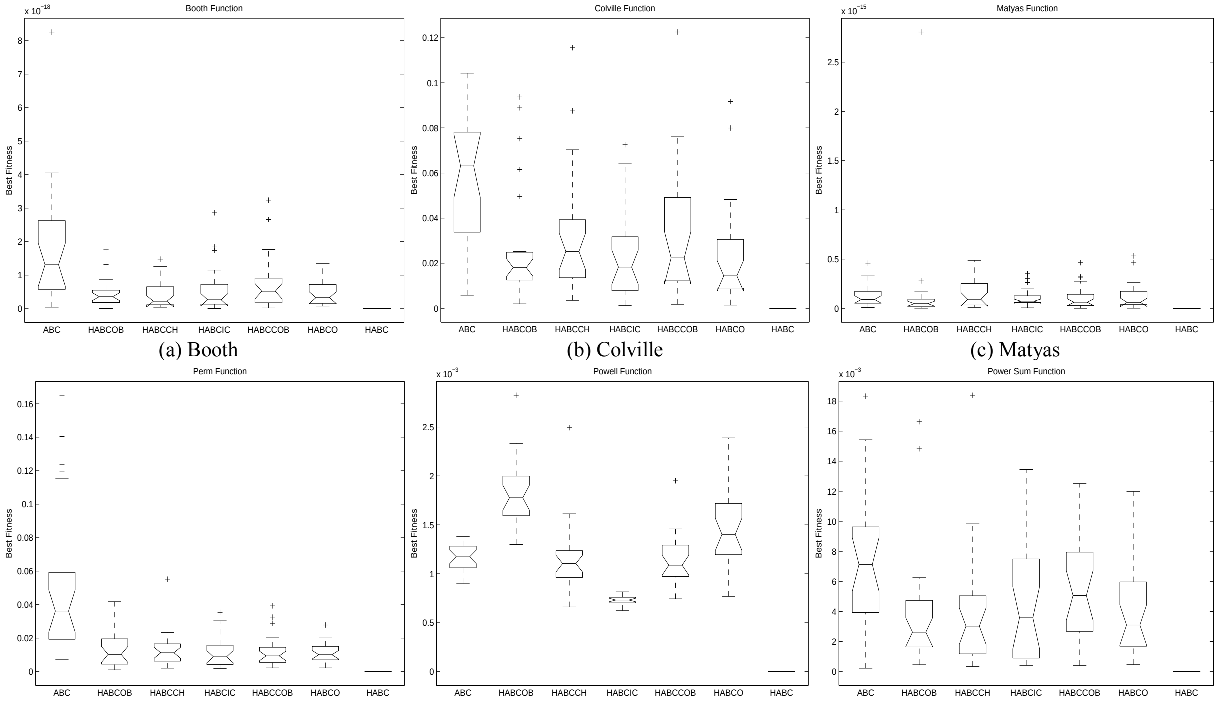
Figure 2. Mean of best function values for test functions with dimension $D = 30$.



(a) Levy (b) Rosenbrock (c) Schwefel

(d) Sphere (e) SumSquares (f) Zakharov

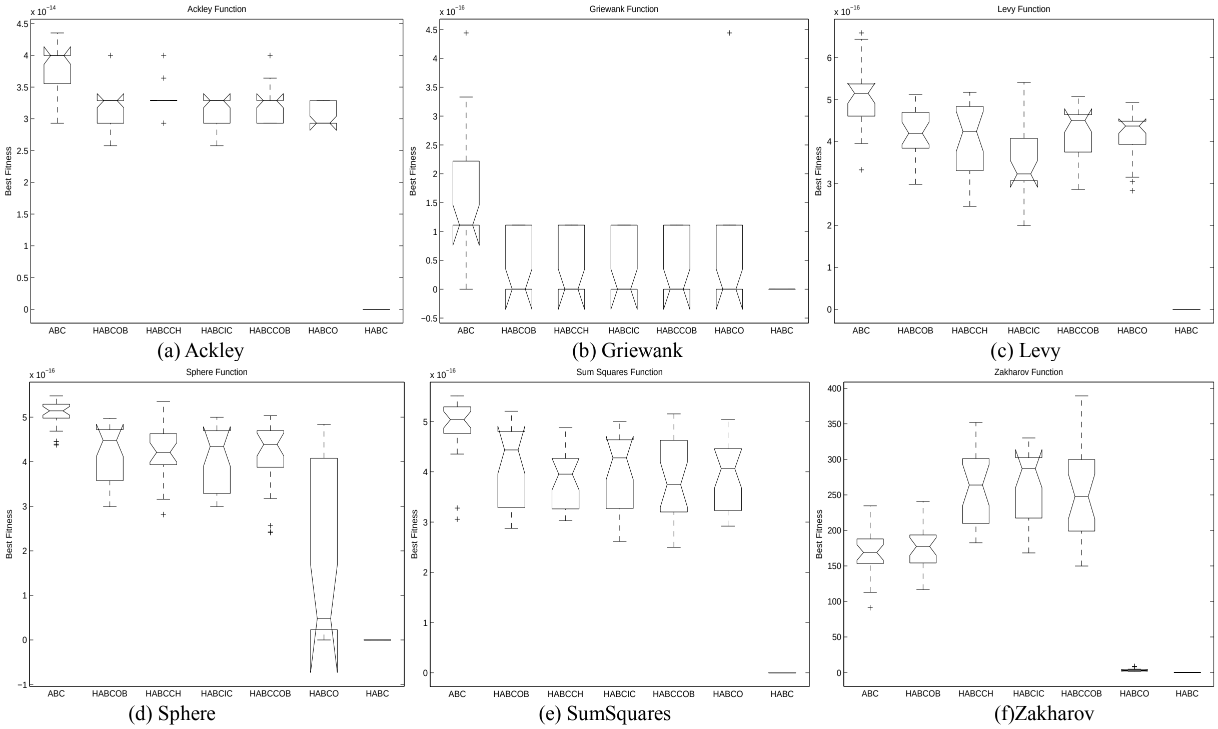
Figure 3. Mean of best function values for test functions with dimension $D = 100$.



(a) Booth (b) Colville (c) Matyas

(d) Perm (e) Powell (f) PowerSum

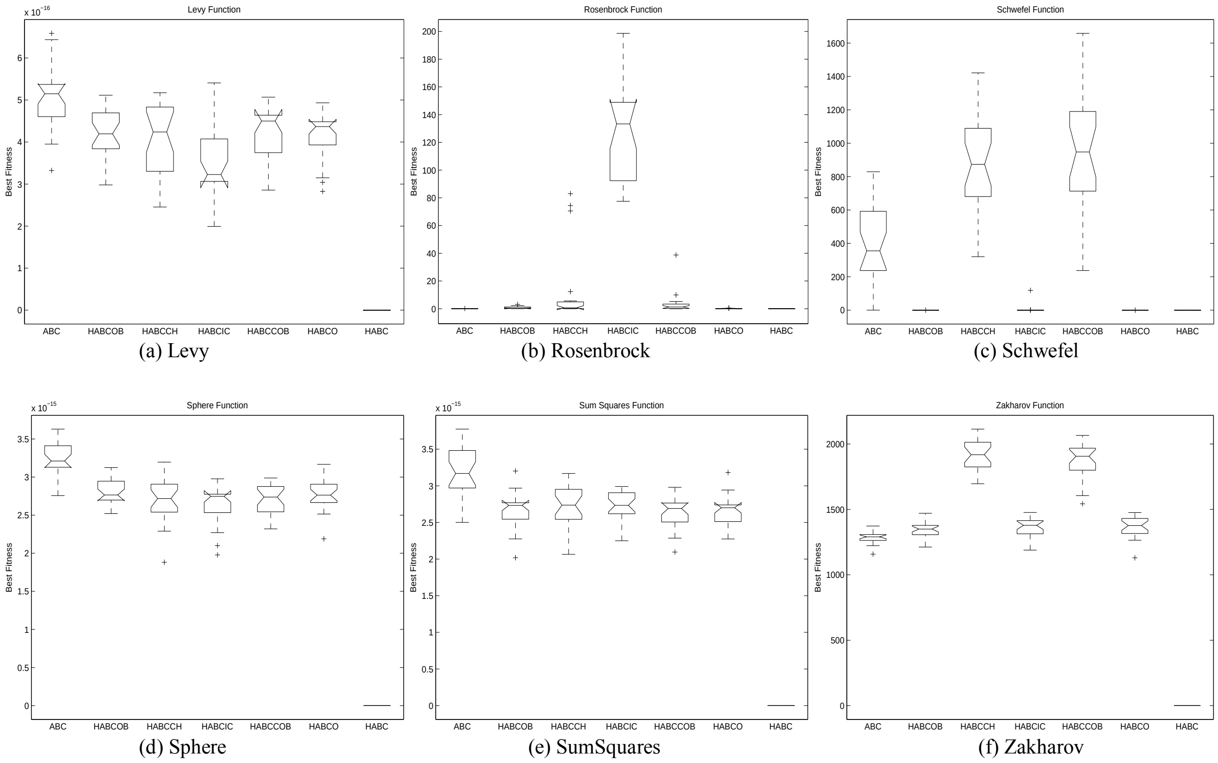
Figure 4. Statistical values of the function values for test functions.



(a) Ackley (b) Griewank (c) Levy

(d) Sphere (e) SumSquares (f)Zakharov

Figure 5. Statistical values of the function values for test functions with dimension $D = 30$.



(a) Levy (b) Rosenbrock (c) Schwefel

(d) Sphere (e) SumSquares (f) Zakharov

Figure 6. Statistical values of the function values for test functions with dimension $D = 100$.

[3] M. Dorigo, T. Stutzle. Ant colony optimization. Cambridge: MAMIT Press, 2004.

[4] R. Storn, K. Price. Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimization, 1997, 11(4):341-359.

[5] J. Kennedy, R. Eberhart. Particle swarm optimization. In IEEE International Conference on Neural Networks, 1995, pp. 1942-1948.

[6] D. Karaboga. An idea based on honey bee swarm for numerical optimization. Technical report-tr06, Kayseri, Turkey: ErciyesUniversity, 2005.

[7] D. Karaboga, B.Basturk. A comparative study of artificial bee colony algorithm. Applied Mathematics and Computation, 2009, 214(1):108-132. [8] A. Singh. An artificial bee colony algorithm for the leaf constrained minimum spanning tree problem. Applied Soft Computing Journal, 2009, 9(2):625-631.

[9] Q. K. Pan, M. F. Tasgetiren, P. Suganthan, T. Chua. A discrete artificial bee colony algorithm for the lot-streaming flow shop scheduling problem. Information Sciences, 2011, 181(12):2455-2468.

[10] F. Kang, J. Li, Q. Xu. Structural inverse analysis by hybrid simplex artificial bee colony algorithms. Computers and Structures, 2009, 87(13):861-870.

[11] R. S. Rao, S. V. L. Narasimham, M. Ramalingaraju. Optimization of distribution network configuration for loss reduction using artificial bee colony algorithm. International Journal of Electrical Power and Energy Systems Engineering, 2008, 1(2):116-122.

[13] C. S. Zhang, D. T. Ouyang, J. X. Ning. An artificial bee colony approach for clustering. Expert Systems with Application, 2010, 37(7):4761-4767.

[14] Z. Hu, M. Zhao. Simulation on traveling salesman problem (TSP) based on artificial bees colony algorithm. Transaction of Beijing Institute of Technology, 2009, 29(11):978-982.

[15] Y. W. Leung, Y. P. Wang. An orthogonal genetic algorithm with quantization for global numerical optimization. IEEE Transactions of Evolutionary Computation, 2001, 5(1):41-53.

[16] B. Alatas. Chaotic bee colony algorithms for global numerical optimization. Expert Systems with Applications, 2010, 37(8):5682-5687.

[17] S. Rahnamayan, et al. Opposition-based differential evolution. IEEE Transaction on Evolutionary Computation, 2008, 12(1):64–79.

[18] W. Gao, S. Liu. A modified artificial bee colony algorithm. Computers and Operations Research, 2012, 39(3):687-697.

[19] Y. W. Leung, Y. P. Wang. An orthogonal genetic algorithm with quantization for global numerical optimization. IEEE Transactions of Evolutionary Computation, 2001, 5(1):41-53.

[20]X. Kong, et al. . Hybrid Artificial Bee Colony Algorithm for Global Numerical Optimization. Journal of Computational Information System, 2012, 8(6):2367-2374.

Xiangyu Kong received his B.A. in Applied Mathematics from Hunan University, Changsha, China, in 2004, the M.S. degree in Applied Mathematics from Xidian University, Xi’an, China, in 2009. Since the year of 2011, he started his learning in Applied Mathematics from Xidian University for his Ph. D. degree. He joined Department of Applied Mathematics, Zhoukou Normal University, China, in 2004 as a research/teaching assistant and then became Instructor in the same department. His main interests include Intelligent Optimization, Theory and Methods of Optimization.

Sanyang Liu is a professor and a tutor of Ph.D. in Xidian University. His current research interests include optimization theory and algorithm, machine learning and pattern recognition.

Zhen Wang lecturer received her Ph.D. degree in Applied Mathematics from Xidian University in 2012. She is currently with Institute of Information and System Computation Science, Beifang University of Nationalities, China. Her main research areas include Mathematical Finance, Portfolio Optimization, and Intelligent Optimization.

Fast Affinity Propagation Clustering based on Machine Learning

Shailendra Kumar Shrivastava[1], Dr. J.L. Rana[2] and Dr. R.C. Jain[3]

**[1] Samrat Ashok Technological Institute
Vidisha, Madhya Pradesh 464 001, India**

**[2] Ex Head of Department, CSE, M.A.N.I.T
Bhopal, Madhya Pradesh, India**

**[3] Samrat Ashok Technological Institute
Vidisha, Madhya Pradesh 464 001, India**

Abstract

Affinity propagation (AP) was recently introduced as an un-supervised learning algorithm for exemplar based clustering. In this paper a novel Fast Affinity Propagation clustering Approach based on Machine Learning (FAPML) has been proposed. FAPML tries to put data points into clusters based on the history of the data points belonging to clusters in early stages. In FAPML we introduce affinity learning constant and dispersion constant which supervise the clustering process. FAPML also enforces the exemplar consistency and one of 'N' constraints. Experiments conducted on many data sets such as Olivetti faces, Mushroom, Documents summarization, Thyroid, Yeast, Wine quality Red, Balance etc. show that FAPML is up to 54 % faster than the original AP with better Net Similarity.

Keywords: *clustering, affinity propagation, exemplar, machine learning, unsupervised learning*

1. Introduction

Clustering is a fundamental task in computerized data analysis. It is concerned with the problem of partitioning a collection of data points into groups/categories using unsupervised learning techniques. Data points in groups are similar. Such groups are called clusters [1][2][3]. Affinity propagation [6] is a clustering algorithm which for given set of similarities (also denoted by affinities) between pairs of data points, partitions the data by passing the messages among the data points. Each partition is associated with a prototypical point that best describes that cluster. AP associates each data point with one such prototype. Thus, the objective of AP is to maximize the overall sum of similarities between data points and their representatives. Affinity propagation clustering algorithm is slow. Fast affinity algorithms for clustering find the clusters in less time as compared to AP. Efforts of Earlier researcher to make AP fast, yielded only limited benefits. Proposed FAPML finds the clusters in much less time as compared to AP and net similarity is much better than AP. We will first discuss the disadvantages of exiting Fast AP. FSAP [9] constructs the sparse similarity matrix by K-nearest neighbor algorithm .The FSAP does not give same result as AP. FSAP uses heuristic approach to find K. This reduces the cluster quality. FAP (based on message pruning) [10] Prunes the unnecessary messages exchange among data points in iterations to compute the convergence. This algorithm requires the extra time to find the necessary and unnecessary messages. Fast affinity propagation clustering (based on sampling of data points) [11] algorithm applies the sampling theorem to choose a small number of representative exemplar whose number is much less than data points but larger than the clusters. Clustering quality is still not as good as AP.

In the Literature[5] Machine Learning is defined as: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Proposed FAPML is based on this definition of machine learning. FAPML does not have disadvantages of earlier reported fast affinity propagation algorithms. FAPML enforces the one of 'N' constraint and exemplar consistency. One of 'N' constraint means that data points belong exactly in one cluster. Exemplar consistency mean if other data points do not choose the given data point as exemplar than given data point cannot choose itself as an exemplar. Proposed FAPML tries to put data points into clusters based on the history of the data points belonging to clusters in early stages. Proposed algorithm has affinity learning constant and dispersion constant. By Affinity learning constant it uses experience in the clustering process to put data points in same clusters and by dispersion constant it uses experience in the process of clustering to put data point into different clusters.

The remainder of this paper is organized as follows. Section 2 gives a brief over view of original Affinity Propagation algorithm, FSAP, Fast algorithm for Affinity propagation (based on message pruning), Fast affinity propagation clustering (based on sampling of data points). Section 3 introduces the main idea and details of our algorithm. Section 4 discusses the experimental results and evaluation. Section 5 provides the concluding remarks and future directions.

2. Related works

Before we go into details of our FAPML approach, we would briefly review some works that are closely related to this paper. FSAP, Fast algorithm for Affinity propagation (based on message pruning) and *Fast* affinity propagation clustering (based on sampling of data points) will be discuss. For the sake of continuity affinity propagation algorithm will be discuses first.

2.1 Affinity Clustering Algorithms

Affinity clustering algorithm [6][9][10] is based on message passing among data points. Each data point receives the availability from others data points (from exemplar) and send the responsibility message to others data points (to exemplar). Sum of responsibilities and availabilities for data points identify the exemplars. After the identification of exemplar the data points are assigned to exemplar to form the clusters. Following are the steps of affinity clustering algorithms.

1. Initialize the availabilities to zero $a(i,k) = 0$
2. Update the responsibilities by following equation. $r(i,k) \longleftarrow (s(i,k) + \max_{k' st\ k' \neq k}\{a(i,k' + s(i,k')\}$ Where $s(i,k)$ is the similarity of data point i and exemplar k.
3. Update the availabilities by following equation
$$a(i,k) \leftarrow min\left\{0, r(k,k) + \sum_{i' s.t. i' \notin \{i,k\}} max\{0, r(i',k)\}\right\}$$
Update self-availability by following equation
$$a(k,k) \longleftarrow \sum \max(\{0, r(i',k)\})$$
4. Compute sum $= a(i,k) + r(i,k)$ for data point i and find the value of k that maximize the sum to identify the exemplars.
5. If Exemplars do not change for fixed number of iterations go to step (6) else go to Step (1)
6. Assign the data points to Exemplars on the basis of maximum similarity to find clusters.

2.2 Fast sparse affinity propagation (FSAP)

Jia et al [9] proposed fast sparse affinity propagation (FSAP) clustering algorithm. First step is construction of sparse similarity matrix. Presume that the data points that are far apart will not choose each other as an exemplar and set the similarity between them as zero. Construct the similarity matrix by K-nearest neighbor algorithm .Second step is iterative edge refinement. Data points that serve as good exemplar locally may be candidate for exemplar globally. Third step uses AP to find exemplar and clusters. Complexity of this algorithm is O(NT). Where N is number of data point and T is number of non-zero entries in sparse matrix. Jia et al apply this algorithm for organizing of image Search results obtained from state-of-the-art image search engines. It discovers exemplars from search results and simultaneously groups the images. The exemplars are delivered to the user as a summary of search results instead of the large amount of unorganized images. The FSAP does not give same result as AP. FSAP uses heuristic approach to find K. The improper value of K reduces the cluster quality.

2.3 Fast algorithm for Affinity propagation (based on message pruning)

Fujiwara et al. [10] proposed Fast algorithm for Affinity propagation (FAP). FAP overcomes the drawback of FSAP. Computational Complexity is $O(N^2 + MT)$ .Where N is number of data points; M is number of entries in similarity matrix. T is the number of iterations. FAP prunes the unnecessary message exchanges among data points in each iteration to compute the convergence (Mathematically Proved that unnecessary pruned messages can be recovered from un-pruned message). Then Computes the convergence values of pruned message from the un-pruned messages. Rest of the algorithm steps is same as AP.

2.4 Fast affinity propagation clustering (based on sampling of data points)

Shang et al. [11] proposed fast affinity propagation clustering (FAP). This Algorithm applies the fast sampling theorem to choose a small number of representative exemplar whose number is much less than data points and larger than the clusters. Secondly the representative exemplar is assigned cluster labels by a density-weighted spectral clustering method. In First step the graph is coarsened by fast sampling algorithm to collapse the neighboring data points into subsets of representative exemplar. In second step density weighted spectral clustering is applied on set of final representative exemplars and last step is to assign cluster membership for each data point corresponding to its representative exemplar. FAP outperforms both spectral clustering and AP in terms of quality, speed, and memory usage.

2.5 Binary Variable Model for affinity propagation clustering

Givoni et al [8] proposed A "Binary Variable Model for affinity propagation clustering". It is a graphical model of AP. This model enforces two constraints. First one of 'N' Constraints $\sum_{i=1}^{N} c_{ij} = 1$ where $\{c_{ij}\}_{i=1,j=1}^{N}$ is the binary variable. One of 'N' constraint ensures that one data point belongs to exactly one exemplar/cluster. Second constraint exemplar consistency ensures that if data point k is chosen as exemplar by other data point i then k must choose itself as an exemplar. We will extend the idea of one of 'N' constraint and exemplar consistency.

3. Proposed FAPML

Fast Affinity Propagation based on machine learning takes as input a collection of real-valued similarities among data points, where the similarity $s(i,k)$ indicates how well the data point with index k is suited to be the class center for data point i. In the process of FAPML availability and responsibility messages are exchanged among data points. Initially all data points can become the candidate exemplar. The responsibility message $r(i,k)$ is sent from data point i to candidate exemplar k. The availability message(i,k), sent from candidate exemplar k to data point i. A responsibility is updated from the modified equations which are as follow.

$$r(i,k) \longleftarrow (s(i,k) + al(i,k) - dl(i,k)) - \max_{k' st\ k' \neq k}\{a(i,k' + (st(i,k') + al(i,k) - dl(i,k)\} \tag{1}$$

Where $al(i,k)$ and $dl(i,k)$ are affinity learning experience and dispersion learning experience between point i and k. In this way we use the machine learning technique learning by experience. Availability message are updated by following equation

$$a(i,k) \leftarrow min\{0, r(k,k) + \sum_{i' s.t. i' \notin \{i,k\}} max\{0, r(i',k)\}\} \tag{2}$$

Equation for updating self-availability

$$a(k,k) \longleftarrow \sum \max(\{0, r(i',k)\}) \tag{3}$$

Next we find the exemplar for point i by finding the value of k (exemplar) that maximizes $r(i,k) + a(i,k)$. Now we enforce the one of 'N' constraint. One of 'N' constraint means each point becomes member of exactly one exemplar/cluster. This uses the array with index data point and its value is exemplar. In array only one value can be store hence each data point has exactly one exemplar.

Next we handle exemplar consistency. Exemplar consistency ensures if data point k is chosen as exemplar by other data point i then k must chose itself as an exemplar. If k does not choose itself as exemplar then assign the similarity between i and k to -∞ . This enforces the exemplar consistence. Repeat above process, if exemplar does not change fixed number iterations or changes in results are below threshold.

FAPML algorithm can be written as following.

1. Initialize the availabilities to zero $a(i,k) = 0$, initialize affinity learning variable $al(i,j) = 0$ and dispersion learning variable$dl(i,j) = 0$.
2. Update the responsibilities by following novel equation.
$r(i,k) \longleftarrow (s(i,k) + al(i,k) - dl(i,k)) - \max_{k' st\ k' \neq k}\{a(i,k' + (st(i,k') + al(i,k) - dl(i,k)\}$ Where $al(i,k)$ and $dl(i,k)$ are affinity learning experience and dispersion learning experience between point i and k.
3. Update the availabilities by following equation
$$a(i,k) \leftarrow min\{0, r(k,k) + \sum_{i' s.t. i' \notin \{i,k\}} max\{0, r(i',k)\}\}$$
Update self-availability by following equation
$$a(k,k) \longleftarrow \sum \max(\{0, r(i',k)\})$$
4. Compute sum = $a(i,k) + r(i,k)$ for data point i and find the value of k that maximizes the sum to identify the exemplars.
5. Increase $al(i,k)$ by $al\ constant$ where i and k are the index of data point and k is the index of exemplar of same cluster. Increase the value of $dl(i,k))\ by\ dl\ constant$ for data point i and exemplar k of different cluster.
6. Check exemplar chosen by other data points in step (4). If exemplar does not choose itself as an exemplar, update similarity of data points (chosen exemplar) to exemplar to minus infinity. This enforces the exemplar consistency.

7. Enforce the one of 'N' constraint.
8. If Exemplars do not change for fixed number of iterations go to step (9) else go to Step (2)
9. Assign the data points to Exemplars on the basis of maximum similarity to find clusters.

4. Experimental Results and Evaluation

In this Section, we present results and evaluation of set of experiments to verify the effectiveness and efficiency of our proposed algorithm for clustering. We conducted experiments on Olivetti faces, Mushroom, Thyroid, Yeast, document summarization, Wine quality red and balance data sets. Details of data sets are as follow:

Table 1

S.No.	Dataset	No. of Instances	Number of Attributes	References
1	Yeast	1484	8	http://archive.ics.uci.edu/ml/machine-learning-databases/yeast/
2	Olivetti faces	900	40	http://www.psi.toronto.edu/
3	Thyroid	215	6	http://archive.ics.uci.edu/ml/machine-learning-databases/Thyroid/
4	Document Summarization	125	6	http://www.psi.toronto.edu/
5	Mushroom	5807	22	http://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/
6	Wine Quality Red	1599	12	http://archive.ics.uci.edu/ml/machine-learning-databases/winequality/
7	Balance	625	3	http://archive.ics.uci.edu/ml/machine-learning-databases/balance-scale/

The measures we use to compare the algorithms are the net similarity/sum of similarities of all non-exemplar data points to their exemplar and number of iterations. AP and FAPML have been run on seven data sets of table 1. Figure 1 to Figure 7 shows the variation in Net similarity with Number of iteration.

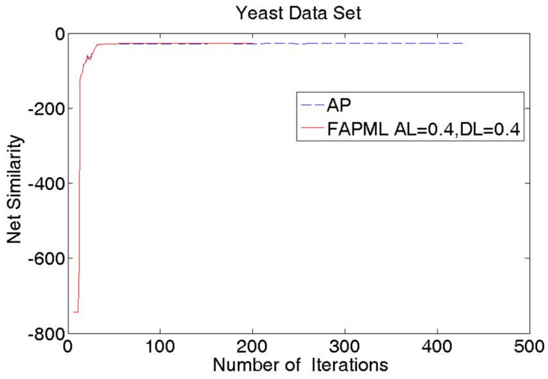


Figure 1

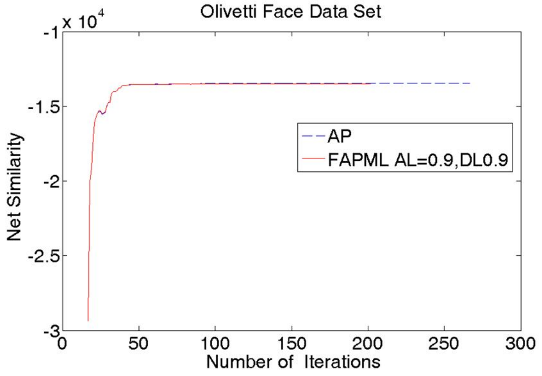


Figure 2

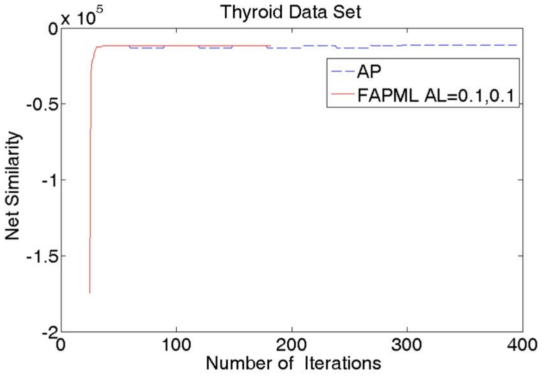


Figure 3

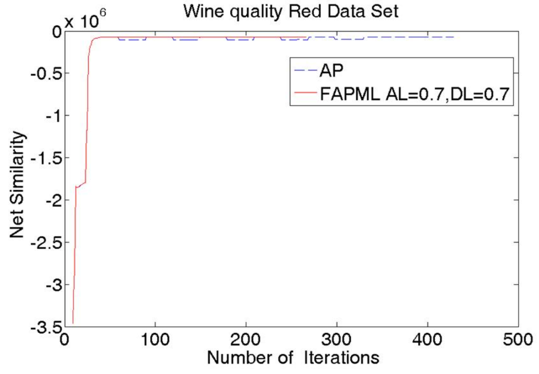


Figure 6

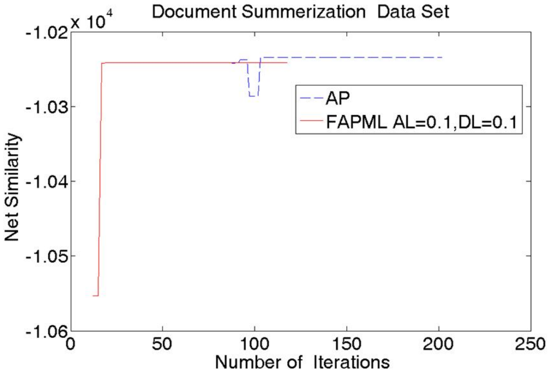


Figure 4

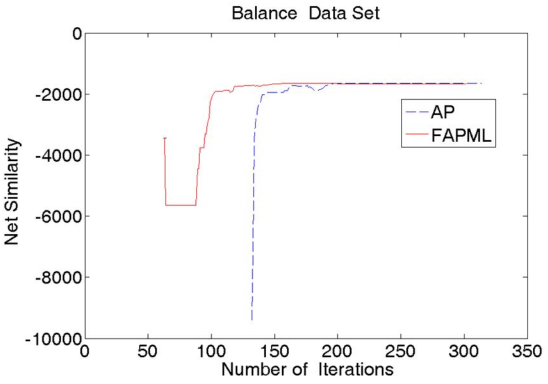


Figure 7

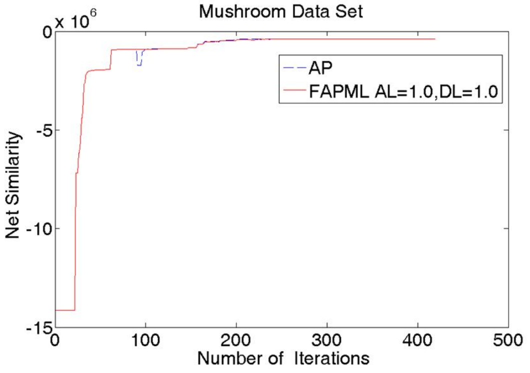


Figure 5

Figure 8 to Figure 14 shows the comparison between AP and FAPML for number of iterations and learning constants.

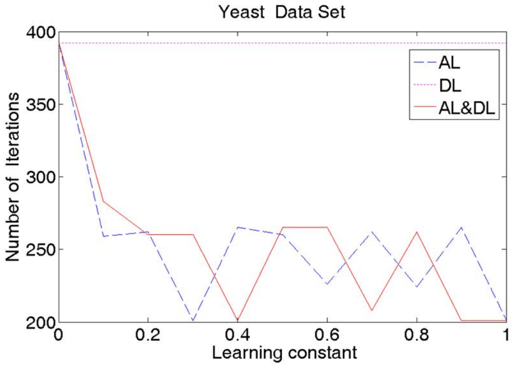


Figure 8

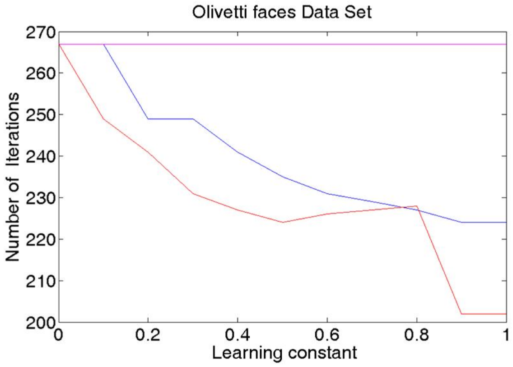


Figure 9

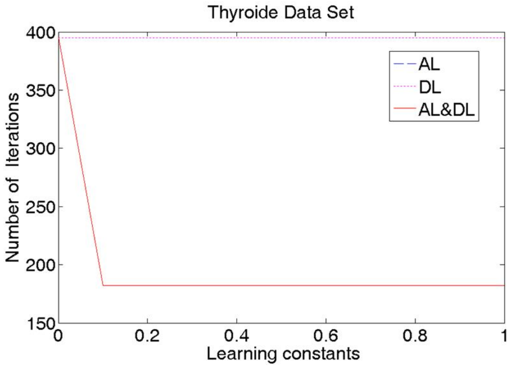


Figure10

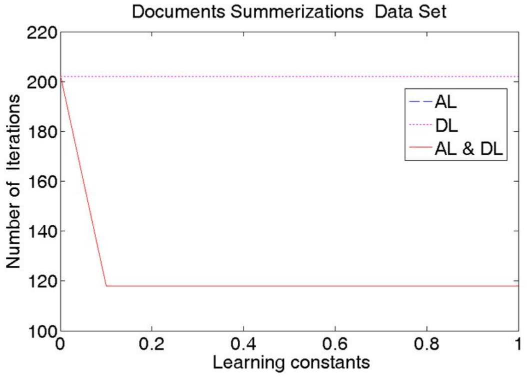


Figure 11

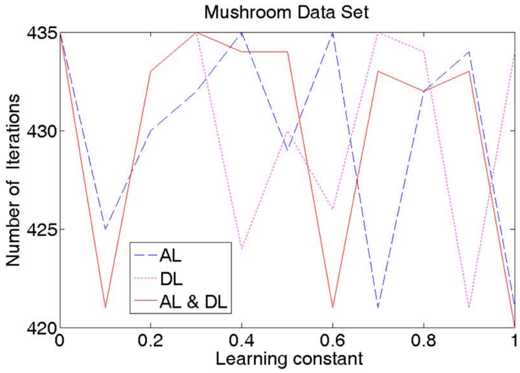


Figure 12

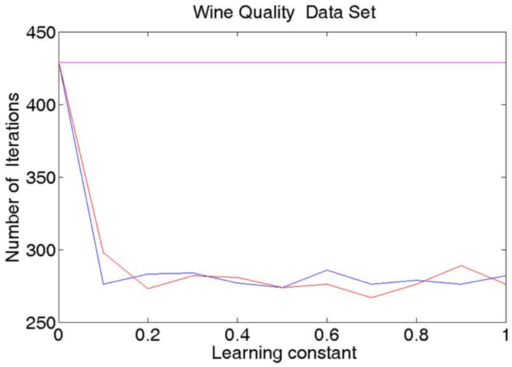


Figure 13

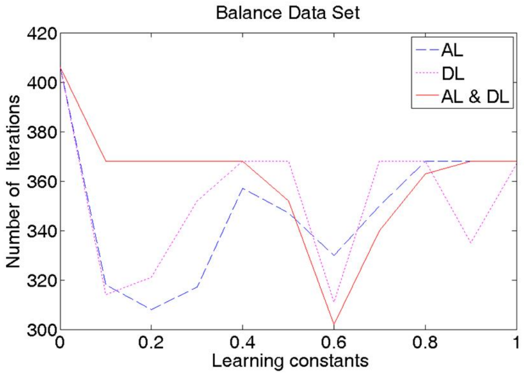


Figure 14

Following tables show the comparison between AP and FAPML.

Table 2

Name of Data Sets	Affinity Propagation			FAPML					Percentage improvement in Results
	Similarities of data points to exemplars	Number of Clusters Identified	Number of Iterations	AL Const.	DL Const.	Similarities of data points to exemplars	Number of Clusters Identified	Number of Iterations	
Yeast	-18.2992	92	392	0.4	0.4	-17.3783	104	201	48.72%
Olivetti faces	-9734.72	62	267	0.9	0.9	-9429.42	68	202	32.17%
Thyroid	-5903	14	395	0.1	0.1	-5607.66	15	182	53.92%
Document summarization	-9607.91	4	202	0.1	0.1	-9582.03	4	118	41.58%
Mushroom	-213315	126	435	1.0	1.0	-213310	126	420	3.4%
Wine Quality Red	-36516.5	36	429	0.7	0.7	-36516.5	37	267	37.76%
Balance	-1643	25	368	0.6	0.6	-1053	31	302	17.93%

Computationally the proposed FAPML algorithm outperformed AP. Net similarity achieved by proposed algorithm is also better than AP. Complexity of FAPML is $O(N^2 T)$, where N is number of data points and T is number of iterations (shown in table2 and figures 1-12). The required number of iterations T is also less than AP, which makes FAPML a Fast affinity propagation algorithm based on machine learning. Clustering quality of FAPML is also better which can be seen from Net similarity/Sum of similarities data points to exemplar. As shown in Table 2 and Figure 1-7, the Net similarity/Sum of similarities of FAPML is higher than AP. Thus the overall performance of FAPML evaluated for net similarity and time is better.

4. Concluding remarks and future directions

Recently introduced Affinity Propagation clustering is slow. In this paper we have proposed a Fast Affinity Propagation algorithm using Machine Learning, based on learning by experience principal of ML. FAPML outperforms AP in terms of speed and clustering accuracy. Extensive experiments on many standard datasets show that the proposed FAPML produces better clustering accuracy in less time.

There are a number of interesting potential avenues for future research. FAPML can be made adaptive, Hierarchical, Partitional, Incremental etc. FAPML can also be applied in Text clustering and clustering based on Heterogeneous Transfer Learning.

References

[1] RuiXu Donald C. Winch, "Clustering" , IEEE Press 2009 ,pp 1-282

[2] Jain, A. and DubesR. "Algorithms for Clustering Data ", Englewood Cliffs, NJ Prentice Hall, 1988.

[3] Jain A.K., Murthy M.N. and Flynn P.J., "Data Clustering: A Review ", ACM Computing Surveys, Vol.31. No 3, September 1999, pp 264-322.

[4] RuiXu, and Donald Wunsch," Survey of Clustering. Algorithms ", IEEE Transactions on Neural Network, Vol 16, No. 3, 2005 pp 645.

[5] EthemAlpaydin , "Introduction to Machine Learning ",Prentice Hall of India Private Limited New Dehli,2006,pp133-150.

[6] Frey, B.J. and Dueck D." Clustering by Passing Messages Between Data Points ", Science 2007, pp 972–976.

[7] Kaijun Wang,„ Junying Zhang, Dan Li, Xinna Zhangand Tao Guo, Adaptive Affinity Propagation Clustering",ActaAutomaticaSinica, 2007 ,1242-1246.

[8] Inmar E. Givoni and Brendan J. Frey,"A Binary Variable Model for AffinityPropagation",Journal Neural Computation,Volume 21 Issue 6, June 2009,pp1589-1600.

[9] Yangqing Jiay, Jingdong Wangz, Changshui Zhangy, Xian-Sheng Hua, "Finding Image Exemplars Using Fast Sparse Affinity Propagation", Proceedings of the 16th ACM International conference on Multimedia,2006 , pp113 – 118.

[10] Yasuhiro Fujiwara, Go Irie and Tomoe Kitahara, "Fast Algorithm for Affinity Propagation",

[11] Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, 2011, pp 2238-2243.

[12] Shang Fanhua,.Jiao L.C, Shi Jiarong, Wang Fei, Maoguo Gong,"Fast affinity propagation clustering: A multi-level approach",Pattern Recognition (Elseevier), 2012 ,pp 474–486.

Shailendra Kumar Shrivastava,B.E.(C.T.),M.E.(CSE) Associate Professor I.T., Samrat Ashok Technological Institute Vidisha. He has more than 23 Years Teaching Experiences. He has published more than 50 research papers in National/International conferences and Journals .His area of interest is machine learning and data mining.

Dr J.L.Rana B.E.M.E.(CSE),PhD(CSE) HE has so many publication in Journal and conferences.

Dr. R.C.Jain PhD .He is the director Samrat Ashok Technological Institute Vidisha(MP).He has published more than 150 research papers in Journals and Conferences.

An Integrated and Improved Approach to Terms Weighting in Text Classification

Jyoti Gautam[1] and Ela Kumar[2]

**[1] School of Information and Communications Technology
Gautam Buddha University
Greater Noida, Uttar Pradesh(India)-201308**

**[2] School of Information and Communications Technology
Gautam Buddha University
Greater Noida, Uttar Pradesh(India)-201308**

Abstract

Traditional text classification methods utilize term frequency (tf) and inverse document frequency (idf) as the main method for information retrieval. Term weighting has been applied to achieve high performance in text classification. Although TFIDF is a popular method, it is not using class information. This paper provides an improved approach for supervised weighting in the TFIDF model. The tfidf-weighting model uses class information to compute weighting of the terms. The model also assumes that low frequency terms are important, high frequency terms are unimportant, so it designs higher weights to the rare terms frequently. So, it uses rare term information along with class information for weighting. So, the paper proposes an improved approach which combines the benefits of the traditional kNN classifiers and Naïve Bayes supervised learning method.

Keywords: *Text classification, tf-idf, term weighting, kNN, feature selection*

1. Introduction

Text classification is the key technique in the data mining (DM) and information retrieval (IR) field and it has got a lot of interest in the recent decades. A lot of research has been done to improve the quality of text representation and develop high quality classifiers. Text classification (TC) is a task to categorize automatically text documents into categories from a predefined set. Most of the machines learning methods treat text documents as bag of words [1].

Vector space model [9, 10] is a classic method in which each document is represented as a vector of its words. Words are regarded as feature vectors. When applied to text categorization, the basic idea is to construct a prototype vector per category using a training set of documents. Given a category, the vectors of documents belonging to this category are given a positive weight, and the vectors of remaining documents are given a negative weight. By summation of the different weighted vectors, the prototype vector of this category is obtained. The method is easy to implement and efficient in computation.

There are two term weighting approaches, i.e. unsupervised and supervised term weighting methods [6] depending on the use of the known information on the membership of training documents. Unsupervised term weighting approaches, such as binary, tf, tfidf, in which binary tells whether a particular term appears in a document, tf indicates how frequently a term appears in a document, and tfidf calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. Words with high TFIDF numbers imply a strong relationship with the document they appear in, suggesting that if that word were to appear in a query, the document could be of interest to the user[5]. The tf part can be regarded as a weight from intra-documents, and idf part is a weight from inter-documents. Unsupervised-based term weighting approach does not use the category information in the training set. Researchers have tried to replace the idf part with feature selection metrics, i.e. information gain, gain ratio, odds ratio and so on. The basis for these is information theory and supervised term weighting approaches. Unlike the case of unsupervised-based term weighting approach, supervised term weighting uses category information. It uses the inter and intra class information. In literature[12], interclass standard deviation(icsd), class standard deviation(csd) and standard deviation, were introduced to tf-idf model, the performance of classification is enhanced.

In yet another method of supervised term weighting,[13] the approach simply thinks low frequency terms are important, high frequency terms are unimportant , so it designs higher weights to the rare terms frequently. The approach presents an effective term weighting to avoid the deficiency of the traditional approach, and make use of KNN classifiers to classify over widely-used benchmark data set Reuters-21578.

In this paper, we propose yet another novel supervised term weighting approach. The approach combines a weighting factor based on the frequency of the number of documents belonging to a category c_i where the term t_k occurs at least once. It also joins the inner and intra class information on a trained document set.

The rest of the paper is organized as follows: Section 2 discusses the traditional term weighting methods .Section 3 presents the shortage of the traditional term weighting scheme. Section 4 presents our improved approach. And finally Section 5 concludes this paper.

2. Term Weighting Approaches: Review and Analysis

The literature [7] provides a text classification method which is an improvement of the previous ones. Most of the traditional text classification methods utilize term frequency (tf) and inverse document frequency (idf) for representing importance of terms and computing weighting of ones in classifying a text document. Term weighting has a significant role to achieve high performance in text classification. The old tf-idf is a popular method, but it does not involve class information of the terms. The paper has provided with an improved tf-idf-ci model to compute weighting of the terms. The intra and inner class information are joined. The role of important and representative terms is raised and the effect of the unimportant feature term to classification is decreased. The F1 based on tf-idf-ci algorithm is higher than based on tf-idf model.

Literature [8] describes a new method which uses term co-occurrence as a measure of dependency between word features. A random-walk model is applied on a graph encoding words and co-occurrence dependencies, resulting in scores that represent a quantification of how a particular word feature contributes to a given text. The new scheme can be used as a text classification method.

The literature [14] provides a method for text categorization when one or more predefined categories are given. The paper reports a study conducted on 20 newsgroup dataset, using TFIDF in the context of document categorization. Feature selection is added to this result to improvise the categorization. The results achieved by this approach are very promising when compared to conventional methods with features chosen on the basis of bag-of-words text.

Literature [13] provides an improved method of term weighting for text classification. Traditional algorithm of term weighting only considers about tf, idf and so on, and this approach simply thinks low frequency terms are important, high frequency terms are unimportant, so it designs higher weights to the rare terms frequently. In this paper, an effective term weighting approach is provided to avoid the deficiency of the traditional approach, and make use of KNN classifiers to classify over widely-used benchmark data set Reuters-21578. The experimental results have proved that the new approach can improve the accuracy of classification.

One of the difficulties in text classification is the high dimensionality of the feature space. How to reduce the dimensionality of the feature space and improve the effectiveness and accuracy of classification has become the main problem to be solved in automatic text classification. So feature selection is often considered a critical step in text classification. Feature selection methods keep a certain number of words with the highest score according to a measure of word relevance. The quality of features will influence accuracy of text classification. According to the literature [2], traditional scoring measures for feature selection have come from the domains of information theory and information retrieval.

This paper [4] by Thorsten Joachims mainly analyses the traditional Rocchio algorithm with TFIDF classifier. The Rocchio classifier, its probabilistic variant with TFIDF classifier and a standard Naïve Bayes classifier are compared. The results provided the information that the probabilistic algorithms are preferable to the heuristic Rocchio classifier. Bag-of –words data is used for analysis. The algorithm proposed in the paper uses feature selection for better processing. Paper proposes a probabilistic classifier based on TFIDF. It makes use of probabilistic indexing paradigm which offers an elegant way to distinguish between a document and the representation of a document. This proposed probabilistic TFIDF classifier offers a theoretical justification for the vector space model and the TFIDF word weighting heuristic for text categorization.

The paper [11] proposes an improved approach named tf.idf.IG to remedy this defect by Information Gain from Information Theory. The paper overcomes the limitations of old tf.idf. The idf can’t well reflect discriminative and importance of feature, weight adjustment method is put forward in which the IDF function is replaced by evaluation function used in feature selection.

Berger and Lafferty in the year 1999[1] proposed a probabilistic framework that incorporates the user’s mindset at the time the query was entered to enhance their approximations. They suggest that the user has a specific information need G, which is approximated as a sequence of words q in the actual query. By accounting for this noisy transformation of G into q and applying Bayes’ Law to equation (1), they show good results on returning appropriate documents given q.

Let us assume that we have a set of documents D, with the user entering a query $q = w_1,w_2,….,w_n$ for a sequence of words w_i .Then we wish to return a subset D* of D such that for each d ϵ D*, we maximize the following probability:

$$P\ (d \mid q, D) \qquad (1)$$

3. The Shortage of Traditional Term Weighting Scheme

The focus of this paper [7] is term-weighting based text classification. So, how to assign appropriate weight to the given feature is of utmost importance. An effective feature can not only represent the content of category belonging to, but also provides discrimination with other categories. The tf-idf approach proposes three basic assumptions. They are:

- Rare terms are no less important than frequent terms- idf assumption.
- Multiple appearances of a term in a document are no less important than single appearance- tf assumption.
- For the same quantity of term matching, long documents are no less important than short documents- normalization assumption.

The Table1 below shows the relation of term t_k and category $C_{i.}$

Table1: relation of term t_k and category $C_{i.}$

	C_i	~~C_i~~
t_k	A	B
~~t_k~~	C	D

A indicates the number of documents belonging to category C_i where the term t_k occurs at least once; B indicates the number of documents not belonging to category $C_{i.}$ where the term t_k occurs at least once; C denotes the number of documents belonging to category $C_{i.}$ where the term t_k does not occur at least once; D denotes the number of documents not belonging to category C_i where the term t_k does not occur at least once.

The tf-idf term weighting scheme assigns higher weights to the rare terms frequently because of idf assumption. Thus, it will influence classification performance. For some a category $c_{i,}$ the terms which are distributed uniformly in the intra-category should be assigned higher weights, but not rare terms. This weighting is not included in the tf-idf approach. For some a category, the higher value of the proportion of A and C is a good feature. So, accordingly, we have to assign appropriate weighting.

In addition to term frequency and inverse document frequency, [8] the weighting is affected by other factors also. The term weighting joins class information also.

- The weighting should be large when the numbers of the classes distributed is more, but the document numbers in one class is far greater than the sum in others.
- The weighting should be small when the numbers of the classes distributed is more, and the document numbers in all classes is large.
- The weighting should be large when the numbers of the classes distributed of the term is very small, even the numbers is one, and the term distributed in the documents of the class is average.
- The weighting should be small when the numbers of the classes distributed of the term is very small, even the numbers is one, but the term distributed only little documents of the class, even one or two documents.

The objective of this paper is to provide an improved weighting algorithm which joins class information along with the weighting assigned to the rare terms.

4. Our Improved Approach

The tf-idf term weighting scheme has been used extensively and has become the default choice in text classification. Our improved approach joins class information combined with the weighting assigned to the rare terms.

$$W(t_{k,}d_{j,}c_i) = (1\text{-}\alpha).\text{tfidf}_{k,j} + \alpha.\ \text{tfidf}_{k,j} \times \text{weighting} \qquad (2)$$

α is called a balance factor, which lies between , $0 \leq \alpha \leq 1$.

When α = 0, equation (2) becomes classic TFIDF approach, and when α = 1, equation (2) becomes our newly improved approach. Using balance factor, we can get better classification results.

Where, the weighting is the class information along with the weighting assigned to the rare terms. The weighting is:

$$\text{Weighting} = \text{CI} \times \frac{Ai}{Ci} \qquad (3)$$

Where, Ai indicates the number of documents belonging to category c_i where the term t_k occurs at least once and Ci denotes the number of documents belonging to category c_i where the term t_k does not occur at least once.

$$\text{Where, the term CI} = \text{Cit}\times\text{Cii} \qquad (4)$$

The class information consists of two parts. One is intra class information, and the other is inner class information. That is:

Where, cit is intra class information, cii is inner class information.

Intra Class Information

$$\text{Cit} = \frac{P(ti|Cj)}{\sum_{k=1,\ k\ddagger j}^{m} P(ti|Ck)} \quad \text{if} \quad \sum_{k=1,\ k\ddagger j}^{m} P(ti|Ck) \ddagger 0$$
(5)

$$\text{Cit} = \frac{P(ti|Cj)}{\beta} \qquad \text{if} \sum_{k=1,\ k\ddagger j}^{m} P(ti|Ck) = 0 \qquad (6)$$

Where, $P(ti|Cj)$ is the probability of documents containing term t_i in the class Cj of the training set. The value of the parameter β is determined through actual situation. Generally, β = 0.001. The number of classes taken is m.

It is quite evident that the Cit is a monotone increasing with the number of documents in the class Cj containing the term t_i increasing. The Cit is increasing with the sum of documents beyond the class Cj containing the term t_i decreasing. Only when the documents of two classes containing the term t_i , and the document numbers of containing the term t_i are almost same the value of Cit approximate to 1. The largest value is achieved when the sum of documents beyond the class Cj containing the term t_i is zero.

Inner Class Information

Inner class divergence can be represented by the term Cii. It is very important to classify when the term t_i appears evenly in the documents of one class.

$$\text{Cii} = \frac{\text{tfavg}\ (\text{ti}\ ,\text{Cj})}{\sum_{k=1}^{N(Cj)}[\,|\ \text{tfik} - \text{tfavg}\ (\text{ti}\ ,\text{Cj})\,|\,]}$$

$$\text{If} \quad \sum_{k=1}^{N(Cj)}[\,|\ \text{tfik} - \text{tfavg}\ (\text{ti}\ ,\text{Cj})\,|\,] \quad \ddagger \quad 0 \qquad (7)$$

$$\text{Cii} = \frac{\text{tfavg}\ (\text{ti}\ ,\text{Cj})}{\gamma}$$

$$\text{If} \sum_{k=1}^{N(Cj)}[\,|\ \text{tfik} - \text{tfavg}\ (\text{ti}\ ,\text{Cj})\,|\,] \ = 0 \qquad (8)$$

Where, γ is a parameter, the value is determined which is based on the actual situation. Generally, γ = 1. The tf_{ik} is the frequency of the term t_i in document k. The tf_{avg} (t_i ,Cj) is the average term frequency of the term t_i in the documents of the class C_j :

$$tf_{avg}(t_i,Cj) = \frac{\sum_{k=1}^{N(Cj)} \text{tfik}}{\text{N(Cj)}} \qquad (9)$$

The largest value of Cii is achieved when the term t_i appears evenly in the documents of the class C_j. If the difference of the term t_i appeared in the documents of the class C_j is larger, then the denominator of the function (6) is larger, then the value obtained of Cii is less. So, it is representative and important for classification purposes when the term t_i appears evenly in the documents of one class.

Normalization

It is normalized for decreasing the high frequency term inhibited to low frequency term. The tf-idf-weighting function is normalized as follows:

$$W_{\text{tf-idf-weighting}}(\text{tij}) \quad = \quad \frac{tf \times idf \times weighting}{\sqrt{\sum_{i=1}^{n}(tf \times idf \times weighting)^2}}$$

(10)

There are different statistics classification and machine learning technologies which are used in text classification. This paper combines the benefits of kNN algorithm and Naiive Bayes algorithm. The kNN algorithm is a method for documents classification based on closet training examples in the feature space. It is amongst the simplest of the machine learning algorithms. And it is an effective method. The core of kNN is, first of all, giving a trained document set, then, for a new preclassified document(that is, a test document), finding most relevant articles of the k documents from the training documents, finally, in accordance with k documents' category information, classifying the test document. Whereas Naiive Bayes is the most effective heuristic learning method.

5. Conclusion and Future Work

Term weighting plays an important role to get high performance in text classification. The traditional tf-idf algorithm is a popular method for document representation and feature selection. But, it is not joining class information. Here, we proposed a supervised term weighting scheme, which makes use of a kind of information ratio to judge a term's contribution for category along with class information. The improved approach of using information ratio has become a new way to compute term's weights to avoid assigning higher weights to rare terms. It has been proved through experiments that the improved approach of using information ratio is an effective solution to improve the performance of text classification. The approach using class information uses intra and inner class information. When using class information, the numbers of the feature term is decreased when the threshold is given. The experimental results showed that the performance is enhanced. The role of important and representative terms is raised and the effect of the unimportant feature term to classification is decreased. So, an improved approach for term weighting which joins information ratio together with class information has been proposed. The approach can be implemented and performance analysis can be done.

References

[1] Berger, A & Lafferty, J., "Information Retrieval as Statistical Translation", Proc. of the 22nd ACM Conference on Research and Development in Information Retrieval (SIGIR), 1999, 222-229.

[2] Elena Montanes, Irene Diaz, Jose Ranilla, Elias F.Combarro, Javier Fernandez, "Scoring and Selecting Terms for Text Categorization" , Journal of IEEE Intelligent Systems, Vol. 20, Issue 3, 2005, pp. 40-47.

[3] G.Salton and M.J.McGill, "An Introduction to Modern Information Retrieval" , McGraw Hill, 1983.

[4] Joachims, Thorsten, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization" , Computer Science Technical Report CMU-CC-96-118, Carnegie Mellon University, 1996.

[5] Juan Ramos, "Using TF-IDF to determine word relevance in document queries" ,Department of Computer Science, Rutgers University.

[6] M.Lan, C.L.Tan, and H.B.Low, "Proposing a new term weighting scheme for text categorization" , Proc. of the Twenty-First National Conference on Artificial Intelligence, 2006, pp. 763-768.

[7] Ma Zhanguo, Feng Jing, Chen Liang, Hu Xiangyi, Shi Yanqin, "An improved approach to terms weighting in text classification" , in proc. of the International Conference on Computer and Management, 2011, pages1-4.

[8] S. Hassan, C. Banea, "Random Walk term weighting for improved text classification" , Proc. of Text Graphs: 2nd Workshop on Graph Based Methods for Natural Language Processing, ACL, 2006, pp. 53-60.

[9] Salton, G., & Buckley, C., "Term-weighting approaches in automatic text retrieval", Journal of Information processing and Management, Vol. 24, Issue No. 5, 1988, pages 513-523.

[10] Salton, G., Wong, A., & Yang, C.S., "A vector space model for automatic indexing" , Communications of the ACM, Vol.18, Issue No. 11, 1975, pages 613-620.

[11] S. Lu, X. Li, S. Bai, S. Wang, " An improved approach to weighting terms in text" , Journal of Chinese Information Processing, 2000, 14(6), pp. 8-13.

[12] V. Lertnattee, T. Theeramunkong, "Effect of term distributions on centroid-based text categorization" , International Journal on Information Sciences – Informatics and Computer science, vol. 158,Issue 1,2004, pp. 89-115.

[13] Xin Hu, Hua Jiang, Ping Li and Shuyan Wang proposed, "An improved method of term weighting for text classification", appears in *International Conference on Intelligent Computing and Intelligent Systems*,Vol. 1, IEEE, 2009, pages 294-298.

[14] Yi-hong Lu, Yan Huang, "Document Categorization with Entropy based TFIDF classifier", in proc. of the WRI Global Congress on Intelligent Systems, vol.4, 2009, pages 269-273.

Jyoti Gautam completed B. Tech. (Instrumentation and Control Engineering) in the year 1997 from Delhi University. Afterwards, she did her M.Tech. (Computer Technology and Applications) in the year 1999 from Delhi University. Currently, pursuing PhD in the area of Semantic Web from Gautam Buddha University, Greater NOIDA. Presently employed as Associate Professor in the Department of Computer Science and Engineering in JSS Academy of Technical Education, NOIDA. One of the published papers titled 'An Improved Framework for Tag-Based Academic Information Sharing and Recommendation System' occurs in the proceedings of the WCE 2012 VOL II.

Dr. Ela Kumar completed B.E.(Electronics and Communication) in the year 1988 from IIT, Roorkee. Afterwards, she did her M.Tech. (Computer Science and Technology) in the year 1990 from IIT, Roorkee. Later, she did Ph.D. (Computer Science and Technology) from Delhi University in the year 2003.Worked as Asstt. Professor at YMCAIE,Faridabad. Presently, working as Dean and Associate Professor in the school of ICT, Gautam Buddha University, Greater Noida. She won Rashtriya Gaurav Award by IIFS society in NOV 2010 for meritorious activities in the field of IT. She has participated as member Advisory Board in various conferences. She has participated in various seminars as Speakers. She has participated as expert for Course Designing. Her publications include 10 research papers in International Referred Journal, 6 in National Referred Journals, 10 in International Conferences and 15 in National Conferences. She has penned 4 books. Her expertise include many other activities.

An Improved Particle Swarm Optimization Algorithm and Its Application

Xuesong Yan[1], Qinghua Wu[2,3], Hanmin Liu[4] and Wenzhi Huang[2,3]

[1] **School of Computer Science, China University of Geosciences**
Wuhan, Hubei 430074, China

[2] **Hubei Provincial Key Laboratory of Intelligent Robot, Wuhan Institute of Technology**
Wuhan, Hubei 430073, China

[3] **School of Computer Science and Engineering, Wuhan Institute of Technology**
Wuhan, Hubei 430073, China

[4] **Wuhan Institute of Ship Building Technology**
Wuhan, Hubei 430050, China

Abstract

In this paper, aim at the disadvantages of standard Particle Swarm Optimization algorithm like being trapped easily into a local optimum, we improves the standard PSO and proposes a new algorithm to solve the overcomes of the standard PSO. The new algorithm keeps not only the fast convergence speed characteristic of PSO, but effectively improves the capability of global searching as well. Compared with standard PSO on the Benchmarks function, the results show that the new algorithm is efficient, we also use the new algorithm to solve the TSP and the experiment results show the new algorithm is effective for the this problem.

Keywords: *Particle Swarm Optimization, Traveling Salesman Problem, Particle, Convergence.*

1. Introduction

Particle Swarm Optimization (PSO) algorithm was an intelligent technology first presented in 1995 by Eberhart and Kennedy, and it was developed under the inspiration of behavior laws of bird flocks, fish schools and human communities [1]. If we compare PSO with Genetic Algorithms (GAs), we may find that they are all maneuvered on the basis of population operated. But PSO doesn't rely on genetic operators like selection operators, crossover operators and mutation operators to operate individual, it optimizes the population through information exchange among individuals. PSO achieves its optimum solution by starting from a group of random solution and then searching repeatedly. Once PSO was presented, it invited widespread concerns among scholars in the optimization fields and shortly afterwards it had become a studying focus within only several years. A number of scientific achievements had emerged in these fields [2-4]. PSO was proved to be a sort of high efficient optimization algorithm by numerous research and experiments [5]. PSO is a meta-heuristic as it makes few or no assumptions about the problem being optimized and can search very large spaces of candidate solutions. However, meta-heuristics such as PSO do not guarantee an optimal solution is ever found. More specifically, PSO does not use the gradient of the problem being optimized, which means PSO does not require that the optimization problem be differentiable as is required by classic optimization methods such as gradient descent and quasi-Newton methods. PSO can therefore also be used on optimization problems that are partially irregular, noisy, change over time, etc.

The traveling salesman problem (TSP) [6] is one of the most widely studied NP-hard combinatorial optimization problems. Its statement is deceptively simple, and yet it remains one of the most challenging problems in Operational Research. The simple description of TSP is: Give a shortest path that covers all cities along. Let $G=(V;E)$ be a graph where V is a set of vertices and E is a set of edges. Let $C=(c_{ij})$ be a distance (or cost) matrix associated with E. The TSP requires determination of a minimum distance circuit (Hamiltonian circuit or cycle) passing through each vertex once and only once. C is said to satisfy the triangle inequality if and only if $c_{ij}+c_{jk}\geq c_{ik}$ for $i,j,k\in V$.

Due to its simple description and wide application in real practice such as Path Problem, Routing Problem and Distribution Problem, it has attracted researchers of various domains to work for its better solutions. Those traditional algorithms such as Cupidity Algorithm, Dynamic Programming Algorithm, are all facing the same obstacle, which is when the problem scale N reaches to a certain degree, the so-called "Combination Explosion" will occur. For example, if $N = 50$, then it will take 5×10^{48} years under a super mainframe executing 100 million instructions per second to reach its approximate best solution.

A lot of algorithms have been proposed to solve TSP [7-12]. Some of them (based on dynamic programming or branch and bound methods) provide the global optimum solution. Other algorithms are heuristic ones, which are much faster, but they do not guarantee the optimal solutions. There are well known algorithms based on 2-opt or 3-opt change operators, Lin-Kerninghan algorithm (variable change) as well algorithms based on greedy principles (nearest neighbor, spanning tree, etc). The TSP was also approached by various modern heuristic methods, like simulated annealing, evolutionary algorithms and tabu search, even neural networks.

This paper improves the disadvantages of standard PSO being easily trapped into a local optimum and proposed a new algorithm called improved PSO (IPSO) which proves to be more simply conducted and with more efficient global searching capability, then use the new algorithm for traveling salesman problem.

2. Basic Particle Swarm Optimization Algorithm

A basic variant of the PSO algorithm works by having a population (called a swarm) of candidate solutions (called particles). These particles are moved around in the search-space according to a few simple formulae. The movements of the particles are guided by their own best known position in the search-space as well as the entire swarm's best known position. When improved positions are being discovered these will then come to guide the movements of the swarm. The process is repeated and by doing so it is hoped, but not guaranteed, that a satisfactory solution will eventually be discovered. Formally, let $f : R^n \to R$ be the cost function which must be minimized. The function takes a candidate solution as argument in the form of a vector of real numbers and produces a real number as output which indicates the objective function value of the given candidate solution. The gradient of f is not known. The goal is to find a solution a for which $f(a) \le f(b)$ for all b in the search-space, which would mean a is the global minimum. Maximization can be performed by considering the function $h = -f$ instead.

PSO was presented under the inspiration of bird flock immigration during the course of finding food and then be used in the optimization problems. In PSO, each optimization problem solution is taken as a bird in the searching space and it is called "particle". Every particle has a fitness value which is determined by target functions and it has also a velocity which determines its destination and distance. All particles search in the solution space for their best positions and the positions of the best particles in the swarm. PSO is initially a group of random particles (random solutions), and then the optimum solutions are found by repeated searching. In every iteration, a particle will follow two bests to renew itself: the best position found for a particle called p_{best}; the best position found for the whole swarm called g_{best}. All particles will determine following steps through the best experiences of individuals themselves and their companions.

For particle id, its velocity and its position renewal formula are as follows:

$$V'_{id} = \omega V_{id} + \eta_1 rand()(P_{idb} - X_{id}) + \eta_2 rand()(P_{gdb} - X_{id}) . \quad (1)$$

$$X'_{id} = X_{id} + V'_{id} \quad (2)$$

In here: ω is called inertia weight, it is a proportion factor that is concerned with former velocity, $0 < \omega < 1$, η_1 and η_2 are constants and are called accelerating factors, normally $\eta_1 = \eta_2 = 2$, $rand()$ are random numbers, X_{id} represents the position of particle id; V_{id} represents the velocity of particle id; P_{id}, P_{gd} represent separately the best position particle id has found and the position of the best particles in the whole swarm.

In formula(1), the first part represents the former velocity of the particle, it enables the particle to possess expanding tendency in the searching space and thus makes the algorithm be more capable in global searching; the second part is called cognition part, it represents the process of absorbing individual experience knowledge on the part of the particle; the third part is called social part, it represents the process of learning from the experiences of other particles on the part of certain particle, and it also shows the information sharing and social cooperation among particles.

The flow of PSO can briefly describe as following: First, to initialize a group of particles, e.g. to give randomly each particle an initial position X_i and an initial velocity V_i, and then to calculate its fitness value f. In every iteration, evaluated a particle's fitness value by analyzing the velocity and positions of renewed particles in formula (1) and (2). When a particle finds a better position than previously, it will mark this coordinate into vector P1, the vector difference between P1 and the present position of the particle will randomly be added to next velocity vector, so that the following renewed particles will search around this point, it's also called in formula (1) cognition component. The weight difference of the present position of the particle swarm and the best position of the swarm P_{gd} will also be added to velocity vector for adjusting the next population velocity. This is also called in formula (1) social component. These two adjustments will enable particles to search around two bests.

The most obvious advantage of PSO is that the convergence speed of the swarm is very high, scholars like Clerc [13] has presented proof on its convergence. In order to verify the convergence speed of the PSO algorithm, we selected four benchmarks function and compared the results with traditional genetic algorithm (GA).

F1: Schaffer function

$$\min f(x_i) = 0.5 - \frac{(\sin^2 \sqrt{x_1^2 + x_2^2} - 0.5)}{[1 + 0.001(x_1^2 + x_2^2)]^2},$$
$$-10 \le x_i \le 10$$

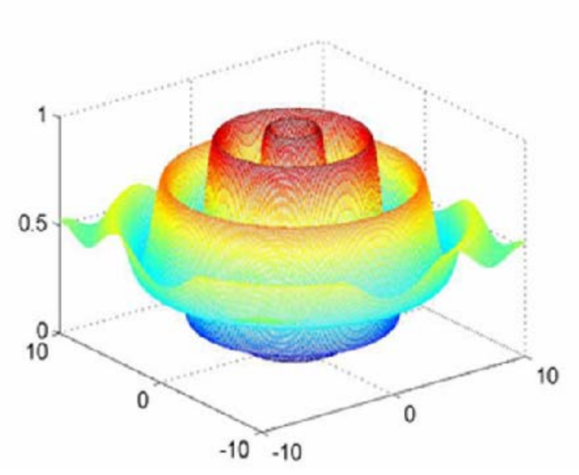


Fig. 1 Schaffer function

In this function the biggest point is in the situation where xi= (0, 0) and the global optimal value is 1.0, the largest in the overall points for the center, and 3.14 for the radius of a circle on the overall situation from numerous major points of the uplift. This function has a strong shock; therefore, it is difficult to find a general method of its global optimal solution.

F2: Shubert function

$$\min f(x, y) = \left\{ \sum_{i=1}^{5} i \cos\left[(i+1)x + i\right] \right\} \times$$
$$\left\{ \sum_{i=1}^{5} i \cos\left[(i+1)y + i\right] \right\}, x, y \in [-10, 10]$$

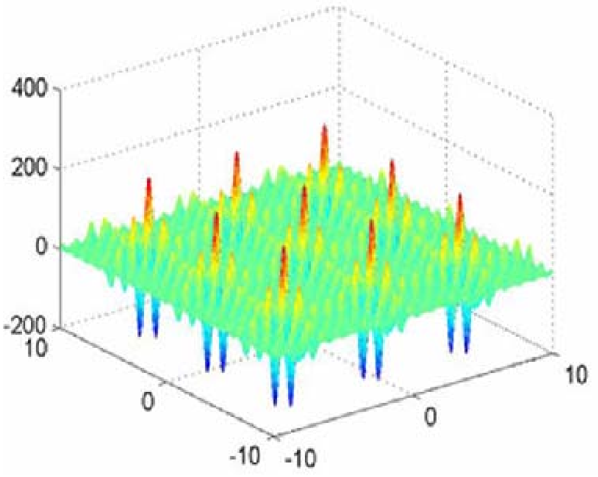


Fig. 2 Shubert function

This function has 760 local minimum and 18 global minimum, the global minimum value is -186.7309.

F3: Hansen function

$$\min f(x, y) = \sum_{i=1}^{5} i \cos((i-1)x + i) \sum_{j=1}^{5} j \cos((j+1)y + j),$$
$$x, y \in [-10, 10]$$

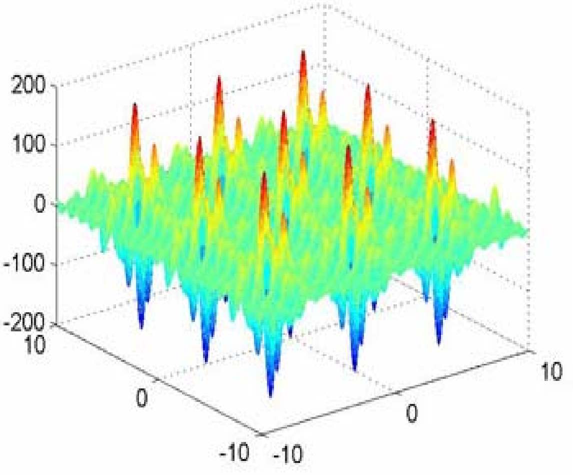


Fig. 3 Hansen function

This function has a global minimum value -176.541793, in the following nine point (-7.589893, -7.708314)、(-7.589893, -1.425128)、(-7.589893, 4.858057)、(-1.306708, -7.708314)、(-1.306708, -1.425128)、(-1.306708, 4.858057)、(4.976478, -7.708314)、(4.976478, -7.708314)、(4.976478, 4.858057) can get this global minimum value, the function has 760 local minimum.

F4: Camel function

$$\min f(x,y)=\left(4-2.1x^2+\frac{x^4}{3}\right)x^2+xy+\left(-4+4y^2\right)y^2,$$

$$x,y\in[-100,100]$$

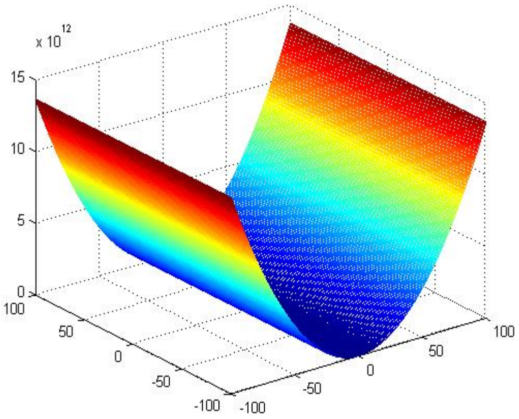


Fig. 4 Camel function

Camel function has 6 local minimum (1.607105, 0.568651)、(-1.607105, -0.568651)、(1.703607, -0.796084)、(-1.703607, 0.796084)、(-0.0898,0.7126) and (0.0898,-0.7126), the (-0.0898,0.7126) and (0.0898,-0.7126) are the two global minimums, the value is -1.031628.

Table 1: Experiment results comparison (100 runs for each case)

Function	*Algorithm*	*Convergence Times*	*Optimal Solution*
F1	GA	72	1.0000000
	PSO	75	1.0000000
F2	GA	75	-186.730909
	PSO	80	-186.730909
F3	GA	85	-176.541793
	PSO	90	-176.541793
F4	GA	23	-1.031628
	PSO	56	-1.031628

In the experiment, each case is repeated for 100 times. Table 1 shows the statistics of our experimental results in terms of accuracy of the best solutions. GA found the known optimal solution to F1 72 times out of 100 runs, found the known optimal solution to F2 75 times out of 100 runs, found the known optimal solution to F3 85 times out of 100 runs, found the known optimal solution to F4 23 times out of 100 runs; PSO algorithm is efficiency for the four cases: found the known optimal solution to F1 75 times out of 100 runs, found the known optimal solution to F2 80 times out of 100 runs, found the known optimal solution to F3 90 times out of 100 runs and found the known optimal solution to F4 56 times out of 100 runs.

3. Improved Particle Swarm Optimization Algorithm

In the standard PSO algorithm, the convergence speed of particles is fast, but the adjustments of cognition component and social component make particles search around P_{gd} and P_{id}. According to velocity and position renewal formula, once the best individual in the swarm is trapped into a local optimum, the information sharing mechanism in PSO will attract other particles to approach this local optimum gradually, and in the end the whole swarm will be converged at this position. But according to velocity and position renewal formula (1), once the whole swarm is trapped into a local optimum, its cognition component and social component will become zero in the end; still, because $0<\omega<1$ and with the number of iteration increase, the velocity of particles will become zero in the end, thus the whole swarm is hard to jump out of the local optimum and has no way to achieve the global optimum. Here a fatal weakness may result from this characteristic. With constant increase of iterations, the velocity of particles will gradually diminish and reach zero in the end. At this time, the whole swarm will be converged at one point in the solution space, if g_{best} particles haven't found g_{best}, the whole swarm will be trapped into a local optimum; and the capacity of swarm jump out of a local optimum is rather weak. In order to get through this disadvantage, in this paper we presents a new algorithm based on PSO. In order to avoid being trapped into a local optimum, the new PSO adopts a new information sharing mechanism. We all know that when a particle is searching in the solution space, it doesn't know the exact position of the optimum solution. But we can not only record the best positions an individual particle and the whole swarm have experienced, we can also record the worst positions an individual particle and the whole swarm have experienced, thus we may make individual particles move in the direction of evading the worst positions an individual particle and the whole flock have experienced, this will surely enlarge the global searching space of particles and enable them to avoid being trapped into a local optimum too early, in the same time, it will improve the possibility of finding g_{best} in the searching space. In the new strategy, the particle velocity and position renewal formula are as follows:

$$V_{id}^{'} = \omega V_{id} + \eta_1 rand()(X_{id} - P_{idw}) + \eta_2 rand()(X_{id} - P_{gdw}) \quad (3)$$

$$X_{id}^{'} = X_{id} + V_{id}^{'} \quad (4)$$

In here: P_{idw}, P_{gdw} represent the worst position particle id has found and the worst positions of the whole swarm has found.

In standard PSO algorithm, the next flying direction of each particle is nearly determined; it can fly to the best individual and the best individuals for the whole swarm. From the above conclusion we may easily to know it will be the danger for being trapped into a local optimum. In order to decrease the possibility of being trapped into the local optimum, the new PSO introduces genetic selection strategy: To set particle number in the swarm as m, father population and son population add up to 2m. To select randomly q pairs from m; as to each individual particle i, if the fitness value of i is smaller than its opponents, i will win out and then add one to its mark, and finally select those particles which have the maximum mark value into the next generation. The experiments conducted show that this strategy greatly reduces the possibility of being trapped into a local optimum when solving certain functions.

The flow of the IPSO is as follows:

Step 1: to initialize randomly the velocity and position of particles;

Step 2: to evaluate the fitness value of each particle;

Step 3: as to each particle, if its fitness value is smaller than the best fitness value P_{idb}, renew the best position P_{idb} of particle id; or else if its fitness value is bigger than the worst fitness value P_{idw}, renew P_{idw};

Step 4: as to each particle, if its fitness value is smaller than the best whole swarm fitness value P_{gdb}, renew the best fitness value P_{gdb} of particle id; or else if bigger than the worst whole swarm fitness value P_{gdw}, renew P_{gdw};

Step 5: as to each particle,

1) To produce new particle t by applying formula (1) (2),

2) To produce new particle t' by applying formula (3) (4),

3) To make a comparison between t and t', then select the better one into the next generation;

Step 6: to produce next generation particles according to the above genetic selection strategy;

Step 7: if all the above steps satisfy suspension needs, suspend it; or turn to Step 3.

In order to verify the improvement of the new algorithm based on PSO, we select the same benchmark function the above have described. We run our algorithm and compare the results with traditional PSO. In the experiment, each case is repeated for 100 times. Table 2 shows the statistics of our experimental results in terms of accuracy of the best solutions. For the four cases PSO algorithm can found the known optimal solution to F1 75 times out of 100 runs, found the known optimal solution to F2 80 times out of 100 runs, found the known optimal solution to F3 90 times out of 100 runs and found the known optimal solution to F4 56 times out of 100 runs. Our new algorithm improved PSO algorithm is efficiency for the four cases: found the known optimal solution to F1 and F2 100 times out of 100 runs, found the known optimal solution to F3 97 times out of 100 runs and found the known optimal solution to F4 65 times out of 100 runs.

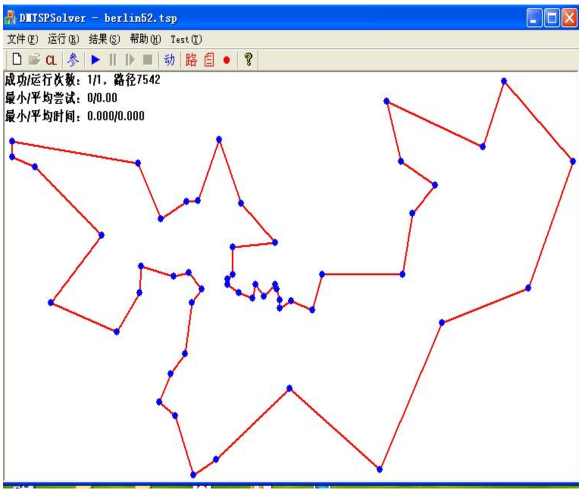
Table 2: Experiment results comparison (100 runs for each case)

Function	***Algorithm***	***Convergence Times***	***Optimal Solution***
F1	PSO	75	1.0000000
	IPSO	100	1.0000000
F2	PSO	80	-186.730909
	IPSO	100	-186.730909
F3	PSO	90	-176.541793
	IPSO	97	-176.541793
F4	PSO	56	-1.031628
	IPSO	65	-1.031628

4. Improved PSO for TSP

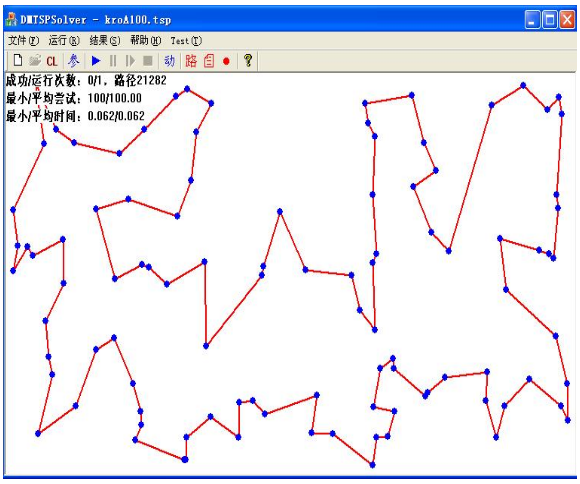
In order to verify the proposed algorithm is useful for the TSP, the experiment test we select 10 TSP test cases: berlin52, kroA100, kroA200, pr299, rd400, ali535, d657, rat783, u1060 and u1432. All experiments are performed on Intel Core(TM)2 Duo CPU 2.26GHz/4G RAM Laptop. In the experiments all test cases were chosen from TSPLIB (http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95), and the optimal solution of each test case is known in the website.

We list the test cases' optimal solutions and compared with traditional genetic algorithm and PSO algorithm [19], the comparison results shown in Table 3. The comparison results demonstrate clearly the efficiency of our algorithm. Note that for the 10 test cases the optimum was found in all ten runs. The number of cities in these test cases varies from 52 to 1432 and Fig. 5 to Fig. 14 is the best solution with IPSO.


DMTSPSolver - berlin52.tsp
文件(F) 运行(R) 结果(S) 帮助(H) Test(T)
CL 参 动 路

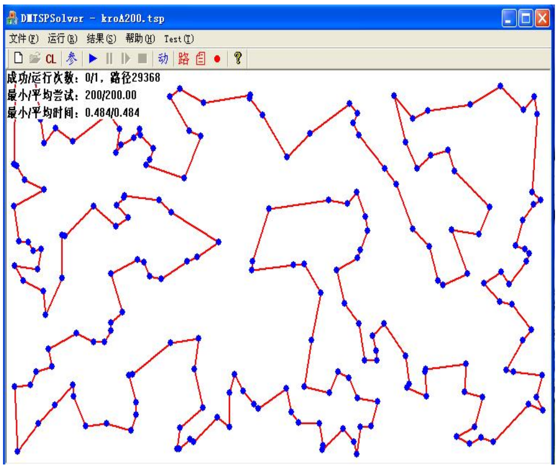
成功/运行次数：1/1，路径7542
最小/平均尝试：0/0.00
最小/平均时间：0.000/0.000

Fig. 5 berlin52


DMTSPSolver - kroA100.tsp
文件(F) 运行(R) 结果(S) 帮助(H) Test(T)
CL 参 动 路

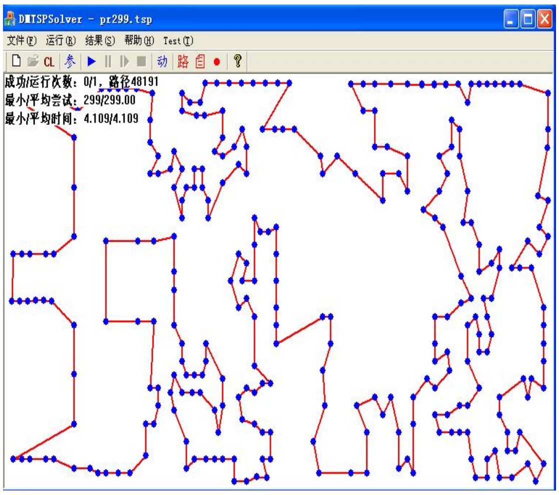
成功/运行次数：0/1，路径21282
最小/平均尝试：100/100.00
最小/平均时间：0.062/0.062

Fig. 6 kroA100


DMTSPSolver - kroA200.tsp
文件(F) 运行(R) 结果(S) 帮助(H) Test(T)
CL 参 动 路

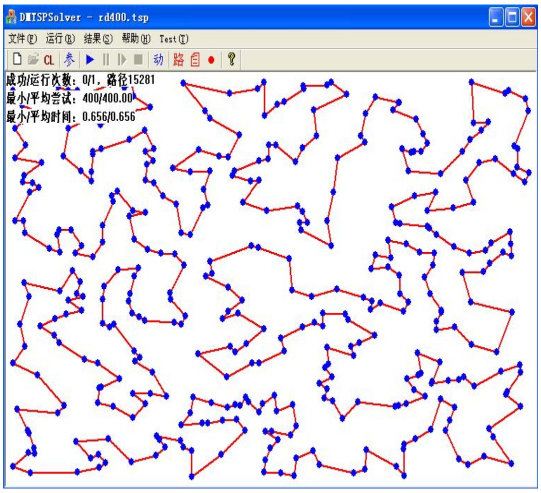
成功/运行次数：0/1，路径29368
最小/平均尝试：200/200.00
最小/平均时间：0.484/0.484

Fig. 7 kroA200



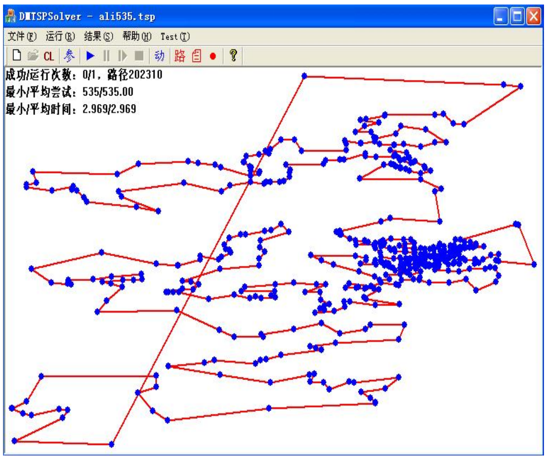
成功/运行次数：0/1，路径48191
最小/平均尝试：299/299.00
最小/平均时间：4.109/4.109

Fig. 8 pr299



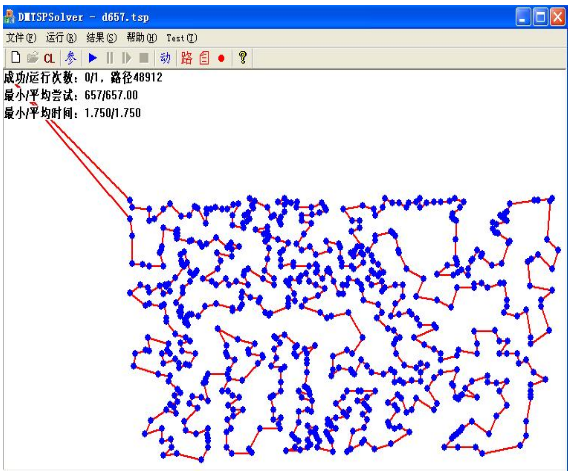
成功/运行次数：0/1，路径15281
最小/平均尝试：400/400.00
最小/平均时间：0.656/0.656

Fig. 9 rd400



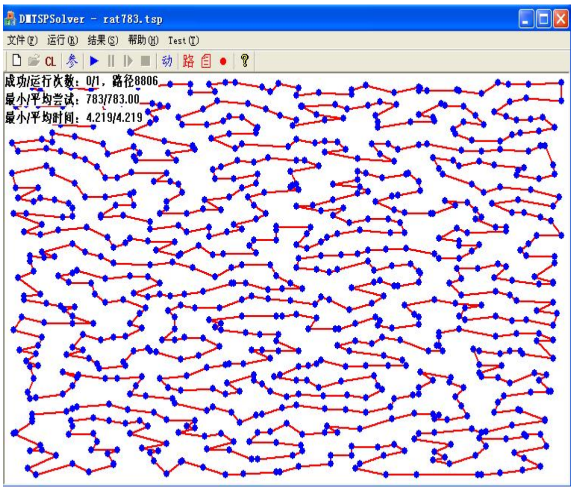
成功/运行次数：0/1，路径202310
最小/平均尝试：535/535.00
最小/平均时间：2.969/2.969

Fig. 10 ali535



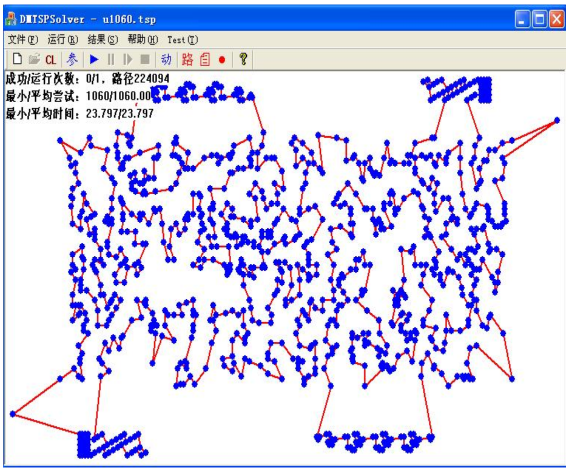
成功/运行次数：0/1，路径48912
最小/平均尝试：657/657.00
最小/平均时间：1.750/1.750

Fig. 11 d657



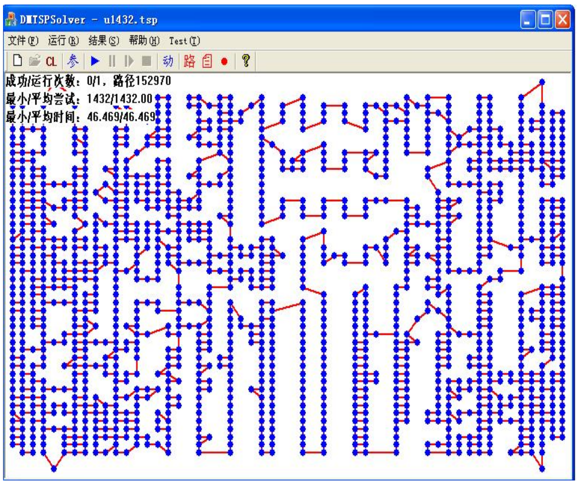
成功/运行次数：0/1，路径8806
最小/平均尝试：783/783.00
最小/平均时间：4.219/4.219

Fig. 12 rat783



成功/运行次数：0/1，路径224094
最小/平均尝试：1060/1060.00
最小/平均时间：23.797/23.797

Fig. 13 u1060



成功/运行次数：0/1，路径152970
最小/平均尝试：1432/1432.00
最小/平均时间：46.469/46.469

Fig. 14 u1432

We also compare our algorithm with other algorithms: Genetic Algorithm (GA), Ant Colony Optimization (ACO) and Ant System-Assisted Genetic Algorithm (ASAGA), the three algorithms are introduced in paper [14]. In the experiment, each case is repeated for 10 times. Table 4 shows the statistics of our experimental results in terms of accuracy of the best solutions. The known optimal solutions are taken from the TSP Library website the above has introduced. ACO failed in reaching the known optimal solution to any case. GA found the known optimal solution to Berlin52 7 times out of ten runs, but could not reach that for the other larger cases. ASAGA found the known optimal solution to berlin52 7 times out of ten runs, found the known optimal solution to kroA100 5 times out of ten runs and found the known optimal solution to kroA200 only one time. Our algorithm is efficiency for the three cases: found the known optimal solution to berlin52 10 times out of ten runs, found the known optimal solution to kroA100 9 times out of ten runs and found the known optimal solution to kroA200 only 8 times.

4. Conclusions

This paper introduce a new algorithm based on the standard PSO algorithm, for the standard PSO algorithm the new algorithm has done two improvements: 1. By introducing a new information sharing mechanism, make particles moved on the contrary direction of the worst individual positions and the worst whole swarm positions, thus enlarge global searching space and reduce the possibility of particles to be trapped into a local optimum; 2. By introducing genetic selection strategy, decreased the possibility of being trapped into a local optimum. Compared with the standard PSO algorithm, the new algorithm enlarges the searching space and the complexity is not high. We use the proposed algorithm for solving the combinatorial problem: TSP, the new algorithm shows great efficiency in solving TSP with the problem scale from 52 to 1432. By analyzing the testing results, we reach the conclusion: in the optimization precision and the optimization speed, the new algorithm is efficiency than the traditional PSO algorithm and the new algorithm is more efficient than traditional algorithms in coping with the TSP.

Acknowledgments

This paper is supported by Natural Science Foundation of China. (No.61272470 and No.61203307), National Civil Aerospace Pre-research Project of China, the Provincial Natural Science Foundation of Hubei (2012FFB04101), the Fundamental Research Founds for National University, China University of Geosciences(Wuhan), Science Research Foundation of Wuhan Institute of Technology (No.12126011) and Youth Foundation of Wuhan Institute of Technology (Q201107).

References

[1] J. Kennedy and R. C.Eberhart, “Particle Swarm Optimization”, IEEE International Conference on Neural Networks, 1995, pp.1942-1948.

[2] Clare M, Kennedy J, “The Particle Swarm - Explosion, Stability, and Convergence in a Multidimensional Complex Space”, IEEE Trans. on Evolution2ary Computation, vol.6(1), 2002, pp.58-73.

[3] C.A.Coello and M.S.Lechuga, Mopso, “A proposal for multiple objective particle swarm optimization”, In IEEE Proceedings World Congress on Computational Intelligence, 2002, pp.1051-1056.

[4] J.Kennedy, “The particle swarm: social adaptation of knowledge”, In Proc. IEEE Conf. on evolutionary computation, 1997, pp.3003-3008.

[5] E. Oscan and C. K.Mohan, “Analysis of A Simple Particle Swarm Optimization System”, Intelligence Engineering Systems Through Artificial Neural Networks, 1998, pp.253-258.

[6] Durbin R, Willshaw D, “An Anlaogue Approach to the Traveling Salesman Problem Using an Elastic Net Approach”, Nature, 326, 6114, 1987, pp.689-691.

[7] Tao Guo and Zbigniew Michalewize, “Inver-Over operator for the TSP”, In Parallel Problem Sovling from Nature(PPSN V), Springer-Verlag press, 1998, pp.803-812.

[8] Zhangcan Huang, Xiaolin Hu and Siduo Chen, “Dynamic Traveling Salesman Problem based on Evolutionary Computation”, In Congress on Evolutionary Computation(CEC’01), IEEE Press, 2001, pp.1283-1288.

[9] Aimin Zhou, Lishan Kang and Zhenyu Yan, “Solving Dynamic TSP with Evolutionary Approach in Real Time”, In Congress on Evolutionary Computation(CEC’03), 2003, pp.951-957.

[10] Hui.Yang, Lishan.Kang and Yuping.Chen, “A Gene-pool Based Genetic Algorithm for TSP”, Wuhan University Journal of Natural Sciences, 8(1B), 2003, pp.217-223.

[11] Xuesong Yan, Lishan.Kang, “An Approach to Dynamic Traveling Salesman Problem”, Proceedings of the Third International Conference on Machine Learning and Cybernetics,2004, pp. 2418-2420.

[12] Xuesong Yan, Aimin Zhou, Lishan Kang, “TSP Problem Based on Dynamic Environment”, Proceedings of the 5th World Congress on Intelligent Control and Automation, 2004, pp.2271-2274.

[13] M.Clerc and J.Kennedy, “The Particle Swarm: Explosion, Stability and Convergence in a Multi-Dimensional Complex Space”, IEEE Trans. on Evolutionary Computation, Vol.6, 2002, pp.58-73.

[14] Gaifang Dong, Xueliang Fu, Heru Xue, "An Ant System-Assisted Genetic Algorithm For Solving The Traveling Salesman Problem", International Journal of Advancements in Computing Technology, Vol. 4, No. 5, 2012, pp.165 -171.

[15] Xuesong Yan, Hui Li, “A Fast Evolutionary Algorithm for Combinatorial Optimization Problems”, Proceedings of the Fourth International Conference on Machine Learning and Cybernetics,2005, pp.3288-3292.

[16] Xuesong Yan, Qing Hua Wu, “A New Optimizaiton Algorithm for Function Optimization”, Proceedings of the 3rd International Symposium on Intelligence Computation & Applications, 2009, pp. 144-150.

[17] Xue Song Yan, Qing Hua Wu, Cheng Yu Hu, Qing Zhong Liang, “Circuit Design Based on Particle Swarm Optimization Algorithms”, Key Engineering Materials, Vols. 474-476, 2011, pp.1093-1098.

[18] Xuesong Yan, Qinghua Wu, Can Zhang, Wei Chen Wenjing Luo, Wei Li, “An Efficient Function Optimization Algorithm based on Culture Evolution”, International Journal of Computer Science Issues, Vol. 9, Issue 5, No. 2, 2012, pp.11-18.

[19] Xuesong Yan, Can Zhang, Wenjing Luo, Wei Li, Wei Chen, Hanmin Liu, “Solve Traveling Salesman Problem Using Particle Swarm Optimization Algorithm”, International Journal of Computer Science Issues, Vol. 9, Issue 6, No. 2, 2012, pp.264-271.

Xuesong Yan associate professor received him B.E. degree in Computer Science and Technology in 2000 and M.E. degree in Computer Application from China University of Geosciences in 2003, received he Ph.D. degree in Computer Software and Theory from Wuhan University in 2006. He is currently with School of Computer Science, China University of Geosciences, Wuhan, China and now as a visiting scholar with Department of Computer Science, University of Central Arkansas, Conway, USA. He research interests include evolutionary computation, data mining and computer application.

Qinghua Wu lecturer received her B.E. degree in Computer Science and Technology in 2000, M.E. degree in Computer Application in 2003 and Ph.D. degree in Earth Exploration and Information Technology Theory from China University of Geosciences in 2011. She is currently with School of Computer

Science and Engineering, Wuhan Institute of Technology, Wuhan, China. Her research interests include evolutionary computation, image processing and computer application.

Hanmin Liu associate professor. He is currently as a Ph.D candidate of School of Computer Science, China University of Geosciences, Wuhan, China. He research interests include evolutionary computation and applications.

Wenzhi Huang Lecturer. She is currently with School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China. Her research interests include image processing and computer application.

Table 3: Optimal results comparison

Test Cases	*Optimal in TSPLIB*	*GA*	*PSO*	*IPSO*
Berlin52	7542	7542	7542	7542
kroA100	21282	21315	21310	21282
kroA200	29368	30168	29968	29368
Pr299	48191	48568	48540	48191
Rd400	15281	15135	15135	15281
Ali535	202310	242310	231120	202310
D657	48912	50912	50612	48912
Rat783	8806	8965	8905	8806
U1060	224094	279094	269908	224094
U1432	152970	182780	177890	152970

Table 4: Experiment results comparison (10 runs for each case)

Test Cases	*Known Optimal*	*GA*	*ACO*	*ASAGA*	*Our Algorithm*
Berlin52	7542	7542(7)	7784	7542(7)	7542(10)
kroA100	21282	21315	21637	21282(5)	21282(9)
KroA200	29368	29694	30143	29638(1)	29368(8)

An Approach to Secure Mobile Enterprise Architectures

Florian G. Furtmüller[1]

[1] **Solution Integration & Architecture Department, Computer Sciences Consulting Austria GmbH
Vienna, AT-1200, Austria**

Abstract

Due to increased security awareness of enterprises for mobile applications operating with sensitive or personal data as well as extended regulations form legislative (the principle of proportionality) various approaches, how to implement (extended) two-factor authentication, multi-factor authentication or virtual private network within enterprise mobile environments to ensure delivery of secure applications, have been developed.

Within mobile applications it will not be sufficient to rely on security measures of the individual components or interested parties, an overall concept of a security solution has to be established which requires the interaction of several technologies, standards and system components. These include the physical fuses on the device itself as well as on the network layer (such as integrated security components), security measures (such as employee agreements, contract clauses), insurance coverage, but also software technical protection at the application level (e.g. password protection, encryption, secure container).

The purpose of this paper is to summarize the challenges and practical successes, providing best practices to fulfill appropriate risk coverage of mobile applications. I present a use case, in order to proof the concept in actual work settings, and to demonstrate the adaptability of the approach.

Keywords: *Mobile Security Architecture, Authentication, Security, Integration Architecture & Interoperability, Privacy & Trust.*

1. Introduction

As a consequence of technological advances, it has become possible to fully integrate various actuator technologies as well as mobile devices into enterprise infrastructures and secure environments. These developments open up a huge amount of innovative interaction scenarios, involving new forms of user communication and behavior. Therefore, enabling companies to manage policy, security and support of mobile devices became a major issue. Enterprise mobile device management (MDM) software is evolving to offer cross-platform device support, vendor independence and keeping the focus on a secure integration layer [17].

This is an opportunity for exploring the potentials and perspectives of mobile enablement of enterprises supporting collaborative work that enables employees to increase their productivity (up to 20 % [20]) and efficiency, as well as flexibility and accessibility. Faster networks and extended battery life offer the support of a hand-held microcomputer as personal assistant. This hand held device collects an enormous amount of personal data, ranging from the user's email address to location, contact list, calendar & photos and tether it to a single unique device ID number [4].

Motivated by these developments, its close integration and interaction between business and private usage this paper can be seen as reference and aims bringing together various challenges (refer to section VI):

--Various data (private or business data)
--Access points (internet, intranet, public Wi-Fi…)
--Sensitive device data (daily financial sales reports)
--Application content (business or private applications)
--Corporate data access (secure or non-secure)
--Malware / Phishing software
--Social media usage
--Bring-your-own-device (BYOD)
--Online & offline availability
--News headlines (malpractice).

Clustering the challenges above will lead to the following security risk domains, physical risk (lost or stolen device), access risk (login or network accessibility of not authenticated persons e.g. man-in-the-middle), usage risk (software bugs, jail-breaking, malware) and memory risk (private or sensitive data stored on device (or even worse stored on removable SD card) [20].

Meeting all the multiple challenges and risks in which a mobile enabled enterprise is confronted with must be discussed and clearly clarified before defining an enterprise mobile security strategy. Providing software for mobile devices, calls for context-, business- and data-dependent analysis already at design phase and requires a framework to manage IT architecture (EAM) [14].

The basic concepts of authentication processes, for two-factor authentication and multi-factor authentication, as

well as for a virtual private network approach (see III.B) are used to meet the requirements given above.
I have identified two main security concepts, which meet the requirements, mentioned in I and II, appropriately. Section III presents the concepts that were applied when designing a communication channel of designated applications. In section III and IV, implementation of the approach and application during design time and in operation are described. Section V shows how the concept has been used in practice. Finally, I conclude with a summary of achieved results and some inspirations for future work.

2. Requirements

The lack of time in development phase and the pressure for releasing new applications - especially in the mobile applications market - are the main reasons why developers neglect security issues. Besides automated code analysis software and sophisticated test procedures, it is recommended to perform security testing during the quality assurance (QA) phase and predefine already approved standard libraries for development.

However, the mentioned approaches in section III, TFA, MFA and VPN in combination with adequate MDM solution allow clarification of problem areas, such as protecting corporate data in transit over public Wi-Fi or cellular networks, encrypting data stored on device, disabling device communication modules and hardware features, authenticating device or user using certificates or domain security credentials, malware protection and intrusion detection, preventing harmful internet downloads and unauthorized software installation or strategies for dealing with lost or stolen devices [23] already at design time.

3. Approach

For the development of our best practices, I have reviewed earlier approaches in terms of concepts and implementation strategies. The concepts, that have influenced the presented approach, are described in the following section:

3.1 Two-Factor Authentication (TFA, MFA)

Two-factor authentication represents the combination of knowledge with possession of an object.

Concept
This approach combines the knowledge of a secret (e.g. password) with the possession of a clearly identifiable object (e.g. token, a special USB key or smart card) or personal characteristics (e.g. biometrics). Extended TFA or multi-factor authentication combines more methods [5].

Example: Maestro debit card, logical knowledge (PIN) and physical object (map).

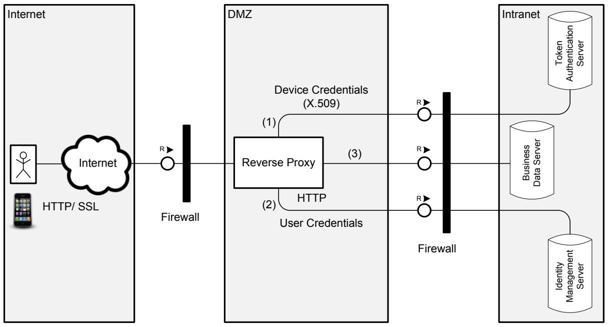


Fig. 1. Two-Factor Authentication (TFA, MFA) Process (refer [4])

Process
As shown in Fig. 1, implementation stack of a two-factor authentication follows the process, (1) authentication of the object (for example via the mobile device installed X.509 certificate), (2) authentication of the property (e.g. password) and as positive result establishes (3) connection to business server. This will lead to strong authentication that meets compliance requirements, provides improved security and minimized risk through a 2nd factor as well as eliminating phishing with one-time passwords and minimized ‘window-of-opportunity’ but does not prevent the possibility of ‘man-in-the-middle’ attacks. Therefore, it will be necessary to establish complex passwords or layers (that will lead to increased training and operational costs) and a contrary waiver of complex passwords requires additional security hardware [19].

3.2 Mobile Virtual Private Network (Mobile VPN)

As already established in various enterprises virtual private networks (VPNs) extends the combination of knowledge with possession of an object.

Concept
In addition to the PIN of the user a second security code (via VPN token) will be required during authentication. This allows establishing a tap- and tamper-proof VPN tunnel as well as encrypting network packets [6].

Example: Home office via VPN, logical knowledge (PIN) and a physical object (SecureID token for One-Time-Passwords)

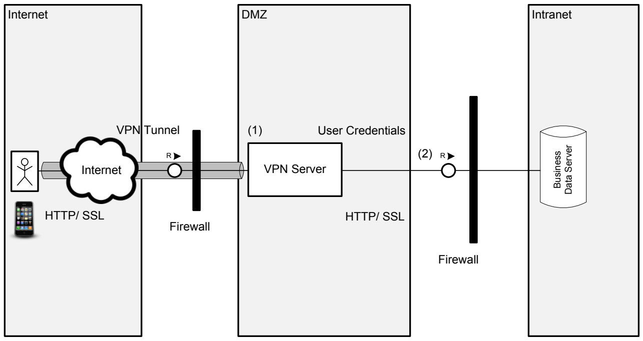


Fig. 2. Virtual Private Network Process (refer [6])

Process

Fig. 2 depicts the process flow of establishing a VPN connection, starting with (1) authentication and setting up a VPN tunnel to the VPN server (e.g. via RSA Token + PIN). After successful authentication the end-user will be (2) connected to the business server (e.g. SAP ERP back-end). VPN incorporates a secure network connection with the possibility of encrypted data packets, scalability (ease of adding or removing users) and standardized by virtual private network consortium (VPNC).

This approach will require additional hardware (e.g. for an RSA token) and encourages users to enter password credentials and token. On the contrary Quality-of-Service (QoS) relies on Internet Service Provider (ISP) which has direct impact on service stability and performance.

More information regarding to user authentication solutions can be found here [26].

4. Implementation & Application

As already defined in the requirements (see, section III and IV), mobile applications are designed to work in secured and non-secured environments. To provide the basic service infrastructure, a mobile device management (MDM) solution should be evaluated upon context and enterprise specific strategic requirements. This kind of middleware embedded in a web-oriented architecture (WOA) provides simple and transparent management of mobile devices over various environments and heterogeneous platforms and features device procurement, asset inventory management, reporting, logging, policy enforcement and management, help desk support, remote configuration and assistance to name but a few.

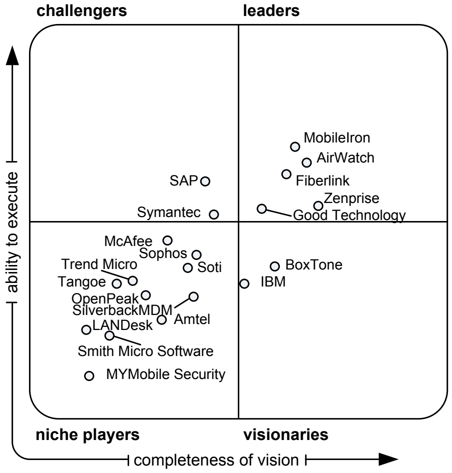


Fig. 3. Magic Gartner Quadrant Mobile Device Management (see [18])

Hence, to customer technical requirements and specific feature requests (see V) I had to evaluate several MDM solutions such as AirWatch, BoxTone MDM, Fiberlink's Maas360 (as pure SaaS), Good Technology's Mobile Control, Matrix42, Enterprise Mobility Management from McAfee, Symantec's Mobile Management and Zenprise an overview of established MDM solutions can be found in Fig. 3, for detailed information regarding to critical capabilities for MDM solutions refer to [17].

To be prepared for upcoming future enterprise requirements and to ensure a viable mobile security infrastructure, a careful selection of corporate MDM solution for implementing mobile security concepts is essential.

4.1 Use of Design Patterns and Frameworks

Object-oriented design patterns (refer to [7]) are commonly accepted means to construct highly structured software that is easy to grasp. Hence, upcoming mobile applications should be developed using established patterns. In this way, existing applications can easily be extended, and its components can easily be modified and exchanged with others (significant reduction of maintenance costs [7]). As an example for the concrete use of frameworks for extended TFA (MFA), Fig. 4 depicts the process flow provided by [29]. To add further factor to the authentication procedure could be realized by simply implementing an extension e.g. dynamic generated password for user login provided by SMS gateway (see Fig. 5).

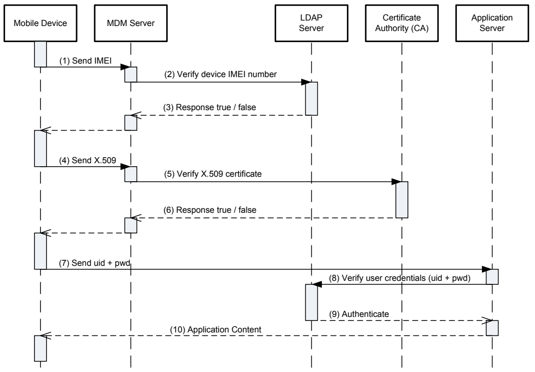

Mobile Device
MDM Server
LDAP Server
Certificate Authority (CA)
Application Server
(1) Send IMEI
(2) Verify device IMEI number
(3) Response true / false
(4) Send X.509
(5) Verify X.509 certificate
(6) Response true / false
(7) Send uid + pwd
(8) Verify user credentials (uid + pwd)
(9) Authenticate
(10) Application Content

Fig. 4. Process Simulation - Extended Security via Static Password

Fig. 4 shows how the authentication process operates, as first factor, device IMEI number (unique number to identify mobile devices) is sent to the MDM solution to verify if device belongs to the group of enterprise managed devices or not. As second factor, a X.509 certificate stored in the web container on the device (see, Fig. 6) will be requested and verified. Finally, as an extended factor, a combination of username and password has to be transmitted to the back-end application server (e.g. SAP). As of this moment, users are authenticated and the back-end service is delivering requested data.

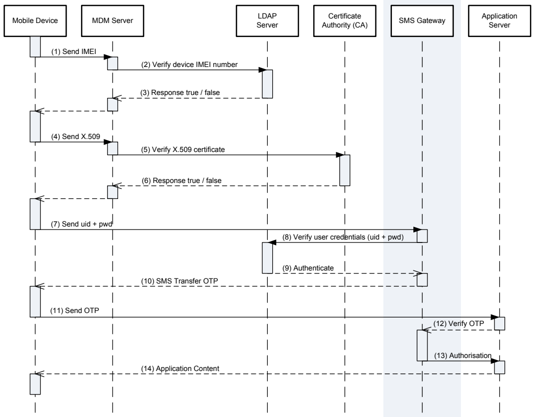


Fig. 5. Process Simulation - Extended Security via One-Time-Password

A reduction of the window-of-opportunity can be achieved by setting up a SMS gateway for issuing One-Time-Passwords (OTP) or Time-based One-Time Passwords (TOTP), as illustrated in Fig. 5. An OTP reference implementation can be found here [15]. Additionally, the HOTP (HMAC-based One Time Password) algorithm represents an open standard for event-based OTP authentication and its community edition is freely available [25].

Alternatives, such as WiKID Strong Authentication System [29] a dual-source, software-based (APIs available for PHP, Java, Ruby, Python and C#) two-factor authentication system (applicable via VPN as well), its flexible design impress with more extensibility than hardware token based solutions.

A popular representative of TFA via TOTP [11] is the open source application *Google Authenticator* available for all Gmail and Google Apps users, details and source can be found on the project page [9].

4.2 Related Work

For the development of our mobile security concepts, I have reviewed earlier approaches in terms of concepts and implementation strategies. The projects - that have influenced the presented approach - are described below:

Project: FACTOR TWO
This approach is built with Sencha Touch mobile framework (see [21]) and provides a simple two-factor authentication mechanism based on SHA256 JS implementation from the *crypto-js* project [10]. FACTOR TWO has been setup for a particular user (via simple PIN exchange between website and mobile application) and a one-time token will be required for each login. If the mobile application has an internet access, the token could be submitted automatically by pushing a single button. Once a valid token had been submitted, the user is able to login into the site with his credentials. The token remains valid for a predefined time interval since it has been submitted (minimize window of opportunity, see section III). If user authentication fails during this time interval, the user will be forced to re-submit a new token. The service utilizes an offline caching mechanism that allows it to run without internet connection. In this mode the application issues a valid token that the users have to enter into login form along with their credentials [24].

Project: QuickSec™ Mobile VPN Client
This project keeps the focus on implementing an application with a specific VPN client, in the way that its VPN connection once established should be only available to the designated application and all other applications, which are installed on the mobile device should use common ISP internet connection.

Android Gingerbread [28] and recent versions, as well as iOS are including a VPN client. Android supports L2TP/IPsec (PPTP) protocols and iOS, SSL VPN, Cisco IPSec, L2TP over IPSec and PPTP. However, split-tunneling is also supported by Android [27] and iOS, this technology ensures that clients are supplied by the existing VPN connection and prevent initiatives to detour data from corporate network [22].

The split-tunneling concept for personal VPN allows users to access a public network commonly the Internet, at the same time that the user is allowed to access resources on the VPN. It avoids bottlenecks and cover bandwidth as the traffic to the Internet does not pass through the VPN server. A drawback of this method is that it essentially renders the VPN vulnerable to attack as it is accessible through the public, non-secure network [1]. This is the case of split-tunneling mechanism of Cisco client Juniper Networks SA2500 SSL VPN appliance. It offers a feature that allows users to work through split tunnel and enable them to route the traffic through different channels. Subsequently, QuickSec Mobile VPN Client offers complete IPsec support, IPv6, IKEv2 with MOBIKE, xAuth, EAP-based authentication and split-tunneling support (ibid.).

However, within this context SSL VPN connection is the most reliable one since features strong encryption and security [16].

In addition to Cisco or QuickSec the following solutions for establishing reliable connections for mobile devices do exist: IAPS Security Services VPN, Hide My Ass! VPN and PureVPN.

5. Proof of Concept

The correct operability and stability of the approach was first tested using simulated components. The practical feasibility was then shown in a real world scenario with an invoice approval process.

5.1 Process Simulation

Hence I developed a process that achieves most of the requirements, mentioned in [I and II]. Therefore I created the following test infrastructure, containing an MDM solution [17] for the central administration of mobile devices (IMEI number will be mapped to user account e.g. LDAP or active directory by the administrator). In a next step, the designated user will receive a download link for the new service or download the application directly from the private application store. Subsequently, during the installation of the app the users will be asked to download and install their personalized X.509 certificate, this X.509 certificate will be stored in the Web Container of the app and cannot be modified or copied by the users.

On the contrary, the Web Container will be administrated by MDM solution (e.g. application updates, wipe-out of stolen or lost devices). As a result the user will now be able to connect to the back-end server. This scenario realized using the multi-factor authentication (MFA) concept (see, section III.A).

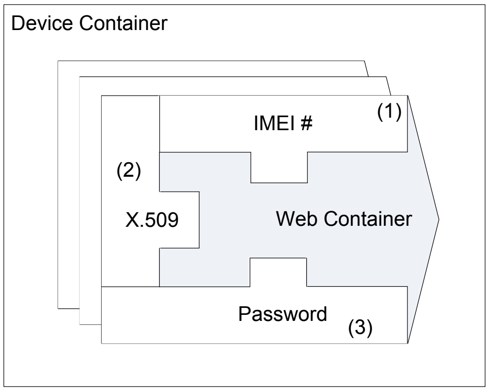


Fig. 6. Simulation – Multi-Factor Authentication (MFA, Extended TFA)

First (1), users must provide device number (IMEI#). As a next step the private user certificate (2) has to be transmitted and verified and finally users must transfer valid logon credentials (3). The logon credentials must be entered by the end user each login time; X.509 (web container) and IMEI# (device container) are stored on the mobile device and will be provided for authentication automatically.

Fig. 6. depicts the following implementation:

1) Device Management (IMEI#) check with registered corporate mobile devices
2) X.509 Certificate User stored by user on mobile device (once a time) check with employee and IMEI#
3) a- Static Password (defined by user) check password e.g. MD5 encrypted
4) b- Dynamic Password (request ‘one-time-password’ (OTP), generated password sent by SMS) e.g. validity 5 min (TOTP).

5.2 An Actual Scenario of Use

Here, an invoice approval process was chosen for demonstration. This process consists of three actors: requestor of the invoice approval (commonly known as accounting assistant) and two approvers (approvers are two signing authorities). The accounting assistant is - among other things - responsible to create, submit, manage and clone requests.

Therefore, it was required to set up a separate security level for the accounting assistant as well as an extended security level for the authorized signatories. Once the invoices have been approved, they will be sent to Accounting for verification and payment (see Fig. 7).

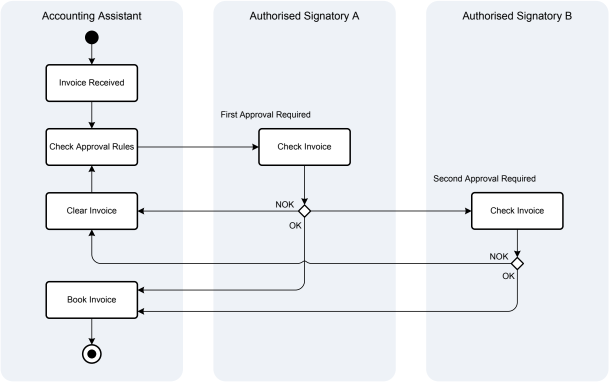


Fig. 7. Scenario of Use - Invoice Approval Process

As a next step it was essential to map approval rules to the enterprise security levels. Therefore, invoices below a minimum limit of € 500.-, recurring invoices or invoices authorized by a contract to which the accounting staff has access do not require approval and can be signed off by the accounting assistant. The amount limit of invoices which can be approved without two authorized signatories are defined with € 100,000.-. To approve an invoice within this range a medium security level was implemented. All invoices above this value must be approved by two authorized approvers, which require a separate security level too - defined as high security level.

Security level descriptions, transformations and technical requirements,

--Low security level -> manage and forward requests, payment sign off for invoices below € 500.- (by accounting assistant) requires IMEI#, X.509 and password (see Fig. 6).

--Medium security level -> one signing authority verification of invoice, approval and payment (above € 500.- to € 100,000.-) requires IMEI#, X.509 and password plus indexed TAN (acting as OTP via SMS Gateway).

--High security level -> two signing authorities verification of invoice, approval and payment (above € 100,000.-) requires IMEI#, X.509 and password plus indexed TAN (acting as OTP via SMS Gateway).

Besides the test for correct operation, a mobile enterprise environment has to be set up. Therefore Fiberlink MaaS360 MDM solution [2] was chosen which allows issuing and verifying X.509 certificates (CA). As a next step, a connection to the corporate LDAP server for provisioning of user objects and user attributes, mapping of IMEI# and UID was established. To achieve a predefined medium and high security level, an OTP [25] service infrastructure (OTP Validation Server and SMS Gateway) [3], which is communicating directly with the mobile application, was implemented (see Fig. 5 & Fig. 8).

The mobile application for invoice approval was built with Sencha Touch [21], a cross-platform HTML5 Mobile application framework, which is based on JavaScript and allows deployment on various mobile devices (tested on Samsung Galaxy, Samsung Galaxy Tab 10, Apple iPhone, Apple iPad 2 and Blackberry Torch) without recompilation.

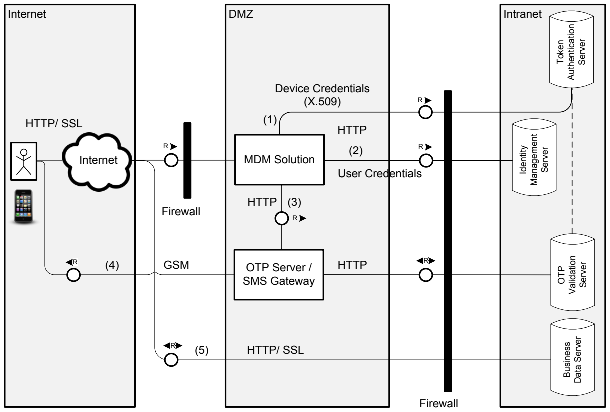


Fig. 8. Solution Architecture - Mobile Invoice Approval Architecture

The integrated environment proved to be successfully orchestrated by the given configuration and allows real-time, asynchronous approving of invoices. In the next stage of the evaluation, I aim at the dynamic extension of the tracked workbench area by attaching further authentication factors (refer to III.A) when necessary.

6. Conclusions

At the very beginning of mobile projects it is recommended to perform a situation analysis, to clarify what is already established in the company, how it can be used, what needs to be procured - how the enterprise can be mobile enabled. It is essential to clearly define requirements (e.g. data quality, integration, single-sign-on, performance, functional and non-functional requirements), perform cost-benefit analysis, verify possible advantages and disadvantages of the planned solution consider legal regulations (e.g. data loss on wipe out of mobile device) and follow the trend.

$$Project\ Duration = \frac{Effort}{Number\ of\ Persons} \qquad (1)$$

Consider Eq. (1), it is commonly accepted but the relation between *'Effort'* and *'Number of Persons'* cannot be seen as valid within software projects and even not for mobile software projects. For this reason it is very important to start a project with prototyping or a simple proof of concept, keep the team small and clearly clarify customer expectations and requirements. In most of the cases mobile projects consists of various infrastructure layers and heterogeneous technologies or components, this will automatically lead to high complexity including steep learning curve and high effort in coordination, communication of all involved parties. 'If you do not actively attack the risks in your project, they will actively attack you' [8].

The presented approach serves to establish several safety standards, especially for mobile business applications within enterprise environments.

All identified requirements could be met both, from a conceptual perspective, and at the implementation level. I was also able to demonstrate user benefits that are as follows: sustainable technology, state-of-the-art security concepts, consideration of legal regulations, combined with years of experience and proven best practices. Additionally, following the presented approaches will minimize implementation risk as well as it will lead to reduce time-to-market of mobile projects.

The concepts mentioned above bring together related topics from security, project management, technology and legal, thus can be seen as cross-divisional procedure or methodology for implementing secured mobile business applications.

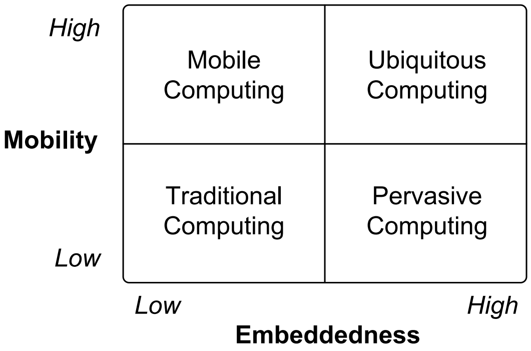


Fig. 9. Future Prospective - Mobile Computing, see [13]

Traditional computing is becoming less important in the future. In contrast, the market is moving increasingly towards ubiquitous computing through to pervasive and mobile computing (see Fig. 9). It is foreseeable that user interfaces will be pushed into the spotlight even more and will provide access to computers for a wide audience (refer to [13]).

In future projects, I will evaluate the approaches technology base and its applicability with VPN split-tunneling in combination with dynamic passwords as Time-based One-Time-Passwords (TOTP) transmitted on SMS gateway infrastructure.
Placing a reverse HTTP proxy in front of the MDM solution that verifies signed OAuth (Open Authorization protocol IETF) requests from mobile devices could also be an option, the claim for universal applicability has still to be validated.

"If you don't know where you are going, any road will take you there." *Lewis Carroll*

Acknowledgments

I would like to thank Jürgen Mairhofer (Department of Biotechnology, University of Natural Resources and Life Sciences, Vienna, Austria), Dietmar Moltner (Computer Sciences Consulting Austria) and Michael Kaufmann (Computer Sciences Consulting Austria) for their input and constructive criticism.

References

[1] An Introduction to IP Security (IPSec) Encryption, Cisco System, 2012. http://www.cisco.com/en/US/tech/tk583/tk372/technologies_tech_note09186a0080094203.shtml (June 2012)

[2] Diodati M., Mobile Device Certificate Enrollment: Are You Vulnerable?, Gartner Blog, 2012. http://blogs.gartner.com/mark-diodati/2012/07/02/mobile-device-certificate-enrollment-are-you-vulnerable (July 2012)

[3] Diodati M., The Evolving Intersection of Mobile Computing and Authentication, Gartner, 2011. http://www.gartner.com/id=1882514 (July 2012)

[4] Etherington D.: Sleepless? Then Stop Taking Your iPhone To Bed, Giga Omni Media, Inc., 2011. http://gigaom.com/2011/05/23/report-mobile-workers-in-bed-with-smartphones (June 2012)

[5] Fadi Aloul, Syed Zahidi, Wassim El-Hajj, Two Factor Authentication Using Mobile Phones, IEEE/ACS 2009.

[6] Ferguson P., Huston G., What is a VPN?, Cisco Systems, 1998. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.972 (July 2012)

[7] Gamma E., Helm R., Johnson R., Vlissides J.: Entwurfsmuster, Elemente wiederverwendbarer objektorientierter Software, Addison-Wesley Verlag, 2004.

[8] Gilb T., Finzi S.: Principles of software engineering management, Addison-Wesley Pub. Co., 1988.

[9] Google Open Source Project, Google Authenticator (TOTP), Google Inc., 2011. http://code.google.com/p/google-authenticator/ (June 2012)

[10] Google Project, JavaScript implementations of standard and secure cryptographic algorithms, Google Inc., 2012 http://code.google.com/p/crypto-js/ (June 2012)

[11] Internet Engineering Task Force, TOTP Time extension of HMAC-based One-Time Password algorithm, 2011. http://www.rfc-editor.org/rfc/rfc6238.txt (June 2012)

[12] Lehner F.: Mobile und drahtlose Informationssysteme: Technologien, Anwendungen, Märkte, Springer, 2003.

[13] Lyytinen K., Yoo, Y.: Issues and Challenges in Ubiquitous Computing, Communications of the ACM, 2002.

[14] Buckow H., Rey S. - Business Needs IT Architecture, McKinsey & Company, 2010.

http://www.mckinseyquarterly.com/Why_business_needs_should_shape_IT_architecture_2563 (June 2012)

[15] OATH Initiative, HOTP Algorithm - One-Time-Password Reference Implementation, 2004. http://rfc-ref.org/RFC-TEXTS/4226/chapter16.html (June 2012)

[16] QuickSec™ Mobile VPN Client for Android, AuthenTec, Inc., 2011. http://www.authentec.com/Products/EmbeddedSecurity/SecurityToolkits/QuickSecVPNAndroid.aspx (June 2012)

[17] Redman P., Basso M., Critical Capabilities for Mobile Device Management, Gartner Inc., 2011. http://www.gartner.com/DisplayDocument?doc_cd=213877 (June 2012)

[18] Redman P., Girard J., Wallin L. O., Magic Quadrant for Mobile Device Management Software, Gartner Inc., 2012. http://www.gartner.com/DisplayDocument?doc_cd=211101 (June 2012)

[19] Schneider B.: Two-Factor Authentication: Too Little - Too Late, Communications of the ACM, 2005.

[20] Schurek A.,Das mobile Industrieunternehmen, IBM, 2012. http://www-05.ibm.com/de/events/swg-solutions/pdf/Das-mobile-Industrieunternehmen-external_2012-05-22.pdf (June 2012)

[21] Sencha Touch 2, A high-performance HTML5 mobile application framework, Motorola Mobility Inc., 2012. http://www.sencha.com/products/touch (June 2012)

[22] Shinder T., Remote Access VPN and a Twist on the Dangers of Split Tunneling, TechGenix Ltd., 2005. http://www.isaserver.org/tutorials/2004fixipsectunnel.html (June 2012)

[23] Symantec Security Analysis of iOS and Android, Symantec Corp., 2011. http://www.symantec.com/about/news/release/article.jsp?prid=20110627_02 (June 2012)

[24] Tchijov A., Project FACTOR TWO: Factor-2 based Authentication built on Sencha Touch Framework), 2011. http://drupal.org/project/factortwo (June 2012)

[25] The Internet Society, Reference HOTP Algorithm - An HMAC-Based One-Time Password Algorithm, 2005. http://www.ietf.org/rfc/rfc4226.txt (June 2012)

[26] VeriSign Benutzerauthentifizierungslösungen, Symantec Corp., 2012. http://www.symantec.com/de/de/user-authentication (June 2012)

[27] Marg C., VPN Konfiguration - VPN unter Android nutzen, TU Clausthal, 2012. https://doku.tu-clausthal.de/doku.php?id=vpn:vpn_unter_android_nutzen (June 2012)

[28] The Android Open Source - Android VPN Profile Reference Implementation, 2009. http://www.java2s.com/Open-Source/Android/android-core/platform-frameworks-base/android/net/vpn/VpnProfile.java.htm (June 2012)

[29] WiKID Strong Authentication System, Open Source Two-Factor Authentication, 2011. http://www.wikidsystems.com/community-version (June 2012)

Florian G. Furtmüller studied business informatics at the Johannes Kepler University of Linz, Austria. He earned his master degree in 2007 and subsequently, started his work as Enterprise Portal Consultant for SAP and IBM systems at IDS Scheer Austria GmbH.

He has more than five years of experience in application integration architectures SOA, WOA, EDA, MDA and design patterns, especially with enterprise portals and components. He is currently working for Computer Sciences Consulting Austria GmbH, department for Solution Integration & Architecture as Senior Consultant and Application Architect.

Mr. Furtmüller is co-author of A Tuple-Space based Middleware for Collaborative Tangible User Interfaces in Proceedings of WETICE 07, IEEE Press, 2007, ISBN 0-7695-2879-1, a collaboration with Stefan Oppl.

An Improved Genetic Algorithm and Its Application in Classification

Xuesong Yan[1,2], Wenjing Luo[1], Wei Li[1], Wei Chen[1], Can Zhang[1] and Hanmin Liu[3]

**[1] School of Computer Science, China University of Geosciences
Wuhan, Hubei 430074, China**

**[2] Department of Computer Science, University of Central Arkansas
Conway, AR 72035, USA**

**[3] Wuhan Institute of Ship Building Technology
Wuhan, Hubei 430050, China**

Abstract

In this paper, based on a simple genetic algorithm and combine the base ideology of orthogonal design method then applied it to the population initialization, using the intergenerational elite mechanism, as well as the introduction of adaptive local search operator to prevent trapped into the local minimum and improve the convergence speed to form a new genetic algorithm. Through the series of numerical experiments, the new algorithm has been proved to be efficiency. we also use this new algorithm in data classification, select 5 benchmark datasets and the experiment results shown the new algorithm can get higher accuracy than k-nearest neighbor method.

Keywords: *Genetic Algorithm, Optimization, Classification, K-Nearest Neighbor, Population.*

1. Introduction

Candidate solutions to some problems are not simply deemed correct or incorrect but are instead rated in terms of quality and finding the candidate solution with the highest quality is known as optimization. Optimization problems arise in many real-world scenarios. Take for example the spreading of manure on a cornfield, where depending on the species of grain, the soil quality, expected amount of rain, sunshine and so on, we wish to find the amount and composition of fertilizer that maximizes the crop, while still being within the bounds imposed by environmental law.

Several challenges arise in optimization. First is the nature of the problem to be optimized which may have several local optima the optimizer can get stuck in, the problem may be discontinuous, candidate solutions may yield different fitness values when evaluated at different times, and there may be constraints as to what candidate solutions are feasible as actual solutions to the real-world problem. Furthermore, the large number of candidate solutions to an optimization problem makes it intractable to consider all candidate solutions in turn, which is the only way to be completely sure that the global optimum has been found. This difficulty grows much worse with increasing dimensionality, which is frequently called the curse of dimensionality, a name that is attributed to Bellman, see for example [1]. This phenomenon can be understood by first considering an n-dimensional binary search-space. Here, adding another dimension to the problem means a doubling of the number of candidate solutions. So the number of candidate solutions grows exponentially with increasing dimensionality. The same principle holds for continuous or real-valued search-spaces, only it is now the volume of the search-space that grows exponentially with increasing dimensionality. In either case it is therefore of great interest to find optimization methods which not only perform well in few dimensions, but do not require an exponential number of fitness evaluations as the dimensionality grows. Preferably such optimization methods have a linear relationship between the dimensionality of the problem and the number of candidate solutions they must evaluate in order to achieve satisfactory results, that is, optimization methods should ideally have linear time-complexity O(n) in the dimensionality n of the problem to be optimized.

Another challenge in optimization arises from how much or how little is known about the problem at hand. For example, if the optimization problem is given by a simple formula then it may be possible to derive the inverse of that formula and thus find its optimum. Other families of problems have had specialized methods developed to optimize them efficiently. But when nothing is known about the optimization problem at hand, then the No Free Lunch (NFL) set of theorems by Wolpert and Macready states that any one optimization method will be as likely as any other to find a satisfactory solution [2]. This is especially important in deciding what performance goals one should have when designing new optimization methods, and whether one should attempt to devise the ultimate optimization method which will adapt to all problems and perform well. According to the NFL theorems such an optimization method does not exist and the focus of this thesis will therefore be on the opposite: Simple optimization methods that perform well for a range of problems of interest.

The most popular evolutionary model used in the current research is Genetic Algorithms (GA), originally developed by John Holland [3]. The GA reproduction operators, such as recombination and mutation, are considered analogous to the biological process of mutation and crossover respectively in population genetics. The recombination operator is traditionally used as the primary search operator in GA while the mutation operator is considered to be a background operator, which is applied with a small probability.

Genetic algorithms have been successfully used in data mining, in order to determine classification rules [4], in order to search for appropriate cluster centers) [5], to select the attributes of interest in predicting the value of a target attribute [6], etc. Classification of instances was performed using some hybrid algorithms based on genetic algorithms and particle swarm optimization [7], respectively Naive Bayes and k-Nearest Neighbors [8]. A few applications in which genetic algorithms were successfully applied to solve classification problems are prints classification, heart disease classification, classification of emotions on the human face, etc.

2. Improved Genetic Algorithm

In general, genetic algorithms are usually used to solve problems with little or no domain knowledge, NP-complete problems, and problems for which near optimum solution is sufficient. The GA methods can be applied only if there exist a reasonable time and space for evolution to take place. But the traditional genetic algorithm has the shortcoming: trapped into the local minimum easily [9].

2.1 Population Initialization

The traditional method of genetic algorithm is randomly initialized population, that is, generate a series of random numbers in the solution space of the question. Design the new algorithm, we using the orthogonal initialization [10] in the initialization phase. For the general condition, before seeking out the optimal solution the location of the global optimal solution is impossible to know, for some high-dimensional and multi-mode functions to optimize, the function itself has a lot of poles, and the global optimum location of the function is unknown. If the initial population of chromosomes can be evenly distributed in the feasible solution space, the algorithm can evenly search in the solution space for the global optimum. Orthogonal initialization is to use the orthogonal table has the dispersion and uniformity comparable; the individual will be initialized uniformly dispersed into the search space, so the orthogonal design method can be used to generate uniformly distributed initial population.

2.2 Elite Select Mechanism

Genetic algorithm is usually complete the selection operation based on the individual's fitness value, in the mechanism of intergenerational elite, the population of the front generation mixed with the new population which generate through crossover and mutation operations, in the mixed population select the optimum individuals according to a certain probability. The specific procedure is as follows:
Step1: using crossover and mutation operations for population P1 which size is N then generating the next generation of sub-populations P2;
Step2: The current population P1 and the next generation of sub-populations P2 mixed together form a temporary population;
Step3: Temporary population according to fitness values in descending order, to retain the best N individuals to form new populations P1.

The characteristic of this mechanism is mainly in the following aspects. First is robust, because of using this selection strategy, even when the crossover and mutation operations to produce more inferior individuals, as the results of the majority of individual residues of the original population, does not cause lower the fitness value of the individual. The second is in genetic diversity maintaining, the operation of large populations, you can better maintain the genetic diversity of the population evolution process. Third is in the sorting method, it is good to overcome proportional to adapt to the calculation of scale.

2.3 Adaptive Search Operator

Local search operator has a strong local search ability, and then can solve the shortcomings of genetic algorithm has the weak ability for the local search. And the population according to the current state of adaptive evolution of the local search space adaptive local search operator will undoubtedly greatly enhance the ability of local search. In the initial stage of the evolution, the current optimal solution from the global optimum region is still relatively far away, this time the adaptive local search operator to require search a large neighborhood space to find more optimal solution, it can maintain the population diversity.

When the population has evolved to the region containing the global optimum, the adaptive local search operator to require a relatively small area to search in order to improve the accuracy of the global optimal solution.

In our algorithm, the adaptive local search operator is the adaptive orthogonal local search operator. Adaptive orthogonal local search operator is aimed at the neighborhood of a point to search, so the key point is to identify a point as the center of the hypercube, the hypercube in the orthogonal test, expect to be better solution.

2.4 Experiment Simulation

We design the experiments to study its convergence speed by comparing with a traditional genetic algorithm (GA). Ten benchmark functions are selected. One of them is a multimodal function, which is a very difficult function (explained in its function description). We choose it because we want to investigate not only their convergence speeds, but also their abilities of finding the optimal solutions. The simple description of each function is given as follows.

F1: Schaffer function

$$\min f(x_i) = 0.5 - \frac{(\sin^2 \sqrt{x_1^2 + x_2^2} - 0.5)}{[1 + 0.001(x_1^2 + x_2^2)]^2},$$

$$-100 \le x_i \le 100$$

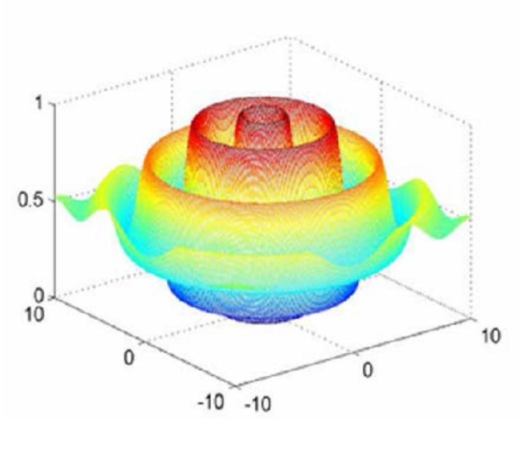


Fig. 1 Schaffer Function.

The global optimal value of this function is 1.0, located at the central point with coordinates (0, 0), and the circle with the radius 3.14 on the overall situation from numerous major points of the uplift. This function has a strong shock. Therefore, it is difficult to find a general method, which can find its global optimal solution.

F2: Shubert function

$$\min f(x, y) = \left\{ \sum_{i=1}^{5} i \cos\left[(i+1) x + i \right] \right\} \times$$

$$\left\{ \sum_{i=1}^{5} i \cos\left[(i+1) y + i \right] \right\}, x, y \in [-10, 10]$$

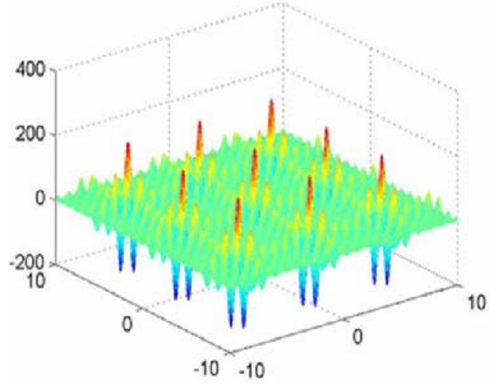


Fig. 2 Shubert Function.

This function has 760 local minima and 18 global ones. The global minimum value is -186.7309.

F3: Hansen function

$$\min f(x, y) = \sum_{i=1}^{5} i \cos((i-1)x + i)$$

$$\sum_{j=1}^{5} j \cos((j+1)y + j), x, y \in [-10, 10]$$

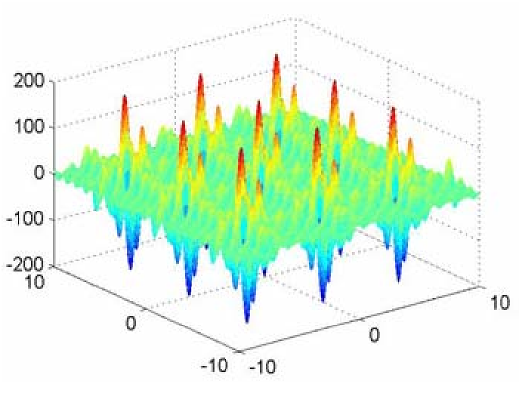


Fig. 3 Hansen Function.

This function has a global minimum value -176.541793, in the following nine points, i.e., (-7.589893, -7.708314), (-7.589893, -1.425128), (-7.589893, 4.858057), (-1.306708, -7.708314), (-1.306708, -1.425128), (-1.306708, 4.858057), (4.976478, -7.708314), (4.976478, -7.708314), and (4.976478, 4.858057). It also has 760 local minima.

F4: Camel function

$$\min f(x,y) = \left(4 - 2.1x^2 + \frac{x^4}{3}\right)x^2 + xy + \left(-4 + 4y^2\right)y^2, x, y \in [-100,100]$$

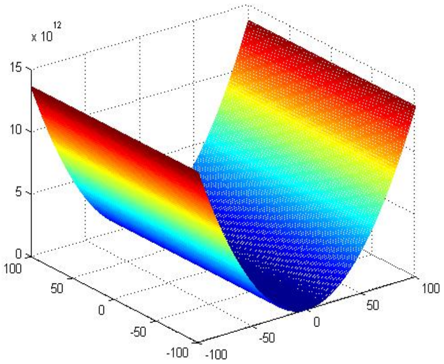


Fig. 4 Camel function.

Camel function has six local minima, i.e., (1.607105, 0.568651), (-1.607105, -0.568651), (1.703607, -0.796084), (-1.703607, 0.796084), (-0.0898, 0.7126) and (0.0898, -0.7126). It has two global minimum points, i.e., (-0.0898, 0.7126) and (0.0898, -0.7126). Its global minimum is -1.031628.

The function 5 (called F5 in this paper) can be stated as follows:

$$\min f(x_1,x_2) = \begin{Bmatrix} 1 + (x_1 + x_2 + 1)^2 \\ (19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1x_2 + 3x_2^2) \end{Bmatrix} \times \begin{Bmatrix} 30 + (2x_1 - 3x_2)^2 \\ (18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2) \end{Bmatrix}, x_1, x_2 \in [-50,50]$$

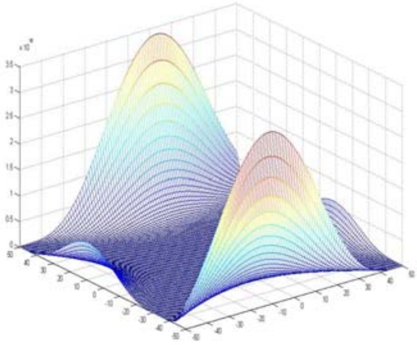


Fig. 5 Function 5.

This function is an eighth-order polynomial with two variables, shown in Figure 5. However, it has four local minima, including a global one, i.e., $f(1.2,0.8) = 840.0$, $f(1.8,0.2) = 84.0$, $f(0.6,0.4) = 30.0$ and $f^*(0,1.0) = 3.0$ (global minimum).

The function 6 (called F6 in this paper) can be stated as follows:

$$\min f(x,y) = -\begin{bmatrix} x\sin(9\pi y) + \\ y\cos(25\pi x) + 20 \end{bmatrix}, x, y \in [-10,10]$$

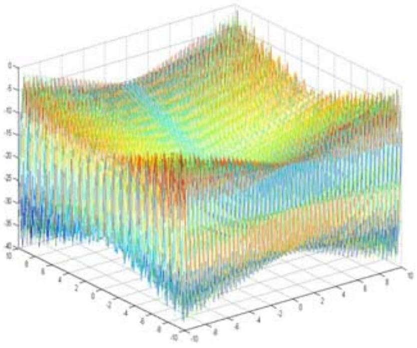


Fig. 6 Function 6.

This is a very complex and difficult function. First, it is a multimodal function. Besides, $\sin(9\pi y)$ and $\cos(25\pi x)$ are high frequency oscillations in the different directions. Furthermore, its peak (or ravine) of the function is intensive at the points when $|x, y| \to 10$. The scene of this function is also very complicated, shown in Figure 6. The global optimal of this function is (-10, 9.9445695) = -39.944506953367.

The function 7 (called F7 in this paper) is defined as follows:

$$\min f(x_1,x_2) = 20 + x_1^2 + x_2^2 - 10 * (\cos 2\pi x_1 + \cos 2\pi x_2)$$

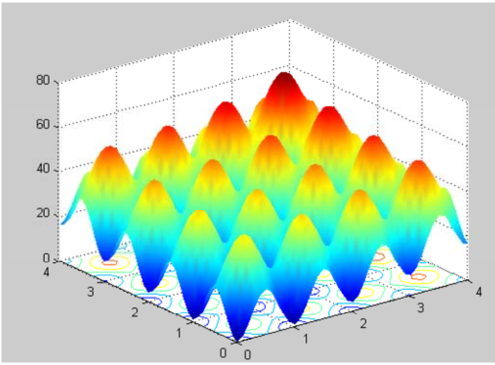


Fig. 7 Function 7.

Its minimum value is 0, as shown in Figure 7.

The function 8 (called F8 in this paper) is defined as follows:

$$\min f(x_1, x_2) = 100*(x_1^2 - x_2)^2 + (1 - x_1)^2,$$
$$x \in [-2.048, 2.048]$$

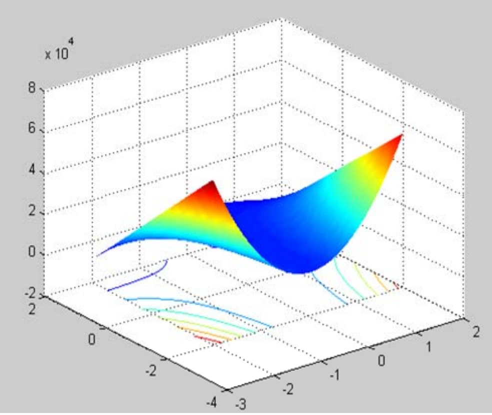


Fig. 8 Function 8.

Its minimum value is 0, as shown in Figure 8.

The function 9 (called F9 in this paper) is defined as follows:

$$\min f(x_1, x_2) = x_1^2 + x_2^2, x_1, x_2 \in [-100, 100]$$

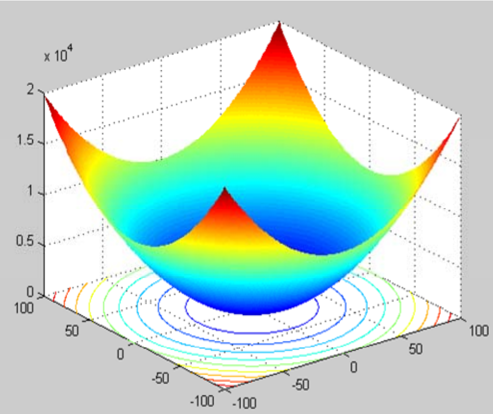


Fig. 9 Function 9.

Its minimum value is 0, as shown in Figure 9.

The function 10 (called F10 in this paper) is defined as follows:

$$\min f(x_1, x_2) = 0.5x_1^2 + 0.5(1 - \cos 2x_2) + x_2^2$$

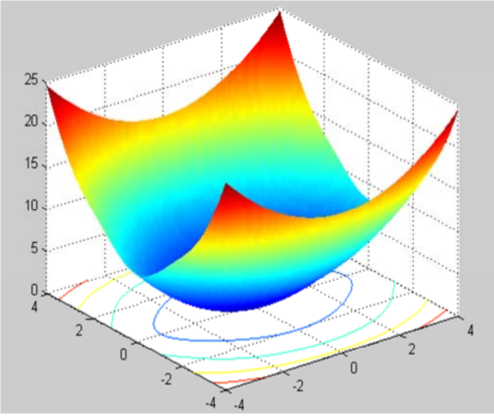


Fig. 10 Function 10.

Its minimum value is 0, as shown in Figure 10.

In order to obtain a solid comparison between GA and improved GA (IGA), we run each algorithm 100 times for the ten functions described above. Our experimental results are shown in Table 1, including the best solution and the number of times of finding the best solution for each function. For example, on the most difficult function F6 among the ten functions, GA could not find its optimal solution (i.e, 0 times out of 100 runs). The best solution GA achieved is -14.786954. However, PSO found its optimal solution (-39.944506) five times out of 100 runs).

Table 1: The number of times of both GA and PSO achieve optimal solutions among 100 runs on the ten functions

Function	***Algorithm***	***Convergence Times***	***Optimal Solution***
F1	GA	72	1.0000000
	IGA	75	1.0000000
F2	GA	75	-186.730909
	IGA	82	-186.730909
F3	GA	85	-176.541793
	IGA	91	-176.541793
F4	GA	23	-1.031628
	IGA	58	-1.031628
F5	GA	16	3.000000
	IGA	25	3.000000
F6	GA	0	-14.786954
	IGA	8	-39.944506
F7	GA	90	0.000000
	IGA	97	0.000000
F8	GA	90	0.000000
	IGA	98	0.000000

F9	GA	100	0.000000
	IGA	100	0.000000
F10	GA	93	0.000000
	IGA	98	0.000000

From Table 1, we can see that IGA achieves optimal solution more frequently than GA does on nine out of the ten functions, except the easiest one (i.e., F9). On the easiest function F9, both of them achieve the optimal solution in all 100 runs. From the experimental results in Table 1, we can conclude that the IGA algorithm has more efficient global searching capability than the GA algorithm. Our experiments verified that IGA converges more quickly than the GA algorithm.

3. K-Nearest Neighbor Classification Algorithm

The nearest neighbor method [11, 12] represents one of the simplest and most intuitive techniques in the field of statistical discrimination. It is a nonparametric method, where a new observation is placed into the class of the observation from the learning set that is closest to the new observation, with respect to the covariates used. The determination of this similarity is based on distance measures.

Formally this simple fact can be described as follows: Let $L = \{(y_i, x_i), i = 1, 2, ..., n_L\}$ be training or learning set of observed data, where $y_i \in \{1, 2, ..., c\}$ denotes class membership and the vector $x_i^{'} = (x_{i1}, x_{i2}, ..., x_{ip})$ represents the predictor values. The determination of the nearest neighbors is based on an arbitrary distance function $d(.,.)$. Then for a new observation (y, x) the nearest neighbor $(y_{(1)}, x_{(1)})$ within the learning set is determined by $d(x, x_{(1)}) = \min_i(d(x, x_i))$ and $\hat{y} = y_{(1)}$ the class of the nearest neighbor is selected as prediction for y. The notation $x_{(j)}$ and $y_{(j)}$ here describes the jth nearest neighbor of x and its class membership, respectively.

For example, such typical distance functions are the Euclidean distance $d(x, x_j) = (\sum_{s=1}^{p}(x_{is} - x_{js})^2)^{\frac{1}{2}}$.

The method has been explained by the random occurrence of the learning set, as described in Fahrmeir et al. [13]. The class label $y_{(1)}$ of the nearest neighbor $x_{(1)}$ of a new case x is a random variable. So the classification probability of x into class $y_{(1)}$ is $P(y_{(1)} | x_{(1)})$. For large learning sets x and $x_{(1)}$ coincide very closely with each other, so $P(y_{(1)} | x_{(1)}) \approx P(y | x)$ results approximately. Therefore the new observation x is predicted as belonging to the true class y with the probability approximately $P(y | x)$.

A first extension of this idea, which is widely and commonly used in practice, is the so-called k-nearest neighbor method. Here not only the closest observation within the learning set is referred for classification, but also the k most similar cases. The parameter k has to be selected by the user. Then the decision is in favor of the class label, most of these neighbors belong to.

Let k_r denote the number of observations from the group of the nearest neighbors, that belong to class r: $\sum_{r=1}^{c} k_r = k$. Then a new observation is predicted into the class l with $k_l = \max_r(k_r)$. This prevents one singular observation from the learning set deciding about the predicted class. The degree of locality of this technique is determined by the parameter k: For $k = 1$ one gets the simple nearest neighbor method as maximal local technique, for $k \to n_L$ a global majority vote of the whole learning set results. This implies a constant prediction for all new observations that have to be classified: Always the most frequent class within the learning set is predicted.

K-Nearest Neighbor (KNN) is one of the most popular algorithms for pattern recognition. Many researchers have found that the KNN algorithm accomplishes very good performance in their experiments on different data sets.
The traditional KNN text classification has three limitations [14]:
1. High calculation complexity: To find out the k nearest neighbor samples, all the similarities between the training samples must be calculated. When the number of training samples is less, the KNN classifier is no longer optimal, but if the training set contains a huge number of samples, the KNN classifier needs more time to calculate the similarities. This problem can be solved in 3 ways: reducing the dimensions of the feature space; using smaller data sets; using improved algorithm which can accelerate to [15];
2. Dependency on the training set: The classifier is generated only with the training samples and it does not use any additional data. This makes the algorithm to depend on the training set excessively; it needs recalculation even if there is a small change on training set;
3. No weight difference between samples: All the training samples are treated equally; there is no difference between the samples with small number of data and huge number

of data. So it doesn't match the actual phenomenon where the samples have uneven distribution commonly.

4. Classification Experiment

String Representation [16]-Here the chromosomes are encoded with real numbers; the number of genes in each chromosome represents the samples in the training set. Each gene will have 5 digits for vector index and k number of genes. For example, if k=5, a sample chromosome may look as follows:
00100 10010 00256 01875 00098

Here, the 00098 represents, the 98th instance and the second gene say that the 1875 instance in the training sample. Once the initial population is generated now we are ready to apply genetic operators. With these k neighbors, the distance between each sample in the testing set is calculated and the accuracy is stored as the fitness values of this chromosome.

The algorithm process step is given as Fig. 11.

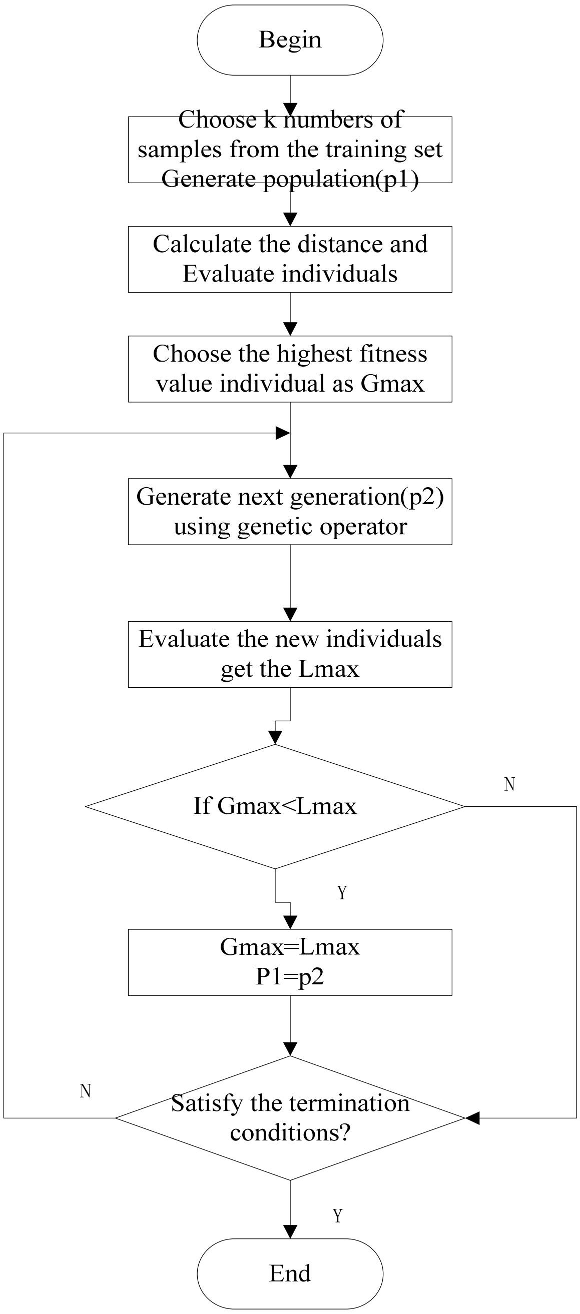

Begin
Choose k numbers of samples from the training set Generate population(p1)
Calculate the distance and Evaluate individuals
Choose the highest fitness value individual as Gmax
Generate next generation(p2) using genetic operator
Evaluate the new individuals get the Lmax
If Gmax<Lmax
N
Y
Gmax=Lmax
P1=p2
N
Satisfy the termination conditions?
Y
End

Fig. 11 Algorithm framework

The performance of the approaches discussed in this paper has been tested with 5 different datasets, downloaded from UCI machine learning data repository. All experiments are performed on Intel Core(TM)2 Duo CPU 2.26GHz/4G RAM Laptop. Each datasets run 10 times with different k values. Table 2 shows the details about the datasets used in this paper.

Table 2: Experiment dataset

Dataset Name	*Total No. of Instances*	*Total No. of Features*
Balance	624	5
Iris	150	4
Sonar	208	60
Glass	214	10
Ionosphere	351	34

Table 3 depicts the performance accuracy of our proposed classifier compared with traditional KNN. From the results it is shown that our proposed method outperforms the traditional KNN method with higher accuracy.

5. Conclusions

This paper introduces a new algorithm based on the traditional genetic algorithm, for the traditional GA algorithm the new algorithm has done some improvements: By introducing genetic selection strategy, decreased the possibility of being trapped into a local optimum. Compared the traditional genetic algorithm, the new algorithm enlarges the searching space and the complexity is not high. By analyzing the testing results of benchmarks functions optimization, we reach the conclusion: in the optimization precision, the new algorithm is efficiency than the traditional genetic algorithm. We also use this new algorithm for data classification and the experiment results shown that our proposed algorithm outperforms the KNN with greater accuracy.

Acknowledgments

This paper is supported by Natural Science Foundation of China. (No.61272470 and No.61203307), National Civil Aerospace Pre-research Project of China, the Provincial Natural Science Foundation of Hubei (2012FFB04101), and the Fundamental Research Founds for National University, China University of Geosciences(Wuhan). The author thanks Dr. Victor Sheng of Department of Computer Science, University of Central Arkansas, U. S. A for his help.

References

[1] R. Bellman, “Dynamic Programming”, Princeton University Press, 1957.

[2] D. H. Wolpert and W. G. Macready, “No free lunch theorems for optimization”, IEEE Transactions on Evolutionary Computation, 1(1), 1997, pp.67-82.

[3] J. Holland, “Adaptation in natural and artificial systems”, University of Michigan press, 1975.

[4] S.Dehuri, A. Ghosh and R. Mall, “Genetic Algorithms for Multi-Criterion Classification and Clustering in Data Mining”, International Jurnal of Computing & Information Sciences, 2006, pp. 143-154.

[5] U. Maulik and S. Bandyopadhyay, “Genetic algorithm-based clustering technique”, Pattern Recognition 33, 2000, pp. 1455-1465.

[6] A.A. Freitas, “A survey of evolutionary algorithms for data mining and knowledge discovery”, Advances in Evolutionary Computation, Springer-Verlag, 2002, pp. 819-845.

[7] R. Ding, H. Dong, X. Feng and G. Yin, “A Hybrid Particle Swarm Genetic Algorithm for Classification”, Proceedings of the 2009 International Joint Conference on Computational Sciences and Optimization, vol.2, 2009, pp. 301.

[8] M. Aci, C. Inan and M. Avci, “A hybrid classification method of k nearest neighbor”, Bayesian methods and genetic algorithm, Expert Systems with Applications, Vol. 37, 2010, pp. 5061-5067.

[9] Xuesong Yan, Qinghua Wu, Can Zhang, Wei Li, Wei Chen, Wenjing Luo, “An Improved Genetic Algorithm and Its Application” TELKOMNIKA Indonesian Journal of Electrical Engineering, vol. 10, No. 5, 2012, pp. 1081-1086.

[10] Leung Yiu-Wing, Wang Yuping, “An orthogonal genetic algorithm with quantization for global numerical optimization”, IEEE Transactions on Evolutionary Computation, 5(1), 2001, pp. 41-53.

[11] E. Fix, and J. Hodges, “Discriminatory analysis Nonparametric discrimination: Consistency properties”, Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[12] T.M. Cover, and P.E. Hart, “Nearest neighbor pattern classification”, IEEE Transactions on Information Theory, 13, pp. 21–27, 1967.

[13] Fahrmeir, Hamerle and Tutz, “Multivariate statistische Verfahren”, Walter de Gruyter & Co Verlag; Berlin, 1996.

[14] W. Yu, and W. Zhengguo, “A fast kNN algorithm for text categorization”, Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, 2007, pp. 3436-3441.

[15] W. Yi, B. Shi, and W. Zhang’ou, “A Fast KNN Algorithm Applied to Web Text Categorization”, Journal of The China Society for Scientific and Technical Information, 26(1), pp. 60-64, 2007.

[16] N. Suguna and Dr. K. Thanushkodi, “An Improved k-Nearest Neighbor Classification Using Genetic Algorithm”, International Journal of Computer Science Issues, Vol. 7, Issue 4, No 2, 2010, pp. 18-21.

Xuesong Yan associate professor received him B.E. degree in Computer Science and Technology in 2000 and M.E. degree in Computer Application from China University of Geosciences in 2003, received he Ph.D. degree in Computer Software and Theory from Wuhan University in 2006. He is currently with School of Computer Science, China University of Geosciences, Wuhan, China and now as a visiting scholar with Department of Computer Science, University of Central Arkansas, Conway, USA. He research interests include evolutionary computation, data mining and computer application.

Wenjing Luo received her B.E. degree in Computer Science and Technology in 2012. She is currently is the M.E. degree candidate with School of Computer Science, China University of Geosciences, Wuhan, China. Her research interests include evolutionary computation.

Wei Li received her B.E. degree in Computer Science and Technology in 2012. She is currently is the M.E. degree candidate with School of Computer Science, China University of Geosciences, Wuhan, China. Her research interests include evolutionary computation.

Wei Chen received him B.E. degree in Computer Science and Technology in 2012. He is currently is the M.E. degree candidate with School of Computer Science, China University of Geosciences, Wuhan, China. Her research interests include evolutionary computation.

Can Zhang received him B.E. degree in Computer Science and Technology in 2011. He is currently is the M.E. degree candidate with School of Computer Science, China University of Geosciences, Wuhan, China. Her research interests include evolutionary computation.

Hanmin Liu associate professor. He is currently as a Ph.D candidate of School of Computer Science, China University of Geosciences, Wuhan, China. He research interests include evolutionary computation and applications.

Table 3: Experiment results comparison

Dataset Name	*K Value*	*Algorithm*	*Best Accuracy*	*Worst Accuracy*	*Mean Accuracy*
Balance	3	IGA	0.904255	0.840426	0.869903
		KNN	0.914894	0.824468	0.866489
	5	IGA	0.882979	0.824468	0.860511
		KNN	0.941489	0.851064	0.875
	7	IGA	0.925532	0.81383	0.863862
		KNN	0.898936	0.819149	0.86117
	9	IGA	0.909574	0.84	0.877272
		KNN	0.920213	0.840426	0.882979
Iris	3	IGA	0.977778	0.866667	0.933333
		KNN	1	0.933333	0.973333
	5	IGA	1	0.933333	0.973867
		KNN	1	0.911111	0.971111
	7	IGA	1	0.911111	0.968889
		KNN	1	0.888888	0.948889
	9	IGA	1	0.96	0.981683
		KNN	1	0.955556	0.977778
Sonar	3	IGA	0.920635	0.825397	0.868173
		KNN	0.904762	0.746032	0.806349
	5	IGA	0.904762	0.714286	0.82618
		KNN	0.857143	0.666667	0.78254
	7	IGA	0.904762	0.714286	0.790363
		KNN	0.888889	0.587302	0.739683
	9	IGA	0.934564	0.698413	0.787446

		KNN	0.873016	0.603175	0.736508
Glass	3	IGA	0.923077	0.643357	0.765542
		KNN	0.861538	0.584615	0.713846
	5	IGA	0.783516	0.676923	0.745714
		KNN	0.784615	0.630769	0.687692
	7	IGA	0.830769	0.630769	0.730769
		KNN	0.753846	0.569231	0.672308
	9	IGA	0.784615	0.553846	0.671057
		KNN	0.753846	0.553846	0.661538
Ionosphere	3	IGA	0.981132	0.871749	0.927245
		KNN	0.943396	0.811321	0.866038
	5	IGA	0.943396	0.867925	0.911003
		KNN	0.915094	0.830189	0.883962
	7	IGA	0.962264	0.792453	0.900066
		KNN	0.896226	0.764151	0.834906
	9	IGA	0.933962	0.839623	0.90321
		KNN	0.858491	0.716981	0.774528

Affective Computing Model for the Set Pair Users on Twitter

Chunying Zhang[1], Jing Wang [2]

**1 College of Science, Hebei United University
Hebei, Tangshan, China**

**[2] College of Science, Hebei United University
Hebei, Tangshan, China**

Abstract

Affective computing is the calculation about sentiment, sentiment generated and the aspects of affecting the sentiment. However, the different factors often cause the uncertainty of sentiment expression of the users. Today twitter as the information media of real-time and timely has become better sentiment expression vector for users themselves. Therefore, in allusion to the diversity of sentiment form of twitter information to express sentiment, this paper constructs affective computing model, starting from the differences of the constituted form of Twitter based on set pair theory to make analysis and calculation for user sentiment, from the text, emoticon, picture information and other multi-angle to analyze the positive, negative and uncertain emotion of the users for the signal twitter, consolidating the weight of various parts in emotional information, building hierarchical set pair affective computing model for twitter users, to offer more favorable data support for the relevant departments and businesses.

Keywords: *Affective computing, hierarchical model, set pair analysis, the constituted form of Twitter, Set Pair Users.*

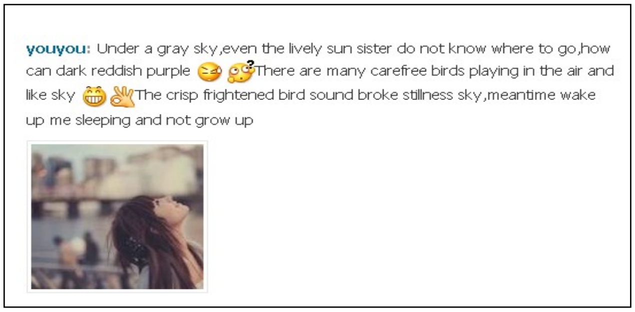
1. Introduction

With the rapid development of the Web2.0, the role of users themselves in the network have undergone a rapid change; initially, from the information recipient only browse the page content to the information producer, publisher and communicator of publishing their own thoughts, viewpoints. The appearance of twitter truly marks the personal Internet age has arrived, by virtue of the autonomous access to information, quickness, breadth and content of dapper has won the favor of people. People on twitter who through their favorite form of text to express their own sentiment, and to share and exchange with the social friends who have the common interests, or one social hot topic of real-time for discussing, which are accustomed to express their own sentiment orientation by means of the social network. In addition, with respect to the relevant departments, they can obtain the "voices" of the people through the sentiment orientation that the user expressed on twitter event. Therefore, they can give back and solve the problem in the shortest time. The affective computing commits to analyze the text to mine the user's sentiment orientation. "Affective computing" [1, 2] was initially proposed in 1995 by the R.Picard professor of MIT Laboratory. He published the monograph - Affective computing in 1997, in the book, his gave the definition that "the affective computing is associated with the sentiment, from sentiment or the calculation that can exert influence on the sentiment". Affective Computing [3-5] is the basis for text orientation analysis. In essence, it is the sentiment analysis and mining for the text, its main purpose is to give the computer ability of understanding and cognition to distinguish the human sentiment and tonality to express sentiment. Currently, in the existing sentiment analysis[6-7] research, most of them are based on the sentiment analysis study of the ordinary text, work[8] divided sentiment into positive sentiment, neutral sentiment, negative sentiment based on the rule method of sentiment dictionary, only by the comparison of the number of sentiment vocabulary of ordinary text to get the user's sentiment orientation, and also mentioned emoticon rule method in the text, the difference of the number of emoticon of different sentiment polarity to get sentiment orientation, one work mentions the latest manifestation on twitter are some pictures and links, but do not start the detailed analysis; The work [9] mentioned a form of expression of the new network word related to the twitter content, but did not expand affective computing analysis according to their characteristics. Now, the sentiment analysis on twitter is only from one particular form of expression to give the sentiment orientation of the users, or just referred to a part, not give the summarized analysis. In fact, with the constantly updated and the development of the network, the picture, new word of network, emoticon, links, etc of representing sentiment is emerging. Therefore, bring the text that be able to express sentiment information integrated analysis is very important, for different events users held views is different, and the level of opinion is not the same, there is uncertainty. Set pair analysis is one method to resolve the uncertainty, the significance of the contact number is that link the number and scope of this number, contacting one specific number and the uncertainty with certainty within its scope to expand trend analysis. This paper integrates the constituted form of twitter that the user can express their own sentiment. Based on set pair analysis theory to construct the model set pair affective computing, so that we can via calculating to get the sentiment orientation for users.

2. The analysis on the constituted form of twitter.

Twitter as a platform which information sharing, dissemination and access to is based on user relationship, through WEB, WAP, and a variety of clients, users can set up personal communities, then update information about 140 words, and greatly promote the dissemination and sharing of the information, its great commercial value began to show and highlight the commercial advantages in the areas of crisis public relations, public opinion speculation and web promotion. With the popularity of the Internet, The constituted form of twitter are constantly update, before there is no word, phrase, sentence as a network of new words appear in network communication, with a circulation speed and fresh are welcomed by the people.

Well, first of all, we give one twitter intercepted on twitter client, the twitter forms of expression including the ordinary text, the new network word, emoticon, and pictures information mentioned in this article. Figure 1 is intercepted from the QQ twitter.



youyou: Under a gray sky,even the lively sun sister do not know where to go,how can dark reddish purple 😆 😵There are many carefree birds playing in the air and like sky 😁 👌The crisp frightened bird sound broke stillness sky,meantime wake up me sleeping and not grow up

Fig. 1 An entry QQ twitter.

Therefore, this paper bring the diversification of the constituted form of twitter information into consideration, set the text messages, emoticon, pictures into together, and design to achieve better computational analysis. We will give a brief description for each component as follows.

2.1 The text information analysis

Now the sentiment analysis method of text mainly has two classification model, on one hand, using the method that combine the sentiment dictionary with rules, the number of positive sentiment words and negative sentiment words contained in the text to make sentiment classification; On other hand, those who use machine learning methods to select some of the characteristics of the text to train and test the set, the main method are classifiers that are Naïve Bayes[10], maximum Entropy, support vector machine. The work [8] contraposes the sentiment analysis of Chinese, based on the sentiment dictionary rules to determine the sentiment polarity of the sentiment word ,it mentioned select the word number of positive affect and negative affect to get sentiment polarity ,the sentiment polarity of the words in certain under any circumstances characterization, but some part of speech is uncertain, the polarity of the sentiment expressed in different contexts is different, so this part uncertain vocabulary polarity analysis is particularly important; The analysis above rarely involved in description and classification for the vocabulary polarity. This paper divides into three categories: positive sentiment vocabulary, negative sentiment vocabulary, uncertainty sentiment vocabulary. As in figure 1, the twitter contains two kinds of text: ordinary text and new network words, the text contain the new network words: dark reddish purple, porridge, drops, and me. The dark reddish purple means like this, the porridge means like. In the calculation of this text, we will convert the new network word into ordinary word to analyze calculation.

2.2 The emoticon analysis

In the paper “Using the to make sentiment classification can be reduce the dependence of the machine learning techniques.”, the foreign scholar Jonathon Read elaborate that one can overcome geographical, subject and time can be as annotations of sentiment, And to achieve the visualization of sentiment states, and this has led to the birth of the emoticon. In the paper, the author used lots of sentiment in Twitter API, thus point out that “smile” usually as a positive text, and “frown” usually show that negative text. And finally using the ways of sentiment to do the sentiment classification, and then reach the accurate rate of 70%. The domestic scholars Xie LiXing has done more hierarchical structure strategy method for the twitter sentiment analysis and applies the emoticon rule method, in view of the expression symbols that sina twitter provides to make classification of the positive and negative emoticon, and extracted the expression symbols of positive and negative in the text with analysis, the difference size of positive and negative expressions symbol as the classification of sentiment polarity. The analysis method of corpus automatic tagging in work, it uses two kinds of sentiment knowledge (emoticon and sentiment words) for large-scale not label samples to conduct automatic tagging and to get the training samples. For the classification of the emoticon also use the way of work [11]: positive sample and negative sample.

For the analysis of a series of emoticons, they all have been given a clear polarity, in fact, the representation of the emoticons are constantly updated, for example, on behalf of positive sentiment : (nod) (excited) (happy), though the affective polarity all represent the positive, the three sentiment belonging to the degree of positive expression is inconsistent, Therefore, in the rest of this analysis are given different weights of the polarity degree of emoticons to reflect such sentiment differences. The cartoon faces in the crawl will display the text meaning "rotation"; the sentiment orientation is different when such category emoticons appear in different contexts. We consider them as uncertain sentiment emoticon to analyze. Based on the above analysis of emoticons, we will give three categories of sentiment symbol, positive emoticon, uncertain emoticon and negative emoticon. In figure 1 included in four emoticon, wherein each emoticon corresponds its homologous text, for example, symbol on behalf of the doubt is an uncertainty; emoticon from the appearance point of view is a sweating state, combined with the text meaning represents a negative sentiment emoticon; growl is pleased state belong to positive emoticon. In the following article, through its text explanation to analyze the generic of the emoticon.

2.3 The picture information analysis

For the picture information from two aspects to consider, on the one hand, the picture is already on the network, users directly upload to enrich the content of their own twitter, a class of pictures accompanied by text attachment, the meaning of the words express the sentiment polarity; Another type is only picture without any text message, abstracting the meaning of the expression from picture content for an sentiment polarity to analyze. The other hand, analyzing the pictures that taken by the user on the ground through the communication tools to express their seen and heard, and this kind of pictures on twitter from users themselves microscopic point. The pictures show a true and more complete world. In discusses the application of twitter on media, the spread of twitter and newspaper illustrates the broad scope of twitter, showing a new trend in the new media environment, and also mentioned the application of expressing their sentiment [12].

For the picture information processing, this paper classify positive picture, uncertain picture, negative picture. For the each picture classification, we adopt the manner consistent with emoticon, which we can also consider the generic weight value of belonging to certain types of sentiment pictures. At the end of figure 1 gives one picture, which can be seen from the analysis of the picture contents, the expression of sentiment orientation in this picture is the positive sentiment pictures.

In view of the above explanation of three parts, the following we utilize the set pair theory to expand affective computing and trend analysis for the form content of twitter.

3. The set pair affective computing and trend analysis on twitter

3.1 The set pair analysis method

Set pair analysis[13-14] was first proposed by Professor Zhao Keqin in 1989 in Baotou that convening the National System Science and Regional Planning Symposium, And this relationship of certainty and uncertainty related influence and mutual restraint, even under certain conditions are met can be mutually converted[15], so the contact can be expressed as:

$$U = A + Bi + Cj \qquad (1)$$

Thereinto A、B、C are non-negative real number, $j = -1, i \in [-1,1]$ depending on the situation to have value in the range, commanding $N = A + B + C$, N is the contact norm, let N divided the two sides of formula (1), then taking $u = \frac{U}{N}, a = \frac{A}{N}, b = \frac{B}{N}, c = \frac{C}{N}$

So

$$u = a + bi + cj \qquad (2)$$

Therefore, the formula (2) is called the contact number expression. Thereinto $a, b, c \in [0,1], a + b + c = 1,\ j = -1, i \in [-1,1]$, a is the same degree, b is the difference degree, c is the confrontation degree, i is the difference coefficient, j is the confrontation coefficient. Thereinto, the trend is the concept that reflecting the size relations order of the same degree (a), the difference degree (b), the confrontation degree (c), and can be divided into the same potential, the balance potential and the anti-potential. In addition, each potential can also be subdivided by the degree of size.

3.2 The set pair affective computing on Twitter

Based on the above analysis on the constituted form of twitter, combined with the characteristics of contents form classification, this paper adopts set pair analysis theory method to calculate from the three aspects: the text, emoticon, pictures information, and the calculation process of each part as follows.

3.21 The set pair affective computing on Twitter

A message that a user issues on twitter is generally constituted by multiples of vocabularies, including general vocabulary and sentiment vocabulary. These vocabularies can be divided into positive sentiment words, negative sentiment words and uncertain sentiment words, according to their parts of speech, then adopt the set pair analysis method to conduct affective computing for the single twitter.

Supposing one twitter TW_k has x_1 positive sentiment vocabularies, x_2 negative sentiment vocabularies, x_3 uncertain sentiment vocabularies, recording as $x_1 + x_2 + x_3 = N_1$, so, the basic model of text set pair affective computing for this twitter, as follows:

$$A_1(TW_k) = \frac{x_1}{N_1} + \frac{x_2}{N_1}i + \frac{x_3}{N_1}j \tag{3}$$

Wherein $j = -1,\quad i \in [-1,1]$ depending on different circumstance to get value, in the formula (3), when $i = -1$, uncertain sentiment vocabulary is converted to the negative sentiment vocabulary, that is, at this time, the number of negative sentiment vocabulary is $x_2 + x_3$, the number of positive vocabulary x_1; when $i = 0$ the twitter does not exist uncertain sentiment vocabulary, the number of positive and negative sentiment vocabulary have no change; when $i = 1$, the uncertainty of the sentiment vocabulary convert into positive sentiment vocabulary, at this time, the number of positive sentiment vocabulary is $x_1 + x_3$, the number of negative sentiment vocabulary is x_2.

In the actual sentiment analysis, the extent of the sentiment orientation of each vocabulary is not identical. We via the weight to improve the above set pair affective computing model.

Supposing one twitter TW_k has x_1 positive sentiment vocabularies, and the weight values is f_1, the Product of vocabulary and its corresponding weights is expressed by a_1 in the contact number; Supposing one twitter TW_k has x_2 negative sentiment vocabularies, and the weight values is f_2, the Product of vocabulary and its corresponding weights is expressed by b_1 in the contact number; Supposing one twitter TW_k has x_3 uncertain sentiment vocabularies, and the weight values is f_3, the product of vocabulary and its corresponding weights is expressed by c_1 in the contact number; Therefore, updating the text set pair affective computing basic model of twitter as follows:

$$\begin{aligned} A_1(TW_k) &= f_1\frac{x_1}{N_1} + f_2\frac{x_2}{N_1}i + f_3\frac{x_3}{N_1}j \\ &= a_1 + b_1 i + c_1 j \end{aligned} \tag{4}$$

Through the sum of the products of the expressions of each word and the corresponding weight to be better reflect the degree of sentiment differences among the different vocabulary, making the analysis results more accurate.

We expand calculation in figure 1 in the text, using equation (4) model to conduct the affective computing:

$$\begin{aligned} A_1(TW_k) &= (\frac{3}{12}\times 0.2 + \frac{1}{12}\times 0.4 + \frac{1}{12}\times 0.6) + (\frac{1}{12}\times 0.2 + \frac{3}{12}\times 0.4)i + (\frac{3}{12}\times 0.4)j \\ &= \frac{2}{15} + \frac{7}{60}i + \frac{1}{10}j \end{aligned}$$

The data obtained from the analysis of the text, we can get $a_1 > b_1 > c_1$, thus the sentiment orientation of the text analysis on twitter is positive.

3.22 The affective computing of the emoticon

Supposing one twitter TW_k has y_1 positive sentiment emoticon, y_2 negative sentiment emoticon, y_3 uncertain sentiment emoticon, recording as $y_1 + y_2 + y_3 = N_2$, so the basic model of emoticon set pair affective computing for this twitter, as follows:

$$A_2(TW_k) = \frac{y_1}{N_2} + \frac{y_2}{N_2}i + \frac{y_3}{N_2}j \tag{5}$$

Wherein $j = -1,\quad i \in [-1,1]$ depending on different circumstance to get value, in the formula (5), when $i = -1$, uncertain sentiment emoticon is converted to the negative sentiment emoticon, that is, at this time, the number of negative sentiment emoticon is $y_2 + y_3$, the number of positive emoticon y_1; when $i = 0$ the twitter does not exist uncertain sentiment emoticon, the number of positive and negative sentiment emoticon have no change; when $i = 1$, the uncertainty of the sentiment emoticon convert into positive sentiment emoticon, at this time, the number of positive sentiment emoticon is $y_1 + y_3$, the number of negative sentiment emoticon is y_2.

For emoticon of strong expressive, and each emoticon belong to each category is different. Therefore, in view of this situation, we have different emoticon calculated to give each a weight value. Giving the updated emoticon connection degree expression is as follows:

$$A_2(TW_k) = w_1\frac{y_1}{N_2} + w_2\frac{y_2}{N_2}i + w_3\frac{y_3}{N_2}j$$
$$= a_2 + b_2 i + c_2 j \qquad (6)$$

In the formula (6), w_1, w_2, w_3 is the corresponding weight value for the each category sentiment emoticon.

3.23 The affective computing of the picture information

This paper adopts the research process that similar with text, emoticon for picture information to conduct affective computing.

Supposing one twitter TW_k has z_1 positive sentiment picture information, z_2 negative sentiment z_3, uncertain sentiment z_3, recording as $z_1 + z_2 + z_3 = N_3$, so the basic model of picture information set pair affective computing for this twitter, as follows:

$$A_3(TW_k) = \frac{z_1}{N_3} + \frac{z_2}{N_3}i + \frac{z_3}{N_3}j \qquad (7)$$

Wherein $j = -1$, $i \in [-1,1]$ depending on different circumstance to get value, in the formula (7), when $i = -1$, uncertain sentiment picture information is converted to the negative sentiment picture information, that is, at this time, the number of negative sentiment picture information is $z_2 + z_3$, the number of positive picture information z_1; when $i = 0$ the twitter does not exist uncertain sentiment picture information, the number of positive and negative sentiment picture information have no change; when $i = 1$, the uncertainty of the sentiment picture information convert into positive sentiment picture information, at this time, the number of positive sentiment picture information is $z_1 + z_3$, the number of negative sentiment picture information is z_2.

Therefore, we give a weight value for the picture information calculating for different pictures. The expression of picture information connection degree is as follows:

$$A_3(TW_k) = g_1\frac{z_1}{N_3} + g_2\frac{z_2}{N_3}i + g_3\frac{z_3}{N_3}j$$
$$= a_3 + b_3 i + c_3 j \qquad (8)$$

In the formula (8), g_1, g_2, g_3 is the corresponding weight value for the each category sentiment picture information.

3.24 The comprehensive affective computing on twitter

Thus we can conclude that the constituted form of twitter mainly including the form of the above three aspects, the proportion of the effect of sentiment expression for each form is not the same, therefore need to give different weight value for each type of the constituted form, respectively is α, β, γ, and the size of three weight value is also depending on the specific circumstances. The specific expression is as follows:

$$A = \alpha A_1 + \beta A_2 + \gamma A_3$$
$$= \alpha(a_1 + b_1 i + c_1 j) + \beta(a_2 + b_2 i + c_2 j) + \lambda(a_3 + b_3 i + c_3 j)$$
$$= (\alpha a_1 + \beta a_2 + \lambda a_3) + (\alpha b_1 + \beta b_2 + \lambda b_3)i + (\alpha c_1 + \beta c_2 + \lambda c_3)j$$
$$= a + bi + cj \qquad (9)$$

The above formula (9) is the analysis calculated combination of the three type constituted form, a is behalf of the positive sentiment part, it includes vocabulary, emoticon, pictures information; b is behalf of the uncertain sentiment part; c is behalf of the negative sentiment part, thereby obtaining comprehensive twitter sentiment formula.

4. Application Example

The twitter of figure 1 contains the text, emoticon, and pictures information as the example to verify the validity of the method, the vocabulary divided into the general vocabulary and sentiment vocabulary, according to the number and weight to summarize the sentiment vocabulary; for emoticon and picture information， under the number and weight to summarize, that are shown in the following Table1:

Consolidating the data in Table I and the weight value of the vocabulary, emoticon, picture information of this twitter to calculate, giving the size is $\beta = 0.4 > \gamma = 0.3 > \alpha = 0.2$ from the expressed result, calculating the size of a, b, c according to the formula (9). Then get the relation is: $a = \frac{68}{300}, b = \frac{19}{300}, c = \frac{24}{300}$, that is $a > c > b$.

Comprehensive analysis by the method of this article, this user expressed sentiment orientation for positive sentiment orientation. We also obviously see that we selected this entry twitter held attitude support by the data that given on table one. Corresponding to the twitter users to express the sentiment state is pleasure, Secondly, we also need to be taken into account for the components of polar uncertain, to some extent these factors are also occupy a certain share ratio on the results of the problem. Through I in the interval of $[-1,1]$ to get the different value to get the different types of sentiment expressions.

.

Table 1: Summarization of Vocabulary, Emoticon, Picture Information

Polarity Weight Range	Positive			Uncertain			Negative		
	v	e	p	v	e	p	v	e	p
$(0-0.2]$	3	1		1					
$(0.2-0.4]$	1		1	3	1		3		
$(0.4-0.6]$	1	1						1	

(In the above table v= vocabulary, e= emoticon, p= picture information)

5. Conclusion and next step

In this paper, to fully consider the constituted form of twitter, which to be updated with the network development. From the angel of sentiment orientation to divide the constituted form of Twitter into text, emoticon, and picture to analyze. And assign weight to solve the magnitude of belonging to a certain type of sentiment polarity; at last, giving the affective computing model of the three constituted form of integration by set pair analysis method to obtain the good results. The next work: (1) According to the above analysis and conduct affective computing trend analysis for the users. (2) Considering the relationship between twitter first posts its thread, affective computing and trend analysis of single twitter extended to multiple twitter.

Acknowledgments

This work is supported by Hebei Province Natural Science Fund Project (No.F2012209019).

References

[1] Pizard R W. Perceptual Computing Technical Report. Affective Computing[R]. Cambridge: MIT Media Lab, 1995, pp. 57-63.

[2] Li Jiayuan. Affective Computing Research Cognitive Dilemma. Journal of Dialectics of Nature, Vol.2, 2010, pp.23-28.

[3] Zhang Chenggong, Liu Peiyu, Zhu Zhenfang, Fang Ming. The sentiment analysis methods based on the polarity dictionary. Shandong University (Natural Science). Vol. 47, No.3, 2010, pp.47-50.

[4] Aparna Trivedi, Apurva Srivastava, Ingita Singh, Karishma Singh and Suneet Kumar Gupta. Literature Survey on Design and Implementation of Processing Model for Polarity Identification on Textual Data of English Language. IJCSI International Journal of Computer Science Issues, Vol. 8, No. 3, 2011, pp.309-312.

[5] Nevin VUNKA JUNGUM, Éric LAURENT. Emotions in Pervasive Computing Environments. IJCSI International Journal of Computer Science Issues, Vol.6, No.1, 2009, pp.8-22.

[6] Xie Lixing, Zhou Ming, Sun Maosong. Based multi-strategy hierarchy of Chinese microblog sentiment analysis and features extraction. The Chinese Information. Vol.26, No.1, 2012, pp. 456-459.

[7] Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Scott D, ed. Proc. of the ACL 2004.Morristown: ACL, 2004, pp.271-278.

[8] Du Weifu. Sentiment dictionary build research of text tendentious analysis technology. Harbin Institute of Technology. 2010.

[9] Akshi Kumar, Teeja Mary Sebastian. Sentiment Analysis on Twitter. IJCSI International Journal of Computer Science Issues. Vol.9, No.3, 2012, pp. 372-378.

[10] Domingos, P, & Pazzani, M. J. On the optimality of the simple Bayesian classifier under zero-one loss [J].Machine Learning.1997, pp.103-130.

[11]Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification[C]//Proceedings of the ACL Student Research Workshop. Association for Computational Linguis-tics, Morristown, NJ, 2005, pp.89-92.

[12]He Xi. Microblog and newspaper picture spread analysis. Business culture (the second half). Vol.5, 2012, pp.236-240.

[13]Zhao Keqin. Set Pair Analysis and its Preliminary Application [M].Science and technology press.Vol.13, No.47, 1994, pp. 67-72.

[14]Zhao Keqin. Uncertainty of description and treatment on set pair. Information and control. Vol.24, No.3, 1995, pp.162-166.

[15]Zhang Chunying, Liang Ruitao, Liu Lu, Wang Jing, 2011,Set pair community mining and situation analysis based on web social network, 2011 International Conference on Advanced in Control Engineering and Information Science, CEIS2011,2011/8/18-2011/8/19,pp.3456-3460,Dali,Yunnam, China(EI).

Chunying Zhang is a PhD in computer application technology of Yanshan University, Master and Bachelor of computer science and technology. She is currently working as a university professor in department of computer science at Hebei United University, and also as the master tutor. She is the members of the experts working group of specialty instruction guidance Committee Sub-Committee of the ministry education college and computer science and technology, a member of the steering committee of computer teaching in Hebei Province; an executive director of the Machine Learning institute in Hebei Province, a director of computer educational research association in Hebei Province, an member of China Computer Federation and other society duties. Her major research directions are: social network analysis, intelligent information processing, and Web mining and other fields. In recent years, she has published more than 50 papers on important domestic, international journals and international conferences, and more than 30 articles are retrieved by the three retrieves.

Jing Wang is doing Master of Science in Applied mathematics from Hebei United University, she has done her bachelor in applied mathematics, and she is currently working in the field of social network analysis.

An Energy Balanced Algorithm of LEACH Protocol in WSN

Chunyao FU[1], Zhifang JIANG[1], Wei WEI[2]and Ang WEI[*3]

[1]**School of Natural Science, Nanjing University of Posts & Telecommunications,
Nanjing 210046, China**

[2]**School of Computer Science and Engineering, XI ' an University of technology,
XI ' an 710048, China**

[*3]**Institute of Advanced Materials, Nanjing University of Posts & Telecommunications,
Nanjing 210046, China.**

Abstract

In wireless sensor networks (WSNs), due to the limitation of nodes' energy, energy efficiency is an important factor should be considered when the protocols are designing. As a typical representative of hierarchical routing protocols, LEACH Protocol plays an important role. In response to the uneven energy distribution that is caused by the randomness of cluster heads forming , this paper proposes a new improved algorithm of LEACH protocol (LEACH-TLCH) which is intended to balance the energy consumption of the entire network and extend the life of the network . The new algorithm is emulated by Matlab simulation platform, the simulation results indicate that both energy efficiency and the lifetime of the network are better than that of LEACH Protocol.

Keywords: *LEACH Protocol; Energy consumption; Network lifetime; Matlab simulation.*

1. Introduction

As a new information acquisition and processing technology, wireless sensor network （WSN）has a wide range of applications in military, environmental monitoring, smart furniture and space exploration and so on [1]. Wireless Sensor Network can be described as an autonomy system consisting of lots of sensor nodes designed to intercommunicate by wireless radio ， and it can collaborate in real time monitoring, perceiving and collecting information of various environmental or monitoring objects and transfer this information to the base station．It does not need a fixed network support，and it has rapid employment, survivability and other characteristics，so it has a good application prospect．

Until now the research on sensor network generally has gone through two stages, the first stage is primarily intended for node, the second one is for network-level issues, the main research works in this stage involve the network layer and MAC layer protocol based on energy optimization, node localization technology, clock synchronization technology and data fusion technology [2]. Study of routing protocols in wireless sensor networks is one of the hot topics at this stage.

LEACH Protocol is the first protocol of hierarchical routings which proposed data fusion, it is of milestone significance in clustering routing protocols. Many hierarchical routing protocols are improved ones based on LEACH protocol [3]. So, when wireless sensor networks gradually go into our lives, it is of great significance to research on LEACH protocol.

2. Brief Introduction to LEACH Protocol

LEACH Protocol is a typical representative of hierarchical routing protocols. It is self adaptive and self-organized. LEACH protocol uses round as unit, each round is made up of cluster set-up stage and steady-state stage, for the purpose of reducing unnecessary energy costs, the steady-state stage must be much longer than the set-up stage. The process of it is shown in Figure 1.

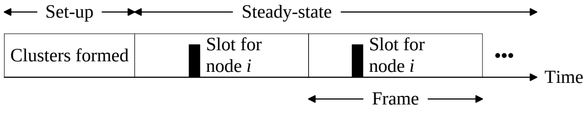


Fig.1 LEACH Protocol process.

At the stage of cluster forming, a node randomly picks a number between 0 to 1, compared this number to the threshold values $t(n)$, if the number is less than $t(n)$, then it become cluster head in this round, else it become common node. Threshold $t(n)$ is determined by the following:

$$\mathbf{t}(\boldsymbol{n}) = \begin{cases} \dfrac{\boldsymbol{p}}{\mathbf{1} - \boldsymbol{p} * (\boldsymbol{r} \bmod \dfrac{\mathbf{1}}{\boldsymbol{p}})} & \boldsymbol{if\ n \in G} \\ \mathbf{0} & \boldsymbol{if\ n \notin G} \end{cases} \tag{1}$$

Where p is the percentage of the cluster head nodes in all nodes, r is the number of the round, G is the collections of the nodes that have not yet been head nodes in the first 1/P rounds. Using this threshold, all nodes will be able to be head nodes after 1/P rounds. The analysis is as follows: Each node becomes a cluster head with probability p when the round begins, the nodes which have been head nodes in this round will not be head nodes in the next 1/P rounds, because the number of the nodes which is capable of head node will gradually reduce, so, for these remain nodes, the probability of being head nodes must be increased. After 1/P-1 round, all nodes which have not been head nodes will be selected as head nodes with probability 1, when 1/P rounds finished, all nodes will return to the same starting line.

When clusters have formed, the nodes start to transmit the inspection data. Cluster heads receive data sent from the other nodes, the received data was sent to the gateway after fused. This is a frame data transmission. In order to reduce unnecessary energy cost, steady stage is composed of multiple frames and the steady stage is much longer than the set-up stage.

3. A new improved algorithm based on LEACH Protocol (LEACH-TLCH)

In LEACH protocol, due to the randomness of clusters forming, the energy of cluster head is very different, so do the distances between cluster heads and base station. Cluster heads are responsible not only for sending data to the base station but also for collecting and fusing the data from common nodes in their own clusters. In the process of data collection and transmission, the energy consumed by data transmission is greater than that of data fusion [4]. If the current energy of a cluster head is less or the distance to base station is much far, then the cluster head will be died quickly because of a heavy energy burden. To address these issues, this article proposes a new improved algorithm on how to balance the energy loads of these cluster heads.

3.1 The idea of improved algorithm

LEACH-TLCH (LEACH Protocol with Two Levels Cluster Head) is an improved one based on LEACH Protocol, the methods of cluster-head selection and clusters forming are same as LEACH protocol. If a cluster head's current energy is less than the average energy, that is $E_{cur}< E_{ave}$,where $E_{ave} = \sum_{1}^{N} E(i)_{cur}$ is the average energy of all nodes in the network, or the distance between the cluster head and base station is longer than the average distance, that is $d > d_{ave}$, where $d_{ave} = \sum_{1}^{N} d_i$ is the average distance of all nodes' distance to base station, then the common node with maximum energy in this cluster will be selected as the secondary cluster head. If $E_{cur}\geq E_{ave}$ and $d \leq d_{ave}$, it is unnecessary to select a secondary cluster head.

In a cluster which has secondary cluster head, the secondary cluster head is responsible for receiving and fusing data collected from the member nodes and sending them to its cluster head, the cluster head is only responsible for transporting data to base station. In a cluster without secondary cluster head，the cluster head is responsible for collecting data from the member node and sending them to base station after the data was fused. It is clear from the first-order energy transfer model (Figure 3) that the energy consumption of data receiving and data fusion are less than that for data transferring [5]，especially for long distance data transferring, so the life of clusters with secondary cluster heads will not be extended a lot so as to bring new energy imbalance of energy consumption of entire network. The network topology of the improved algorithm is shown in Figure 2.

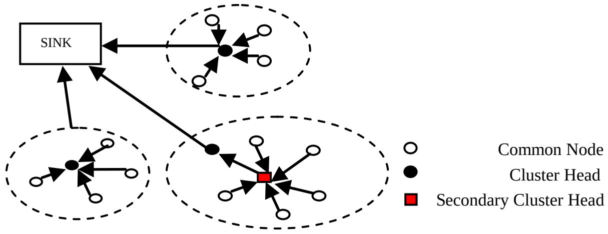


Fig.2 Network topology

3.2 First-order wireless transmission model

This article uses first-order wireless communication model, it is shown in Figure 3.

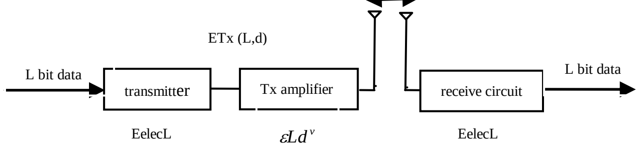


Fig. 3 The wireless communication model

The total energy consumed in Figure 3 is calculated by formula (2) and (3) [6],

$$E_{Tx}(L,d)=\begin{cases} LE_{elec}+L\varepsilon_{fs}d^2, d\le d_0 \\ LE_{elec}+L\varepsilon_{mp}d^4, d>d_0 \end{cases} \tag{2}$$

$$E_{Rx}(L)=LE_{elec} \tag{3}$$

Where E_{elec} represents the energy consumed to transmit or receive 1 bit message; ε_{fs} is the amplification coefficient of free-space signal and ε_{mp} is the multi-path fading signal amplification coefficient, their value depend on the circuit amplifier model; d represents the distance between transmitter and receiver; L is the bit amount of sending information.

3.3 The optimal number of cluster heads

In LEACH Protocol, all nodes are divided into n clusters randomly, if the value of n is too small, each cluster head burdens so heavily that some clusters will die earlier due to energy draining, this will affect the network lifetime; If the value of n is too large, this also results in some unnecessary overhead because clusters need to send broadcast messages to all nodes. Suppose N nodes are randomly distributed within the square area of the edge length M, assuming that the base station locates in the centre of region, and the distance of each node or cluster-head to the base station is less than or equal to d_0,where $d_0=\sqrt{\frac{\varepsilon_{fs}}{\varepsilon_{mp}}}$, we know by references [6] that the optimal number of cluster heads should be

$$n_{opt}=\sqrt{\frac{N}{2\pi}}\frac{M}{d_{toBS}} \tag{4}$$

If some nodes' distance to base station is greater than d_0, we can also get formula (5) by the same method which was used in references [6] and [7].

$$n_{opt}=\sqrt{\frac{N}{2\pi}}\sqrt{\frac{\varepsilon_{fs}}{\varepsilon_{mp}}}\frac{M}{d_{toBS}} \tag{5}$$

So the optimal probability for nodes to become cluster heads is

$$p_{opt}=\frac{n_{opt}}{N} \tag{6}$$

By the formula (4) and (5) we know that the optimal number of cluster heads only relates to the number of network nodes N, the regional side length M, as well as the location of the base station. We can set these parameters in the network initialization. In this article, the optimal probability for nodes to become cluster heads is chosen as 7% according to formula (6) and the parameters we have set.

3.4 The description of improved algorithm

The parameters need to be used in description of algorithm are as following: Threshold value, as shown in formula (1)； Average energy of all nodes is $E_{ave}=\sum_{1}^{N}E(i)$； Average distance between nodes and base station is $d_{ave}=\sum_{1}^{N}d(i)$.

- The stage of cluster forming

First, a node choose a number between 0 to 1, if the number is less than $T(n)$, then the node becomes cluster head, else, normal nodes it becomes. Cluster heads broadcast their own information to other nodes, the other nodes will listen to the broadcasting messages. All normal nodes determine which cluster they should join in this round based on the strength of the signal they received. After determining which cluster they should belong, CSMA Protocol will be used to send a confirmation message to their cluster heads. At this point, the clusters forming stage is finished.

- The selecting of secondary cluster head

Each cluster head decides whether to set a secondary cluster head according to the current energy itself and the distance to the base station, if $E(i)<E_{ave}$ or $d(i)>d_{ave}$, then these kinds of cluster heads should choose the node with maximum energy as secondary cluster head in its cluster, otherwise, the secondary cluster head is not required.

- To create a transport schedule

All clusters are divided into two categories, in clusters with secondary cluster heads, the secondary cluster head broadcasts message of being secondary cluster head to the other ordinary nodes and builds a schedule (uses TDMA access channel, a time slot is assigned to each node), informs the schedule to the other nodes. In clusters without secondary cluster head, the cluster heads distribute sending time slot to the others after get the join information of normal nodes. The stable stage begins when each node have gotten its sending time slot.

- Data transferring

When clusters have formed and the TDMA schedule is determined, the nodes start to transfer the monitoring data. The secondary cluster heads receive data from the other nodes and fuse these data, these fused data was sent to the cluster heads, then cluster heads send these data to base station by single-hop method. In those clusters without secondary cluster head, the cluster heads receive the information from other nodes, fuse them and send them to base station.

4. Simulation of improved algorithm

This article uses Matlab7.0 as simulation platform to emulate LEACH protocol and the improved protocol (LEACH-TLCH), the improved algorithm aims at balancing the total energy consumption of nodes and extending the network's survival time. So we measure the improved protocol performance from two aspects: the lifetime and the total energy consumption of the network. The lifetime of network means the time from the beginning of simulation to the time when the last node died. As the energy of WSN is limited, so the energy consumption in its lifetime is a meaningful indicator to measure the performance of it.

4.1 Simulation parameters

Simulation scenarios in this article are:
1. Sensor nodes are randomly distributed in a square region;
2. Sensor nodes are homogeneous and have a unique ID number throughout the network, nodes energy is limited. The node’s location is fixed after deployed;
3. The base station is in the center of region with fixed-location;
4. Nodes communicate with base station via single-hop or multi-hop;
5. The wireless transmitter power is adjustable.
Specific parameters are shown in table 1.

Table1: Simulation environment parameters

parameters		parameters	
area	200*200	Packet size	4000bits
Nodes number	200	Eelec	50nJ/bit
Initial energy	0.5J	ε_{fs}	10pJ/bit/m2
CH proportion	p=7%	ε_{mp}	0.0013pJ/bit/m4
BS location	(100,100)	EDA	5nJ/bit

4.2 Analysis of simulation results

200 nodes randomly distribute within the square area of the 200m*200m, the base station is located in the centre of the region, the base station coordinates is (100,100). It can be seen from the figure 4 that the nodes’ distribution are not very evenly.

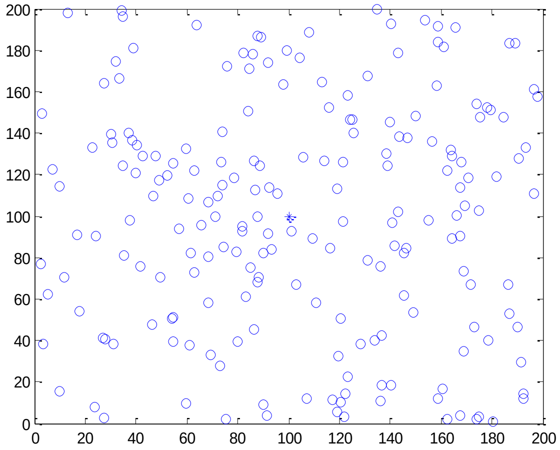


Fig. 4 Randomly distributed nodes

4.2.1 The network lifetime

The network lifetime in this article is defined as the time from the beginning of the simulation to the time when the last node died. In WSN, the network life is divided into stable and unstable period [6]. Stable period usually means the time from the beginning of the simulation to the time when the first node dies, the unstable period refers to the time from the death of first node to the end of simulation.

If it happened that some nodes begin to die, the network operation may become unstable and unreliable data transferring will occur. Therefore, the longer the stable period is, the better the performance of the network. In LEACH Protocol, cluster heads are responsible not only for communicating with the base station, but for the data fusing. Randomly distributing the nodes and randomly selecting the cluster heads causes some cluster heads die earlier because of the low energy or the long distance to base station. Secondary cluster heads are set for these clusters to be responsible for the communication with common nodes and data fusing, this balances the energy load of cluster heads and avoids premature death of these cluster heads, so the stable period of network lifetime will be prolonged.

Figure 5 is network lifetime in simulation, simulation results indicates that the network lifetime of the improved protocol and LEACH Protocol are about the same, the first node died in LEACH Protocol in round 561, the first node died in the improved Protocol in round 857. When 90% nodes died, the network reliability is extremely reduced and the running is almost meaningless. We may as well to define the time from the simulation beginning to the time 90% nodes died as effective lifecycle, analyzing from figure 5, we know that the effective lifecycle of the improved algorithm is longer 9% than that of LEACH protocol. The percentage of stable period of lifecycle in LEACH Protocol is 28%, the one in the improved protocol is 43%, The percentage of stable period of lifecycle in improved algorithm increases 15%. This indicates that the running performance of improved protocol is much better than that of LEACH Protocol. The analysis of simulation results is consistent with the theoretical analysis.

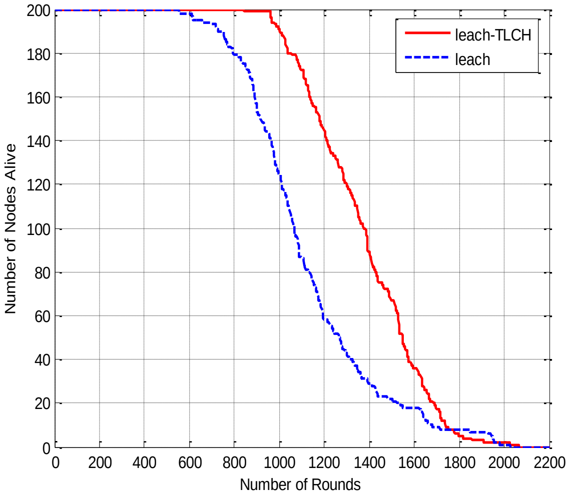


Fig. 5 The network lifetime

4.2.2 The total energy consumption

Figure 6 is the energy consumption curve. Improved algorithm reduced the energy consumption of few cluster heads which has low energy or is far away to base station by setting secondary cluster heads reasonably. This balanced the energy consumption of the whole networks, extended the lifetime of cluster heads which may die earlier and optimized the performance of the network thereby reduced the total energy consumption of the effective lifecycle.

From the analysis of Figure 6, we know that in the whole running of the network, the energy consumption of improved algorithm is much lower than that of LEACH Protocol at the same round of simulation. These results are consistent with the design purposes of improved algorithm.

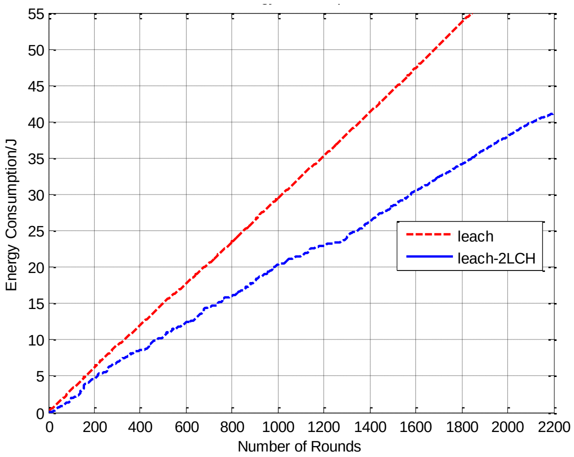


Fig.6 The total energy consumption

5. Conclusions

Electing cluster head randomly in LEACH protocol causes that the current energy of some cluster heads are less or their distances to base station are far, because of the heavy energy burden, these cluster heads will soon die. For this issue, this article proposed a new improved algorithm of LEACH protocol which is aim at balancing energy consumption of the whole network and extending the network lifetime by balancing the energy consumption of these cluster heads. The new improved algorithm is emulated by Matlab platform, the simulation results indicate that the energy efficiency and the lifetime of network are both better than that of LEACH Protocol.

References

[1] Akyildiz LF, Su W, Sankarasubramaniam Y, Cayirci E. A survey on sensor networks. IEEE Communications Magazine, 2002, 40(8): 102~114.Vol.25, No.4: 114-124.

[2] R.A.Roseline and Dr.P.Sumathi, Energy Efficient Routing Protocol and Algorithms for Wireless Sensor Networks-A Survey. Global Journal of Computer Science and Technology, vol.11, December 2011.

[3] Ian F. Akyildiz, Weilian Su, Yogesh Sankarasubramaniam, et al. Wireless Sensor Network: A survey [J]. Computer Networks, 2003, 38(4): 393-422.

[4] Heinzelman W, Chandrakasan A, Balakrishnan H. Energy Efficient Communication Protocol for Wireless Microsensor Networks. In Proceedings of the 33rd Hawaii International Conference on System Sciences. Maui: IEEE Computer Society, 2000, Vol.2: 3005-3014.

[5] Cui Li Ju, Hailing, Miao Yong, Li Tianpu, Liu Wei and Zhao Ze,Overview of Wireless Sensor Networks[J];Journal of Computer Research and Development;2005-01

[6] Smaragdakis G. Matta I. Bestavros A. A Stable Election Protocol for Clustered Heterogeneous Wireless Sensor

Networks, Proceedings of the 2nd International Workshop on SANPA 2004,Massachusetts, U.S, 2004:1-11

[7] WANG Jin-wei, 2,SUN Hua-zhi, SUN De-bing, Research on the Number of Optimal Cluster Heads of Wireless Sensor Networks Based on Energy Consumption. Journal of Computer Research and Development, 2008,03.

[8] Estrin D, Govindan R, Heidemann J, Kumar S. Next century challenges: Scalable coordinate in sensor network. In Proceedings of the 5th ACM/IEEE International Conference on Mobile Computing and Networking. Seattle: IEEE Computer Society, 1999, 263~270.

Chunyao FU received her BS degree in Mathematics from Shaanxi Normal University and MS degree in computer science from Nanjing University of Posts and Telecommunications. She is now an assistant professor in College of Science in Nanjing University of Posts and Telecommunications. Her main research work is related to computer technology and mathematics.

Zhifang JIANG is an associate professor in College of Science in Nanjing University of Posts and Telecommunications. She received her BS degree in Mathematics from East China Normal University and MS degree in applied mathematics from Nanjing University. Her main research is algebra.

Wei WEI received the PhD and MS degrees from Xi’an Jiaotong University in 2011 and 2005, respectively. He is now an assistant professor at Xi’an University of Technology. His academic interests in the following areas: Wireless Networks and Wireless Sensor Networks Application, Mobile Computing, Wireless Network Security, Image Processing, Distributed Computing.

Ang WEI received the PhD degree from Fudan University in 2006. He is an associate professor at Nanjing University of Posts and Telecommunications. His academic interests are sensor device, nano materials and wireless sensor networks.

A parameter optimized approach for improving credit card fraud detection

A.Prakash [1], Dr.C.Chandrasekar [2]

[1] Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India

[2] Department of Computer Science, Periyar University, Salem, Tamil Nadu, India

Abstract

The usage of credit cards has highly increased due to high-speed innovation in the electronic commerce technology. Since credit card turns out to be the majority well-liked manner of payment for mutually online as well as habitual purchase, cases of fraud correlated through it are as well increasing. In normal Hidden Markov Model the problem of cannot find an optimal state sequence for the underlying Markov process also this observed sequence cannot be viewed as training a model to best fit the observed data. In this research, the main aim is to model the sequence of observations in credit card transaction processing using an Advanced Hidden Markov Model (AHMM) and show how it can be utilized for the exposure of frauds. In this process an AHMM is initially trained with the regular manners of a cardholder. If an incoming credit card transaction is not recognized by the trained AHMM with adequately high probability, it is believed to be fraudulent. This proposed work desire to regulate the model parameters to best fit the observations. The ranges of the matrices (N and M) are fixed but the elements of A, B and π are to be decided, focus to the rank stochastic condition. The information that can efficiently re-estimate the model itself is one of the more incredible features of HMMs this referred here as AHMM.

Key Words: *Hidden Markov Model (HMM), Advanced Hidden Markov Model (AHMM), Hill Climb, and credit card fraud detection*

1. Introduction

An unauthorized account movement by a person for whom the account was not be set to can be referred as credit card fraud. Preparedly, this is an event for which action can be taken to discontinue the misuse in steps forward and integrate risk executive applies to defend alongside comparable acts in the future. In straightforward expressions, Credit Card Fraud is described as when an individual exploits another individual's credit card on behalf of personal causes though the proprietor of the card and the card issuer are not conscious of the information that the card is being used. In addition to the persons using the card has not at all having the association with the cardholder or the issuer and has no purpose of making the repayments for the acquire they done. The anticipate user behavior in economic systems can be utilized in many situations. Forecasting client relocation, public associations can accumulate a lot of wealth and other assets. One of the most motivating pastures of forecast is the fraud of credit stripes, especially credit card expenditure. Positively, all transactions deals with financial records of known abuse are not authoritative. However, there are transactions which are officially suitable, but knowledgeable people can advise that these transactions are probably misused, caused by stolen cards or fake merchants. So, the assignment is to avoid a fraud by a credit card transaction previous to it is known as “illegitimate”. By means of growing number of transactions people can no longer manage all of them. As a solution, one might hold the experience of the experts and put it interested in an expert system. This habitual approach has the disadvantage that the expert's knowledge, yet as it can be mined unambiguously, alters quickly with novel manners of prepared attacks and models of credit card fraud. So as to keep track with this, no predefined fraud models but routine learning algorithms are needed.

HMM-based applications are ordinary in an assortment of areas such as speech recognition, bioinformatics, and genomics. Ourston et al have projected the application of HMM in identifying multistage credit card attacks. Hoang et al projected a new method for abnormality exposure using HMM. The main idea is to construct a multilayer model of transaction behaviors based on HMMs and specifying methods for anomaly detection. In recent days, Joshi and Phoba have examined the capabilities of HMM in credit card fraud detection. Cho and Park recommended an HMM-based credit card fraud detection system that advances the time to model and routine by allowing for only the right transition streams based on the province knowledge of assaults. Lane has examined HMM to model human behavior. On one occasion human behavior is properly formed, any sensed departure is a reason for concern because an attacker is not predictable to have a behavior similar to the genuine user. For this reason, in this research work the Advanced Hidden Markov Model is formed to find the credit card fraud detection. In HMM one more drawback is that for the dynamic programming approach the optimal observation sequence would not be found. With this the best fit point should modeled. In AHMM the drawbacks will be resolved with α and β value. The contribution of the works as follows:

1. In training phase obtain the card holder profile and calculate the probability for each transaction.

2. Using the AHMM creates the observation model with best fit observation states and regulates the model parameters (α and β) to best fit the observations.

3. In testing phase the detection of fraud is obtained If both probability value from multiple observation are same it will be a normal customer else there will be fraud signal will be provided.

The remainder of this paper is as follows: The related work is discussed in section2. The hidden markov model is explained in section 3. In section 4 the Advanced Hidden Markov Model is explained. The section 5 is deal with experimental results and discussion. In section 6 the conclusion of the paper is described.

2. Related Work

A lot of research notice and a number of techniques have developed to the Credit card fraud detection, with unique prominence on data mining and neural networks, have been proposed. Recently, Syeda et al [4] have employed parallel granular neural networks (PGNNs) for civilizing the rapidity of data mining and knowledge discovery process in credit card fraud detection. Aleskerov et al [7] developed CARDWATCH, a database mining system employed for credit card fraud detection which is based on a neural learning module, offers a crossing point to the variety of commercial databases. Ghosh and Reilly [2] have proposed credit card fraud detection with a neural network which is trained on a large illustration of labeled credit card account transactions. These transactions enclose instance fraud cases due to gone cards, stolen cards, purpose fraud, forged fraud, mail-order fraud and non-received issue fraud. Where a drawback is that a complete system has been implemented for this purpose which is time consuming. Stolfo et al [5] recommend a credit card fraud detection system (FDS) by means of meta-learning techniques to discover models of fraudulent credit card transactions.

A general strategy that provides a means for combining and integrating a number of separately built classifiers or models is called as Metalearning. This will be trained correlation of the predictions of the base classifier. The equivalent collection has also labored for fraud and intrusion detection on a cost-based model [6]. Kim and Kim [8] have recognized twisted distribution of data and mix of legal and fraudulent transactions as the two main motivations for the complexity of credit card fraud detection. Supported on this observation, they utilize fraud density of real transaction data as a confidence value and produce the weighted fraud score to diminish the number of misdetections. Fan et al [9] suggested the application of distributed data mining in credit card fraud detection. Brause et al [10] have extended an approach that entails advanced data mining techniques and neural network algorithms to attain high fraud coverage. Chiu and Tsai [11] have proposed web services and data mining techniques to initiate a collaborative scheme for fraud detection in the banking industry.

By means of this system, participating banks partition knowledge about the fraud patterns in a heterogeneous and distributed environment. Phua et al [12] have done a research based a widespread survey of existing data mining supported fraud detection systems and published an inclusive information. Prodromidis and Stolfo [13] exploited an agent based move toward with distributed learning for detecting frauds in credit card transactions. For achieving higher accuracy it is supported on artificial intelligence and inductive learning algorithms and Meta learning methods is combined. Phua et al [16] proposed the use of Meta classifier similar to [5] in fraud detection troubles. They believe naïve Bayesian, C4.5 and Back Propagation neural networks as the base classifiers. Vatsa et al [17] have recently planned a game theoretic approach to credit card fraud detection. They model the communication among an attacker and a fraud detection system as a multi-stage game between two players, each trying to take full advantage of his bribe. The difficulty with most of the above-mentioned approaches is that they want labeled data for both authentic as well as fraudulent transactions to train the classifiers. Receiving real world fraud statistics is one of the major harms connected with credit card fraud detection. In addition, these approaches cannot discover new categories of frauds for which branded data is not accessible.

3. Credit Card Fraud Detection Using Hmm

The credit card fraud detection system is based on Hidden Markov Model, which does not require fraud signatures and still it is capable to perceive frauds just by bearing in mind a cardholder's spending habit. The specifics of purchased items in single transactions are generally unidentified to any Credit card Fraud Detection System organization either at the bank that issues credit cards to the cardholders or at the commercial site where goods is going to be obtained. As business processing of credit card fraud detection system runs on a credit card issuing bank site or merchant site. Every arriving transaction is submitted to the fraud detection system for verification intention. The fraud detection system recognize the card details such as credit card number, cvv number, card type, expiry date and the amount of items acquire to validate, whether the transaction is genuine or not.

The accomplishment techniques of Hidden Markov Model in order to notice fraud transaction through credit cards, it generate clusters of training set and identify the spending profile of cardholder. In that process the number of items purchased by customers, types of items that are bought in a particular transaction deliberates on the amount of item acquired and use for further processing that are not known to the Fraud Detection system completely. It supplies higher amount of dissimilar data transactions in form of clusters depending on transaction amount which will be moreover in low, medium or high value assortments. It tries to discover out any discrepancy in the transaction based on the spending behavioral profile of the cardholder, shipping address, and billing address and so on. Based on the expenditure behavioral profile of card holder the probabilities of initial set have been selected and bring together a series for additional processing. If the fraud detection system generates sure that the transaction to be of fake, it raises an alarm and the issuing bank refuses the transaction.

For the protection purpose, the refuge information module will get the information features and its store's in database. To identify the safety measures information if the card missing then the security information module structure arises. The security form has a number of safety questions like account number, date of birth, mother name, other personal question and their answer, etc. where the abuser has to respond it correctly to move to the transaction division in which all those information must be known by the card holder only and can continue only by the card holder. It has informational confidentiality and informational self strength of mind that are tackled consistently by the novelty giving people and entities a trusted means to user, protected, search, process, and exchange personal and/or secret information.

The system and tools for pre-authorizing commerce offered that a relations tool to a trader and a credit card proprietor. By communicating to a credit card number, card type with expiry date and storing it into database, a unique portion of information that describes a fastidious transaction to be complete by a trustworthy user of the credit card at a later occasion the cardholder will be initiating a credit card transaction procedure. The particulars are conventional in the type of system data in the database only when if a correct individual recognition code is used with the statement the cardholder can precede with further steps with the credit card. Because the transaction is pre-authorized, the merchant does not require observing or transmitting an accurate individual recognition code.

1. The number of states in the model is N. The set of states is $S = \{S_1, S_2, \ldots S_N\}$, where $\mathrm{S_i}$, $i = 1, 2, \ldots, N$ is an individual state. The state at time instant t is denoted by $\mathrm{q_t}$.

2. The number of distinct observation symbols per state is M. The set of symbols is $V = \{V_1, V_2, \ldots V_3\}$, where V_i, $i = 1; 2; \ldots; M$ is an individual symbol.

3. The state transition probability matrix $A = [a_{ij}]$ where $a_{ij} = P(q_t + 1 = S_j \,|\, q_t = S_i)$; $1 \leq i \leq N$; $1 \leq j \leq N$; $t = 1$; 2; $\ldots$: where $a_{ij} > 0$ for all i, j. Also, $\sum_{j=1}^{N} a_{ij} = 1{,}1 \leq i \leq N$.

4. The observation symbol probability matrix $B = [b_j(k)]$, where $b_j(k) = P(V_k \,|\, S_j){,}1 \,| \leq j \leq N, 1 \leq k \leq M$ and $\sum_{k=1}^{M} b_j\,(k) = 1{,}1 \leq j \leq N$

5. The initial state probability vector $\pi = [(\pi_i)]$, where $\pi_i = P(q_1 = S_i)$,$1 \leq j \leq N$, such that $\sum_{k=1}^{M} \pi_i = 1$

6. The observation sequence $O = O_1, O_2, O_3 \ldots O_R$, where each observation $\mathrm{O_t}$ is one of the symbols from V, and R is the number of observations in the sequence.

3.1 HMM Model for Credit Card Transaction Processing

First begin through deciding the observation symbols in our model which is to record the credit card transaction processing function in terms of an HMM. Then quantize the acquisition values x into M worth ranges V_1, $V_2 \ldots V_M$ structuring the observation symbols at the issuing depository. The concrete outlay range for every symbol is configurable based on the expenditure routine of individual credit card holders. These worth ranges can be found dynamically through applying a clustering method on the values of each cardholder's transactions. Let assume V_k, $k = 1, 2, \ldots M$ to stand for both the observation symbol as well as the equivalent charge assortment. A credit cardholder constructs diverse types of purchases of unlike amounts more than a period of time. Single prospect is to believe the sequence of transaction amounts and look for divergences in them.

On the other hand, the sequence of kinds of purchase is additional constant contrasted to the series of transaction quantities. The motive here is that, a cardholder precedes purchases depending on his require for procuring diverse types of items greater than a period of time. Consecutively, produces a series of transaction quantities. The kinds of each acquire are linked to the row of business of the equivalent trade. The kind of purchase of the cardholder is hidden from the FDS. The position of all probable categories of purchase and consistently, the position of every one potential lines of commerce of merchants structures the position of concealed states of the HMM. The line of business of the commercial is identified to the acquiring bank which should be noted at this stage that, since this information is furnished at the time of registration of a merchant. As well, a number of merchants might be trade in various types of merchandise. Such kinds of line of business are judged as Miscellaneous and there is no need to determine the authentic types of items purchased in these transactions.

A few assumptions as regards accessibility of this information with the issuing depository and therefore with the FDS are not matter-of-fact hence, would not have been suitable. In the consequences part shows the cause of choice of the number of states on the method performance. Subsequent to deciding the state and symbol illustrations, after that have to find out the probability matrices A, B and π thus the representation of the HMM is inclusive. These three model parameters are found in a training phase. Hence, they should be chosen carefully through preliminary selection of parameters influences the performance of the algorithm. A method based on Hidden Markov Models (HMMs) is a stochastic method, which can be very useful for some applications involves the making of auditory models of program that build use of temporal information. The HMMs can model a lesser unit of the statement. The HMMs can be analysed as fixed state machines, wherever every unit of time, a state transition happens, and all state produces an auditory vector with a connected likelihood density function. So as to is, in every state, a GMM (Gaussian mixture model) is second-hand to exemplify an auditory vector experiential.

4. AHMM For Credit Card Fraud Detection

Here to alter the model parameters to best fit the observations. The ranges of the matrices (N and M) are fixed however the elements of A, B and π are to be determined, focus to the strip stochastic condition. The actuality that can professionally re-estimate the model itself is one of the more astonishing aspects of HMMs. Let assume $\leftthreetimes = (A, B, \pi)$ be a given model and series of observations $O = (O_0, O_1, \dots O_{T-1})$. For $t = 0,1, \dots . T-2$ and $i, j \in \{0,1, \dots, N-1\}$, define "$di - gamma$" as

$$\gamma_t(i,j) = P(x_t = q_i, x_{t+1} = q_j | 0, \leftthreetimes)$$

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\leftthreetimes)}$$

Table 1: Notations

SYMBOL	REPRESENTATION
T	Observation sequence length
N	Number of states in the model
M	Number f observation symbols
O	Observation sequence $(O_0, O_1, \dots O_{T-1})$
Q	Markov process fo distinct states $\{q_o, q_1 \dots q_{N-1}\}$
V	Set of possible observations $\{0,1 \dots . M - 1\}$
A	Probability for each state transition
π	Probability matrix of observation sequence

Then $\gamma_t(i,j)$ is the probability of being in state q_i at time t and transiting to state q_j at time $t + 1$. The di-gamma will be formed with the terms taken as α , β, A and B as:

$$\gamma_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O|\leftthreetimes)}$$

In this we should re-estimate this with the parameter $\beta_t(i)$ which measures the relevant probability after time t

$\beta_t(i) = \frac{\gamma_t(i) P(O|\leftthreetimes)}{\alpha_t(i)}$ which is represented also as:

$$\beta_t(i) = \sum_{j=0}^{N-1} a_{ij} b_j(O_{t+1})\, \beta_{t+1}(j)$$

Where $\beta_t(i) = P(O_{t+1}, O_{t+2} \dots O_{T-1} | x_t = q_i, \leftthreetimes)$

Denote the $\beta_t(i,j) = P(x_t = q_i, x_{t+1} = q_j | O_{t+1}, \leftthreetimes)$,define the $di - Betas$ as

$$\beta_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \gamma_{t+1}(j)}{P(O_{t+1}|\leftthreetimes)}$$

Where $\beta_{t+1}(j) = \frac{\gamma_t(i,j) P(O|\leftthreetimes)}{\alpha_t(i) a_{ij} b_j(O_{t+1})}$. The $P(O|\leftthreetimes)$ is obtained by summing $\alpha_{T-1}(i)$ over i. From the defenition of $\beta_t(i)$it follows the most likely state at time t is the state q_i for which $\beta_t(i)$ is maximum,where the maximum is taken over the index i.

$\beta_t(i)$ and $\beta_t(i,j)$ are related by

$$\beta_t(i) = \sum_{j=0}^{N-1} \beta_t(i,j)$$

Given with the β and di- Betas verify the model $\leftthreetimes = (A, B, \pi)$ can be re-estimated as follows:

1. For $i = 0,1, \dots . N-1$
2. For $i = 0,1, \dots . N-1$ and $j = 0,1, \dots . N-1$ compute

$$a_{ij} = \frac{\sum_{t=0}^{T-2} \beta_t(i,j)}{\sum_{t=0}^{T-2} \beta_t(i)}$$

The numerator of re-estimated a_{ij} can be observed to give the supposed number of transitions from state q_i to state q_j and the denominator denotes the expected number of transition from the state q_i to any state. Then the ratio is the probability of transiting as of state to q_i state q_j, which is the desired value of a_{ij}.

3. For $j = 0,1, \dots . N-1$ and $k = 0,1, \dots . M-1$ compute

$$b_j(k) = \frac{\sum_{t \in \{0,1 \dots T-2\} O_t = k} \beta_t(i)}{\sum_{t=0}^{T-2} \beta_t(i)}$$

The numerator of the re-estimated $b_j(k)$ is the anticipated number of times the model is in state q_j with observation k, at the same time as the denominator is the estimated number of times the model is in state q_j. The ratio is the probability of observing symbol k, given that the model is in state q_j, which is the desired value of $b_j(k)$.

Re-estimation is an iterative process. Foremost, we initialize $\leftthreetimes = (A, B, \pi)$ through a best guess or, if no logical guess is obtainable, choose with arbitrary values such that $\pi_i \approx 1/N$and $a_{ij} \approx 1/N$ and $b_j(k) \approx 1/M$. It's vital that A, B and π be randomized, because precisely consistent ideals will consequence in a confined maximum from which the model cannot Hill climb. As constantly, π, A and B must be row stochastic. The AHMM process can be summarized as follows.

1. Initialize the model, $\leftthreetimes = (A, B, \pi)$

2. Evaluate $\alpha_t(i), \gamma_t(i),\ \beta_t(i)$and $\beta_t(i, j)$

3. Re-estimate the model$\leftthreetimes = (A, B, \pi)$.

4. If $P(O|\leftthreetimes)$increases, goto 2.

Certainly, it may be enviable to end if $P(O|\leftthreetimes)$does not increase by at any rate various predestined threshold and/or to locate a maximum amount of iterations.

6. Experimental Results And Discussion

5.1 Precision accuracy

This graph shows the precision rate of existing and proposed system based on two parameters of precision and the number of Dataset. From the graph we can see that, when the number of number of Dataset is advanced the precision also developed in proposed system but when the number of number of Dataset is improved the precision is reduced somewhat in existing system than the proposed system. From this graph we can say that the precision of proposed system is increased which will be the best one. The values are given in Table 1:

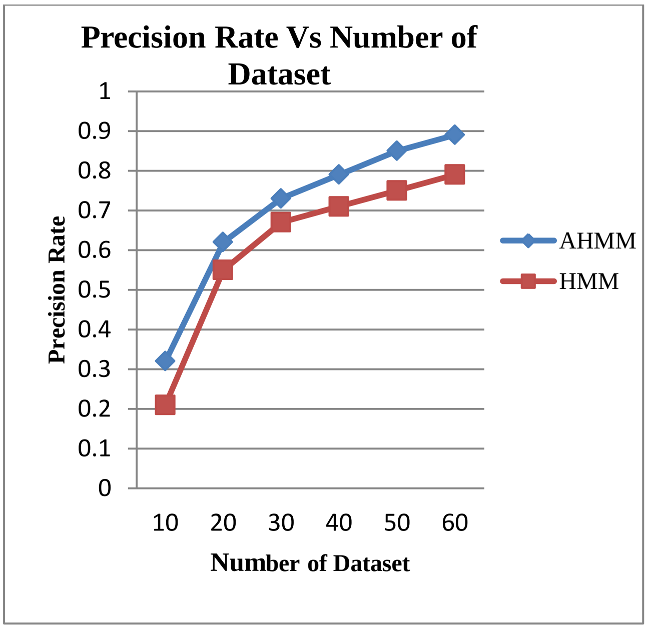


Fig 1: Precision vs. Number of Dataset

Table 2: Precision vs. Number of Dataset

SNO	Number of Dataset	AHMM	HMM
1	10	0.32	0.21
2	20	0.62	0.55
3	30	0.73	0.67
4	40	0.79	0.71
5	50	0.85	0.75
6	60	0.89	0.79

In this graph we have chosen two parameters called number of Dataset and precision which is help to analyze the existing system and proposed systems. The precision parameter will be the Y axis and the number of dataset parameter will be the X axis. The blue line represents the existing system and the red line represents the proposed system. From this graph we see the precision of the proposed system is higher than the existing system. Through this we can conclude that the proposed system has the effective precision rate.

5.2 Recall vs. Number of Dataset

This graph shows the recall rate of existing and proposed system based on two parameters of recall and number of Dataset. From the graph we can see that, when the number of number of Dataset is improved the recall rate also improved in proposed system but when number of number of Dataset is improved the recall rate is reduced in existing system than the proposed system. From this graph we can say that the recall rate of proposed system is increased which will be the best one. The values of this recall rate are given below:

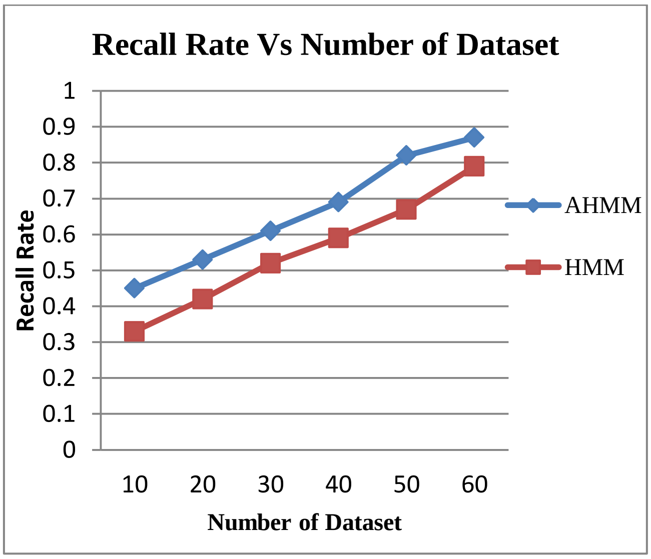


Fig 2: Recall vs. Number of Dataset

Table 3: Recall vs. Number of Dataset

SNO	Number of Dataset	AHMM	HMM
1	10	0.87	0.78
2	20	0.82	0.72
3	30	0.76	0.63
4	40	0.64	0.54
5	50	0.56	0.45
6	60	0.46	0.34

In this graph we have chosen two parameters called number of Dataset and recall which is help to analyze the existing system and proposed systems on the basis of recall. In X axis the Number of dataset parameter has been taken and in Y axis recall parameter has been taken. From this graph we see the recal rate of the proposed system is in peak than the existing system. Through this we can conclude that the proposed system has the effective recall.

5.3 Fmeasure vs. Number of Dataset

This graph shows the Fmeasure rate of existing and proposed system based on two parameters of Fmeasure and number of Dataset. From the graph we can see that, when the number of number of Dataset is improved the Fmeasure rate also improved in proposed system but when the number of number of Dataset is improved the Fmeasure rate is reduced in existing system than the proposed system. From this graph we can say that the Fmeasure rate of proposed system is increased which will be the best one. The values of this Fmeasure rate are given below:

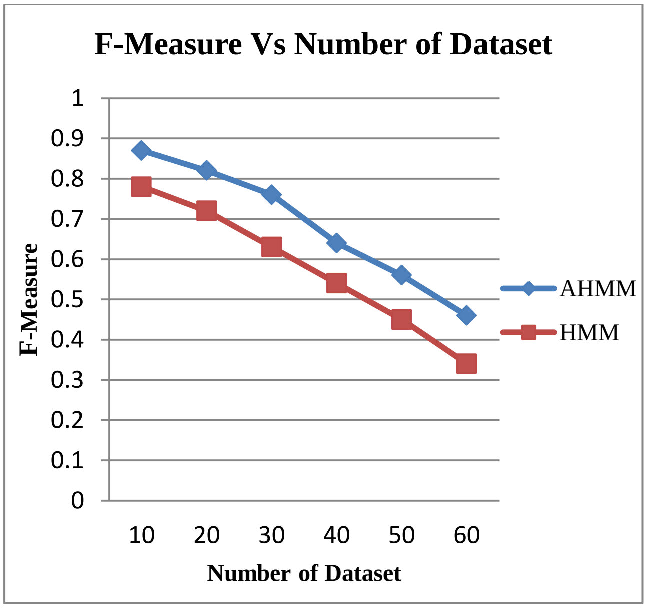


Fig 6: Fmeasure vs. Number of Dataset

Table 3: Fmeasure vs. Number of Dataset

SNO	Number of Dataset	AHMM	HMM
1	10	0.87	0.78
2	20	0.82	0.72
3	30	0.76	0.63
4	40	0.64	0.54
5	50	0.56	0.45
6	60	0.46	0.34

In this graph we have chosen two parameters called number of Dataset and recal which is help to analyze the existing system and proposed systems on the basis of Fmeasure. In X axis the Number of dataset parameter has been taken and in Y axis Fmeasure parameter has been taken. From this graph we see the Fmeasure of the proposed system is in peak than the existing system. Through this we can conclude that the proposed system has the effective Fmeasure.

6. Conclusion

The credit card transaction method is examined as the basic stochastic process of an (Advanced Hidden Markov Model) AHMM. The variety of transaction quantity considered as the observation symbols, while the kinds of item have been deemed to be states of the AHMM. In addition to comprise recommended a technique for decision the spending profile of cardholders is authorized or not. As well as purpose of this knowledge in deciding the value of observation symbols and initial estimate of the model parameters with the best fit observation is that providing an effective credit card fraud detection system. It has also been enlightened how the HMM vary with the AHMM can detect whether an arriving transaction is fake or not. Experimental results show the performance and effectiveness of AHMM system and show the efficiency of knowledge the spending profile of the cardholder in AHMM system.

REFERENCES

[1]. L.R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.

[2]. S. Ghosh and D.L. Reilly, “Credit Card Fraud Detection with a Neural-Network,” Proc. 27th Hawaii International Conference on System Sciences: Information Systems: Decision Support and Knowledge-Based Systems, vol. 3, pp. 621-630, 1994.

[3]. S. Axelsson, “The Base-Rate Fallacy and the Difficulty of Intrusion Detection,” ACM Transactions on Information and System Security, vol. 3, no. 3, pp. 186-205, 2000.

[4]. M. Syeda, Y. Q. Zhang, and Y. Pan, "Parallel Granular Networks for Fast Credit Card Fraud Detection," Proc. IEEE International Conference on Fuzzy Systems, pp. 572-577, 2002.

[5]. S.J. Stolfo, D.W. Fan, W. Lee, A.L. Prodromidis, and P.K. Chan, "Credit Card Fraud Detection using Meta-Learning: Issues and Initial Results," Proc. AAAI Workshop on AI Methods in Fraud and Risk Management, pp. 83-90, 1997.

[6]. S.J. Stolfo, D.W. Fan, W. Lee, A. Prodromidis, and P.K. Chan, "Cost-based Modeling for Fraud and Intrusion Detection: Results from the JAM Project," Proc. DARPA Information Survivability Conference and Exposition, vol. 2, pp. 130-144, 2000.

[7]. E. Aleskerov, B. Freisleben, and B. Rao, "CARDWATCH: A Neural Network Based Database Mining System for Credit Card Fraud Detection," Proc. IEEE/IAFE: Computational Intelligence for Financial Engineering, pp. 220-226, 1997.

[8]. M.J. Kim and T.S. Kim, "A Neural Classifier with Fraud Density Map for Effective Credit Card Fraud Detection," Proc. International Conference on Intelligent Data Engineering and Automated Learning, Lecture Notes in Computer Science, Springer Verlag, no. 2412, pp. 378-383, 2002.

[9]. W. Fan, A.L. Prodromidis, and S.J. Stolfo, "Distributed Data Mining in Credit Card Fraud Detection," IEEE Intelligent Systems, vol. 14, no. 6, pp. 67-74, 1999.

[10]. R. Brause, T. Langsdorf, and M. Hepp, "Neural Data Mining for Credit Card Fraud Detection," Proc. IEEE International Conference on Tools with Artificial Intelligence, pp. 103-106, 1999.

[11]. C. Chiu and C. Tsai, "A Web Services-based Collaborative Scheme for Credit Card Fraud Detection," Proc. IEEE International Conference on e-Technology, e-Commerce and e-Service, pp. 177-181, 2004.

[12]. C. Phua, V. Lee, K. Smith, and R. Gayler, "A Comprehensive Survey of Data Mining-based Fraud Detection Research," available on-line at http://www.bsys.monash.edu.au/people/cphua/, 07 March 2007.

[13]. S. Stolfo and A.L. Prodromidis, "Agent-based Distributed Learning applied to Fraud Detection," Technical Report, CUCS-014-99, Columbia University, USA, 1999.

[14]. D.J. Hand, G. Blunt, M.G. Kelly, and N.M. Adams, "Data Mining for Fun and Profit," Statistical Science, vol. 15, no. 2, pp. 111–131, 2000.

[15]. Sushmito Ghosh and Douglas L. Reilly, "Credit Card Fraud Detection with a Neural-Network." Nestor, Inc. IEEE (1994).

[16]. C. Phua D. Alahakoon, and V. Lee, "Minority Report in Fraud Detection: Classification of Skewed Data," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 50-59, 2004.

[17]. V. Vatsa, S. Sural, and A.K. Majumdar, "A Game-theoretic Approach to Credit Card Fraud Detection," Proc. 1st International Conference on Information Systems Security, Lecture Notes in Computer Science, Springer Verlag, pp. 263-276, 2005.

A.Prakash done M.Sc (CT), from Periyar University Salem in 2001, completed M.Phil.(CS), from Manonmaniam Sundaranar University,Tirunelveli in 2003. Received MCA, from Periyar University Salem in 2011. Currently working as a Asst. Professor in Dept. of Computer Applications, Hindusthan College of arts and science, Coimbatore. His research area is data mining.

Dr. C. Chandrasekar received his Ph.D. degree from Periyar University, Salem, TN, India. He has been working as Associate Professor at Dept. of Computer Science, Periyar University, Salem – 636 011, Tamil Nadu, India. His research interest includes Wireless networking, Mobile computing, Computer Communication and Networks. He was a Research guide at various universities in India. He has been published more than 50 research papers at various National / International Journals.

Applying Encryption Algorithm for Data Security and Privacy in Cloud Computing

Mohit Marwaha[1], Rajeev Bedi[2]

[1] **Computer Science And Engineering, Punjab Technical University, Beant College of engineering and Technology
Gurdaspur, Punjab, India**

[2] **Computer Science And Engineering, Punjab Technical University, Beant College of engineering and Technology
Gurdaspur, Punjab, India**

Abstract

Cloud computing is the next big thing after internet in the field of information technology; some say it's a metaphor for internet. It is an Internet-based computing technology, in which software, shared recourses and information, are provided to consumers and devices on-demand, and as per users requirement on a pay per use model. Even though the cloud continues to grow in popularity, Usability and respectability, Problems with data protection and data privacy and other Security issues play a major setback in the field of Cloud Computing. Privacy and security are the key issue for cloud storage. Encryption is a well known technology for protecting sensitive data. Use of the combination of Public and Private key encryption to hide the sensitive data of users, and cipher text retrieval. The paper analyzes the feasibility of the applying encryption algorithm for data security and privacy in cloud Storage.

Keywords: *Cloud Storage, Cipher text retrieval, encryption algorithm***.**

1. Introduction

Cloud computing is a flexible, cost- effective and proven delivery platform for providing business or consumer IT services over the Internet. Cloud computing supports distributed service oriented architecture, multi-users and multi-domain administrative infrastructure, it is more prone to security threats and vulnerabilities. At present, a major concern in cloud adoption is its security and Privacy. Intrusion prospects within cloud environment are many and with high gains. Security and Privacy issues are of more concern to cloud service providers who are actually hosting the services. In most cases, the provider must guarantee that their infrastructure is secure and clients' data and applications are safe by implementing security policies and mechanisms. While the cloud customer must ensure that provider has taken proper security measures to protect their information. The issues are organized into several general categories: trust, architecture, identity management, software isolation, data protection, availability Reliability, Ownership, Data Backup, Data Portability and Conversion, Multiplatform Support and Intellectual Property.

2. Cloud Computing Framework

Service Models: These three are the most widely used service models of cloud computing.

2.1 Software as a service.

Software-as-a-Service (SaaS): It is also referred as software available on demand, it is based on multi-tenant architecture. Software like word processor, CRM (Customer Relation Management), etc. or application services like schedule, calendar, etc. are executed in the "cloud" using the interconnectivity of the internet to do manipulation on data. Custom services are combined with 3rd party commercial services via Service oriented architecture to create new applications. It is a software delivery for business applications like accounting, content delivery, Human resource management (HRM), Enterprise resource planning (ERP) etc on demand on pay-as-you go model[1].

2.2 Platform as a Service.

Platform-as-a-Service (PaaS): This layer of cloud provides computing platform and solution stack as service**.** Platform-as-a-Service provides the user with the freedom of application design, application development, testing, deployment and hosting as well as application services such as team collaboration, web service integration and database integration, security, scalability, storage, persistence, state management, application versioning, without thinking about the underlying hardware and software layers by providing facilities required for completion of project through web application and services via Internet.

2.3 Infrastructure as a Service.

Infrastructure-as-a-Service (IaaS): Infrastructure as a service delivers a platform virtualization environment as a service. Instead of purchasing servers, software, data center space or network equipment, clients can buy these resources as outsourced service. In other words the client uses the third party infrastructure services to support its operations including hardware, storage, servers and networking components.

3. Cloud Deployment Models

There are three types cloud Deployment models that widely used are:

3.1 Public.

It is referred as external cloud or multi-tenant cloud, this model represents an openly accessible cloud environment in this cloud can be accessed by general public. Customer can access resources and pay for the operating resources. Public Cloud can host individual services as well as collection of services

3.2 Private.

It is also known as internal cloud or on-premise cloud, a private cloud provides a limited access to its resources and services to consumers that belong to the same organization that owns the cloud. In other words, the infrastructure that is managed and operated for one organization only, so that a consistent level of control over security, privacy, and governance can be maintained.

3.3 Hybrid.

A hybrid cloud is a combination of public and private cloud. It provides benefits of multiple deployment models. It enables the enterprise to manage steady-state workload in the private cloud, and if the workload increases asking the public cloud for intensive computing resources, then return if no longer needed.

3.4 Community.

This deployment model share resources with many organizations in a community that shares common concerns (like security, governance, compliance etc). It typically refers to special-purpose cloud computing environments shared and managed by a number of related organizations participating in a common domain or vertical market [12].

4. Issues in Cloud Data Storage.

Cloud Computing moves the application software and databases to the large data centers, where the management of the data and services may not be fully trustworthy. This unique attribute, however, poses many new security challenges which have not been well understood. In this article, we focus on cloud data storage security, which has always been an important aspect of quality of service. To ensure the correctness of users' data in the cloud.

A. Trust: Trust is defined as reliance on the integrity, strength, ability and surety of a person or thing. Entrusting your data on to a third party who is providing cloud services is an issue. Recent incidents like In April of 2012 Amazon's Elastic Compute Cloud service crashed during a system upgrade, knocking customers' websites off-line for anywhere from several hours to several days. That same month, hackers broke into the Sony PlayStation Network, exposing the personal information of 77 million people around the world. And in June a software glitch at cloud-storage provider Dropbox temporarily allowed visitors to log in to any of its 25 million customers' accounts using any password or none at all. These issues have certainly created doubts in mind of cloud consumers and damaged the trust ability of Consumers [4].

B. Privacy: Different from the traditional computing model, cloud computing utilizes the virtual computing technology, users' personal data may be scattered in various virtual data center rather than stay in the same physical location, even across the national borders, at this time, data privacy protection will face the controversy of different legal systems. On the other hand, users may leak hidden information when they accessing cloud computing services. Attackers can analyze the critical task depend on the computing task submitted by the users [9].

C. Security: Cloud service providers employ data storage and transmission encryption, user authentication, and authorization. Many clients worry about the vulnerability of remote data to criminals and hackers. Cloud providers are enormously sensitive to this issue and apply substantial resources to mitigate this problem.

D. Ownership: Once data has been relegated to the cloud, some worry about losing their rights or being unable to protect the rights of their customers. Many cloud providers address this issue with well-skilled user-sided agreements. According to the agreement, users would be wise to seek advice from their favourite legal representative [10].

E. Performance and Availability: Business organizations are worried about acceptable levels of performance and availability of applications hosted in the cloud.

F. Legal: There are certain apprehensions for a cloud service provider and a client receiving the service like location of the cloud provider, infrastructure and physical location of the data and outsourcing of the cloud provider's services etc.

G. Multiplatform Support: More an issue for IT departments using managed services is how the cloud-based service integrates across different platforms and operating systems, e.g. OS X, Windows, Linux and thin-clients. Usually, some customized adaption of the service takes care of any problem. Multiplatform support requirements will ease as more user interfaces become web-based.

H. Intellectual Property: A company invents something new and it uses cloud services as part of the invention. Is the invention still patentable? Or there can be issues like cloud service provider can make claim for that invention or leak the information to the competitor.

I. Data Backup: Cloud providers employ redundant servers and routine data backup processes, but some people worry about being able to control their own backups. Many providers are now offering data dumps onto media or allowing users to back up data through regular downloads.

J. Data Portability and Conversion: Some people have concerns like, switching service providers; there may be difficulty in transferring data. Porting and converting data is highly dependent on the nature of the cloud provider's data retrieval format, particular in cases where the format cannot be easily revealed. As service competition grows and open standards become established, the data portability issue will ease, and conversion processes will become available supporting the more popular cloud providers. Worst case, a cloud subscriber will have to pay for some custom data conversion.

These are certain areas in which cloud computing requires to excel and solve problem related to it. Out of all the problems Security, Privacy and Intellectual property put the major threats on growth of cloud computing that are needed to be worked upon.

5. OVERVIEW OF OUR APPROACH

Our goal is to build up a repository to facilitate the data integration and sharing across cloud along with preservation of data confidentiality. For this we will be using an encryption technique to provide data security on data storage [16].

Objective of our System.

1. To develop a system that will Provide Security and Privacy to Cloud Storage

2. To Establish an Encryption Based System for protecting Sensitive data on the cloud and Structure how owner and storage Service Provider to operate on encrypted Data

3. To Create a System where the user store its data on the cloud the data is sent and stored on the cloud in encrypted form As in normal cases in cloud computing when a user login to the cloud and they store data on cloud storage device the data stored on the server cloud is not much secure as it can be readable to anyone which have permission to access and Leaving data vulnerable,

4. To Develop a retrieval System in which the data is retrieved by the user in encrypted form and is decrypted by the user at its own site using a public and private key encryption both the keys working at the user level.

6. Conclusion

Our research indicates that that Security and Privacy are the major issues that are needed to be countered, efforts are being made to develop many efficient System That can Provide Security and privacy at the user level and maintain the trust and intellectual property rights of the user. Our method States Encryption is one such method that can provide peace of mind to user and if the user have control over encryption and decryptions of data that will boost consumer confidence and attract more people to cloud platform.

References

[1]http://en.wikipedia.org/wiki/Cloud_computing.

[2] Rich Maggiani, solari communication. "Cloud computing is changing how we communicate".

[3] Randolph Barr, Qualys Inc, "How to gain comfort in losing control to the cloud".

[4] Greg Boss, Padma Malladi, Dennis Quan, Linda Legregni, Harold Hall, HiPODS, www.ibm.com/ developerworks/websphere/zones/hipods

[5] http://www.roughtype.com.

[6] Tharam Dillon, Chen Wu, Elizabeth Chang, 2010 24th IEEE International Conference on Advanced Information Networking and Applications ,"Cloud computing: issues and challenges".

[7]June13,2009,http://server.zol.com.cn/183/1830464.html.

[8] Elinor Mills, January 27,2009. "Cloud computing security forecast: clear skies".

[9] Jianchun Jiang, Weiping Wen, "Information security issues in cloud computing environment", Netinfo Security,doi:10.3969/j.issn.1671-1122.2010.02.026.

[10] Jianchun Jiang, Weiping Wen, "Information security issues in cloud computing environment", Netinfo Security,doi:10.3969/j.issn.1671- 1122.2010.02.026. of virtual machines" In Proc. Of NSDI'05, pages 273-286, Berkeley CA, USA, 2005. USENIX Association.

[11] Eucalyptus Completes Amazon Web Services Specs with Latest Release.
[12] Open Cloud Consortium.org.
[13] July 27,2009. Available from http://fx.caixun.com/.
[14] Jack Schofield. Wednesday 17 June 2009 22.00 BST, http://www.guardian.co.uk/technology/2009/jun/17/cloud-computingjack- schofield.
[15] Gartner. "Seven cloud-computing security risks".
[16] Ranjita Mishra "A Privacy Preserving Repository for Securing Data across the Cloud".

First Author Mohit Marwaha completed BTech from Beant College of Engineering and Technology in 2008 Pursuing MTech from Beant College of Engineering and Technology I have published two papers one in an international jouranal and other in an international conference and is presently working with Beant college of Engineering and Technology as Assistant Professor. Area of Research is security on cloud computing.

Second Author Rajeev Bedi completed B.Tech Computer Science and Engineering in 2000 and M.Tech. Computer Science and Engineering in 2008 from Punjab Technical University, Jalandhar and Pursuing PhD from CMJ University Shillong. Currently Working as Assistant Professor in Beant College of Engineering and Technology, Gurdaspur since 2004. I am Reviewer of IJCSIT journal. I have 13 publications in different Internationl, National Journals and Conferences. My current research interest is Cloud Computing.

A Method of LSB substitution based on image blocks and maximum entropy

Mohamed RADOUANE[1], Tarik BOUJIHA[1], Rochdi MESSOUSSI[1], Nadia IDRISSI[2], Ahmed ROUKH[2]

**[1] Department of physics, Faculty of Science, Ibn tofail University,
B.P 242, Kénitra, Morocco**

**[1] Ecole Nationale des Sciences Appliquées, Ibn tofail University,
Kénitra, Morocco**

**[2] Department of physics, Faculty of Science, Moulay ismail University
B.P 11201 zitoune Meknes, Morocco**

Abstract

In this paper we introduce an algorithm of digital watermarking based on embedding watermark into sub images with LSB technique. The watermark is embedded into specifics blocks of the host image, the selection of blocks are based on entropy value.
The simulation results show that the visual quality of the watermarked image and the extracted watermark is good, this result is presented and proved by a high PSNR value.

Keywords: *Information security, digital image watermarking, Entropy, Least Significant Bit (LSB), Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR).*

1. Introduction

In the last years, due to the advancement in technologies and the increase rapidly of data transmission, most people prefers to use the internet as the essential medium to transfer data. The data transmission is made very simple, fast and accurate using the internet.

However, the protection and enforcement of intellectual property copyrights has become an important issue in the digital world.
There are several approaches, methods and techniques have been developed to protect our information during transfer data from source to destination like Cryptography, Steganography and digital image Watermarking.
Fundamentally, watermarking can be described as a method for embedding information into another signal.

In case of digital images, the embedded information can be either visible or hidden from the user. A host image used to hide the secret data is called the host image [5] or the carrier image. After embedding the secret data into the host image, the resultant image is called the watermarked image.

2. Digital watermarking

Digital watermarking technology is an emerging field in computer science, cryptography, signal processing and communications. Digital Watermarking is intended by its developers as the solution to the need to provide value added protection on top of data encryption and scrambling for content protection. Like other technology under development, digital watermarking raises a number of essential questions as follows. [1]
Digital watermarking is defined as a process of embedding data (watermark), into a multimedia object to help to protect the owner's right to that object. The embedding data (watermark) may be either visible or invisible. In case of visible watermarking, the watermark is embedded into the host image such that the watermark is intentionally perceptible to a human observer; whereas in the case of invisible, the embedded image data that is not perceptible, but may be extracted by a computer program. [2]

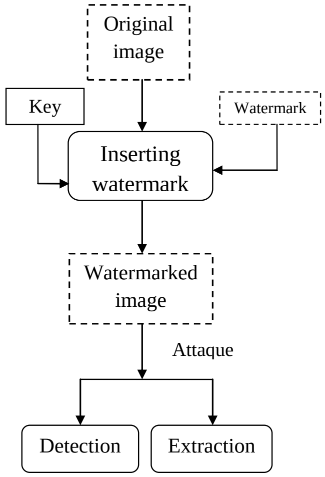


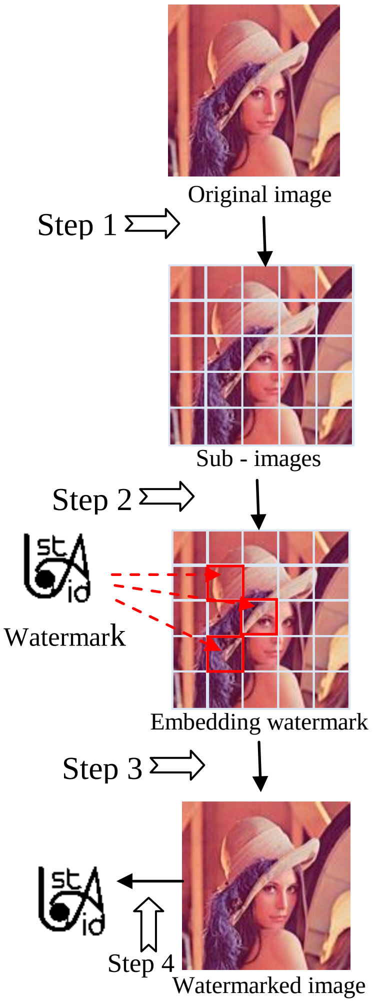
Figure 1 : Scheme of watermarking

3. Least Significant Bit (LSB)

The least significant bit (LSB) technique is used to embed information in a cover image. The LSB technique of a cover image is described by changing pixels by bits of the secret message. These changes cannot be perceived by the human visibility system. This technique was originally designed to work with gray-scale images but is easily extended to color images by treating each color plane as a single plane in which data is inserted in the LSB. [3][4]

The process of embedding data is described with these different steps:

Step 1: devising image into different blocks (sub-images).
Step 2: calculate the entropy of each block
Step 4: inserting watermark image into sub images which have the maximum entropy and calculate the PSNR.
Step 5: recognizing the image devised to have the original image.
Step 6: retrieving watermark with extraction method.



4. Entropy of grayscale image:

Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. Entropy is defined as:

E=-sum(p.*log2(p))

Where p contains the histogram counts returned from imhist. By default, entropy uses two bins for logical arrays and 256 bins for uint8, uint16, or double arrays.[6]

Results:

We embed watermark in blocks of original image using maximum entropy to select the suitable blocks.

We notice that there is no difference between the original and watermarked images and there is no distortion occurs for these watermarked images.

We got the result after we calculated the Peak signal-to-noise ratio (PSNR).
The PSNR is a better test since it takes the signal strength into consideration. The values were used to evaluate the quality of the watermarked images.

This equation describes how this value is obtained:

$$PSNR = 10 \log_{10}\left[\frac{R^2}{MSE}\right]$$

Where R represents maximum fluctuation or value in the image, its value is 255 for 8 bit unsigned number.

The MSE represents the cumulative squared error between the compressed and the original image.
To compute the PSNR, the mean squared error is first calculated using the following equation:

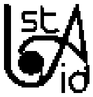
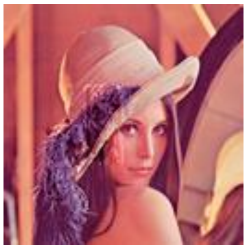
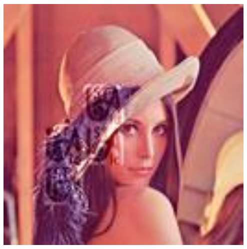
$$MSE = \frac{\sum_{M,N}[I1(m,n) - I2(m,n)]^2}{M * N}$$

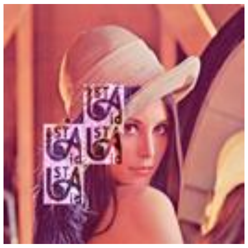
Where M and N are the number of rows and columns in the input images, respectively and I1 (m, n) is the original image, I2 (m, n) is the Watermarked image.

Number Of blocks	**Method**	**PSNR**
1 block	1st Bit Substitution (LSB)	65.7059
	2nd Bit Substitution	59.7120
	3rd Bit Substitution	53.7899
	4th Bit Substitution	47.5984
	5th Bit Substitution	41.2750
	6th Bit Substitution	35.4789
	7th Bit Substitution	29.5223
	8th Bit Substitution (MSB)	23.9433
2 blocks	1st Bit Substitution (LSB)	62.7029
	2nd Bit Substitution	56.6401
	3rd Bit Substitution	50.7461
	4th Bit Substitution	44.5714
	5th Bit Substitution	38.3996
	6th Bit Substitution	32.5842
	7th Bit Substitution	26.5419
	8th Bit Substitution (MSB)	20.7166
3 blocks	1st Bit Substitution (LSB)	60.9924
	2nd Bit Substitution	54.8748
	3rd Bit Substitution	48.9145
	4th Bit Substitution	42.8616
	5th Bit Substitution	36.7401
	6th Bit Substitution	30.8495
	7th Bit Substitution	24.7399
	8th Bit Substitution (MSB)	18.8826
4 blocks	1st Bit Substitution (LSB)	59.7102
	2nd Bit Substitution	53.6238
	3rd Bit Substitution	47.6575
	4th Bit Substitution	41.6151
	5th Bit Substitution	35.5435
	6th Bit Substitution	29.6488
	7th Bit Substitution	23.6147
	8th Bit Substitution (MSB)	17.4579

Table1: PSNR values for the blocks method

These figures demonstrate the embed watermark in four blocks of original image.

 Figure1: logo to be embedded in image	 Figure2:watermarked image with LSB
 Figure3:Watermarked Image 2nd bit substitution	 Figure3: Watermarked Image 3rd bit substitution
 Figure4: Watermarked Image 4th bit substitution	 Figure5: Watermarked Image 5th bit substitution
 Figure6: Watermarked Image 6th bit substitution	 Figure7: Watermarked Image 7th bit substitution

 Figure8: Watermarked Image 8th bit substitution	 Figure9: extracted watermark

Conclusions

This paper proposed a method of LSB digital watermarking scheme based on combination of LSB and maximum entropy.

The experimental result shows that the proposed method maintains the quality of the watermarked image.

This method is also tested using PSNR and the result of PSNR is compared with different insertion blocks of the host image.

For the future research, we will focus on the studies comparison of different watermarking schemes based on different LBP (Local Binary Pattern).

References

[1] Saraju P. Mohanty, “Digital Watermarking: A Tutorial Review”, Department of Computer Science and Engineering, University of South Florida, 1999.

[2] R. Chandramouli, Nasir Memon, and Majid Rabbani "Digital Watermarking", Encyclopedia of Imaging Science and Technology, 2002, 10.1002/0471443395.img010.

[3] Dr. Ekta Walia a, Payal Jainb, Navdeepc, An Analysis of LSB & DCT based Steganography, Global Journal of Computer Science and Technology ,Vol. 10 Issue 1 (Ver 1.0), April 2010.

[4] D. Biswas, S. Biswas, P.P. Sarkar, D. Sarkar, S. Banerjee, A. Pal, comparison and analysis of watermarking algorithms in color image-image security paradigm, International Journal of Computer Science & Information Technology Vol 3, No 3, June 2011.

[5] Katzenbeisser, S. and Petitcolas, F(1999).: Information hiding techniques for steganography and digital watermarking. Artech House Books.

[6] Acta Applicandae Mathematicae, Zouhir Mokhtari, Khaled Melkemi, A New Watermarking Algorithm Based on Entropy Concept, Volume 116, Issue 1, pp 65-69, Octobre 2011.

Assessing Pareto Software Reliability Using SPC

Satya Prasad R[1], Sita Kumari K[2] and Sridevi G[3]

**[1]Department of CSE, Acharya Nagarjuna University,
Guntur, Andhra Pradesh, India**

**[2]Department of IT, V.R.Siddhartha Engineering College,
Kanuru, Vijayawada, Andhra Pradesh, India.**

**[3]Department of CSE, Nimra Women's College of Engineering,
Vijayawada, Andhra Pradesh, India.**

Abstract

Software reliability is one of the most important characteristics of software quality and is very much essential for producing reliable software systems. The reliability of software can be monitored efficiently using Statistical Process Control (SPC). It helps to identify when the failure takes place during the software development process. In this paper we proposed a control mechanism based on the cumulative observations of the time domain data using the mean value function of Pareto type II distribution, which is based on Non-Homogenous Poisson Process (NHPP). To estimate the unknown parameters of the model, maximum likelihood estimation method is used. The failure data is analyzed with the proposed mechanism and the results are exhibited through control charts.

Keywords: *Control Charts, Mean Value Function, NHPP, Pareto type II distribution, Statistical Process Control, Time domain data.*

1. Introduction

Software Reliability is an important quality characteristic of a software which can evaluate and predict the operational quality of software system during its development. Software Reliability is the probability of failure free operation of software in a specified environment for a specified period of time [5], [6]. Software Process Control (SPC) concepts and methods are used for improving the software reliability by identifying and eliminating the human errors in the software development process. SPC is an important tool for monitoring and controlling manufacturing processes. SPC can be used to monitor the performance of a software process over time in order to verify that the process remains in the state of statistical control. It helps in finding assignable causes, long term improvements in the software process. Software quality and reliability can be achieved by eliminating the causes or improving the software process or its operating procedures [1].

SPC is a powerful tool to optimize the amount of information needed for use in making management decisions. Statistical techniques provide an understanding of the business baselines, insights for process improvements, communication of value and results of processes, and active and visible involvement. SPC provides real time analysis to establish controllable process baselines; learn, set, and dynamically improve process capabilities; and focus business areas needing improvement. The early detection of software failures will improve the software reliability. The selection of proper SPC charts is essential to effective statistical process control implementation and use. The SPC chart selection is based on data, situation and need [2]. An advantage of SPC over other methods of quality control, such as "inspection", is that it emphasizes early detection and prevention of problems, rather than the correction of problems after they have occurred.

Control charts are the key tools which are used in SPC to monitor the quality. Basically there are two types of Control charts that can be used depending on the characteristics to be monitored. There are two main categories of Control Charts, those that display *attribute data*, and those that display *variables data*.

Attribute Data -- This category of Control Chart displays data that result from counting the number of occurrences or items in a single category of similar items or occurrences. These "count" data may be expressed as pass/fail, yes/no, or presence/absence of a defect.

Variables Data -- This category of Control Chart displays values resulting from the measurement of a continuous variable. Examples of variables data are elapsed time, temperature. The univariate control chart is a graphical display of one quality characteristic and the multivariate control chart is a graphical display of statistics that represents more than one quality characteristic. The control chart is one of the seven tools for quality control. The control limits for the chart are defined in such a manner that the process is considered to be out of control when the time to observe exactly one failure is less than LCL or greater than UCL. Our aim is to monitor the failure process and detect any change of the intensity parameter [2].

The Non-Homogeneous Poisson Process (NHPP) based models are the most important models because of their simplicity, convenience and compatibility. The NHPP based software reliability growth models are proved quite successful in practical software reliability engineering [5]. The NHPP model represents the number of failures experienced up to certain time. The main issue in the NHPP model is to determine an appropriate mean value function to denote the expected number of failures experienced up to a certain time point [3]. The Maximum likelihood estimation (MLE) is the most useful technique for deriving the point estimators. Parameter estimation is of primary importance in software reliability prediction. Once the analytical solution for m (t) is known for a given model, parameter estimation is achieved by applying a technique of Maximum Likelihood Estimate (MLE). The failure data is collected in time domain data. The idea behind maximum likelihood parameter estimation is to determine the parameters that maximize the probability (likelihood) of the sample data. The method of maximum likelihood is considered to be more robust (with some exceptions) and yields estimators with good statistical properties. In other words, MLE methods are versatile and apply to most models and to different types of data. Although the methodology for maximum likelihood estimation is simple, the implementation is mathematically intense. In our proposed model the parameters are estimated using MLE. The Newton Raphson method is used for obtaining the parameter values.

This paper presents Pareto type II model for analyzing the reliability of a software system using time domain data. The layout of the paper is as follows: Section II gives the interpretation of the model for the underlying NHPP, Section III describes the proposed Pareto type II software reliability growth model, Section IV discusses parameter estimation of Pareto type II model based on time domain data. Section V describes the control charts that are used for analysing the live data for software failures and finally Section VI gives the Conclusion.

2. Model Formulation

Software reliability growth models can be used as an indication of the number of failures that may be encountered after the software has shipped and thus as an indication of whether the software is ready to ship. These models use system test data to predict the number of defects remaining in the software. There are essentially two types of software reliability models - those that attempt to predict software reliability from design parameters and those that attempt to predict software reliability from test data. The first type of models are usually called "defect density" models and use code characteristics such as lines of code, nesting of loops, external references, input/outputs, and so forth to estimate the number of defects in the software. The second type of models is usually called "software reliability growth" models. These models attempt to statistically correlate defect detection data with known functions such as an exponential function. If the correlation is good, the known function can be used to predict future behaviour. Software reliability growth models are the focus of this report. Most software reliability growth models have a parameter that relates to the total number of defects contained in a set of code. If we know this parameter and the current number of defects discovered, we know how many defects remain in the code (see Figure 1).Knowing the number of residual defects, it can be decided whether or not the code is ready to ship and how much more testing is required if we decide the code is not ready to ship. It gives us an estimate of the number of failures that our customers will encounter when operating the software. This estimate helps us to plan the appropriate levels of support that will be required for defect correction after the software has shipped and determine the cost of supporting the software.

Software reliability growth models are a statistical interpolation of defect detection data by mathematical functions. The functions are used to predict future failure rates or the number of residual defects in the code. There are different ways to represent defect detection data as discussed in Section 2.1. There are many types of software reliability growth models as described in Section 2.2, and there are different ways to statistically correlate the data to the models as discussed in Section 2.3. Current software reliability literature is inconclusive as to which data representation, software reliability growth model, and statistical correlation technique works best. The advice in the literature seems to be to try a number of the different techniques and see which works best in your environment. In Section 3, we describe the application of the techniques in the Tandem environment.

There are numerous software reliability growth models available for use according to probabilistic assumptions. The Non Homogenous Poisson Process (NHPP) based software reliability growth models are proved quite successful in practical software reliability engineering. NHPP model formulation is described in the following lines.

A software system is subject to failures at random times caused by errors present in the system. Let {N(t), t >0 } be the cumulative number of software failures by time ‘t’, where t is the failure intensity function, which is proportional to the residual fault content.

Let m (t) represents the expected number of software failures by time't'. The mean value function m (t) is finite valued, non-decreasing, non-negative and bounded with the boundary conditions.

$$m(t) = 0, t = 0$$
$$= a,\ t \to \infty$$

Where a is the expected number of software errors to be eventually detected.

Suppose N (t) is known to have a Poisson probability mass function with parameters m (t) i.e.,

$$P\{N(t) = n\} = \frac{[m(t)]^n . e^{-m(t)}}{n!}, n = 0,1,2 \dots \infty$$

Then N (t) is called an NHPP. Thus the stochastic behaviour of software failure phenomena can be described through the N (t) process. Various time domain models have appeared in the literature that describes the stochastic failure process by an NHPP which differ in the mean value functions m (t).

3. The Proposed Pareto Type II SRGM

In this paper we consider m (t) as given by

$$m(t) \ = a\,[1 \ - \tfrac{c^b}{(t+c)^b}\Big] \qquad (3.1)$$

Where [m (t)/a] is the cumulative distribution function of Pareto type II distribution (Johnson et al, 2004) for the resent choice.

$$P\{N(t) = n\} = \frac{[m(t)]^n . e^{-m(t)}}{n!}$$

$$\lim_{n\to\infty} P\,\{N(t) = n\} = \frac{e^{-a}.a^n}{n!}$$

This is also a Poisson model with mean 'a'.

Let N (t) be the number of errors remaining in the system at time't'

$$N(t) = N(\infty) - N(t)$$

$$E[N(t)] \ = E[N(\infty)] - E[N(t)]$$
$$= a - m(t)$$
$$= a - \ a\left[1 - \tfrac{c^b}{(t+c)^b}\right]$$
$$= \tfrac{ac^b}{(t+c)^b}$$

4. Parameter Estimation Based On Time Domain Data

In this section we develop expressions to estimate the parameters of the Pareto type II model based on time domain data. Parameter estimation is of primary importance in software reliability prediction.

A set of failure data is usually collected in one of two common ways, time domain data and time domain data. In this paper parameters are estimated from the time domain data.

The mean value function of Pareto type II model is given by

$$m(t) = a\left[1 - \tfrac{c^b}{(t+c)^b}\right], \qquad t \geq 0 \qquad (4.1)$$

In order to have an assessment of the software reliability, a, b and c are to be known or they are to be estimated from software failure data. Expressions are now delivered for estimating 'a', 'b' and 'c' for the Pareto type II model [7].

We conduct an experiment and obtain N independent observations, t_1, t_2…, tn. The likelihood function for time domain data [8] is given by

$$Log\ l = -a\left[1 - \left(\tfrac{c}{t+c}\right)^b\right] +$$

$$\sum_{i=1}^{n} [\log a + \log b + b\ logc - (b+1)\log(t_i + c)]$$

$$(4.2)$$

Accordingly parameters 'a','b' and 'c' would be solutions of the equations.

$$\frac{\partial \log L}{\partial a} = 0$$

$$a = \tfrac{n(t+c)^b}{(t+c)^b - c^b} \qquad (4.3)$$

The parameter 'b' is estimated by iterative Newton Raphson Method using

$$b_{n+1} = b_n - \tfrac{g(b)}{g'(b)}$$

Where $g(b) and\ \ g^{'}(b)$ are expressed as follows.

$$g(b) = \frac{\partial LogL}{\partial b} = 0$$

$$\frac{\partial Log\ L}{\partial b} = \frac{nlog\left(\frac{1}{t+1}\right)}{(t+1)^b - 1} + \frac{n}{b} - \textstyle\sum_{i=1}^{n} log(t_i + 1)$$

$$(4.4)$$

$$g'(b) = \frac{\partial^2\ Log\ L}{\partial b^2} = 0$$

$$\frac{\partial^2 \, Log \, L}{\partial b^2} = -nlog\left(\frac{1}{t+1}\right)\left[\frac{(t+1)^b \log{(t+1)}}{\left[(t+1)^b - 1\right]^2}\right] - \frac{n}{b^2} \tag{4.5}$$

The parameter ‘c’ is estimated by iterative Newton Raphson Method using

$$c_{n+1} = c_n \; - \frac{g(c_n)}{g\prime(c_n)}$$

Where $g(c)$ *and* $g'(c)$ are expressed as follows.

$$g(c) = \frac{\partial LogL}{\partial c} = 0$$

$$\frac{\partial Log \, L}{\partial c} = \frac{n}{(t+c)} + \frac{n}{c} - \sum_{i=1}^{n} \frac{2}{t_i + c} \tag{4.6}$$

$$g'(c) = \frac{\partial^2 LogL}{\partial c^2} = 0$$

$$\frac{\partial^2 logL}{\partial c^2} = \frac{-n}{(t+c)^2} - \frac{n}{c^2} + \sum_{i=1}^{n} \frac{2}{(t_i+c)^2} \tag{4.7}$$

The values of ‘b’ and ‘c’ in the above equations can be obtained using Newton Raphson Method. Solving the above equations simultaneously yields the point estimates of the parameters b and c. These equations are to be solved iteratively and their solutions in turn when substituted in equation (4.3) gives value of ‘a’.

5. Data Analysis

In this section, we present the analysis of software failure data set. The set of software errors analysed here is borrowed from software development project as published in Pham (2005) [3].

The data named as NTDS data are summarized in the below table.

Table 1. NTDS Data

Failure Number n	Time Between Failure (x_k) days	Cumulative Time
1	9	9
2	12	21
3	11	32
4	4	36
5	7	43
6	2	45
7	5	50
8	8	58
9	5	63
10	7	70
11	1	71
12	6	77
13	1	78
14	9	87
15	4	91
16	1	92
17	3	95
18	3	98
19	6	104
20	1	105
21	11	116
22	33	149
23	7	156
24	91	247
25	2	249
26	1	250
Test Phase		
27	87	337
28	47	384
29	12	396
30	9	405
31	135	540
User Phase		
32	258	798
Test Phase		
33	16	814
34	35	849

Solving equations by Newton Raphson Method for the NTDS test data, the iterative solutions for MLEs of a, b and c are

$$a = 55.01871$$

$$b = 0.998899$$

$$c = 278.6101$$

Using ‘a’ and ‘b’ and ‘c’ values we can compute m(t). Now the control limits are calculated by the following equations taking the standard values 0.00135, 0.99865 and 0.5.

Table 2. Successive differences of Cumulative mean values

Failure number	Cumulative failures	Mean values	Successive differences
1	9	1.7198	2.132428
2	21	3.852228	1.810059
3	32	5.662288	0.626838
4	36	6.289126	1.059468
5	43	7.348594	0.294291
6	45	7.642885	0.720064
7	50	8.362949	1.107632
8	58	9.470581	0.66594
9	63	10.13652	0.90024
10	70	11.03676	0.125665
11	71	11.16243	0.739154
12	77	11.90158	0.120775
13	78	12.02235	1.057264
14	87	13.07962	0.453377
15	91	13.533	0.111816
16	92	13.64481	0.331858
17	95	13.97667	0.326574
18	98	14.30324	0.637793
19	104	14.94104	0.10436
20	105	15.0454	1.113071
21	116	16.15847	2.995794
22	149	19.15426	0.577016
23	156	19.73128	6.103281
24	247	25.83456	0.110506
25	249	25.94507	0.05494
26	250	26.00001	------

$$T_u = [1 - \frac{c^b}{(t+c)^b}] = 0.99865$$

$$T_c = [1 - \frac{c^b}{(t+c)^b}] = 0.5$$

$$T_l = [1 - \frac{c^b}{(t+c)^b}] = 0.00135$$

These limits are converted to $m(t_u)$,$m(t_c)$ and $m(t_l)$ form. They are used to find whether the software process is in control or not by placing the points in Mean value chart shown in figure 1.

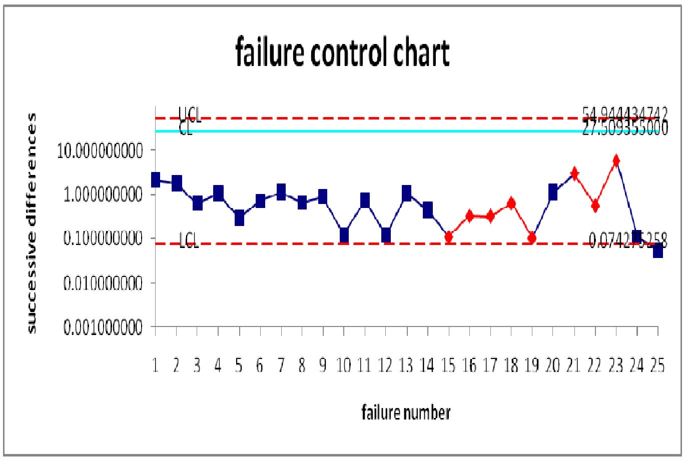


Fig 1. Mean Value Chart

A point falling below the control limit $\boldsymbol{m(t_l)}$ indicates an alarming signal. A point above the control limit $m(t_u)$ indicates the better quality. If the points are falling within the control limits it indicates that the software process is in stable. The mean value chart shows all the successive differences. No failure data fall outside $m(t_u)$. It does not indicate any alarm signal. The values of control limits are as follows.

$$m(t_U) = 54.944434742$$

$$m(t_C) = 27.509355000$$

$$m(t_L) = 0.074275258$$

By placing the failure cumulative data shown in table 2 on y axis and failure week on x axis and the values of control limits are placed on Mean Value chart, we obtained Figure2. The Mean Value chart shows that the 25th failure data has fallen below $m(T_l)$ which indicates the failure process is identified. It is significantly early detection of failures using Mean Value chart.

6. Conclusion

Software reliability is an important quality measure that quantifies the operational profile of computer systems. In this paper we proposed Pareto type II software reliability growth model. This model is primarily useful in estimating and monitoring software reliability, viewed as a measure of software quality. Equations to obtain the maximum likelihood estimates of the parameters based on time domain data are developed.

This analysis of NTDS data shows out of control signals i.e., below the LCL. We conclude that our method of estimation and the control chart are giving a +ve recommendation for their use in finding out preferable

By observing the Mean Value control chart we have identified that the failure situation is detected at 25th point of Table-2. Hence our proposed Mean Value Chart detects out of control situation. This is a simple method for model validation and is very convenient for practitioners of software reliability.

The early detection of software failure will improve the software reliability. The methodology adopted in this paper is better than the methodology adopted by Xie et al, [2002] .Therefore; we may conclude that this model is the best choice for an early detection of software failures.

7. Acknowledgements

Our thanks to Department of Computer Science and Engineering, Acharya Nagarjuna University for providing necessary details to carry out the research work.

8. References

[1] Kimura, M., Yamada, S., Osaki, S., 1995. ”Statistical Software reliability prediction and its applicability based on mean time between failures”. Mathematical and Computer Modelling Volume 22, Issues 10-12, Pages 149-155.

[2] MacGregor, J.F., Kourti, T., 1995. “Statistical process control of multivariate processes”. Control Engineering Practice Volume 3, Issue 3, March 1995, Pages 403-414 .

[3] Pham. H., 2006. “System software reliability”, Springer.

[4] Goel, A.L., Okumoto, K., 1979. Time-dependent error detection rate model for software reliability and other performance measures. IEEE Trans. Reliab. R-28, 206-211.

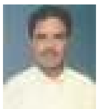
[5] Musa J.D, Software Reliability Engineering MCGraw-Hill, 1998.

[6] Wood, A(1996), “Predicting software Reliability”, IEEE Computer,2253-2264.

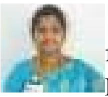
[7] Satya Prasad R and Geetha Rani N (2011), “Pareto type II Software Reliability Growth Model”. International Journal of Software Engineering, Volume 2, Issue(4) 81-86.

[8] Musa J.D., Iannino, A.,Okumoto, K.,1987. Software Reliability: Measurement Prediction application. MC Graw Hill, NewYork.

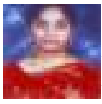
9. Authors Profile



Dr. R. Satya Prasad received Ph.D.degree in computer science in the faculty of Engineering in 2007 from Acharya Nagarjuna University, Andhra Pradesh. He received gold medal from Acharya Nagarjuna University for his outstanding performance in master’s degree. He is currently working as Associate Professor in the department of Computer Science &Engg., Acharya Nagarjuna University. His current research is focused on Software engineering. He has published several papers in National & International Journals.



Mrs. K.Sita kumari received M.Sc degree from Acharya Nagarjuna University, Guntur and M.Tech degree from Dr.MGR university, Chennai. She is currently pursuing her Ph.D from Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh , India. Presently she is working as an Associate Professor in the department of Information Technology , V.R.Siddhartha Engineering College, Kanuru, Vijayawada.



Mrs. G. Sridevi received M.Sc. and M.Tech degree from Acharya Nagarjuna University. She is currently pursing Ph.D at Department of Computer Science and Engineering ,Acharya Nagarjuna University, Guntur, Andhra Pradesh, India. She is currently working as a Vice-Principal and Associate professor in the Department of Computer Science, Nimra Women’s College of Engineering, Jupudi, Ibrahimpatnam, Vijayawada, Andhra Pradesh. Her research interests lies in Data Mining and Software Engineering.

A Survey on Software Testing Techniques using Genetic Algorithm

Chayanika Sharma[1], Sangeeta Sabharwal[2], Ritu Sibal[3]
Department of computer Science and Information Technology, University of Delhi, Netaji Subhas Institute of Technology
Azad Hind Fauz Marg, Dwarka, Sector -3, New Delhi - 110078, India

Abstract

The overall aim of the software industry is to ensure delivery of high quality software to the end user. To ensure high quality software, it is required to test software. Testing ensures that software meets user specifications and requirements. However, the field of software testing has a number of underlying issues like effective generation of test cases, prioritisation of test cases etc which need to be tackled. These issues demand on effort, time and cost of the testing. Different techniques and methodologies have been proposed for taking care of these issues. Use of evolutionary algorithms for automatic test generation has been an area of interest for many researchers. Genetic Algorithm (GA) is one such form of evolutionary algorithms. In this research paper, we present a survey of GA approach for addressing the various issues encountered during software testing.

Keywords: *Software testing, Genetic Algorithm*

1. Introduction

Testing is primarily done on software as well as in web for testing client and server architecture. Software testing is one of the major and primary techniques for achieving high quality software. Software testing is done to detect presence of faults, which cause software failure. However, software testing is a time consuming and expensive task [29], [20], [28]. It consumes almost 50% of the software system development resources [3], [20]. Software testing can also be defined as process of verifying and validating software to ensure that software meets the technical as well as business requirements as expected [16].

Verification is done to ensure that the software meets specification and is close to structural testing whereas validation is close to the functional testing and is done by executing software under test (SUT) [18]. Broadly, testing techniques include functional (black box) and structural (white box) testing. Functional testing is based on functional requirements whereas structural testing is done on code itself [13] [10] [24]. Gray box testing is hybrid of white box testing and black box testing [8].

Testing can be done either manually or automatically by using testing tools. It is found that automated software testing is better than manual testing. However, very few test data generation tools are commercially available today [14]. Various techniques have been proposed for generating test data or test cases automatically. Recently, lot of work is being done for test cases generation using soft computing techniques like fuzzy logic, neural networks, GA, genetic programming and evolutionary computation providing keys to the problem areas of software testing.

Evolutionary testing is an emerging methodology for automatically producing high quality test data [10]. GA is well known form of the evolutionary algorithms conceived by John Holland in United States during late sixties [6] [25]. In [21], evolutionary black box testing is also applied on embedded systems to test its functional and non-functional properties. GA has been applied in many optimization problems for generating test plans for functionality testing, feasible test cases and in many other areas [5] [15]. GA has also been used in model based test case generation [3] [23] [26] [27]. In object oriented unit testing as well as in the black box testing, GA is used for automatic generation of test cases [23], [10], [15]. Concerning testing of web applications, many tools, new techniques and methods have been developed to address issues like maintainability, testability, security, performance, correctness and reliability of web application [8]. Web applications are composed of web pages and components and interaction between them executes web servers, HTTP, browser (the client side) and networks. A web page is information viewed on the client side in a single browser window [16]. In [30], user session data of web application is used to generate test cases by applying GA.

In this research paper, a survey of different software testing techniques where GA is efficiently used is presented. This paper is divided into 4 sections. Section 2 describes briefly the working of a GA. In section 3, applications of GA in different types of software testing is described. Section 4 concludes the paper and gives an overview of our future work.

2. Genetic Algorithm: A Brief Introduction

In the past, evolutionary algorithms have been applied in many real life problems. GA is one such evolutionary algorithm. GA has emerged as a practical, robust optimization technique and search method. A GA is a search algorithm that is inspired by the way nature evolves species using natural selection of the fittest individuals.

The possible solutions to problem being solved are represented by a population of chromosomes. A chromosome is a string of binary digits and each digit that makes up a chromosome is called a gene. This initial population can be totally random or can be created manually using processes such as greedy algorithm. The pseudo code of a basic algorithm for GA is as follows [6]:-

```
Initialize (population)
Evaluate (population)
While (stopping condition not satisfied)
{
Selection (population)
Crossover (population)
Mutate (population)
Evaluate (population)
}
```

GA uses three operators on its population which are described below:-

- **Selection**: A selection scheme is applied to determine how individuals are chosen for mating based on their fitness. Fitness can be defined as a capability of an individual to survive and reproduce in an environment. Selection generates the new population from the old one, thus starting a new generation. Each chromosome is evaluated in present generation to determine its fitness value. This fitness value is used to select the better chromosomes from the population for the next generation.

- **Crossover or Recombination**: After selection, the crossover operation is applied to the selected chromosomes. It involves swapping of genes or sequence of bits in the string between two individuals. This process is repeated with different parent individuals until the next generation has enough individuals. After crossover, the mutation operator is applied to a randomly selected subset of the population.

- **Mutation**: Mutation alters chromosomes in small ways to introduce new good traits. It is applied to bring diversity in the population.

3. Using Genetic Algorithm in Software Testing

In this section we will discuss in detail the applications of GA in different areas of testing like test planning [5], minimization of test cases in regression testing [11], model based testing [3] [23] [26] [27] and web testing [30].

3.1 Applications of GA in White Box Testing

Structural testing can be done in the form of data flow testing or path testing. Path testing involves generating a set of paths that will cover every branch in the program and finding the set of test cases that will execute every path in this set of program path [16] [18]. In data flow testing, the focus is on the points at which variables receive values and the points at which these values are used [2]. Next, we will discuss briefly some of the research work regarding the applications of GA in white box testing.

3.1.1 Data Flow Testing

M.R. Girgis

Girgis [7] has proposed a structural oriented automatic test data generation technique that uses a GA guided by the data flow dependencies in the program to fulfil the all-uses criterion. The program to be tested is converted into a Control Flow Graph (CFG) where each node represents a block in a program and the edges of the flow graph depict the control flow of the statements.

Variables in a program under test are divided into 'c-uses' and 'p-uses' variables. c-uses variables are those which are used in computations or as a predicates in a program whereas p-uses variables are associated with edges of the flow graph. In order to fulfil the all-uses criteria, the def-clear path (a path containing no new definition of a current variable) from each definition of a variable to each use of that variable need to be determined. To find out the set of paths satisfying all-uses criteria, it is necessary to determine def c-use(dcu) and def p-use(dpu) of a variable i.e. the def- clear paths to their c-use at node i and def-clear paths to their p-use at edge (i, j).

Using the location of a variable defs and uses in a program under test, combined with the 'Basic state reach algorithm', the sets dcu(i)and dpu(i,j) are determined. From the 'Basic state reach algorithm' two sets reach (i) and avail (i) are determined where reach (i) is the set of all variable defs that "reach" node i and set avail (i) is the set of all "available" variable defs at node i i.e. the union of set of global defs at node i and the set of all defs that reach this node.

$$dcu(i):\ \ reach(i)\cap c-use(i) \tag{1}$$

$$dpu(i,j):\ \ avail(i)\cap p-use(i,j) \tag{2}$$

List of all dcu and dpu sets in the procedure calls satisfying the all-uses criterion are determined along with killing nodes (nodes containing other definition of a variable in a current path) that must not be included in the current path. In this approach, GA accepts instrumented version of the program under test, the list of def-use sets to be covered, the number of input variables, and the domain and the precision of each input variables as an input. A binary vector is used to represent a chromosome. The length of the input is determined by the domain and the precision. The domain is represented by $D_i = [a_i, b_i]$ where each variable in a program takes values from the range $[a_i, b_i]$. Each domain D_i should be cut into $(b_{i\_}a_i).10^{d_i}$ equal size ranges if decimal places d_i is desired for a variable to achieve precision. If m_i is an integer denoting the length of a chromosome or a string such that $(b_i - a_i).10^{d_i} \leq 2^{m_i} - 1$ then a binary string denoting a variable of length m_i fulfil the precision requirement. The mapping from the binary string i, into a real number from the range $[a_i, b_i]$ is performed by the following formula:-

$$x_i \quad a_i + x_i'.\frac{b_i - a_i}{2^{m_i} - 1} \tag{3}$$

Where, $x_i^{'}$ represents the decimal value of the binary string i.

$$x_i \quad a_i + \text{int}(x_i'.\frac{b_i - a_i}{2^{m_i} - 1}) \tag{4}$$

By applying $d_i = 0$, the above formula can be used to map binary string i into an integer number from the range $[a_i, b_i]$. Each chromosome represents a test case for a program which is represented by a binary string of specified length. Each chromosome is then represented by a decimal number by using (3) or (4).

The fitness value *eval* (v_i) for each chromosome v_i (i = 1...., *pop_size*) is calculated as follows:-

$$eval\ (v_i) = \frac{no.of\ def - use\ paths\ covered\ by\ v_i}{total\ no.of\ def - use\ paths} \tag{5}$$

A test case, v_i is effective if its fitness value *eval* $(v_i) > 0$. Each test case or chromosome is evaluated and the program is executed to record the def – use paths in the program that are covered by the test cases as its input.

All the test cases are selected with effective eval (v_i) or good fitness value. Selection is done by roulette wheel selection and proposed random selection method. The effective test cases then become parents of the new population. If none is effective then all the individuals are chosen as the parents.

Discussion

The test case generation by the proposed GA is more effective as compared to the random testing technique. The proposed selection method generates better results than the roulette wheel selection method. However, the proposed selection method has not been evaluated and compared with other selection methods like stochastic, uniform and tournament selection methods.

3.1.2 Path Testing

P.R Srivastava et al.

In [20], P.R. Srivastava and Tai have presented a method for optimizing software testing efficiency by identifying the most critical path clusters in a program. The SUT is converted into a CFG. Weights are assigned to the edges of the CFG by applying 80-20 rule. 80 percentage of weight of incoming credit is given to loops and branches and the remaining 20 percentage of incoming credit is given to the edges in sequential path. The summation of weights along the edges comprising a path determines criticality of path. Higher the summation more critical is path and therefore must be tested before other paths. In this way by identifying most critical paths that must be tested first, testing efficiency is increased.

Another test generation approach proposed by P.R Srivastava is based on path coverage testing [19]. The test data is generated for Resource Request algorithm using Ant Colony Optimization algorithm (ACO) and GA. *Resource request algorithm* is deadlock avoidance algorithm used for resource allocation by operating system to the processes in execution cycle [10]. The ACO algorithm is inspired from behaviour of real ants where ants find closest possible route to a food source or destination. The ants generate chemical substance called pheromones which helps ants to follow the path. The pheromone content increases as more ants follow the trail. The possible paths of CFG are generated having maximum number of nodes. Using ACO, optimized path ensuring safety sequence in resource request algorithm is generated covering all edges of CFG.

Using GA, suitable test data set is generated which covers the need for each process. The backbone of genetic process is the fitness function which counts number of times a particular data enters and continues the resource request algorithm. Higher the value of count, higher is chances of avoiding a deadlock. The test data with higher values of count is taken and genetic crossover and mutation is applied to yield better results. Simultaneously, poor test data is removed each time.

Discussion

The experimental results shows that success rate of ACO are much better than GA. In weighted CFG approach [20], experiments were done on small examples and need to be done on larger commercial examples. Moreover, method can be further improved.

Dr. Velur et al.

The approach for test cases generation from directed graph has been proposed by Dr. Velur [29]. In this work, directed graph of intermediate states of system under test is created to exhibit the expected behaviour of system as shown in Fig 1.

A directed graph is represented by G = [V, E], where V represents state or vertices and edges represents flow of control. Thereafter, a graph containing n nodes is represented by an incidence matrix of order n * n where an entry '1' in the matrix represents edge between nodes and '0' represents no edge or connection between them (see Fig 1). By using the nodes of graph as the base population, pair of nodes are generated which are selected as parents by applying the dual graph generation technique proposed as shown in Fig 2.

The population is initialized by random selection of graphs of size 43 and 250 individuals are generated. The tournament selection method is used, where two individuals are chosen randomly and individual with the maximum fitness is chosen for crossover. Fitness is calculated by using the '*Current maximum clique algorithm*' and *Approximation algorithm*'. Fitness is assigned by finding the clique of size 5 and the number of independent sets of size 5 in the population which comprise of number of graphs in the population. The graph with 0 fitness value indicates the clique of size 0 and no independent sets of size 5 in the graph. Nodes which are already visited are discarded and GA cycle continues till all the nodes of the graph are visited once.

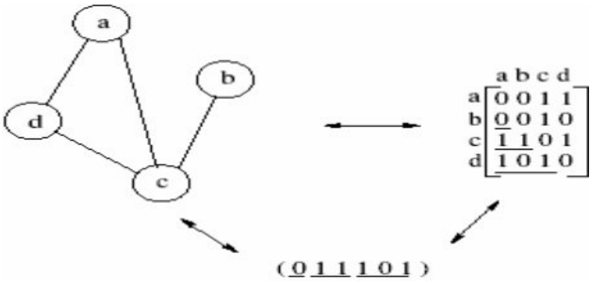


Fig. 1 Graph representation as a binary string [29]

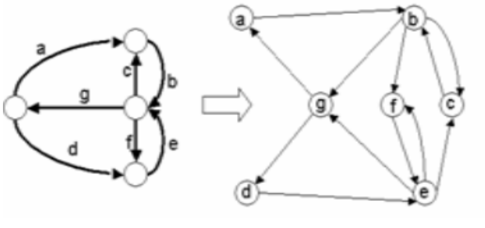


Fig. 2 Dual graph generation [29]

The graph is first converted into a binary string as shown in Fig 1. Next, the arcs of an original graph are converted into nodes as shown in Fig 2. For example, if an edge1 is an incoming to some node and the edge 2 is outgoing edge for the same node then an edge is created from edge 1 to edge 2 which acts as nodes in corresponding dual graph. The dual graph is then eulerized by duplicating the arcs for balancing the node polarities. As dual graph is traversed, all possible two links combination in dual graph for example bc, bf, bg ... are noted down. All the dual combinations are then encoded in 0 and 1 format as genetic population.

Discussion

This technique will be more suitable for network testing and system testing where predictive model based tests are not optimized to generate the outputs. The approach uses tournament selection method only. This approach has not been compared with new proposed approaches for generating test cases from the CFG. Moreover, the effectiveness of the fitness evaluation criteria has not been justified.

Maha Alzabidi et al.

Maha Alzabidi [14] has proposed automatic structural test case design using evolutionary testing. Software structural testing is done by taking path coverage for testing. For path testing, CFG is used in their work for representing a program where the nodes are the basic blocks and the edges between the nodes indicate the flow of the program. Meaningful paths are extracted from CFG and are selected as a target path for testing. Test cases are generated to trace the new path which leads to the target path. The test result is evaluated to determine that the testing objective criteria are satisfied by executing the selected path.

The fitness function is named as a Shifted–Modified-Similarity (SMS) which is a modification to the hamming distance. The symmetric difference or hamming distance is calculated for cascading edges for target path and current path. Similarities are then normalized and summed associated with a weighting factor. This value is used as an objective function to evaluate the individuals in the population.

Discussion

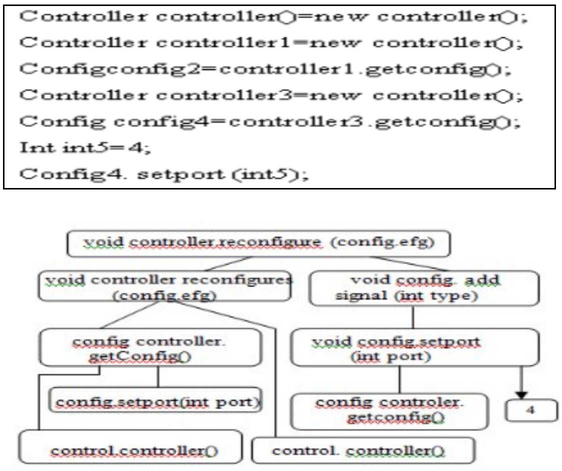
The approach improves the fitness function. Performance of different GA parameters is studied in this paper. Parameters have been applied on different test programs and results have shown that double crossover is more effective than single crossover applied on a test program. The approach is applied on the small programs but has not been evaluated on the complex programs involving loops, arrays and linked lists using different data types.

Jose Carlos et al.

The structural testing of object oriented programs requires traversing the complex control flow paths, resulting in the complex test cases generation which defines the elaborate state scenarios. Jose has proposed the methodology for evaluating the quality of both feasible and unfeasible test cases for structural oriented unit testing of object oriented java programs [10]. The test cases that are terminated with a call to a method and are completely executed are termed as feasible test cases whereas the test cases which abort prematurely are termed as unfeasible test cases. In this work, the test cases are represented as strongly typed genetic programming (STGP) individuals where each individual contains number of STGP trees equal to the number of arguments in the current method or method under test (MUT). The tree is traversed by depth first traversal algorithm which generates the sequence of method calls or scenarios.

Traversing the trees by depth first traversal generates a linear sequence of computations or method call sequence (MCS) of MUT which is represented by a CFG as shown in Fig 3. The nodes in a CFG are assigned the weights and the fitness of test cases is computed. The fitness of feasible test case is computed on the basis of number of times a particular CFG node was exercised by the test cases of previous generations whereas unfeasible test cases are measured by the method calls that threw the exceptions or distance between the runtime exception indexes which results in prematurely termination of a test case.The CFG nodes weights are evaluated at the beginning of every generation according to formula proposed by Carlos:-

$$W_{ni} \quad (\alpha W_{ni})(\frac{hitC_{ni}}{|T|}+1)(\frac{\sum_{x\varepsilon} N_s^{ni} W_x}{|N_s^{ni}| * \frac{W_{init}}{2}}) \qquad (6)$$



```
Controller controller0=new controller0;
Controller controller1=new controller0;
Configconfig2=controller1.getconfig0;
Controller controller3=new controller0;
Config config4=controller3.getconfig0;
Int int5=4;
Config4. setport (int5);
```

void controller.reconfigure (config.efg)
void controller reconfigure (config.efg)
void config. add signal (int type)
config controller. getConfig()
void config.setport (int port)
config.setport(int port)
config controler. getconfig()
4
control.controller()
control. controller()

Fig. 3 Example of a STGP tree and corresponding MCS [10]

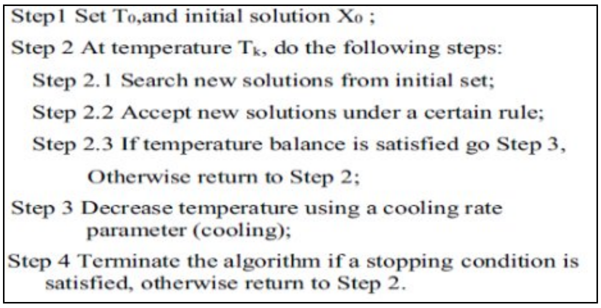
Where the hitC$_{ni}$ parameter is the "Hit count" representing the number of times a CFG node was traversed by the test cases of the previous versions, T represents the set of test cases in previous generation and α represents weight decrease constant value which ranges from 0 to 1. With this approach unfeasible test cases are considered at certain stages of evolutionary testing, thereby enhancing the diversity and full structural coverage.

Discussion

The weights of the CFG are dynamically revaluated each generation. The technique finds a good balance between intensification and diversification of the search by fine tuning the evolutionary operators.

Bo Zhang and Chen Wang

Bo zhang and Chen wang [2] used simulated anneal algorithm into GA to generate test data for path testing. A simulated annealing algorithm is inspired by the annealing of metals. In this method, solid is heated from high temperature and then cooled down slowly to maintain thermodynamic equilibrium of system.



```
Step1 Set T0,and initial solution X0 ;
Step 2 At temperature Tk, do the following steps:
  Step 2.1 Search new solutions from initial set;
  Step 2.2 Accept new solutions under a certain rule;
  Step 2.3 If temperature balance is satisfied go Step 3,
        Otherwise return to Step 2;
Step 3 Decrease temperature using a cooling rate
        parameter (cooling);
Step 4 Terminate the algorithm if a stopping condition is
     satisfied, otherwise return to Step 2.
```

Fig. 4 Simulated Annealing Algorithm [2]

The steps in simulated annealing algorithm are shown in Fig 4. The Adaptive Genetic Simulated Annealing Algorithm (AGSAA) is proposed by Zhang to automatically generate test data. The CFG is used for path coverage testing. The following section shows the fitness function, crossover, mutation and other modifications applied in the GA procedure and elements by Bo Zhang.

Fitness Function: - The fitness function used is named SIMILARITY proposed by Lin and Yeh [9]. The SIMILARITY function is modification to hamming distance. It is used to get distance between two paths. The hamming distance is derived from the symmetric difference in set theory. As stated in [9], symmetric difference between two sets α and β is denoted by $\alpha \oplus \beta$. The symmetric difference between two sets α and β is the set containing the elements either in α or β but not in both. The Bo Zhang expressed the SIMILARITY as stated in equation 7.

$$\text{SIMILARITY}_{i\text{-}j}=M^1_{i\text{-}j}\times W_1+M^2_{i\text{-}j}\times W_2+\ldots+M^n_{i\text{-}j}\times W_n$$

Where,

$$M^n_{i\text{-}j}=1-|S^n_i\oplus S^n_j|/|S^n_i\cup S^n_j|$$

$$W_n=W_n\times|S^{n-1}_i|$$

$$S^n_i\oplus S^n_j=(S^n_i\cup S^n_j)-(S^n_i\cap S^n_j)$$

$$W_1=1 \tag{7}$$

Using the target path and the current path in the CFG, the fitness function is calculated.

Adaptive Selection: - The adaptive power selection strategy is used by Bo Zhang and Chen Wang. The power function used is shown as:-

$$P_s=\{Y_i=X_j(t)\}=n(X_j(t))J^{a(t)}(X_j(t))/\sum_{K=1}^{N}J^{a(t)}(X_k(t)) \tag{8}$$

Where, Y_1, Y_2,....., Y_n is new population. $P_s = X_j$ (t) is the probability of selection for X_j (t). N (X_j(t)) is number of X_j (t) in current population, J (X_j(t)) is fitness of X_j (t) , a(t) is a monotonously increasing sequence of positive real and ∑J (X_k(t)) is sum of all individuals in the current population.

Elitist Preservation:-The good individuals having good fitness value are protected from being modified.

Adaptive Crossover and Mutation: - In general, GA uses constant crossover and mutation probability but in this work crossover probability and mutation probability changes are according to the fitness value. If fitness value of parent is bigger than average fitness value, probability of crossover is smaller and the parent will be protected from being modified whereas when fitness value is smaller than average fitness value, probability of crossover is large and parent will be died out. For random number r where r Є [0, 1], if r < crossover probability, then parent individuals are selected for crossover. In mutation, the good fitness chromosomes are considered to have smaller mutation probability while bad chromosomes have high mutation probability.

Simulated Annealing: -Simulated annealing is used to decide whether a chromosome is better than the original one and based on those criteria chromosome is accepted or rejected.

Discussion

As a case study, Zhang used triangle classification problem for the experiment. The target path is selected from the structural code. In the experiment, initial population size is 100 individuals. Maximum number of generations is 20. The results show that AGSAA performs better than GA in terms of covering the objective path quickly and the rate of coverage.

3.2 Applications of GA in Black Box Testing

In functional or black box testing, a program is considered to be a function that maps values from its input domain to values in the output range [18]. In other words, black box testing first concentrates on test to pass and then test to fail. This section describes work on black box testing for test case generation using GA.

3.2.1 Functional Testing

Francisca Eanuelle et al.

Francisca Eanuelle [5] has presented GA based technique to generate good test plans for functionality testing in an unbiased manner to avoid the experts interference. The motivation behind this work is to prove that the GA is able to generate good test plan although the best sequence of test plan is unknown.

The test plan or test sequence totally relies on the experts or the people who understand the application well. The emphasis is given on the fact that "*an error in a program is not necessarily due to the last operation executed by the user but may have been due to a sequence of previously executed operations that leads an application in an inconsistent state*". In other words, as a sequence of operations is executed, the state of inconsistency is non-decreasing or a problem in a software application is directly proportional to the level of inconsistency of the state in which application is. In this work, the operation of large granularity has been chosen so that the sequence of operation that leads application to inconsistent state can be identified. The transitions of an operation l_i yield a new operation l_{i+1} which leads the system into a new state as shown in Table 1.

As shown in Table 1, the objective is to find the sequence of operations which leads the system in an inconsistent state. Fitness value for the Table 1 is calculated as:-

$$f_p \quad \sum_{i}^{k-1} t(l_i \rightarrow l_{i+1}) \tag{9}$$

Table 1: Representation of the assigned values for the inconsistency added by each transition for instance $t(l_2 \rightarrow l_3) = v_{2,3}$ [5]

	l_1	l_2	l_3	……	l_n
l_1	$v_{1,1}$	$v_{1,2}$	$v_{1,3}$	……	$v_{1,n}$
l_2	$v_{2,1}$	$v_{2,2}$	$v_{2,3}$	……	$v_{2,n}$
l_3	$v_{3,1}$	$v_{3,2}$	$v_{3,3}$	……	$v_{3,n}$
:	:	:	:		
l_n	$v_{n,1}$	$v_{n,2}$	$v_{n,3}$	……	$v_{n,n}$

Where p = l_1, l_2,...., l_k is a test plan or sequence of operations and t is a transition function for converting one operation l_i to the next operation l_{i+1} in a sequence or in a new state. Larger the value of fitness function, better the sequence is considered which is likely to take the application to an inconsistent state. The GA is applied on the table of size 30*30 with randomly generated transitions values as shown in Table 1. The results have shown that the GA improves the quality of the test plans.

Discussion

The fitness function is defined which is used to determine the inconsistency of the application. In this work, a test plan having highest contribution to the inconsistency of the application is considered as a good test plan. This approach can eliminate the bias from the plans generated by experts. A technique based on GA proposed by Francisca, generates good test plans in an unbiased way but this requires computer applications to be tested more thoroughly. The approach does not use the structure of the application or the program flow. Moreover, approach does not deal with data and focus is only at the level of macro operations in functional testing.

Ruilian Zhao et al.

In [22], Zhao used the neural network and GA for the functional testing. Neural network is used to create a model that can be taken as a function substitute for the SUT. The emphasis is given on the outputs which exhibit the important features of SUT than inputs. In that case, test cases should be generated from the output domain rather than input domain. The feed forward neural network and back propagation training algorithm is used for creating a model. Neural network is trained by simulating the SUT. The outputs generated from the created model are fed to the GA which is used to find the corresponding inputs so that automation of test cases generation from output domain is completed.

In this paper, inputs to the GA are the function model generated from neural network, number of input variables n, range of input variables i.e. upper [n] and lower [n], population size, maximum iteration number, goal output g, maximum fitness function f_{max}, crossover probability and mutation probability. The fitness function is defined as:-

$$f = \begin{cases} \frac{1}{|c-g|} & c \neq g \\ f_{\max} & |c-g| \leq 10^{-8} \end{cases} \tag{10}$$

Where c is the actual output and the g is the goal output of the SUT. Population is evaluated by applying GA operations such as reproduction, crossover and mutation. The current individuals generated are considered as test inputs if the fitness value reaches or exceeds f_{max}.

$$\begin{aligned} y_1 &= x_1 - r(x_2 - x_1) \\ y_2 &= x_1 + r(x_2 - x_1) \\ y_3 &= x_2 + r(x_2 - x_1) \\ y_4 &= (1-r)x_{\min} + rx_1 \\ y_5 &= (1-r)x_{\max} + rx_2 \end{aligned} \tag{11}$$

Zhao has proposed new strategy for the crossover operation as shown in equation 11. In equation 12, x_1 and x_2 are the chosen parent individuals and y_1, y_2, y_3, y_4, y_5 are new individuals by applying crossover operation and r is a random number generated in (0, 1).

The difference between goal output and actual output of SUT using neural network is used for calculating fitness value of the individuals in the population. If fitness value exceeds or reaches the maximum fitness value, then search stops and the current individual is taken as the test inputs for the corresponding outputs.

Discussion

The test cases are generated from the output domain. Results have shown that this approach can generate test cases from output domain with high efficiency. The experiments were conducted only on small programs. The effectiveness of this approach can be further evaluated on large size programs. The actual outputs of created model are closed to the correct outputs. To minimize this difference, fitness function can also improve.

3.2.2 Mutation Testing

Mark Last et. al.

Mark Last [13] used the fuzzy based extension of GA (FAexGA) approach for test case generation. The aim is to find minimal set of test cases that are likely to expose faults using mutated versions of the original program. In FAexGA approach, crossover probability varies according to the age intervals assigned during lifetime. The crossover probability of young and old individuals is assigned low while for other age interval this probability is high. The very young offsprings crossover probability is low thus enabling exploration capability. Old offsprings have also less crossover probability and eventually dying out would help avoiding a local optimum or premature convergence. On the other hand, middle-age offsprings are frequently used for crossover operation.

Fuzzy logic controller (FLC) is used for determining probability of crossover. The FLC state variables include the age and lifetime of chromosomes (parents). The emphasis of this work is on the exploration and exploitation of individuals. The fuzzification interface of FLC includes variables that determine the age of an offspring. FLC assigns every parent values Young or Middle-age or Old. These values determine the membership for each rule in FLC rule base. The fuzzification interface of FLC defines for each parents the truth value of being Young, middle-age and old as shown in Table 2.

Table 2: M. Last's fuzzy rule for crossover probability [13]

Parent 1 / Parent 2	"Young"	"Middle-age"	"old"
"Young"	Low	Medium	High
"Middle-age"	Medium	High	Medium
"old"	Low	Medium	Low

The fuzzy rule base used in this experiment is presented in Table2. Each cell defines a single fuzzy rule. For example, "If Parent 1 is *old* and Parent 2 is *old* then crossover probability is *Low*". The centre of gravity (COG) is used as a defuzzification method which computes crisp value for the crossover probability based on values of the linguistic labels as shown in Table 2.

The test cases relate to the inputs of tested software and are represented as a vector of binary or continuous values. The test cases are initialized randomly in the search space of possible input values. Genetic operators are applied and the test cases are evaluated based on the fault – exposing - capability using mutated versions of original program.

The Boolean expression composed of 100 Boolean attributes and three logical operators: AND, OR and NOT (correct expression) is taken as the case study. The expression was generated randomly and to define an evaluation function for each test case an erroneous expression is generated. The chromosomes are 1-dimensional binary strings of 100 bit length. The value of the evaluation function F is calculated as follows:-

$$F(T) = \begin{cases} 1, & if\ Eval\_Correct \neq Eval\_Erroneous(T) \\ 0, & respectively \end{cases} \qquad (12)$$

Where T is a 100 bit 1-dimensional binary chromosome representing a single test case and Eval_Correct (T) or Eval_Erroneous (T) are binary results of applying chromosome T to the correct or erroneous expression.

Discussion

FAexGA has not been evaluated on the real programs. Moreover, sophisticated and continuous evaluation functions need to be developed.

3.2.3 Regression Testing

Liang You and YanSheng Lu

In [11], redundant test cases in the regression test suite are deleted and the total running time of remaining test cases are minimized by applying GA. The satisfaction matrix S_{ij} is used to represent relationship between requirements and the test cases. The rows in the satisfaction matrix represent requirements and column represents test cases. $S_{ij} = 1$ represents that jth test case t_j satisfy the ith requirement else $S_{ij} = 0$.The time aware regression testing reduction problem is defined as:-

$$Minimize \sum_{j\ 1}^{n} C_j x_j \qquad (13)$$

The fitness function shown in equation 14 and represents total running time of remaining test cases after eliminating redundant test cases. C_j represents the running time of test case $T_{j,}$ x_j is the vector of the test case T_j. $x_j = 1$ represents that test case T_j exists in T_{min} and $x_j = 0$ represents that test case T_j does not exists in T_{min}. If $T_1 = \{t_1, t_3, t_5\}$, then Require $(T_1) = \{r_1, r_2, r_3, r_4\}$ = R = Require (T). Using greedy algorithm, $T_{min} = T_1 = \{t_1, t_3, {}_{t5}\}$. T_{min} is the minimal regression testing suite consisting of seven test cases i.e. T [7]. X = $\{x_1, x_2, \ldots, x_n\}$ is the reduction regression testing suite T_{min}. For example, X= {1, 0, 1, 0, 1, 0, 0} represents $T_{min} = \{t_1, t_3, t_5\}$. The X = $\{x_1, x_2, \ldots, x_n\}$ is used as a bit string to represent chromosome X. The repair operator is used to transform infeasible solution into feasible solution. The repair operator, crossover and mutation are same as CHU's [17] GA. In CHU's GA, uniform crossover operator is used and mutation rate equals to two bits per string.

$$R_1 = \left(1 - \frac{number\ of\ reduced\ test\ suite}{number\ of\ unreduced\ test\ suite}\right) * 100\%$$

$$R_2 = \left(1 - \frac{total\ running\ time\ of\ reduced\ test\ suite}{total\ running\ time\ of\ unreduced\ test\ suite}\right) * 100\%$$

(14)

In equation 14, R1 represents reduction rate of number of test cases and R2 represents reduction rate of total running time of test cases. The paper compares GA based reduction with vector based reduction on all test cases (VA).

Conclusion

Results show that GA is better than VA reduction strategy in 7 out of 8 case studies and the saving time is greater than saving size of reduced test cases i.e. R2 > R1. Thus, GA is an effective technique for minimizing test cases in a test suite.

3.2.4 Model Based Testing

Chartchai Doungsa et al.

State diagrams or state chart diagrams are used to help the developer better understand any complex functionality or depict the dynamic behaviour of the entire system, or a sub-system, or even a single object in a system. GA can be used to generate the test data using UML state chart diagram as described by Doungsa [3]. Sometimes after coding developers don't have time to test the software. Generating test cases from UML state chart diagram can solve this problem by generating them before the coding. Then the test cases can be generated as per the specifications of the software. Specifications can be in the form of UML diagrams, formal language specifications or natural language description.

Sequence of triggers for UML state diagram can be used as a chromosome. The sequence of triggers is an input for the state diagram which acts as test cases for a program to be tested. Each trigger is examined to check for the transitions which lead to a new state. Each trigger checks for state and transition coverage. If the trigger can generate new state from the current state then next trigger is checked. If the trigger in a sequence cannot generate a new state then tracing for the state coverage will be stopped and the state and the transition coverage are recorded without taking rest of the sequence to consider. As each trigger is traced, new states and transitions are recorded. The process continues for fixed number of generations until all the states and transitions are covered. Fitness of the chromosome is evaluated by using objective function as follows:

$$a W + b X + c Y + Z \tag{15}$$

Where, a, b, c are constant value and a = 0 when there is no guard condition in selected transition. W are a number of states in test cases where value of attribute in that state make guard condition to be true. X is a number of transitions which is covered by this test but have not been covered by previous test set. Y is a number of states that can be reached by test case to reachable transition source. Z is a number of state and path coverage for the test case. Test cases are selected based on their fitness function. Test case with best fitness value is selected as parents. Based on the fitness function the selection operator is used to apply crossover and mutation operator to the sequence of triggers. Crossover operator is then applied to the sequence of triggers. This operator then generates new states and transitions. After a new generation is created, UML state diagram is then executed again to check for new chromosomes.

Discussion

In this work, test cases are generated from UML state diagram so that test data can be generated before coding. The effectiveness of test cases generated from the proposed fitness function is not evaluated with other test case generation techniques from the UML.

Sangeeta Sabharwal et. al.

In this work, software testing efficiency is optimized by identifying critical path clusters [26]. The test case scenarios are derived from activity diagram. The activity diagram is converted into CFG where each node represents an activity and the edges of the flow graph depict the control flow of the activities. Path testing involves generating a set of paths that will cover every branch in the program and finding the set of test case scenarios that will traverse every activity in these scenarios. It may be very tedious expensive and time consuming to achieve this goal due to various reasons. For example, there can exist infinite paths when a CFG has loops.

In this approach, critical path are identified that must be tested first using the concept of information flow (IF) metric and GA. The IF metric is adopted in this work for calculating the IF complexity associated with each node of the activity diagram. According to the basic IF model, IF metric are applied to the components of system design. In this work, the component is taken as a node in the CFG. The IF is calculated for each node of CFG. For example, IF of node A i.e. IF (A) is calculated using equation given below:-

$$IF(A) \quad [FANIN(A) \times FANOUT(A)] \tag{16}$$

Where FANIN (A) is a count of number of other nodes that can call, or pass control to node A and FANOUT (A) is a number of nodes that are called by node A. IF is calculated for each node of a CFG. The weighted nodes in the path are summed together and the complexity of each path is calculated.

In [27], to take care of the software requirements change and to improve software testing efficiency a stack based approach is adopted for assigning weights to the nodes of an activity diagram and state chart diagram. The nodes of CFG and intermediate graph of state chart diagram i.e. state dependency graph (SDG) are prioritized using stack based memory allocation approach and IF metrics.

In the stack based memory allocation approach, data or info is pushed or popped only at one end called top of stack. The stack uses last in first out (LIFO) approach. Node pushed first is removed last from the stack. The top is incremented when node is inserted and decremented when node is deleted. In the proposed technique, each node of CFG or SDG is assigned a weight w based on number of operations to access element in the stack. To access or modify the node (data), all the data above it is popped. Higher the number of operations required to access the node, higher is the weight or complexity of the node. If the weight of the node or number of operations to access the node increases, the cost of modifying the node also increases.

The IF is calculated for each node of a CFG and SDG. The weighted nodes in the path are summed together and the complexity of each path is calculated. Therefore, the sum of the weight of a node by stack based weight assignment approach and IF complexity contributes to the total weight of a node of CFG and SDG. The fitness value of each chromosome is calculated by using the formula given below:-

$$F \quad \sum_{i \quad 1}^{n} w_i \qquad (17)$$

Where, w_i is weight of i^{th} node in a path under consideration and n is number of nodes in a current path. Weight of i^{th} node is the sum of IF complexity and stack based complexity given by equation given below.

$$w_i \quad IF(i) + STACKBASED\,WEIGHT\ (i) \qquad (18)$$

Discussion

An approach is proposed for identifying the test path that must be tested first. Test paths or scenarios are derived from activity diagram. The approach makes use of IF model and GA to find path to be tested first.

3.2.5 Usage Testing

Robert M. Patton et al.

Robert M. Patton [23] has used the usage models that depict the usage scenarios of the system. They are used in test planning, to generate a sample of test cases that represent usage scenarios, and to test results. In system testing, determining the nature and the location of the errors can be difficult which later on can be the problem for the developers to fix the errors. The system testing contains only small information from the usage scenarios of the system. Due to the limited information, generalizing testing results could be difficult.

To solve these problems, GA accepts the domain data generated by the usage model and the results of system test as two inputs. A set of test cases are the initial population generated from a usage model. Each individual in the population represents a single test case. The two objectives were taken to determine the fitness of individuals. The first objective is 'Likelihood of occurrence' which represents the possible usage scenarios of what the user will do with the system. The second objective is 'failure intensity' which is the capability of an individual to exhibit failures or problems in the system. The individuals maximizing these two objectives are selected for mating.

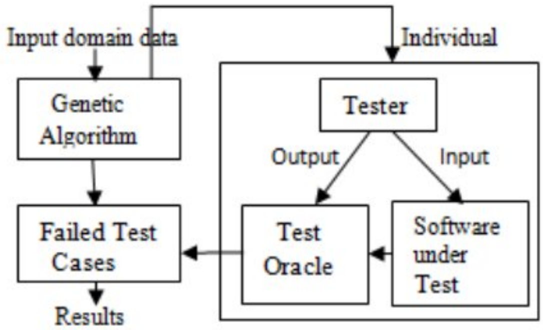


Fig. 5 R. Patton's GA approach to Focused software usage testing [23]

As shown in Fig 5, each individual represents a single test case and is sent to the tester and is then applied to SUT. The SUT processes this input and generates the output that is later analysed by the Test Oracle. The Test Oracle will then determine if the output is correct or incorrect or if the SUT failed or crashed. The GA uses this result along with the likelihood that it would occur as defined by the usage model for determining the overall fitness of the individual.

Discussion

The R. Patton's strategy shows that GA helps to identify failures that are more severe and are likely to cause faults in the software.

3.2.6 GUI testing

Abdul Rauf et. al.

In this work [1], GA is used to apply coverage criteria on GUI (Graphical User Interface). GUI is event based testing where the test cases consist of GUI events. GA is used to find optimized test suite for GUI testing. GUI testing is divided into three phases:-

1. Test Data Generation
2. Path Coverage Analysis
3. Optimization of Test Path

The Notepad, MS Word and Word Pad are used for the event based test data generation.

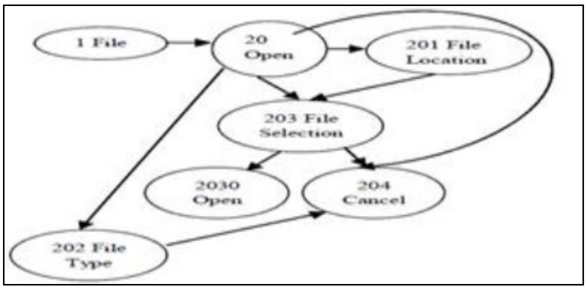


Fig. 6 Path generation for Open in Notepad [1]

In Fig 6, events generation while opening a file is shown. The nodes represent the objects in a notepad like File, Print, Edit and the sequence of operations between the nodes is shown as path between them. The test paths are optimized using GA. The size of chromosome in GA is measured as maximum length of test path in a Notepad. The fitness function is defined as number of paths covered by chromosome shown in equation 19.

$$Accuracy \quad \frac{TestPathsCoveredbyChromosome}{TotalNumberofChromosome} \qquad (19)$$

By applying crossover and mutation operator, new off springs are obtained having higher fitness value.

Discussion

The results show 85% coverage after 500 generation. Since the probability of crossover and mutation is unknown, the results might be superior if the crossover and mutation rate is changed.

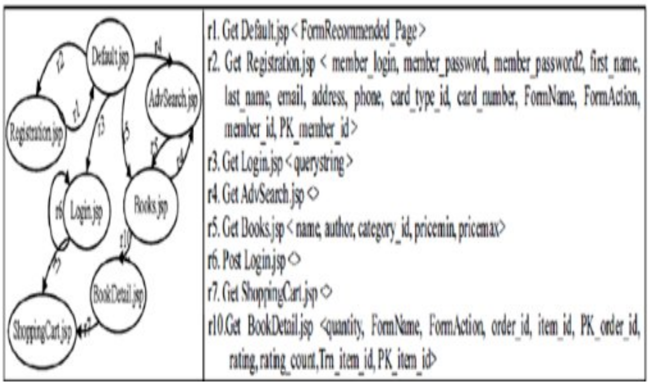
3.3 Applications of GA in Gray box testing

In gray box testing, test cases are designed using both black box testing and white box testing. The software is tested against its specification but using some knowledge of internal working [8] [30].

3.3.1 User Session Testing

Xuang Peng et. al.

In [30], Xuan Peng et. al. used the user session data in their request dependence graph (RDG) to generate test cases by applying GA. The structural analysis of web application is done using RDG construction. The request dependence is shown as the relationship between the components or nodes of the web application. The edge represents the request dependence between two pages. The request labelled in the graph is formatted as: - "GET/POST PAGE < P1, P2,......,Pn > where P1, P2,....., Pn are the possible parameters in the request. All the parameter- value pair of requests is enumerated without values. The RDG and their corresponding labelling are shown in Fig 7.



r1. Get Default.jsp < FormRecommended_Page >
r2. Get Registration.jsp < member_login, member_password, member_password2, first_name, last_name, email, address, phone, card_type_id, card_number, FormName, FormAction, member_id, PK_member_id >
r3. Get Login.jsp < querystring >
r4. Get AdvSearch.jsp <>
r5. Get Books.jsp < name, author, category_id, pricemin, pricemax>
r6. Post Login.jsp <>
r7. Get ShoppingCart.jsp <>
r10.Get BookDetail.jsp <quantity, FormName, FormAction, order_id, item_id, PK_order_id, rating, rating_count, Trn_item_id, PK_item_id>

Fig. 7 Partial Request Dependence Graph and the Labelled Requests [30]

The test cases should cover as many relationships as possible in RDG. The user session data is taken as the initial population in GA. A gene is encoded as combination of requests and pages. User session is identified as a sequence of requests and parameter values to describe the user's requests for web services. Each session is represented as transition relation. e.g. "Request -> Page -> Request -> ....->Page -> Request". A chromosome is encoded as a transition relationship between page and request. The fitness value is computed as: -

$$Fitness \quad (\propto *|CDTR|+|CLTR|)/(\propto *|DTR|+|LTR|) \qquad (20)$$

Where, CDTR is number of data dependence transitions covered in the chromosome, CLTR is number of link dependence transitions covered in the chromosome. DTR is number of data dependence and LTR is number of link dependence transition relations in the web application.

Discussion

Results show that user session (US) – RDG performed well for web application testing. Test suite reduction and fault detection results are quite satisfactory.

A comparative study of all existing techniques discussed in our work is shown in Table 3 which shows the values of GA parameter used in different types of software testing.

Table 3: GA Parameters used in different types of software testing (C.R = Crossover Rate, M. R = Mutation Rate, N.O.G = Number of Generation)

Author	Testing Technique	C. R	M. R	Crossover Method	Mutation Method	Selection	Initial Pop. size	N.O.G	Encoding	Result
D. J Berndt, 2005	High Volume Testing	--	--	--	--	--	--	500	Real numbers	GA is being mined with decision tree to generate test cases.To explore the GA, an autonomic vehicle simulation is being developed. Robots are built with limited resources and seeded with errors. On long term execution, robot simulation shows the promising results in the high volume testing.
Doungsa	Gray-box testing	0.4	0.3	Single one point	Random	Random	6	2	Value encoding	GA succesfully generates test data from the UML state diagram.
Dr. Velur Rajappa, 2008	System testing or network testing	50%	50%	Single one point	--	Tournament selection	250	--	Tree encoding	The solution is proposed using graph theory and genetic algorithm.
Francisca Emanuelle [35], 2006	Functional Testing	80%	1%	Point Crossover	Flip	Random	150	50	--	Test plans showing highest inconsistency of application.
Jose Carlos, 2008	Unit testing	0.1, 0.8, 0.33	0.1, 0.8, 0.34	Random point	> 7.5 average generations (Combination of C.R and M. R and r =[0.1, 0.8, 0.33]	Tournament selection	--	--	Tree encoding	Weighted CFG is used where the weights of CFG are dynamically revaluated to determine the qulaity of test cases.
Maha Alzabidi, 2009	Path testing	1.0 (Tri class program) and Max-Min Program)	0.005 (Tri class) and 0.05 for Max- Min program	Double point for Tri class program and Single -point for Max- Min	Flip	Random and roulette wheel	500 for Tri - class and 500 for Max-Min	50 for Tri - class and 10 for Max-Min	Binary	1. In double point there are good chances to double exchange the generation than single point crossover. 2. Mutation o.005 gave better results 3. Generation of next population according to their fitness value generates better offspring than random selection.
Mark Last, 2006	Black Box	Adaptive	0.01	One point	Flip	Random	100	200	Binary	FAexGA is efficient than simple GA and GA with varying population size in terms of probability of finding an error in tested software, faster rate of finding error and number of distinct solution.
Moheb R. Girgis, 2005	Data flow testing	0.8	0.15	--	Flip	Random selection and Roulette wheel selection	4	--	Binary	Propsed Random selection technique requies less number of generation to cover data flow dependencies than roulette wheel selection.
Nirmal Kumar Gupta, 2008	Unit testing	--	--	--	Random	Random or execution traces	--	--	Tree encoding	Test case generation for java classes and encoding and decoding of test program into changeable data structure.
P. R Srivastava, 2009	Path Testing	For r = [0,1], r < 0.8	For r = [0, 1], r < 0.3	Pairwise	Flip	Random	4	3	Binary	Increase in Testing efficiency by testing critical path in CFG
R. Krishnamoorthy , 2009	Regression testing	for r = [0,1], r < user value	1%	Random point	Random	Roulette Wheel	60	25	Real numbers	Using GA, the proposed time aware coverage based prioritization technique shows 120% improvement in APFD over other prioritization.
Robert M. Patton, 2003	Usage testing	--	--	One point	Random One - point	Fitness proportionate	100	30	Real numbers	GA is used in focused sofware usage testing by identifying the nature and locations of errors thereby improving the quality of software and efficiency of debugging activity.
Ruilian Zhao , 2008	Black Box	0.8	0.15	--	--	Roulette Wheel	--	> 500	Real	1. Improved GA is better than faster evolutionary speed 2. Approach can generate test cases with high efficiency.
Stefan Wappler, 2006	Unit Testing	--	--	Point crossover	Point mutation & Real Mutation	Stochastic universal Sampling & Tournament Selection	10	<10	--	Full branch coverage achieved for all test objects
Vahid Garousi, 2008	Stress Testing	50% & 70%	--	--	--	--	--	< 100 , for C.R = 50% and > 100 for C. R = 70 %	--	1. For C. R = 70%, fitness value increased by 80 % 2. GA can reach max. plateau even size of a component in SUT is large 3. Maximum search time delays the convergence across GA to find the best chromosome.

4. Conclusion and Future Work

In this paper, applications of GA in different types of software testing are discussed. The GA is also used with fuzzy as well as in the neural networks in different types of testing. It is found that by using GA, the results and the performance of testing can be improved. Our future work will involve applying GA for regression testing in web based applications. In future, we plan to use GA along with other soft computing techniques like fuzzy logic or neural networks for test case generation from UML diagrams. We also plan to use GA in integration testing for finding optimal test order.

References

[1] Abdul Rauf et. al., "Automated GUI test coverage analysis using GA", Seventh international conference on information technology. IEEE, 2010, pp. 1057-1063.

[2] Bo Zhang, Chen Wang, "Automatic generation of test data for path testing by adaptive genetic simulated annealing algorithm", IEEE, 2011, pp. 38 – 42.

[3] Chartchai Doungsa et. al., "An automatic test data generation from UML state diagram using genetic algorithm",http://eastwest.inf.brad.ac.uk/document/publication/Doungsa-ard-SKIMA.pdf, Accessed on 25.10.2012.

[4] D.J Berndt, A. Watkins, "High volume software testing using genetic algorithms", Proceedings of the 38t h International Conference on system sciences (9), IEEE, 2005, pp. 1- 9.

[5] Francisca Emanuelle et. al., "Using Genetic algorithms for test plans for functional testing", 44th ACM SE proceeding, 2006, pp. 140 - 145.

[6] Goldberg, D.E, Genetic Algorithms: in search, optimization and machine learning, Addison Wesley, M.A, 1989.

[7] Girgis, "Automatic test generation for data flow testing using a genetic algorithm", Journal of computer science, 11 (6), 2005, pp. 898 – 915.

[8] Giuseppe A. et. al., "Testing Web –applications: The State of Art and Future Trends". Information and Software Technology. Elsevier, 2006, pp. 1172-1186.

[9] Jin- Cherng Lin, Pu- Lin Yeh, "Automatic test data generation for path testing using Gas", International journal of information sciences. Elsevier, 2000, pp. 47- 64.

[10] Jose Carlos et. al., "A strategy for evaluating feasible and unfeasible test cases for the evolutionary testing of object-oriented software", AST' 08. ACM, 2008, http://www.cs.bham.ac.uk/~wbl/biblio/cache/http___jcbribeiro.googlepages.com_ast12-ribeiro.pdf, Accessed on 6.11.2012.

[11] Liang You, YanSheng Lu, "A genetic algorithm for the time – aware regression testing reduction problem", International conference on natural computation, IEEE, 2012, pp. 596 – 599.

[12] McMinn, "Search based software test generation: A survey", Software testing, Verification and reliability 14 (2), 2004, pp. 105-156.

[13] Mark Last et. al., "Effective black-box testing with genetic algorithms", Lecture notes in computer science, Springer, 2006, pp. 134 -148.

[14] Maha alzabidi et. al., "Automatic software structural testing by using evolutionary algorithms for test data generations", International Journal of Computer science and Network Security 9 (4), 2009, pp. 390 – 395.

[15] Nirmal Kumar Gupta, Dr. Mukesh Kumar Rohil, "Using Genetic algorithm for unit testing of object oriented software", First International conference on Emerging trends in Engineering and technology, 2008, pp. 308 - 313.

[16] Naresh Chauhan, Software Testing: Principles and Practices, Oxford University Press, 2010.

[17] P.C. Chu, J.E. Beasley, "A genetic algorithm for the multidimensional knapsack problem" Journal of heuristics, 1998, pp. 63- 86.

[18] Paul C. Jogersen, Software testing: A craftsman approach. 3rd edition, CRC presses, 2008.

[19] Praveen Ranjan Srivastava et. al., "Generation of test data using Meta heuristic approach" IEEE, 2008, pp.19 - 21.

[20] Praveen Ranjan Srivastava and Tai-hoon Kim, "Application of genetic algorithm in software testing" International Journal of software Engineering and its Applications, 3(4), 2009, pp.87 – 96.

[21] Peter M. Kruse et. al., "A Highly Configurable test systems for evolutionary black box testing of embedded systems" GECCO. ACM, 2009, pp.1545 – 1551.

[22] Ruilian zhao, shanshan lv, "Neural network based test cases generation using genetic algorithm" 13th IEEE international symposium on Pacific Rim dependable computing. IEEE, 2007, pp.97 - 100.

[23] Robert M .Patton et. al. "A genetic algorithm approach to focused software usage testing" Annals of software engineering,http://www.cs.ucf.edu/~ecl/papers/03.rmpatton.pdf. Accessed on 29. 07. 2012.

[24] Sthamer "The automatic generation of software test data using genetic algorithms". Phd thesis, university of Glamorgan, Pontyprid, wales, Great Britain, 1996.

[25] Stefan Wappler, Frank Lammermann, "Using evolutionary algorithms for unit testing of object oriented software" GECCO. ACM, 2005, pp.1925 - 1932.

[26] Sangeeta sabharwal et. al., "Prioritization of test case scenarios derived from activity diagram using genetic algorithm". ICCCT . IEEE, 2010, 481- 485.

[27] Sangeeta sabharwal et. al., "Applying genetic algorithm for prioritization of test case scenarios derived from UML diagrams" International journal of computer science issues 8 (3), 2011, 433 - 444.

[28] Timo Mantere, "Automatic software testing by Genetic Algorithms" Phd thesis, University of Vaasa, Finland, 2003.

[29] Velur Rajappa et. al., "Efficient software test case generation Using genetic algorithm based graph theory" International conference on emerging trends in Engineering and Technology, IEEE, 2008, pp. 298 - 303.

[30] Xuan Peng, Lu Lu, "A new approach for session - based test case generation by GA". IEEE, 2011, pp. 91-96.

An Insight into Spectrum Occupancy in Nigeria

Bara'u Gafai Najashi[1*], Feng Wenjiang[2] Choiabu Kadri[3]

[1]College of Communication Engineering, Chongqing University,
Chongqing, China, Postal Code 400044
174 Shazheng Street, Shapingba District, Chongqing

[2]College of Communication Engineering, Chongqing University,
Chongqing, China, Postal Code 400044
174 Shazheng Street, Shapingba District, Chongqing

[3]College of Communication Engineering, Chongqing University,
Chongqing, China, Postal Code 400044
174 Shazheng Street, Shapingba District, Chongqing

Abstract

The rapid evolution in wireless communication which has led to the development of several standards has also brought about a perceived spectrum scarcity. Studies have shown that contrary to popular belief concerning spectrum scarcity, most of the allocated spectrum is heavily underutilized. This has led to several spectrum occupancy measurements mostly in the US, Europe and recently Asia to ascertain the utilization level of the allocated spectrum. These measurements will help in determining which bands will be suitable for the deployment of cognitive radio technology. In this paper, an indoor spectrum occupancy measurement conducted within the region of 700 MHz to 2.5 GHz in Abuja, Nigeria is presented. The results obtained indicate that large portion of the allocated spectrum is underutilized which could be considered for the deployment of cognitive radio paradigm in the near future.

Keywords: *Cognitive radio, spectrum occupancy, wireless communication, Abuja*

1. Introduction

The rapid development of wireless standards and bandwidth hungry technologies has led to a perceived shortage of spectrum. In order to satisfy the growing demand for new spectrum, the spectrum management policy needs to be changed. Currently, the administrative or command and control approach to spectrum management has proven to be ineffective [10]. Several spectrum occupancy measurements carried out all over the world [2-9] mostly in USA and Europe with very few found in Africa have buttressed the fact that the assigned spectrum bands are underutilized. The idea of sharing the spectrum between a primary user (PU, the entity the spectrum was assigned to) and a secondary user (SU, the user that uses the spectrum opportunistically without interference)by using dynamic spectrum access techniques is what mainly cognitive radio is all about[7].The concept of Cognitive radio was first introduced by Mitola [1] in 1999. He described cognitive radio as a smart radio that has a full understanding of its environment and can also adapt its operating parameters to adapt to changes in the environment. The importance of spectrum occupancy measurements cannot be underestimated. The knowledge of the utilization level of the spectrum bands will help researchers to conceive (devise) spectrum models that could be used to predict future utilization. It could also help policy makers in determining which bands have low occupancy: this will go a long way in determining the bands that are suitable for dynamic spectrum access. Spectrum usage is dependent on location, time, measurement conditions, and equipment [2]. While spectrum monitoring tends to provide detailed information on conformity with laid down rules pertaining spectrum usage which a spectrum planner uses to ascertain the level of compliance and also to confirm the effectiveness of current planning system, spectrum measurements on the other hand tend to quantify the performance of the measuring band. It should be noted that the utilization level of a particular band cannot be extended to other bands or other locations [11]. In [3], Kishor Patil *et. al* conducted a spectrum occupancy measurements in the frequency band from 700MHz to 2700MHz in an outdoor scenario in suburban Mumbai, it was found that the spectrum occupancy of the entire band to be roughly 6.62%. In [12], a measurement campaign was performed in Guangdong province in china which also showed a low occupancy. Shared Spectrum Company has performed many measurement campaigns in several American cities which also showed low spectrum occupancy. [4] De Francisco, R et al. and Miguel Lopez-Benitez et. al performed spectrum occupancy measurement in Netherlands and Spain respectively. The results were similar to what was obtained in other campaigns. [5,6]. Other measurements conducted [7-9] also proved that the spectrum utilization level is low.

In this paper, a spectrum occupancy measurement was performed in the region of 700MHz-2.5GHz in a predominantly residential district of Abuja, Nigeria. The measurements were conducted indoors and the bands with low utilization levels were identified. Each band was monitored for 12 hours daily. From 9 am to 9 pm. The statistics obtained can also prove valuable to policy makers in managing this precious resource.

The remainder of the paper is organized as follows; in section II the equipment used in the measurement setup are described. In section III we shall present the results obtained from the measurements and analysis of those results. Section IV concludes this paper and suggests potential future research areas.

2. MATERIALS AND METHOD

2.1 Measurement Setup

The measurement setup used consists of an Aaronia AG HF-6060 V4 spectrum analyzer with a range of 10MHz-6GHz, an Aaronia AG OmniLOG 90200 antenna with a range of 700MHz to 2.5GHz, a laptop system that is connected to the spectrum analyzer via a USB cable, and an MCS software specially designed to run on Aaronia AG spectrum analyzers. The setup is connected as shown in figure 1. MATLAB software package was used to process and analyze the data and the results presented in later sections.

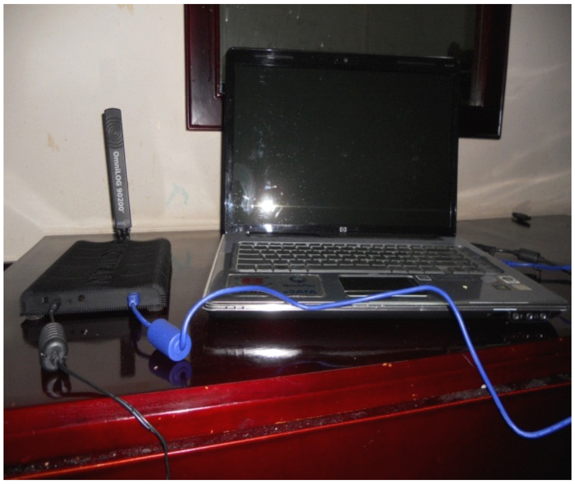


Figure 1: Setup used for measurement

2.2 Location

The measurement was conducted indoors at Gwarinpa District a primarily residential district in Abuja, Nigeria. These measurements were conducted indoors as part of larger measurement campaign which we hope will provide an insight into the utilization level of the spectrum in Nigeria. Abuja is located at with a population of about 770,000 as at 2006. The measurements were taken from 9 am to 9 pm a duration of roughly 12 hours. The spectrum analyzer settings employed in this work are given in the table 1

PARAMETER	VALUE
Span	200MHz/300 MHz
Timing	500 ms
Resolution(samples)	51
Bandwidth Filter	200 KHz
Video Filter	200KHz

Table 1: Spectrum Analyzer Parameters used in conducting the measurement

2.3 Decision Threshold

In spectrum occupancy measurements, determining the decision threshold upon which a particular channel can be deemed as free or busy is very important especially when energy detection is employed. In energy detection, no prior knowledge of the signal is known therefore it's very important to correctly determine the threshold for accurate readings. Setting the threshold metric too high will lead to under estimation of the spectrum while low decision metric will lead to over estimation of the spectrum. The normal convention is to keep the decision metric some certain dB above the equipment's noise floor level. The noise floor for the setup was obtained by replacing antenna with a 50ohm resistor we can similarly measure the noise floor by removing the antenna and not replacing it with anything[]. In this work, a threshold of -76 dBm was used after the two methods were tested.

2.3 Spectrum allocation in Nigeria

Nigeria can be said to be arguably the leading country in Africa as far as spectrum deregulation and licensing are concerned. Since 1992, over 350 broadcasting licenses have been issued by the Nigerian Broadcasting Corporation (NBC) and over 300 licenses issued in the telecommunication sector by the Nigerian Communication Commission (NCC) [16]. Policy formulation and management of spectrum in Nigeria is determined by the several bodies. These include National Frequency Management Council, the Ministry of Information and Communication, Nigerian Communication Commission, and the Nigerian Broadcasting Corporation:

National Frequency Management Council: The National Frequency Management Council (NFMC) is the apex body for radio frequency spectrum management in Nigeria. Established by Section 26 of the Nigerian Communications Act 2003 and located within the Ministry of Information & Communications, NFMC is the primary sponsor and influence on the Government's frequency spectrum policies and legislation. The Council is responsible for the planning, coordination and bulk trans-sectoral allocation of radio spectrum to the regulatory bodies, namely the National Communications Commission, the National Broadcasting Commission and the Ministry, and acts as the focal coordinator of all frequency spectrum activities in Nigeria. The Council also advises the Minister on Nigeria's representation at international and multi-lateral frequency spectrum bodies. NFMC is chaired by the Minister of Information & Communications and consists of high-level representatives of the Ministries of Aviation, Transport, Science & Technology, NCC, NBC and the Security Services, and meets at least four times in a year [17]

Nigerian Communications Commission: NCC is the regulator of the telecommunications industry and has wide discretionary powers to plan, manage, assign and monitor the use of spectrum by commercial users of telecommunications services. The roles of NCC also includes: the encouragement of competition; the removal of market entry barriers; interconnection of new operators with incumbents; the monitoring of tariffs and quality of service; the protection of consumer rights; and the overall promotion of affordable telecommunications services. The Commission develops and publishes radio frequency regulations and standards for the industry.

National Broadcasting Commission: The Commission derives its powers from the NBC Act 38 of 1992 as amended by the National Broadcasting Commission Act 55 of 1999 and is the sole body charged with regulating the broadcast industry, setting broadcast standards and upholding equity and fairness in broadcasting. NBC assigns broadcast frequencies it receives from NFMC to private & public radio & TV stations, monitoring for compliance with administrative procedures, the broadcast code and technical standards. NBC processes applications for the ownership of all types of radio and television stations and has licensed over 350 operational stations in several categories including private, public, satellite, network, campus and community radio & TV stations. The Commission regulates broadcasting through 27 state & zonal offices and regularly publishes updates of the radio frequencies it assigns on its website [16].

Ministry of Information & Communications: The Ministry, through the Department of Spectrum Management, is responsible for the formulation and monitoring of communications policies, international treaties and national representation in international organizations, including the International Telecommunication Union (ITU), International Civil Aviation Organization (ICAO), International Telecommunication Satellite Organization (ITSO), International Maritime Organization (IMO), among others. With the establishment and increased legislative empowerment of both the NCC and NBC, MoIC's function has gradually been limited to the management and assignment of frequencies to Government and non-commercial users including the military, security services, diplomatic missions, voluntary organizations and non-profit groups. The Ministry raises revenue for the Government through the sale of amateur radio communication license application forms, issuance and renewal of licenses, and type-approval testing of radio communication equipment. MoIC is the secretariat of NFMC and acts as the custodian of all frequencies in Nigeria [16].

BAND	FREQUENCY RANGE	USAGE AND PLAN
800 MHz	790 – 806	Trunk Radio Services
900 MHz	890-960 MHz	GSM
1 GHz	1.35-1.525,1.579-1.772 1.805-1.91,1.96-1.99GHz	Rural Telecoms, GSM, Oil coy. Satellite broadcast, radio navigation services.
2 GHz	1.99-2.11,2.2-2.285, 2.305-2.32,2.345-2.36, 2.4- 2.5GHz	3G mobiles ,wireless local loop, satellite up/downlink, scientific, industrial and medical applications

Table 2: Spectrum Allocation Table for some services in Nigeria.

RESULTS AND DISCUSSION

The first band considered was the 700-1000MHz. It comprises the 800MHz band used for trunk radio services, emergency services, CDMA (fixed), 900 MHz for GSM and also the 470-960 MHz for analogue television broadcasting. [13]. In the VHF band there are 12 channels where as the UHF band consists of 49 channels making a total of 61 channels [14]. This band has the highest utilization level experienced at 26% due to the activities of the analogue broadcasting (part of it to be precise) GSM operations and the radio trunk services. CDMA technology employs the use of spread spectrum where by the signal power is very low almost the same with noise power. This makes the signal difficult to detect by the spectrum analyzer. The utilization level could be much more than the obtained value due to this factor.

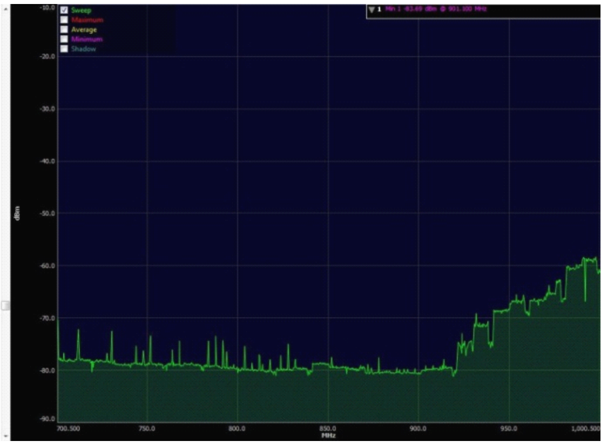


Figure 2: Power level in dBm versus frequency (700-1000 MHz)

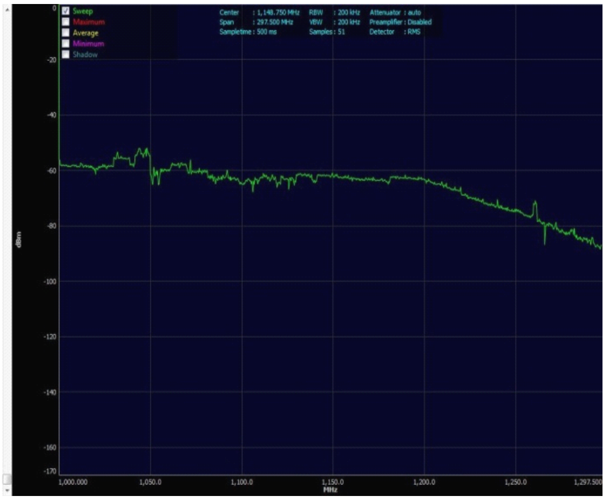


Figure 3: Power level in dBm versus frequency (1.1-1.3GHz)

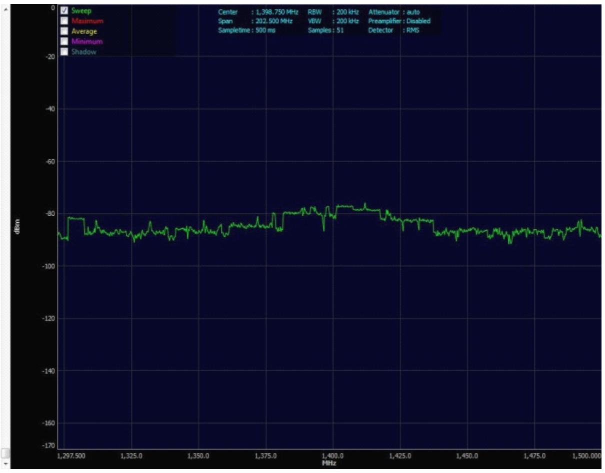


Figure 4: Power level in dBm versus frequency (1.3-1.5GHz)

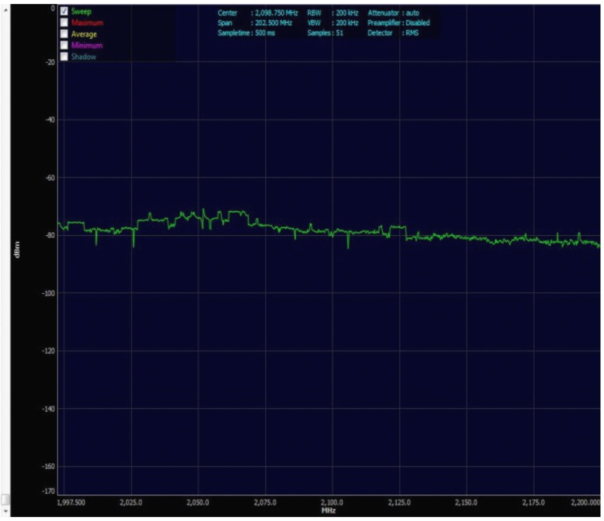


Figure 5: Power level in dBm versus frequency (2-2.2GHz)

The 1000- 1500 MHz band is mostly used for microwave point to point communication [1350-1550MHz], government agencies and oil companies in the Niger delta region and Lagos [18].The 1000-1300 MHz and 1300-1500 MHz with a utilization level of 2.13 and 1.85 respectively are among the bands with the lowest utilization level. Apart from microwave point to point transmission observed around 1350-1450MHz; there is virtually no activity at all.

Above 1.5 GHz, majority of the utilization can be observed in the 3G mobile standards. With a utilization level of around 25.1% it has one of the highest utilization level amongst the bands considered for this work. In Nigeria, there are currently five mobile companies delivering 3G mobile services in Nigeria: MTN, Globacom, Airtel, Etisalat and Starcomms [15]. Networks employing UMTS use WCDMA technology as stated above, the spread spectrum nature of the signals where by the signals are modulated over a wide bandwidth thus making them having a noise-like character due to the very low transmission power makes them difficult to detect. This makes it difficult for the spectrum analyzer to determine such signals. Similarly, since the measurements' were conducted indoors, the ability of the antenna to receive signals might be hindered.

Above 2.42 GHz, with 17% utilization, the ISM band shows considerable utilization but it could also provide some opportunity for secondary usage. As the measurement was done indoors it was able to detect much of the signals due to the short nature of signals in this band. Some activity on the satellite uplink and downlink bands were also detected at a frequency of 2.305-2.32 MHz and 2.335-2.36 MHz

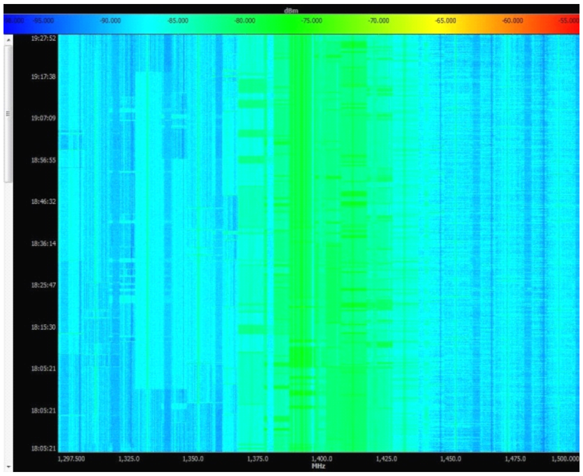


Figure 6: Waterfall of 1.3-1.5 GHz range

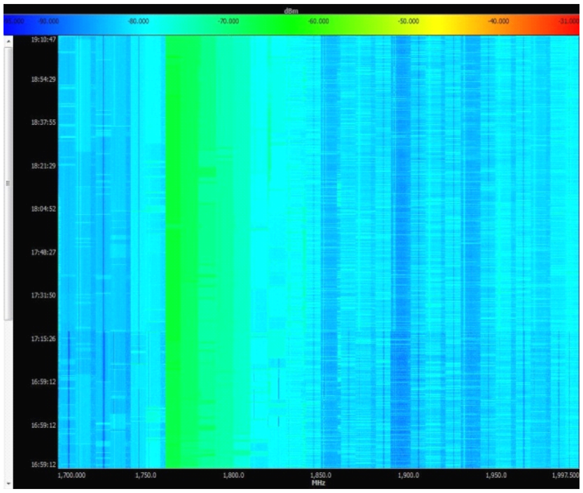


Figure 7: Waterfall of 1.7-1.9 GHz range

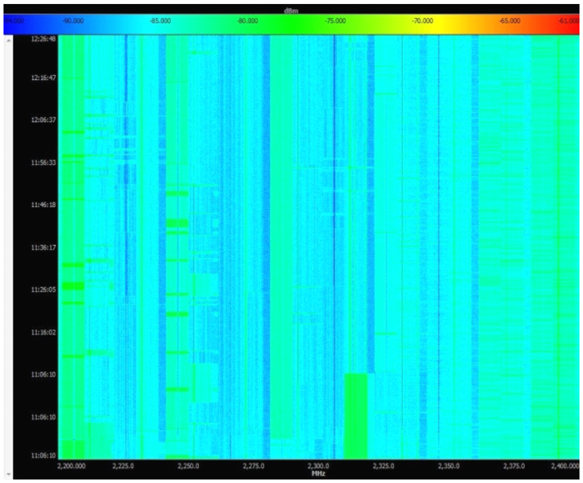


Figure 8: Waterfall of 2.2-2.4 GHz range

BLOCK	FREQUENCY RANGE(MHz)	DUTY CYCLE
1	700-1000	26%
2	1000-1297.5	2.13%
3	1297.5-1500	1.85%
4	1500-1700	12.7%
5	1700-1997	25.56%
6	1997-2200	0.45%
7	2200-2400	17.42%

Table 3: Summary of Spectrum Occupancy

CONCLUSION

In this work, an attempt was made to get the spectrum utilization level in Abuja the capital city of Nigeria. The 700-2500 MHz band was considered. The measurements were conducted indoors as part of wider planned measurements to be conducted. The results indicate abundant potential for CR deployment.. The cellular communication bands can be said to possess the highest utilization level at 26% and 25.56% for the bands containing the 2G and 3G cellular standards, while the 1000-1500 MHz and 2000-2200 MHz bands have the lowest utilization level as the results in table 3 indicate. These results will hopefully aid the academia and policy makers in formulating future policies for spectrum management. In the near future, further measurements would be taken to obtain a realistic picture of the utilization level. Similarly, measurements would be conducted in other locations both indoors and outdoors over a longer period of time to get a more detailed picture of spectrum usage in Nigeria

REFERENCES

1. J. Mitola and G. Maguire, “Cognitive Radio: Making Software radios more Personal”, IEEE Personal Communications, 13:70(1999), 13-18.

2. Soraya Contreas et. al, “An investigation into the Spectrum Occupancy in Japan in the Context of TV White Space Systems”, 6th International Conference on Cognitive Radio oriented Wireless Networks and Communications, 2011.

3. Kishor Patil, Ramjeed Prasad et. al, “Spectrum occupancy measurement statistics in the context of cognitive radio”. IEEE 14th International Symposium on Wireless Personal Multimedia Communications (WPMC), 2011.

4. Shared Spectrum Company, “Spectrum Occupancy Measurements”, Shared Spectrum Company reports. Available at http://www.sharedspectrum.com/measurements/.

5. De Francisco, R “Spectrum Occupancy in the 2.36-2.4 GHz band: Measurement and Analysis”, European Wireless Conference (EW), IEEE 2010.

6. Miguel Lopez-Benitez et. al. “Evaluation of Spectrum Occupancy in Spain for Cognitive Radio Application”, IEEE 69th Conference on vehicular Technology. 2009

7. Salim A. Hanna et. al, “Spectrum metrics for 2.4GHz ISM Band Cognitive Radio Applications”, IEEE 22nd International Symposium on Personal, Indoor, and Mobile Radio Communications. 2011

8. Yanfeng Han et al. “Spectrum Occupancy Measurement: Focus on the TV Frequency”, International Conference on Signal Processing System, 2010.

9. Oliver Holland et. al, ”Spectrum Power Measurements in 2G and 3G Cellular Phone Bands during the 2006 Football World Cup in Germany”, New Frontiers in Dynamic Spectrum Access Networks, 2007.

10. Linda E. Doyle, “Essentials of Cognitive Radio”, New York: Cambridge University Press, 2009.

11. Zhe Wang, Sana Salous,”Spectrum Occupancy Statistics and Time series models for cognitive radio”, Journal of Signal Processing System, Springer, 2009.

12. Dawein Chen, Sixing Yin et. al. “Mining Spectrum Usage Data: a large-scale Spectrum Measurement Study”, IEEE Transactions on Mobile Computing, 2009.

13. Gbenga Ilori, “Development of VHF and UHF Spectrum Optimization for digital services in selected states of Nigeria”, University of Ilorin PhD thesis 2010.

14. Gbenga-Ilori et al. “Nigerian Broadcast Spectrum Usage in the analogue and digital domain”, Proceedings of the Nigerian Institute of Electrical and Electronics Engineers (NIEEE), 5th Annual National Conference, 2009.

15. www.ncc.gov.ng

16. Fola Odufuwa, “Open Spectrum for Development Nigeria Case Study”, Association for Progressive Communication, 2010.

17. www.nfmc.gov.ng

18. Nigerian Communication Commission, “Commercial Frequency Management Policy and Technical Guidelines” January 2007.

Ant-Crypto, a Cryptographer for Data Encryption Standard

Salabat Khan, Armughan Ali and Mehr Yahya Durrani

**Dept. of Computer Science, COMSATS Institute of Information Technology,
Attock Campus, Pakistan**

Abstract

Swarm Intelligence and Evolutionary Techniques are attracting the cryptanalysts in the field of cryptography. This paper presents a novel swarm based attack called Ant-Crypto (Ant-Cryptographer) for the cryptanalysis of Data Encryption Standard (DES). Ant-Crypto is based on Binary Ant Colony Optimization (BACO) i.e. a binary search space based directed graph is modeled for efficiently searching the optimum result (an original encryption key, in our case). The reason that why evolutionary techniques are becoming attractive is because of the inapplicability of traditional techniques and brute force attacks against feistel ciphers due to their inherent structure based on high nonlinearity and low autocorrelation. Ant-Crypto uses a known-plaintext attack to recover the secret key of DES which is required to break/ decipher the secret messages. Ant-Crypto iteratively searches for the secret key while generating several candidate optimum keys that are guessed across different runs on the basis of routes completed by ants. These optimum keys are then used to find each individual bit of the 56 bit secret key used during encryption by DES. Ant-Crypto is compared with some other state of the art evolutionary based attacks i.e. Genetic Algorithm and Comprehensive Binary Particle Swarm Optimization. The experimental results show that Ant-Crypto is an effective evolutionary attack against DES and can deduce large number of valuable bits as compared to other evolutionary algorithms; both in terms of time and space complexity.

Keywords: *Ant-Crypto, Binary Ant Colony Optimization, Comparison of Optimization Techniques, Cryptanalysis of Data Encryption Standard.*

1. Introduction

Most important and precious element in any Information/ Communication system is DATA. Apart from giving us information and knowledge about past events/ activities and patterns, analysis of data can also help us in decision making process, keeping in view the objectives to be achieved in future. There are numerous techniques to store, retrieve and mine the data in databases and data warehouses but in this competitive world where adversaries can illegally access the data, the only way to survive and compete the adversaries is to keep the valuable data, safe and secure. The data cannot be kept secure using classical security techniques e.g. locks; either physically or electronically. In the literature, two inevitable categories of attacks are described; one is passive attack and the other is active attack. In the passive attack, an attacker get access to the communication system and find information contained within secret data. These attacks are difficult to intercept because the attacker do not change the contents of the original data. On the other hand, in active attack an attacker not only gets access to the data but also disrupt the original data. The active attacks are easily detectable but difficult to recover.

Organizations cannot rely on the original form of their secret data and they even don't want any attacker to launch the passive attack (active attack is more harmful) against their communication/ information system. So, they use encryption schemes usually known as cipher (encryption algorithm) in the field of cryptography. Some ciphers e.g. Data Encryption Standard (DES), Advance Encryption Standard (AES) uses secret keys to encrypt the secret data/ message or plaintext. Ant-Crypto is a novel swarm based attack for the cryptanalysis of DES. Cryptanalysis is about the techniques in cryptography that tries to recover the original message or plaintext from an encrypted message, without knowing the secret key used during encryption phase. It includes the study of mathematical techniques e.g. linear cryptanalysis and differential cryptanalysis for attacks against communication/ information system security.

There are two types of ciphers based on the unit of a plaintext that goes under processing; first, the Block ciphers and second, the Stream ciphers. Block ciphers are modern ciphers and operates on a block or chunk of the original plaintext using fixed transformation based on the combination of substitution and permutation. Stream ciphers process a single byte of a message at a time when en/decrypting. DES is based on feistel block cipher. Substitution ciphers are easily breakable due to their weedy encryption process [13]. The length of the key is the main indicator of how difficult it would be to break a cipher. DES with a 56 bit key length makes brute force attack infeasible as it would take several years to find the secret key even if the original plaintext is known. In the next section, we will review the related work in the domain of DES cryptanalysis.

2. Related Work

Cryptanalysis of DES is an interesting problem for the researchers in the field of cryptography. The effectiveness of optimization techniques in cryptanalysis is apparent with the research carried out on classical as well as modern block ciphers (that are more resistant to attacks). Spillman et al. [1] and Castro et al. [2] used Genetic Algorithm in their research to break different ciphers. A fair amount of detailed analysis of how different optimization techniques can be used in the field of cryptography is provided in the research of Clark [3]. Further, Clark et al. [4] have investigated the automated cryptanalysis of classical ciphers which is also considers as an extensive effort. In their thesis [5], they also investigated the effectiveness of simulated annealing, tabu search and genetic algorithm for the cryptanalysis of substitution ciphers. Garici et al. [6] used a population based approach for the automated cryptanalysis of substitution ciphers.

All the papers described above are considered effective [7] for classical ciphers because the complexity of these ciphers is low and there is some inherent linear relationship that may be exploited by an attacker to break them easily. Modern ciphers are complex and highly resistant to any known attack for classical ciphers e.g. character frequency analysis and information of digram and trigram etc. DES also experiences the avalanche effect; a property of any encryption algorithm such that a small change in either the plaintext or the key produces a significant change in the ciphertext. The strength of DES is discussed in detail [8] against differential cryptanalysis attacks. Matsui [9] presented the first experimental cryptanalysis of DES using an improved version of linear cryptanalysis technique. Bafghi et al. [10] proposed weighted directed graph model to find the differential characteristics of a block cipher using Ant Colony Optimization based on shortest path. Laskari et al. [11] used Particle Swarm Optimization technique for the cryptanalysis of simplified version (four-rounded) of DES. J. Song et al. [12], [7] used Genetic Algorithm for the cryptanalysis of two and four-rounded DES. Waseem Shahzad et al. [13] used comprehensive learning binary particle swarm optimization for the cryptanalysis of four rounded DES. In this paper, we use for the first time, Ant Colony Optimization algorithm for the cryptanalysis of four-rounded DES (block cipher). The remainder of this paper is organized as follows.

In the next section, we review the working of DES. In Section 4, we present the basics of ant colony optimization meta-heuristic. In Section 5, architecture, design and detail of the proposed solution is given. Subsequently, in Section 6, we present some experimental results of Ant-Crypto compared with other techniques to show the promising ability of our approach. Finally, Section 7 will conclude this work.

3. Four Rounded DES

In feistel ciphers, transformations are usually carried out as a combination of substitution and permutation. A mapping function is applied repeatedly several times in an iterative manner; iteration is usually called a round. DES is one of the most famous feistel ciphers and has enjoyed widespread use internationally during last few decades. It consists of sixteen rounds and operates over a 64 bits data block using a 56 bits key. DES is a symmetric cipher as encryption and decryption is almost same but the same secret key is applied in reverse order during decryption. Four-Rounded DES is a restricted form of the original DES in which only four rounds are used during encryption/ decryption.

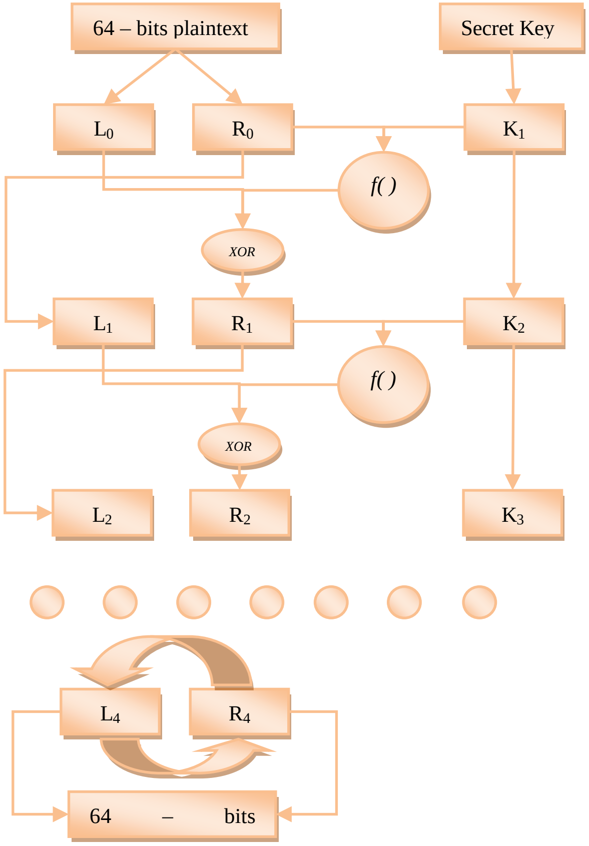


Fig. 1 Four Round DES Working

In DES, 64 bits data is divided into left and right halves. The key is stored as 64 bits but reduced to 56 bits after applying a permutation table. This 56 bits key is then divided in two parts each of 28 bits. After that 16 sub keys are created after applying circular left shifts and permutations according to the given shift and permutation tables, respectively. In each round of DES, a main function is applied to the right half of data and a subkey of 48 bits. During this process, eight S-boxes are used which convert each 6-bit block into a 4-bit block generating 32-bit data. Finally, the left half of the data is XORed with 32-bit output of the main function. In each round, two mapping equation are used; first $L_i = R_{i-1}$ and second $R_i = L_{i-1} \oplus f(R_{i-1}, K_i)$ for above-mentioned process. Where 'i' denotes the round number and $\oplus$ denotes the XOR operation. Li and Ri are left and right halves, respectively. Ki is the ith subkey used in round 'i'. In four-rounded DES, maximum value of 'i' will be four. The readers are directed to [14], [15] for more detailed description of Data Encryption Standard (DES).

4. Ant Colony Optimization

Suppose, we have a connected graph G = (V, E) where |V| denotes the total number of nodes/ vertices and |E| total number of connecting edges in graph. The simple ant colony optimization meta-heuristic can be used to find the shortest path between a given source node 'Vs' and a given destination node 'Vd' in the graph 'G'. Each edge of the graph connecting the nodes 'Vi' and 'Vj' has a variable (artificial pheromone), which is modified by the ants when they visit the nodes [19].

From a node, when an ant decides which node to move next, it uses two parameters to calculate the probability of moving to a particular node; first, distance to that node and second, amount of pheromone on the connecting edge. Let $d_{i,j}$ be the distance between the nodes 'i' and 'j', the probability that the ant chooses 'j' as the next node after it has arrived at node 'i' where 'j' is in the set 'S' of nodes that have not been visited [19] is:

$$p_{i,j} = \frac{[\tau_{i,j}]^{\alpha} . [\eta_{i,j}]^{\beta}}{\sum_{k \epsilon S} [\tau_{i,k}]^{\alpha} . [\eta_{i,k}]^{\beta}} \qquad (1)$$

Where $\tau_{i,j}$ is the pheromone/ trail value on edge and $\eta_{i,j}$ is a heuristic value calculated as $1/d_{i,j}$. The parameters α and β are influencing factors of pheromone value and heuristic value, respectively. The pheromone on edges is modified using equation (2) as:

$$\tau_{i,j} = \tau_{i,j} + (Q/L) \qquad (2)$$

Where 'Q' is some constant and 'L' is the length of the tour, small the value of 'L' high the pheromone value added to the previous pheromone value on an edge. With time, concentration of pheromone decreases due to diffusion affects; a natural phenomenon known as evaporation. This also ensures that old pheromone should not have a too strong influence on the future. Evaporation can be performed using equation (3).

$$\tau_{i,j} = \tau_{i,j} . \rho \quad (where\ \rho\ is\ between\ 0\ and\ 1) \qquad (3)$$

5. Proposed Technique

In the following subsections, each and every stage of Ant-Crypto is further discussed in a fair amount of detail:

5.1 Search Space (A directed graph)

The cores of our approach include "structure of search space" and "calculation of heuristic value". The search space modeled in the article is generic in nature; as it can very easily be used for the cryptanalysis of other ciphers including but not limited to DES and AES. The search space consists of two layers of vertices. One layer at top and second at bottom, both consists of '*n*' vertices where '*n*' is the length of secret key. In our case, the key length is 64 as used by DES. For the cryptanalysis of AES, the '*n*' will be 128. The top layer vertices are labeled as '1' where the bottom layer vertices are labeled as '0'; thus called Binary Ant Colony Optimization.

In order to precisely describe the constraints on the movement of ant, let us see the search space from another point of view. Search space is a grid of two rows and '*n*' columns. Every vertex in a column is connected to all the vertices in the next column through directed edges except the vertices in last column, so, the total number of edges are *4*(n-1)* and total number of vertices are *(n*2)*. An ant starts it tour from a node at left most column by choosing the node label '0' or '1', randomly. An ant can only move from left to right and its tour is finished at the *nth (i.e. the last)* column. In a column, an ant can only select a single vertex during a particular tour. At the end, when the tour is completed, it will consist of '*n*' vertices labels which in turn form an *n*-bit long binary string. This binary string is a candidate or guessed key that will be applied to the original plaintext and a candidate cipher text is calculated.

5.2 Initialization

At start of ACO, edges in the search space are required to be initialized with some small values of pheromone. Usually this is done randomly but in our case, initialization of pheromone is not random. A seeding population is generated based on the equation (4). Seeding population is then used to initialize the pheromone values.

$$M_k \oplus C_k \quad (for\ k = 1,2,3.....n) \tag{4}$$

Where '*k*' is the *kth* bit of plaintext *M* and its corresponding ciphertext is denoted as *C*. Seeding population is generated based on random multi plaintext-ciphertext pairs. XOR operation is basically performed on a plaintext and its correspondence ciphertext. This will speed up the evolution process [7] and is adopted from the approximate expression in linear cryptanalysis [20, 21]. Note that equation (4) will result in *n*-bit long binary string denoting '*n*' vertices and the edges between these vertices will be initialized with some small pheromone values. In order to keep diversity in our search space, we used seeding population of size 100. We used four ants as original swarm size and the parameters 'α' (pheromone influence factor) is set to '1.5' and 'β' (heuristic influence factor) is set to '1' in our experiments.

5.3 Fitness Function

Let '*n*' is the key length, C_{si} and C_{ti} are the *ith* bits of the original ciphertext generated using original secret key and the candidate ciphertext generated using trial key; then fitness function '*f*' is defined as:

$$f(Cs, Ct) = \frac{\sum_{i=1}^{n} h(Cs_i,\ Ct_i)}{n}$$

$$h(Cs_i, Ct_i) = \begin{cases} 1, & \text{if } Cs_i = Ct_i \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

The possible range of fitness value is between 0 – 1. The tour of an ant nearest to the real key is supposed to have higher fitness value, see e.g. [13].

5.4 Heuristic Value

The transition probability equation (1) needs a heuristic value calculation method from the problem domain as an efficient search methodology. Ant-Crypto uses no heuristic value in the first iteration (discussed in section 5.5). After *1st* iteration, four ants results in four candidate keys. The candidate key with the best fitness value using equation (5) is saved as a global best ant. Now, in the subsequent iterations, at every decision point, ant uses heuristic value which is calculated as follows:

$$\eta_{i,j} = f(Original\_Key, Concatenate(\lambda_{|1\ to\ i|},\ j, \Omega_{|i+2\ to\ n|}))$$

The 'Concatenate' is a function that returns an *n*-bit binary string after concatenating the given three binary string parameters. The '$\lambda_{|1\ to\ i|}$' is a binary string denoting the partial tour of an ant where '*i*' is the vertex in the *ith* column (in the search space) at which ant has to decide which node to move next. The '*j*' is the vertex in next column where an ant can move, only two values are possible i.e. either '0' or '1'. The '$\Omega_{|i+2\ to\ n|}$' is the best ant binary substring from index '*i+2*' to '*n*'. So, concatenated binary string becomes a guessed key which is evaluated using equation (5) to be used as a heuristic value in equation (1). Note that '*n*' is the key length as discussed, previously.

5.5 Proposed Algorithm

Seeding population is generated and initialization is done as discussed in Section 5.2. The ants complete their tours by making decisions using equation (1). Each completed tour represents a trial/ candidate key to the problem. Pheromone values are updated using equation (6), only best ant in a particular iteration is allowed to update the pheromone values on the edges constituting the tour. The ants also update the best ant information based on their tours fitness values. Evaporation is performed after an iteration using equation (3). The pheromones over the edges constituting the tour of an ant is updated using equation (6), so larger the fitness value, the greater is the amount pheromone concentrated, and the more attractive the edges become for subsequent ants:

$$\tau_{i,j} = \tau_{i,j} + \frac{tour\ fitness}{log_2 n} \tag{6}$$

We used 500 runs (R) and in each run there are 1000 iterations (N). In a run during an iteration, if we found the fitness value of the best ant greater than or equal to a threshold value 'γ', we declare the tour (64 bits binary string) an optimum key. Once the optimum key is found next run is started. Several optimum keys are generated across multiple runs. For all optimum keys, we count the sum of 1 and 0 for all bit positions and each sum is then divided by *R* (denoting total candidate optimum keys) [13]. Now, if the sum after division is higher than 'Ç' (a threshold already set by user) then these bits can be deduced as '0' or '1'. We fix these bits in the seeding population and start next run. The algorithm runs again and again until all bits are deduced.

Table 1: Comparison of ACO with PSO and GA for four rounded DES

Rounds	**γ**	**Ϛ**	**Opt.rate (GA)**	**Opt.rate (PSO)**	**Opt.rate (Ant-Crypto)**	**Suc.bits (GA)**	**Suc.bits (PSO)**	**Suc.bits (Ant-Crypto)**
Four	0.78	0.60	0.87	0.93	**0.94**	4	4	**05**
Four	0.80	0.70	0.38	0.47	**0.50**	13	19	**19**
Four	0.80	0.75	0.38	0.47	**0.49**	8	16	**17**
Four	0.82	0.65	0.12	0.32	**0.35**	9	13	**15**

Suppose that we have three candidate keys; Key 1: (11010101), Key 2: (11001000) and Key 3: (10001110), ‘R’ = 3 and the value of ‘Ϛ’ is 0.80. So, the count (1) is (3,2,0,1,2,2,1,1) and count (0) is (0,1,3,2,1,1,2,2). Now, Count (1)/ R = (**1**, 0.67, 0, 0.33, 0.67, 0.67, 0.33, 0.33) and Count (0)/ R = (0, 0.33, **1**, 0.67, 0.33, 0.33, 0.67, 0.67). For all the bit positions in the first vector Count(1)/ R and in second vector Count(0)/ R if values are greater than or equal to ‘Ϛ’, they are fixed as 1 and 0, respectively. In the above example, bit number one of the seeding population is fixed as 1 and bit number three is fixed as 0. In the successive runs, these deduce bits will initialize the edges with high pheromone values and help in guessing other valuable bits of the complete secret key. Note that however, deduction may be either true or false; if true, subsequent runs will produce good results and more valuable bits of the secret key will be found. Otherwise, ants will be attracted to local optimum peaks because of this deceptive deduction. The pseudo code of the algorithm is given below. This algorithm is run again and again until all the bits of the secret key are found.

Proposed Algorithm

1. Generate seeding population from multi plaintext-ciphertext pairs and perform initialization
2. Complete the tours of ants by making the decisions using probability equation (1)
3. Calculate fitness value for the tours of ants according to equation (5)
4. Update best ant information. Update pheromone using equation (6)
5. Perform evaporation using equation (3)
6. Repeat this process until an optimum key is found or maximum number of iterations (*N*) have been reached, if found optimum key, deduce bits and fixed these bits in seeding population
7. Repeat the steps from 1 – 6 for (*R*) times run

6. Simulation Results

The performance of the proposed algorithm is compared with the cryptanalysis of four-rounded DES using Genetic Algorithm [7] and Binary Particle Swarm Optimization [13]. We have implemented the Ant-Crypto in Visual Studio 2008 C#. We used the optimum rate (the ratio of the optimum keys in all solution) and number of success bits (the bits in guessed key that are matched with the original key) as the performance metrics.

Table 1. shows the parameter values used during experiments. It also summarizes the performance of our approach against other evolutionary attacks against DES. The fixed number of generations ‘N’ was 1000 and run ‘R’ was 500 in our experiments. Variations of the values of parameters ‘γ’ and ‘Ϛ’ are also shown in Table 1. The reason to use these parameter values is to compare the Ant-Crypto with GA [7] and Binary PSO [13] where the same parameters values were used. We can see that the Ant-Crypto performs better than the results obtained using GA and Binary PSO. We obtained higher number of optimum keys and success bits on different setting of the parameters values. We found largest number of optimum keys when we settled small value of ‘γ’; as the target for ants was easy to find. We can see from Table 1. that when ‘γ’ is 0.80 and ‘Ϛ’ is 0.70, we have maximum number of success bits i.e. 19 in total. The algorithm when run four times with these parameter values, finds all bits of the secret key.

Table 2: Experimental results for 1, 2 and 3 Rounded DES

Rounds	**γ**	**Ϛ**	**N**	**R**	**Opt.rate (Ant-Crypto)**	**Suc.rate (%) (Ant-Crypto)**
One	0.70	0.25	10,000	100	**0.97**	**0.99**
Two	0.75	0.30	10,000	150	**0.96**	**1.00**
Three	0.76	0.30	10,000	200	**0.96**	**0.99**

Table 2 summarizes the results obtained for one, two and three rounded DES. The ‘N’ i.e. total numbers of generations are 10,000 in order to get the optimum key with high probability during a run as we are using small numbers of runs ‘R’, this time. The values of ‘γ’ and ‘Ϛ’ in experiments are used such that the results can be found quickly with good possible quality. From Table 2, it can be observed that the approach is performing very well in case of optimum success rate, as DES with one, two and three rounds is easy to break. Each round in DES adds more complex mapping between plaintext and ciphertext. So, it will be a kind of a complex challenge to use this approach further for the cryptanalysis of more rounds of DES as compared to less complex four-rounded DES.

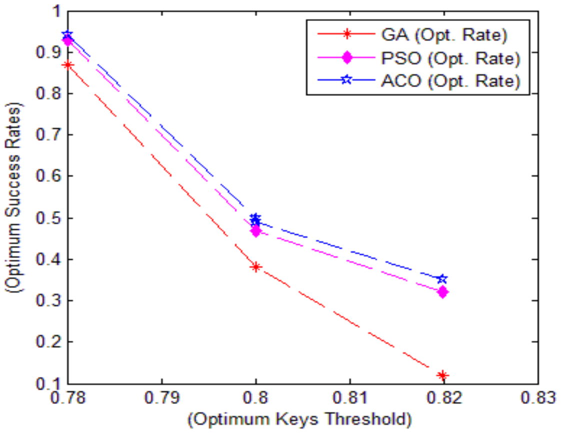


Fig. 2 Comparison of GA, B-PSO and B-ACO based on optimum success rate

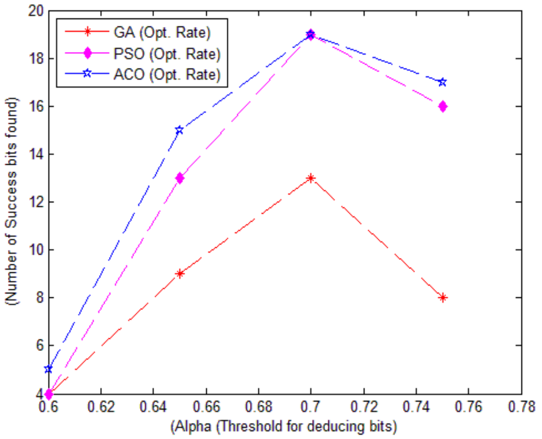


Fig. 3 Comparison of GA, B-PSO and B-ACO based on success bits

In Fig. 2, the comparison of GA, PSO and Ant-Crypto (ACO) is given for optimum success rate. X-axis shows the optimum key threshold and Y-axis shows the optimum success rate. Graph depicts that the Ant-Crypto performs better than GA as well as PSO based cryptanalysis of four-rounded DES in all the cases. Fig. 3 shows the comparison of GA, PSO and Ant-Crypto for finding the success bits. X-axis shows the threshold of guessing a bit and Y-axis shows the number of success bits. Comparison results show that Ant-Crypto is effective and robust technique for cryptanalysis of modern block cipher such as DES as compared with other optimization techniques e.g. GA and PSO. Fig. 2 depicts that Ant-Crypto obtained highest optimum success rate with optimum key threshold is 0.78 and Fig. 3 depicts that Ant-Crypto has highest number of success bits with Alpha (threshold for deducing bits) is equal to 0.70.

7. Conclusions

This article proposed a new version of cryptanalysis algorithm for four rounded DES using binary ant colony optimization. Ant-Crypto is compared with other two well known evolutionary algorithms (GA and PSO) used for cryptanalysis of four rounded DES. We compared the results on that basis of optimum rate and number of success bits found. The experimental results show that Ant-Crypto is an efficient and effective method for the cryptanalysis and it achieves higher optimum rate and number of success bits when compared with other evolutionary approaches used for the cryptanalysis of four rounded DES.

There are several important avenues for future research; the search space structure and/ or heuristic value calculation may be changed for acquiring other valuable findings. It will be a good idea to use a hybrid solution e.g. combining Genetic Algorithm and Ant Colony Optimization (if possible) etc. and have some new experiments. There are several parameters that are required to be well tuned in order to get these results and thus results may be improved further for example, the values of parameters alpha and beta can be tuned in an effort to find better values than those currently used in our experiments (i.e. $\alpha = 1.5$, $\beta = 1$). Furthermore, the effect of pheromone evaporation rate ‘ρ’ needs to be studied in search of an optimal value. Currently we are, somewhat arbitrarily, using $\rho = 0.15$. There are different variants of the original ACO algorithm that can be used to obtain better results. This approach can also be applied to cryptanalysis of some other block ciphers e.g. AES.

References

[1] R. Spillman, M. Janssen, B. Nelson, and M. Kepner. Use of A Genetic Algorithm in the Cryptanalysis of simple substitution Ciphers, April 1993, Vol. 17(1), pp. 31-44.

[2] Julio César Hernández Castro, José María Sierra, Pedro Isasi and Arturo Ribagorda. Genetic Cryptoanalysis of Two Rounds TEA. ICCS2002. Lecture Notes In Computer Science; Vol. 2331 2002. pp. 1024-1031.

[3] Clark. Modern Optimization Algorithms for Cryptanalysis. Proceedings of Second IEEE Australian and New Zealand Conference on Intelligent Information Systems. 1994.

[4] Clark and Ed Dawson. Optimization Heuristics for the Automated Cryptanalysis of Classical Ciphers. In Journal of Combinatorial Mathematics and Combinatorial Computing, Papers in honour of Anne Penfold Street, 1998, vol. 28, pp. 63-86.

[5] Andrew John Clark. Optimization Heuristics for Cryptology, PhD thesis, 1998.

[6] Mohamed Amine Garici, Habiba Drias. Cryptanalysis of Substitution Ciphers Using Scatter Search. IWINAC 2005. Lecture Notes in Computer Science Vol. 3562. pp. 31-40.

[7] J. Song, H. Zhang, Q. Meng and Z. Wang. Cryptanalysis of Four-Round DES Based on Genetic Algorithm. International Conference on Wireless Communications, Networking and Mobile Computing WiCom 2007. Issue. 21-25. Sept. 2007, pp. 2326 – 2329.

[8] Don Coppersmith. The Data Encryption Standard (DES) and its strength against attacks. IBM Journal of Research and Development. Vol. 38 Issue 3. May 1994. pp. 243– 250.

[9] M. Matsui. First Experimental Cryptanalysis of the Data Encryption Standard. Advances in Cryptology—CRYPTO '94, Lecture Notes in Computer Science, Vol. 839 pp 1-11.

[10] Abbas Ghaemi Bafghi, Babak Sadeghiyan. Finding Suitable Differential Characteristics for Block Ciphers with Ant Colony Technique. Proceedings of Ninth International Symposium on Computers and Communications 2004 (ISCC"04), Vol. 2 pp .418-423.

[11] E.C. Laskari, G.C. Meletiouc, Y.C. Stamatiou, M.N. Vrahatis. Evolutionary Computation based Cryptanalysis: A first study. Nonlinear Analysis, 2005, pp. 823-830.

[12] Jun Song, Huanguo Zhang, Qingshu Meng, Zhangyi Wang. Cryptanalysis of Two-Round DES using Genetic Algorithm. ISICA 2007, LNCS 4683, pp. 583–590, 2007.

[13] Waseem Shahzad, Abdul Basit Siddiqui, Farrukh Aslam Khan. Cryptanalysis of Four-Rounded DES using Binary Particle Swarm Optimization. GECCO'09, Montréal Québec, Canada. ACM, pp. 2161-2166, 2009.

[14] National Bureau of Standards, Data Encryption Standard, FIPS-Pub.46. National Bureau of Standards, U.S. Department of Commerce, Washington D.C., January 1977.

[15] Whitfield Diffie, Martin Hellman, "Exhaustive Cryptanalysis of the NBS Data Encryption Standard", IEEE Computer 10(6), pp74-84, June 1977.

[16] M. Dorigo. Optimization, "Learning and Natural Algorithms". PhD thesis, 1992.

[17] L.M. Gambardella and M. Dorigo. Ant-Q: A Reinforcement Learning Approach to the TSP. In Proceedings of Twelfth International Conference on Machine Learning, pp. 252-260, 1995.

[18] L.M. Gambardella and M. Dorigo. Solving Symmetric and Asymmetric TSPs by Ant Colonies. IEEE International Conference on Evolutionary Computation, pp. 622627, 1996.

[19] S. Khan, Mohsin Bilal, M. Sharif, Malik Sajid, Rauf Baig. Solution of n-Queen Problem Using ACO. International Multitopic Conference, Islamabad, IEEE, pp. 1-5, 2009.

[20] Mitsuru Matsui: Linear Cryptanalysis Method for DES Cipher, pp. 386-397, 1993.

[21] Mitsuru Matsui: The First Experimental Cryptanalysis of the Data Encryption Standard. CRYPTO, pp. 1-11, 1994.

[22] S. Khan et. al., Cryptanalysis of Four-Rounded Data Encryption Standard using Binary Ant Colony Optimization", ICISA, IEEE, pp. 1-7, 2010.

Cramer-Rao Lower Bound for NDA SNR Estimation from Linear Modulation Schemes over Flat Rayleigh Fading Channel

Monia Salem, Slaheddine Jarboui and Ammar Bouallegue

Laboratory of Communication Systems, National School of Engineers of Tunis
Tunis, 1002, Tunisia

Abstract

In this contribution, Cramer-Rao lower bound (CRLB) for signal-to-noise ratio (SNR) estimation from linear modulation signals over flat Rayleigh fading channel is addressed. Therefore, we derive the analytical expressions of Fisher information matrix entries that assess the optimal variance of any unbiased SNR estimator. Based on statistical Monte Carlo computing method, simulation results are drawn from several constellation densities and observation window sizes. For the linear modulation schemes used here, it is shown that the lower bound is as higher as the modulation order increases. The derived bound provides an efficient standard for evaluating the performance of any unbiased non-data aided (NDA) SNR estimator from linear modulation signals over flat Rayleigh fading channel (FRFC).

Keywords: *Cramer-Rao lower bound, signal-to-noise ratio, non-data aided estimation, FRFC, complex AWGN.*

1. Introduction

Modern communication systems often require the knowledge of the SNR level at the receiver side in order to qualify the performance of the received signal quality. Then accurate SNR estimate is required for measuring the channel quality for adaptive modulation schemes as well as for soft decoding procedures as shown in [1], [2] and [3]. In addition to low-complexity requirement, it is essential to assess the truthfulness of SNR estimators in term of their statistical variances. For this purpose, the well-known CRLB is a prominent benchmark to evaluate the statistical variance performance of unbiased estimators.

Actually both data aided (DA) and non-data aided (NDA) trends are considered for either performance bounds derivation or estimation algorithms. Data aided approach, which relies on the transmission of known data streams such as training sequences and also pilot symbols, should expedite and ease the estimation process. Unfortunately, this approach limits the system through-put in the sense that adding known pilot symbols to the data stream should drop down the spectral efficiency of the communication system. Hence NDA SNR estimation approach receives substantial attention in recent literature. CRLB for NDA SNR estimation is derived in [4] from both BPSK and QPSK modulated signals with AWGN channel. Derived bounds are compared to those obtained for DA estimation. In [5], a straightforward approximation of the CRLB for NDA SNR estimation from BPSK modulated signals over AWGN channel is presented in efficient form that avoids tedious numerical integration. Authors, in [6], derive a lower bound for SNR estimation from general M-ary one/two dimensional modulation signals with axis/half plane symmetry over AWGN channel. Exact analytical CLRB of unbiased NDA SNR estimation from square QAM signals using I/Q received signal model is addressed in [7], where a generalization of the elegant CLRB expressions presented in [4] is also introduced.

In addition to AWGN channel, derivation of SNR estimates CRLB for fading channels deserves great attention regarding its significance for modern wireless communication systems. Hence, derivation of the CRLB for SNR estimation is addressed in [8] over a time-varying channel based on polynomial-in-time model according to Taylor’s theorem. Recent works presented in [9] and [10] deal with the CRLB derivation for carrier phase and frequency estimation assuming transmission over fading channel. On this basis, the present work is devoted for analytically deriving the CRLB for NDA SNR estimation from linearly modulated signals over flat Rayleigh fading channel (FRFC) where the transmitted signal is scaled by a non-constant fading gain during the estimator observation window. Noise power and also signal amplitude are assumed as completely unknown at the receiver side. The lower bounds derived hereafter offer an efficient standard to assess NDA SNR estimator performance over FRFC.

2. System Model

Consider the transmission of linearly modulated signal over FRFC corrupted by a complex AWGN (CAWGN). In absence of carrier phase and frequency offsets and also under the assumption of ideal timing recovery, the complex sample at the output of the receiver matched filter x_k can be written as:

$$x_k = \rho_k a_k + \omega_k \qquad ; \qquad k = 0,\ldots,N-1 \tag{1}$$

where, a_k is the transmitted symbol and N is the observation window size. Note that the transmitted symbols $a_0,\ldots,a_{N-1}$ are assumed as independent and identically distributed. ω_k is the CAWGN sample. The vector $\boldsymbol{\omega} = \{\omega_0,\ldots,\omega_{N-1}\}$ is a set of randomly drawn samples from independent zero-mean complex Gaussian process with uncorrelated real and imaginary parts having equal variances σ^2. ρ_k is a Rayleigh distributed positive random variable, where the well-known Rayleigh probability density function (PDF) is given by [11]:

$$f(\rho,\sigma_0) = \frac{\rho}{\sigma_0^2} e^{-\frac{\rho^2}{2\sigma_0^2}} \tag{2}$$

The signal-to-noise ratio (SNR) is then given by:

$$SNR_{FRFC} = \gamma = \frac{\sigma_0^2}{\sigma^2} \tag{3}$$

We expect to estimate γ based on the observation samples vector $\boldsymbol{x} = \{x_0,\ldots,x_{N-1}\}$. Then two parameters are involved in this estimate. For convenience, we note:

$$\alpha = \sigma_0^2 \qquad ; \qquad \beta = \sigma^2 \tag{4}$$

Let us define a parameter vector $\boldsymbol{\theta}$ such that:

$$\boldsymbol{\theta} = [\alpha \quad \beta] \tag{5}$$

While the estimated SNR unit is usually the decibel, thus we consider the following function:

$$\boldsymbol{g}(\boldsymbol{\theta}) = 10\log(\frac{\alpha}{\beta}) \tag{6}$$

The CRLB of the SNR estimation is given by [11, pp.45-46]:

$$CRLB(\gamma) = \frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \boldsymbol{I}^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^T \tag{7}$$

where $\boldsymbol{I}(\boldsymbol{\theta})$ is the 2×2 Fisher information matrix (FIM) defined as:

$$\boldsymbol{I}(\boldsymbol{\theta}) = \begin{bmatrix} -E_x\left(\frac{\partial^2 \mathrm{Ln}(p(\boldsymbol{x}\,|\,\boldsymbol{\theta}))}{\partial \alpha^2}\right) & -E_x\left(\frac{\partial^2 \mathrm{Ln}(p(\boldsymbol{x}\,|\,\boldsymbol{\theta}))}{\partial \alpha \partial \beta}\right) \\ -E_x\left(\frac{\partial^2 \mathrm{Ln}(p(\boldsymbol{x}\,|\,\boldsymbol{\theta}))}{\partial \beta \partial \alpha}\right) & -E_x\left(\frac{\partial^2 \mathrm{Ln}(p(\boldsymbol{x}\,|\,\boldsymbol{\theta}))}{\partial \beta^2}\right) \end{bmatrix} \tag{8}$$

and $\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is the 1×2 Jacobian matrix given by:

$$\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left[\frac{10}{Ln(10)\alpha} \quad -\frac{10}{Ln(10)\beta}\right] \tag{9}$$

3. CRLB Derivation for SNR Estimation

To derive the CRLB expressions, we have to evaluate the probability $p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$ given in (8). The PDF $p(x_k\,|\,\boldsymbol{\theta},a_i,\rho_k)$ for a single received sample x_k parameterized by $\boldsymbol{\theta}$, a_i and ρ_k is given by:

$$p(x_k\,|\,\boldsymbol{\theta},a_i,\rho_k) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}|x_k - \rho_k a_i|^2} \tag{10}$$

Then the PDF $p(x_k\,|\,\boldsymbol{\theta},a_i)$ parameterized by $\boldsymbol{\theta}$ and a_i is computed by integration over the Rayleigh fading gain ρ_k as follows:

$$p(x_k\,|\,\boldsymbol{\theta},a_i) = \int_0^{+\infty} f(\rho,\sigma_0) p(x_k\,|\,\boldsymbol{\theta},a_i,\rho) d\rho \tag{11}$$

After several algebraic handling, we obtain the following expression from (11):

$$p(x_k\,|\,\boldsymbol{\theta},a_i) = \frac{C_k(\boldsymbol{\theta})}{4(A_i(\boldsymbol{\theta}))^{3/2}} F(A_i(\boldsymbol{\theta}), B_{k,i}(\boldsymbol{\theta})) \tag{12}$$

where:

$$F(u,v) = 2\sqrt{u} + \sqrt{\pi} v e^{\frac{v^2}{4u}}\left(1 + erf\left(\frac{v}{2\sqrt{u}}\right)\right) \tag{13}$$

$$A_i(\boldsymbol{\theta}) = \frac{1}{2\sigma_0^2} + \frac{|a_i|^2}{2\sigma^2} \tag{14}$$

$$B_{k,i}(\boldsymbol{\theta}) = \frac{\Re(x_k a_i^*)}{\sigma^2} \tag{15}$$

$$C_k(\boldsymbol{\theta}) = \frac{e^{-\frac{|x_k|^2}{2\sigma^2}}}{2\pi\sigma^2\sigma_0^2} \tag{16}$$

and erf(.) is the error function defined by:

$$erf\left(x\right)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \tag{17}$$

We consider that transmitted symbols $\{a_i\}$ fit in an M-ary constellation C , then the PDF $p(x_k \mid \boldsymbol{\theta})$ can be expressed as follows:

$$p(x_k \mid \boldsymbol{\theta})=\sum_{a_i\in C}p(a_i)p(x_k \mid \boldsymbol{\theta},a_i) \tag{18}$$

Assuming that the received samples $\{x_k\}$ are independent random variables and also that the transmitted symbols $\{a_i\}$ are equally likely random variables $(p(a_i)=p(a))$, then the probability $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ is given by:

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{\theta}) &= \prod_{k=0}^{N-1} p(x_k \mid \boldsymbol{\theta}) \\ &= \prod_k \sum_{a_i\in C} p(a_i)p(x_k \mid \boldsymbol{\theta},a_i) \qquad (19) \\ &= \prod_k p(a)\sum_{a_i\in C} p(x_k \mid \boldsymbol{\theta},a_i) \qquad (20) \end{aligned}$$

We inject (12) in (20), then we obtain:

$$p(\boldsymbol{x} \mid \boldsymbol{\theta})=\prod_k p(a)\frac{C_k(\boldsymbol{\theta})}{4}\sum_{a_i\in C}\left(A_i(\boldsymbol{\theta})\right)^{-\frac{3}{2}}F\left(A_i(\boldsymbol{\theta}),B_{k,i}(\boldsymbol{\theta})\right) \tag{21}$$

Taking the logarithm of (21) and retaining the $\boldsymbol{\theta}$ dependent terms only, we obtain the following expression:

$$\begin{aligned} &\mathrm{Ln}\left\{\prod_k\left(\frac{C_k(\boldsymbol{\theta})}{4}\sum_{a_i\in C}\left(A_i(\boldsymbol{\theta})\right)^{-\frac{3}{2}}F\left(A_i(\boldsymbol{\theta}),B_{k,i}(\boldsymbol{\theta})\right)\right)\right\} \\ &=\sum_k \mathrm{Ln}\left(\mathrm{P_k}(\boldsymbol{\theta})\right) \end{aligned} \tag{22}$$

where:

$$\boldsymbol{P_k}(\boldsymbol{\theta})=\frac{C_k(\boldsymbol{\theta})}{4}\sum_{a_i\in C}\left(A_i(\boldsymbol{\theta})\right)^{-\frac{3}{2}}F\left(A_i(\boldsymbol{\theta}),B_{k,i}(\boldsymbol{\theta})\right) \tag{23}$$

Then, the first diagonal element of the Fisher information matrix $\boldsymbol{I}(\boldsymbol{\theta})$ can be expressed as follows:

$$\begin{aligned} \left[\boldsymbol{I}(\boldsymbol{\theta})\right]_{11} &= -E_{\boldsymbol{x}}\left[\frac{\partial^2 Ln\left(p(\boldsymbol{x} \mid \boldsymbol{\theta})\right)}{\partial\alpha^2}\right] \\ &= \sum_{k=0}^{N-1}\left\{\frac{\frac{\partial^2\boldsymbol{P_k}(\boldsymbol{\theta})}{\partial\alpha^2}}{\boldsymbol{P_k}(\boldsymbol{\theta})}-\left(\frac{\frac{\partial\boldsymbol{P_k}(\boldsymbol{\theta})}{\partial\alpha}}{\boldsymbol{P_k}(\boldsymbol{\theta})}\right)^2\right\} \end{aligned} \tag{24}$$

In order to derive $\frac{\partial^2 Ln\left(p(\boldsymbol{x} \mid \boldsymbol{\theta})\right)}{\partial\alpha^2}$, we compute $\frac{\partial\boldsymbol{P_k}(\boldsymbol{\theta})}{\partial\alpha}$ and also $\frac{\partial^2\boldsymbol{P_k}(\boldsymbol{\theta})}{\partial\alpha^2}$.

The first partial derivative $\frac{\partial\boldsymbol{P_k}(\boldsymbol{\theta})}{\partial\alpha}$ may be written as:

$$\begin{aligned} &\frac{\partial\boldsymbol{P_k}(\boldsymbol{\theta})}{\partial\alpha}=\frac{C_k(\boldsymbol{\theta})}{4}\times \\ &\sum_{a_i\in C}\left(A_i(\boldsymbol{\theta})\right)^{-\frac{3}{2}}\left\{\begin{array}{l} F\left(A_i(\boldsymbol{\theta}),B_{k,i}(\boldsymbol{\theta})\right)\left[\left(A_i(\boldsymbol{\theta})\right)^{-\frac{3}{2}}G_i(\boldsymbol{\theta})-\frac{1}{\sigma_0^2}\right] \\ +H_{k,i}(\boldsymbol{\theta}) \end{array}\right\} \end{aligned} \tag{25}$$

$G_i(\boldsymbol{\theta})$ and $H_{k,i}(\boldsymbol{\theta})$ are given in appendix A.

The expression of the second derivative $\frac{\partial^2\boldsymbol{P_k}(\boldsymbol{\theta})}{\partial\alpha^2}$ is given by:

$$\begin{aligned} &\frac{\partial^2\boldsymbol{P_k}(\boldsymbol{\theta})}{\partial\alpha^2}=\frac{C_k(\boldsymbol{\theta})}{4}\times \\ &\sum_{a_i\in C}\left\{\begin{array}{l} \left[\frac{2}{\sigma_0^4}\left(A_i(\boldsymbol{\theta})\right)^{-\frac{3}{2}}+\dot{G}_i(\boldsymbol{\theta})\right]F\left(A_i(\boldsymbol{\theta}),B_{k,i}(\boldsymbol{\theta})\right) \\ +2G_i(\boldsymbol{\theta})H_{k,i}(\boldsymbol{\theta})+\left(A_i(\boldsymbol{\theta})\right)^{-\frac{3}{2}}\dot{H}_{k,i}(\boldsymbol{\theta}) \end{array}\right\} \end{aligned} \tag{26}$$

$\dot{G}_i(\boldsymbol{\theta})$ and $\dot{H}_{k,i}(\boldsymbol{\theta})$ denote the first derivatives of $G_i(\boldsymbol{\theta})$ and $H_{k,i}(\boldsymbol{\theta})$ with respect to α , respectively. Their expressions are detailed in appendix B.

Applying the same procedure, then the derivation of the remaining elements of $\boldsymbol{I}(\boldsymbol{\theta})$ is given by:

$$\begin{aligned} \left[\boldsymbol{I}(\boldsymbol{\theta})\right]_{12} &= \left[\boldsymbol{I}(\boldsymbol{\theta})\right]_{21} = -E_{\boldsymbol{x}}\left[\frac{\partial^2 Ln\left(p(\boldsymbol{x} \mid \boldsymbol{\theta})\right)}{\partial\alpha\partial\beta}\right] \\ &= \sum_{k=0}^{N-1}\left\{\frac{\frac{\partial}{\partial\alpha}\left(\frac{\partial\boldsymbol{P_k}(\boldsymbol{\theta})}{\partial\beta}\right)}{\boldsymbol{P_k}(\boldsymbol{\theta})}-\frac{\frac{\partial\boldsymbol{P_k}(\boldsymbol{\theta})}{\partial\alpha}\times\frac{\partial\boldsymbol{P_k}(\boldsymbol{\theta})}{\partial\beta}}{\left[\boldsymbol{P_k}(\boldsymbol{\theta})\right]^2}\right\} \end{aligned} \tag{27}$$

$$[\boldsymbol{I}(\boldsymbol{\theta})]_{22} = -E_{\boldsymbol{x}}\left[\frac{\partial^2 Ln(p(\boldsymbol{x}\,|\,\boldsymbol{\theta}))}{\partial\beta^2}\right] = \sum_{k=0}^{N-1}\left\{\frac{\frac{\partial^2 \boldsymbol{P_k}(\boldsymbol{\theta})}{\partial\beta^2}}{\boldsymbol{P_k}(\boldsymbol{\theta})} - \left(\frac{\frac{\partial \boldsymbol{P_k}(\boldsymbol{\theta})}{\partial\beta}}{\boldsymbol{P_k}(\boldsymbol{\theta})}\right)^2\right\} \qquad (28)$$

After derivation of several elements of both the matrix $\boldsymbol{I}(\boldsymbol{\theta})$ and $\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, the CRLB for SNR estimation is given by:

$$CRLB(\gamma) = \frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \boldsymbol{I}^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^T \qquad (29)$$

4. Simulation Results

We use Monte Carlo simulation techniques to evaluate the statistical expectation in (24), (27) and (28) with respect to the N-dimensional vector $\boldsymbol{x}$. Note that once $\{x_k\}$ are statistically independent variables, we estimate the value of this expectation by generating a sequence of random samples at each SNR value, then computing the average of $\frac{\partial^2 \mathrm{Ln}(p(\boldsymbol{x}\,|\,\boldsymbol{\theta}))}{\partial\alpha\partial\beta}$ for each sample. Note that a minimum of 1000 trials is considered to ensure that the estimate stemming from Monte Carlo integration converges to the statistical expectation value. Figures 1 and 2 depict the CRLB curves from M-QAM signals for an observation window size N=100 and N=1000. It is shown that CRLB values decrease as far as the observation window size increases. Moreover, CRLBs for large modulation order take higher values at low SNR range.

The method described here stand useful to determine the CRLB for larger M-QAM constellation densities and also general linear modulation schemes.

5. Conclusions

True Cramer-Rao lower bound for NDA signal-to-noise ratio estimation from linearly modulated signals over FRFC with CAWGN is derived. At low SNR range, simulation results show that CRLB values decrease as far as the modulation order increases. For high SNR levels, CRLBs almost coincide either for various modulation orders or various observation window sizes. The method introduced here represents a standard for NDA SNR estimator over FRFC from linearly modulated signals.

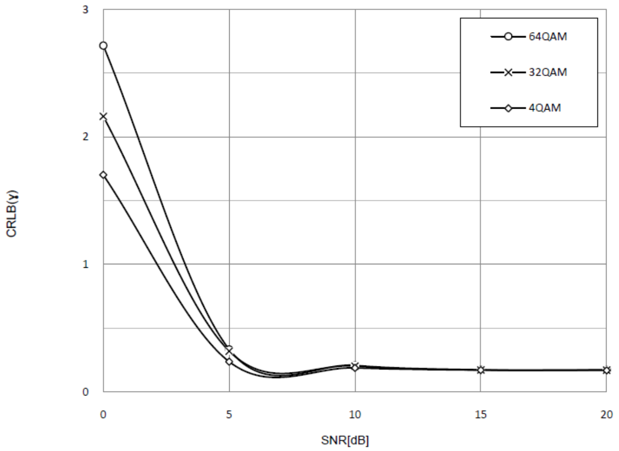


Fig. 1 CRLB versus SNR for 4, 32 and 64-QAM and N=100.

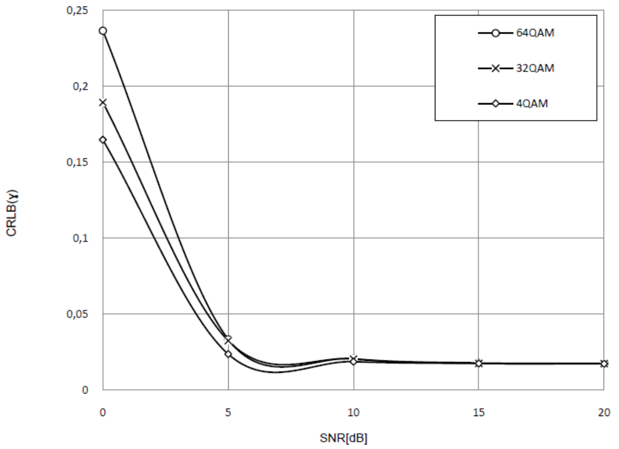


Fig. 2 CRLB versus SNR for 4, 32 and 64-QAM and N=1000.

Appendix A

The first derivative of $A_i(\boldsymbol{\theta})$ with respect to α is given by:

$$\frac{\partial}{\partial\alpha}(A_i(\boldsymbol{\theta})) = -\frac{1}{2\sigma_0^4} \qquad (A.1)$$

The expressions of $G_i(\boldsymbol{\theta})$ and $H_{k,i}(\boldsymbol{\theta})$ are given by:

$$G_i(\boldsymbol{\theta}) = -\frac{3}{2}\frac{\partial}{\partial\alpha}(A_i(\boldsymbol{\theta}))(A_i(\boldsymbol{\theta}))^{-\frac{5}{2}} \qquad (A.2)$$

$$H_{k,i}(\boldsymbol{\theta}) = \frac{\partial}{\partial\alpha}\left(F\left(A_i(\boldsymbol{\theta}), B_{k,i}(\boldsymbol{\theta})\right)\right) \qquad \text{(A.3)}$$

$$= A_i(\boldsymbol{\theta})^{-\frac{1}{2}} \frac{\partial}{\partial\alpha}\left(A_i(\boldsymbol{\theta})\right)$$

$$-\frac{\sqrt{\pi}}{4} A_i(\boldsymbol{\theta})^{-2} \frac{\partial}{\partial\alpha}\left(A_i(\boldsymbol{\theta})\right) B_{k,i}(\boldsymbol{\theta})^3 e^{\frac{B_{k,i}(\theta)^2}{4A_i(\theta)}}\left(1+erf\left(\frac{B_{k,i}(\boldsymbol{\theta})}{2\sqrt{A_i(\boldsymbol{\theta})}}\right)\right)$$

$$-\frac{1}{2} A_i(\boldsymbol{\theta})^{-\frac{3}{2}} \frac{\partial}{\partial\alpha}\left(A_i(\boldsymbol{\theta})\right) B_{k,i}(\boldsymbol{\theta})^2$$

(A.4)

Appendix B

$$\dot{G}_i(\boldsymbol{\theta}) = -\frac{3}{2} A_i(\boldsymbol{\theta})^{-\frac{5}{2}} \times$$

$$\left(\frac{\partial^2}{\partial\alpha^2}\left(A_i(\boldsymbol{\theta})\right) - \frac{5}{2}\left(\frac{\partial}{\partial\alpha}\left(A_i(\boldsymbol{\theta})\right)\right)^2 A_i(\boldsymbol{\theta})^{-1}\right) \qquad \text{(B.1)}$$

$$\dot{H}_{k,i}(\boldsymbol{\theta}) = \frac{\partial}{\partial\alpha} T_{11} + \frac{\partial}{\partial\alpha} T_{12} + \frac{\partial}{\partial\alpha} T_{13} \qquad \text{(B.2)}$$

$$\frac{\partial}{\partial\alpha} T_{11} = A_i(\boldsymbol{\theta})^{-\frac{1}{2}} \frac{\partial^2\left(A_i(\boldsymbol{\theta})\right)}{\partial\alpha^2} - \frac{1}{2} A_i(\boldsymbol{\theta})^{-\frac{3}{2}}\left(\frac{\partial\left(A_i(\boldsymbol{\theta})\right)}{\partial\alpha}\right)^2 \qquad \text{(B.3)}$$

$$\frac{\partial}{\partial\alpha} T_{12} =$$

$$\frac{\sqrt{\pi}}{2} A_i(\boldsymbol{\theta})^{-3}\left(\frac{\partial\left(A_i(\boldsymbol{\theta})\right)}{\partial\alpha}\right)^2 B_{k,i}(\boldsymbol{\theta})^3 e^{\frac{B_{k,i}(\theta)^2}{4A_i(\theta)}}\left(1+erf\left(\frac{B_{k,i}(\boldsymbol{\theta})}{2\sqrt{A_i(\boldsymbol{\theta})}}\right)\right)$$

$$+\frac{\sqrt{\pi}}{16} A_i(\boldsymbol{\theta})^{-4}\left(\frac{\partial\left(A_i(\boldsymbol{\theta})\right)}{\partial\alpha}\right)^2 B_{k,i}(\boldsymbol{\theta})^5 e^{\frac{B_{k,i}(\theta)^2}{4A_i(\theta)}}\left(1+erf\left(\frac{B_{k,i}(\boldsymbol{\theta})}{2\sqrt{A_i(\boldsymbol{\theta})}}\right)\right)$$

$$-\frac{\sqrt{\pi}}{4} A_i(\boldsymbol{\theta})^{-2} \frac{\partial^2\left(A_i(\boldsymbol{\theta})\right)}{\partial\alpha^2} B_{k,i}(\boldsymbol{\theta})^3 e^{\frac{B_{k,i}(\theta)^2}{4A_i(\theta)}}\left(1+erf\left(\frac{B_{k,i}(\boldsymbol{\theta})}{2\sqrt{A_i(\boldsymbol{\theta})}}\right)\right)$$

$$+\frac{1}{16} A_i(\boldsymbol{\theta})^{-\frac{5}{2}}\left(\frac{\partial\left(A_i(\boldsymbol{\theta})\right)}{\partial\alpha}\right)^2 B_{k,i}(\boldsymbol{\theta})^4$$

(B.4)

$$\frac{\partial}{\partial\alpha} T_{13} = \frac{3}{4} A_i(\boldsymbol{\theta})^{-\frac{5}{2}}\left(\frac{\partial\left(A_i(\boldsymbol{\theta})\right)}{\partial\alpha}\right)^2 B_{k,i}(\boldsymbol{\theta})^2$$

$$-\frac{1}{2} A_i(\boldsymbol{\theta})^{-\frac{3}{2}} \frac{\partial^2\left(A_i(\boldsymbol{\theta})\right)}{\partial\alpha^2} B_{k,i}(\boldsymbol{\theta})^2 \qquad \text{(B.5)}$$

$$\frac{\partial^2\left(A_i(\boldsymbol{\theta})\right)}{\partial\alpha^2} = \frac{1}{\sigma_0^6} \qquad \text{(B.6)}$$

References

[1] N. Celandroni, E. Ferro, and F. Potorti, “Quality estimation of PSK modulated signals,” IEEE Commun. Mag., pp. 50–55, July 1997.

[2] K. Balachandran, S. R. Kadaba, and S. Nanda, “Channel quality estimation and rate adaptation for cellular mobile radio,” IEEE J. Select. Areas Commun., vol. 17, pp. 1244–1256, July 1999.

[3] T. A. Summers and S. G.Wilson, “SNR mismatch and online estimation in turbo decoding,” IEEE Transactions on Communications, vol. 46, no. 4, pp. 421–423, April 1998.

[4] N. S. Alagha, “Cramer–Rao Bounds of SNR Estimates for BPSK and QPSK Modulated Signals”, IEEE Comm. Letters, vol. 5, no. 1, pp. 10-12, January 2001.

[5] A. Wiesel, J. Goldberg and H. Messer, “Non-Data-Aided Signal-to-Noise-Ratio Estimation”, ICC, vol. 1, pp. 197-201, 2002.

[6] W. Gappmair, “Cramer-Rao Lower Bound for Non-Data-Aided SNR Estimation of Linear Modulation Schemes”, IEEE Trans. on Commun., vol. 56, no. 5, pp. 689-693, May 2008.

[7] F. Bellili, A. Stéphenne and S. Affes, “Cramér-Rao Lower Bounds for NDA SNR Estimates of Square QAM Modulated Transmissions”, IEEE Trans. on Commun., vol. 58, pp. 3211-3218, November 2010.

[8] A. Wiesel, J. Goldberg and H. Messer-Yaron, “SNR Estimation in Time-Varying Fading Channels”, IEEE Trans. on Commun., vol. 54, no. 5, pp. 841-848, May 2006.

[9] S. Jarboui, M. Salem and A. Bouallegue, “Cramér–Rao Lower Bound for Non-Data-Aided Carrier Phase Estimation from General M-Ary Phase-Shift Keying Modulation Signals over Flat Rayleigh Fading Channel”, IET Communications, vol. 6, Iss. 13, pp. 2108-2113, Sept 2012.

[10] M. Salem, S. Jarboui and A. Bouallegue, “True Cramer-Rao Lower Bound for NDA Carrier Frequency Estimation from General M-QAM Modulated Signals over Flat Rayleigh Fading Channel”, ICCIT, pp. 357-362, June 2012.

[11] S. M. Kay, "Fundamentals of statistical signal processing: estimation theory”, Prentice Hall 1993.

An approach for an optimized web service selection based on skyline

[1]Mohamed Ali Bouanaka,[2]Naceredine Zarour

[1] LIRE laboratory, Department of Software Technologies and Information Systems, University of Constantine 2
Constantine, Algeria

[2] LIRE laboratory, Department of Software Technologies and Information Systems, University of Constantine 2
Constantine, Algeria

Abstract

Nowadays, considering Web Services has become one of the hot issues in the area of computer science that makes an ability to collect capabilities and components in a unique interface to fulfil the user's requirements. Sometimes two or more services are discovered in available list of services; therefore, there should be a possibility for selecting the best services from discovered list which can satisfy the user's goal. The selected services should optimize the overall QoS of the composed application, while satisfying all the constraints specified by the client on individual QoS parameters. In this paper, we propose an approach based on the notion of skyline to effectively and efficiently select services for composition and reducing the number of candidate services. At the end of the paper we evaluate our approach experimentally.

Key words: *Web Services selection, Skyline, Service Composition, QoS, Optimization.*

1. Introduction

With the large amount of information stored and manipulated by users, the customer requirements, the diversity of needs, public infrastructure and private information technologies sector want to communicate more easily and without need to focusing on each of their transaction to interpret their various data, they also want to remove the isolation of their systems, for this reason they are more and more multiplatform, multivendor distributed for large scale, which has led to a great development of information systems which become more complex and heterogeneous.

To be adapted to this new situation, technologies enabling communication and data exchange between heterogeneous applications and systems and their distributed environments called Web services have emerged. They aim to ensure interoperability through a standardized presentation of services and a communications protocol standard for structuring exchanged messages between software components. They also provide a specification of publication and localization of services. The particularity of Web services is that they use the Internet as an infrastructure technology for communication between software components. Migration of companies information systems to services (making their services available to other companies as software components) has increased as a result of a rapid increasing of web services. The problem with this increase is how to select the best web service especially if satisfaction of a request requires the intervention of several of them.

Sometimes two or more services are discovered in available list of services. Therefore, there should be a step or process for selecting the best services from discovered list that can satisfy the user's goal. When more than one Web Service which meets functional requirements is available, the Web Service Selection uses some criteria to select the best candidate service. The value of non-functional properties in these matching Web Services may be different, but essentially they should have minimum requirements.

The selection criteria may have an interdependent relationship. A number of methods for decision making are addressed in Web Service Selection because of the complication that exists during the selection process. [1]. Two significant tasks in the process of using services are selection and ranking in which every solution for them is affected directly on description of services. During describing a service, three items have to be considered: behaviour, functional, and non-functional. The non-functional properties of the services are used as criteria for selecting services.

The basic idea in our approach is to make a selection of the best services needed to be composed to have a better composition without having to test all combinations, so to allow compositions to respond to customers' requests in real time even if the number of available web services is important.

In the following sections we will first discuss related studies and works in this field which is web service selection. Section 3, exposes our conceptual methodology and discusses the choices made. Section

4 shows our experimental study followed by a comparative evaluation. Finally, perspectives of the accomplished work are presented in the conclusion of the paper.

2. Related work

The problem of QoS-based web service selection and composition has received a lot of attention during the last years. In [2] and [3], authors proposed approaches on policy languages. Coding in policy language or in a QoS policy model is used for defining the non-functional requirements. The policy based designs the QoS policy model as a textual document. Preferences and non-functional limitations of the service requestor are shown in the content of the policy model. For showing the non-functional criteria's relations, a matrix is used, also it is applied for their aggregation [4]. The problem with these approaches is that only a limited number of non-functional properties is accepted. In addition, it is difficult for users to understand the matrix aggregation function, because of the complexity that it have. A based user feedback technique is treated in [11]. Authors propose an extensible QoS computation model for a QoS fair management. Nevertheless, the problem of QoS-based composition is not addressed. In [12] Zeng at al, focuses on a dynamic and a quality-driven selection of services. They use to find the best and the optimal selection of component services a mixed linear programming techniques. A similar work is proposed by Ardagnaet al. [14]. In this work authors extend a linear programming model to include local constraints. Linear programming methods suffer from poor scalability when the size of the problem is big, because of the exponential time complexity of the applied search algorithms [15]. In [16], heuristic algorithms are used to efficiently find a near-to-optimal solution. The authors propose two models for the QoS-based service composition problem: (a) a combinatorial model and (b) a graph model. A heuristic algorithm is introduced for each model. The time complexity of the heuristic algorithm for the combinatorial model (WS HEU) is polynomial, whereas the complexity of the heuristic algorithm for the graph model (MCSP-K) is exponential. In [17], a method for semantic Web service composition is presented. The proposed idea is based on Genetic Algorithms and using both semantic links between I/O parameters and QoS attributes. Despite the significant improvement of these algorithms compared to exact solutions, both algorithms do not scale with respect to the number of candidate web services, and hence are not suitable for real-time service composition. In [29] authors proposed a QoS-based web service selection approach. The approach adopts genetic algorithm to select the most suitable web service with user's QoS requests. Another approaches based on semantic web service are addressed to define non-functional models for selection of web services, such as WSMO [5], OWL-S[6], and SAWSDL[7]. In [8] [9] and [10], approaches based on UDDI extensions are proposed. For example in [9], authors added a Quality broker as an additional component in the SOA architecture between repository service requestor and UDDI. The problem with this approach is that only three non-functional properties are addressed. Another defects noticed, is that these kind of methods are based on a limited number of criteria because, it is not extensible for any new service quality.

The proposed Skyline based method in this paper is complementary to these solutions as it can be used as a pre-processing step to prune non-interesting candidate services even if the their number is huge and hence to reduce the computation time of the applied selection algorithm (even if number of constraints is big).

3. The proposed approach

The issue here is how to make a selection of a big number of services that provide the same functionality, but with different qualities, the composition becomes a problem of decision making on the selection of services, taking into account the functional and non-functional needs. According to the SOA paradigm, the composition of the applications is carried out as abstract processes composed of a set of abstract services. Then, at the time of implementation for each abstract service, a real web service is selected and used. This ensures a low coupling and flexibility in the design. The parameters of Quality of Service (QoS) (for example the reactivity, availability, throughput ...) play a major role in determining the success or failure of the composition. A composition of web services based on QoS aims to find the best combination of web services that satisfies a client request. Do an exhaustive search to find the best combination that meets a certain level of composition is not practical in this scenario, because the number of possible combinations can be quite large, depending on the number of sub-tasks that composes the process and the number of alternative services (with the same functionality) for each subtask. In our work, we address this problem using dominance relationships between web services basing on their quality of service attributes. We focus only on web services that belong to the Skyline set [18], i.e. which are not dominated by any other service with an equivalent functionality. These services are good candidates for composition.

We propose a method with two-steps to select services for an optimal composite web service. The web services are classified according to their functional cores to reduce the search time. We assume a set **S** of classes of services, which classify all available web services according to their functionality. Each service class **Sj** contains all services that provide the same functionality, but differ in terms of non-functional potential properties (QoS). Service providers provide the same web service with different quality levels: different response times, different prices...etc. The first step consists on selecting a subset of services, only the best ones. This first selection function is called Skyline. We used two algorithms adapted specially for this problem which are: Branch and Bound and Block Nested Loop. Different combinations of services of each class must be taken into consideration. However, all services are not candidates for the final solution. The basic idea in our approach is to perform a function on the Skyline services of each category to distinguish between services being candidates for a possible composition, and those who cannot possibly be part of the final solution can actually be pruned to reduce the search space [19].

For this, there are two types of algorithms:
- Approximate algorithms: used for classes with a large number of services. In our work, we chose as an approximate algorithm Branch and Bound (BB).
- Exact algorithms: Used for classes that contains a limited number of services. As exact algorithm we chose Nested Loop Block (BNL).

In [20], authors made a comparative study between these two classes of algorithms, the results showed that BB is the most performed algorithm followed by BNL. BNL is faster than BB but only in data spaces with a reduced size. However BB is the most efficient algorithm when dealing with huge data spaces.

The second step uses the Skyline services to construct a vector which will remain in main memory. The first step will be executed only one time because it is very costly in time calculation especially if we have several classes with a big number of web services. The second step was added because Web services can be inserted, removed and updated continuously, so it is impossible to perform the first step every time. On the other hand checking if a web service belongs to a set of Skyline services does not require much time because of the particular organization of Skyline services in memory.

3.1 Skyline

Given a set of points in a d-dimensional space, a Skyline function selects points which are dominated by other points. A point $\mathbf{a_i}$ (In our case points are web services and the dimensions are the qualities of services) dominates another point $\mathbf{a_j}$, if $\mathbf{a_i}$ is lower (or higher if we look for maximum) or equal to $\mathbf{a_j}$ in all dimensions and strictly less (above) in at least one dimension. Intuitively, a Skyline function selects the best points or the most interesting points in all dimensions.

In this paper, we define and exploit the dominance relationships between web services based on their quality of service attributes. This function is used to identify and reduce the number of services in a class that are dominated by other services in the same class. Determining Skyline services of a class of service requires pair's comparisons of QoS vectors of web services. This process can be expensive in computation time if the number of services is important. More efficient algorithms have been proposed for calculating Skyline. Therefore, we use two existing methods for determining the Skyline services in offline to accelerate the process of services selection after the client request [19]. Skyline operation aims to reduce data space and reduce calculation time. This data filtering is used to sort the present objects by putting those chosen in the central memory and those rejected to a buffer file.

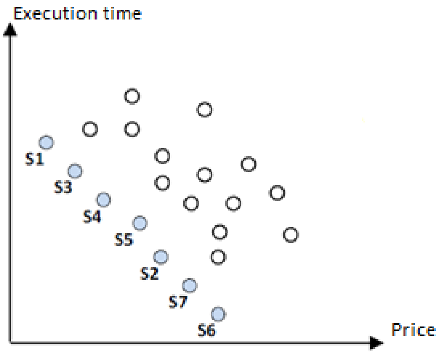


Fig 1. Skyline points (in Blue)

3.2 Utility function

A utility function is used to evaluate the overall multidimensional qualities of a given service, as an example, determine the classification by aggregating the vector of QoS in a single value. For this, we use the technique of Simple Additive Weighting technique [21]. The utility function calculation consists of using the attribute values of QoS to enable a uniform measure of QoS regardless of their units and their ranks. In this technique, each value of QoS is transformed into a value between 0 and 1, by comparing the minimum and the maximum value possible in based on available information about QoS of alternative web services (the other web services that provide the same functionality).
Qmin (j, k) and Qmax (j, k) are respectively the minimum and the maximum values of the *k-th* attribute of QoS of the class **Sj** of web services.

$$Qmin(j,k) = \min_{\forall s \in S_j} q_k(s)$$

$$Qmax(j,k) = \max_{\forall s \in S_j} q_k(s)$$

The function **U** is an aggregate function (Utility function). Now the utility of a web service s belonging to **Sj** is calculated as follows:

$$U(s) = \sum_{k=1}^{r} \frac{Qmax(j,k) - q_k(s)}{Qmax(j,k) - Qmin(j,k)} \cdot w_k$$

and the overall utility of a composite service is computed as follows :

$$U'(CS) = \sum_{k=1}^{r} \frac{Qmax'(k) - q'_k(CS)}{Qmax'(k) - Qmin'(k)} \cdot w_k$$

wk ∈ R^{+}_{0} and $\sum_{k}^{r} = 1$, the weight of the QoS *qk* represents the user priority, it is like a coefficient. If all attributes have the same priority (a user can prefer time execution than the price of a web service, in this case time weight will be greater than price weight), then all *wk* will be equals : w1 = w2 = w3 =…..= wr.

3.3 QoS calculation of composite web services

The value of a QoS of a composite service is determined by the QoS values of its sub services and the structure of the used composition (for example sequential, parallel, conditional and / or loops). Here, we focus on the sequential composition model. Other models can be reduced or transformed into sequential model [22]. The QoS vector of a composite service CS = {s1, ..., sn} is defined as follows:

Qcs = {q_1 (cs), ..., q_r (cs) or q} in which q_i(cs) is the estimated value of the attribute ***i*** of QoS and can be calculated by aggregating the corresponding values of the sub services. Typical functions of QoS aggregation are addition, multiplication, and minimum. Examples are given in Table 1.

Table 1: Examples of QoS aggregation functions

Type	Examples	Function
summation	response time, price	$q'(CS) = \sum_{j=1}^{n} q(s_j)$
	reputation	$q'(CS) = 1/n \sum_{j=1}^{n} q(s_j)$
multiplication	availability, reliability	$q'(CS) = \prod_{j=1}^{n} q(s_j)$
minimum	throughput	$q'(CS) = \min_{j=1}^{n} q(s_j)$

In what follows, we present the second step of our approach, which is a method of selecting the representative Skyline services to remedy to the situation where the number of Skyline web services of a certain class remains too large and cannot be treated effectively. The main challenge raised is how to organize a set of Skyline services that represent the best compromise of all QoS parameters, so it will be possible to find a solution that satisfies all constraints and also has a high utility score. The main idea is to put all Skyline services in a sorted vector. Services will be sorted according to their utility function.

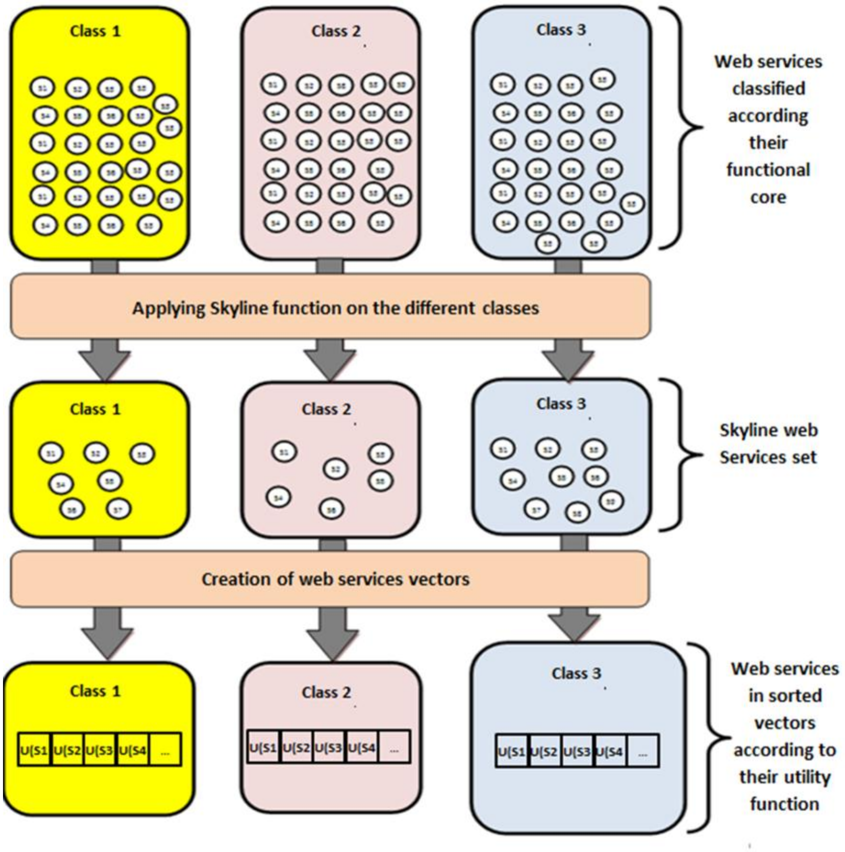

Class 1
Class 2
Class 3
Web services classified according their functional core
Applying Skyline function on the different classes
Class 1
Class 2
Class 3
Skyline web Services set
Creation of web services vectors
Class 1
Class 2
Class 3
U(S1
U(S2
U(S3
U(S4
...
Web services in sorted vectors according to their utility function

Fig 2. Web services process selection for a composition

At runtime, when a request for a web service composition is treated, the system starts with the first element (the first element contains the service with the highest value of the utility function).

To resume the proposed approach (Figure 2):

- Web services are classified in classes with the same functionality.
- A Skyline function is applied on these classes to select only the best candidates for the future composition.
- Two algorithms are used for Skyline selection: Branch and Bound, and Best Nested Loop. The first algorithm is an approximate method used for large data space. The second one is known for its exactitude but only in small data spaces. We took profits of these two qualities and we used BB when the class size is important to avoid long calculation time. And we used BNL with classes which have a limited size because we are sure that execution time remains acceptable with an optimal an extract results.
- After Skyline sets are created, real-time update, composition, insertion… will be possible.

4. Experimental study

In this section, we present an experimental evaluation of our approach, taking into account the measurement of the effectiveness in terms of execution time. We started our experience by creating databases with different dimensions according to QoS attributes with a measure of their values (Utility function), and also by increasing the number of services of each dimension in order to test our approach with a larger number of services and different distributions.

Firstly, we worked on a database having as a dimension: execution time and price, with a price value between [0,100] (MU), and time between [0,1] (MU). We filled the database with 10000 services, with different values for each attribute discussed before. Two algorithms which are well known with their limitations are adapted to our approach so it is up to us to customize them for our case like the specification of the size of the endpoints in the R-tree (the data space is organize with a specific method using R-trees technique [23]). The generated results of the used two algorithms (BNL and BB) will be put in sorted vector that contains only Skyline web services; each element contains the utility value of a service, and other information that allows its identification. Figure 3 is a graphical representation of the results of the first step. The red points represent Skyline services in the opposite of the black ones. We can notice that the most representative services represent a very limited number, what facilitates the real time web services composition. Tables 2 and 3 below contain the execution time of the two algorithms BB and BNL with of the number of web services. It should be noted that this test was carried out on a computer with an Intel (R) Core (TM) i3 M 370 2.40 GHz and 4GB RAM

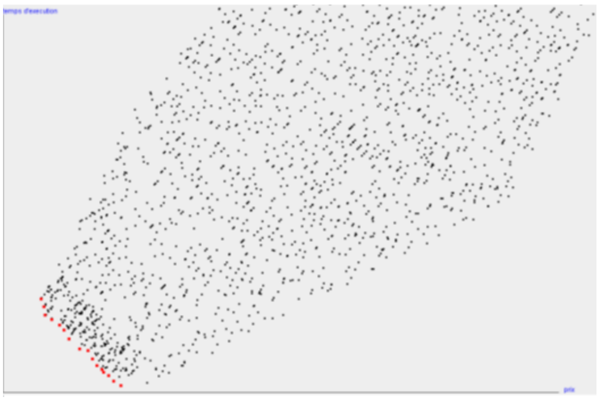


Fig 3. Graphical representation of Skyline web services

These results confirm the comparative study done in [20]. We notice that the time taken by the BB algorithm is significantly lower than time taken by BNL. But it does not preclude the use of BNL in some cases where the initial number of web services is not significant (neglected execution time and maximum efficiency).

Table 2. Execution time of BNL algorithm in 2d data space

Number of web services	*Execution time (in ms)*
100	73
1000	76
5000	1480
10000	2068

Table 3. Execution time of BB algorithm in 2d data space

Number of web services	*Execution time (in ms)*
100	1
1000	2
5000	4
10000	6

Computation time taken by the system during the creation of QoS sorted vectors in the second of our approach, are listed in the table 4. Because the system uses only Skyline services, we notice that the computation time is almost equals to zero (some milliseconds).

Table 4. Execution time to create Skyline vectors

Number of Skyline web services	*Execution time (in ms)*
10	≈0
20	≈0
50	≈0
100	8
200	9
500	10
1000	12
2000	13

Table 5. Comparison of some proposed QoS based selection methods

	Criteria				Overall result
Approaches	Performance	Hierarchical Properties	Automatic	Scalability	
Improve Protocol [9]	Low	No	Average	Low	Weak
Semantic [6][25]	Average	Yes	Average		Good
Policy [3][26]	Low	No	Low	Low	Weak
Trust and Reputation [27]	Average	Yes	Low		Average
Preference [13][28][29]	N/A	No	Low	Low	Weak
Our approach	High	No	High	High	Very Good

The results of this experimental study show a significant gain in terms of performance compared to existing approaches. The same thing is noticed for the management of web services i.e. when the system receives a new web services, or when a web service leaves the system, the required time for updating web services classes is also almost equals to zero, thanks to the selected data structure, what results a faster update.

5. Comparative evaluation

The comparison of Web Service Selection approaches that is described in section 2 is based on the following criteria:

5.1 Performance

Performance refers to time needed for web service selection and composition.

5.2 Hierarchical Properties

Ordering properties in an hierarchical structure is meaningful when most interested non-functional properties are at a lower level. Also, this is helpful to sort the properties and group them together, for example by their broader or their domain. While privacy and security are both safety aspects, speed and quality properties are performance aspects. By using this structure, users can show their preferences in the higher level and offerings will be represented by service providers, in useful detail. Therefore, new criteria will be added to mechanism of ranking, also a benefit is provided by considering aggregation of results into final score of ranking.

5.3 Automatic

Last step of service selection is performed by a human; it means, in a registry, user find appropriate services and decides to choose one of them. Providing fully automatic processes for service selection, is still one of the important things for researchers especially for service selection based on non-functional criteria. The essential thing in automatic service selection is in the last step when a service is available service designer specify data for it and user can specify the requirement but in performing service selection human doesn't participate. A situation that needs automation is the aggregation function selection; and the other one is for specific criteria, evaluation functions selection.

5.4 Scalability

Scalability is, once an approach considers lots of properties and also many ranking processes are occurred concurrently. Obviously in this situation the important thing is the accuracy of the result. Therefore by increasing the number of criteria or properties in selection mechanism the accuracy should not be affected.

Table 5 shows a comparison based on these criteria. We notice that the proposed approach is better than many existing solution in literature, in terms of performance, scalability…

6. Conclusion

In this paper we presented an effective method with two steps to make web services composition possible in real time, even if the number of the initial set is important. In the first step, we identify the Skyline services in terms of their QoS values to select representative services for a composition.

To deal with cases where the size of the Skyline is still large compared to the initial dataset, in the second step we organized services in sorted vectors according to their utility function. These vectors will remains in main memory for a quick access. When a client request arrives, only these vectors are used to make a limited number of combinations to find the best composition. The required time in this case is very limited as we showed it in the experimental study.

Finally, the results of the experimental evaluation indicate a significant performance gain in comparison to existing approaches. Our experiments have shown also that the performance of our skyline-based methods is affected by the difficulty of the composition problem, in terms of the number of the specified QoS constraints.

In the future work, we plan to develop a method for estimating the difficulty of each composition problem. We intend to use this method to deal with other problems like composition problem in cooperative information systems.

References

[1] G. Manish, S. Rajendra, and M. Shrikant, Web Service Selection Based on Analytical Network Process Approach,in Proceedings of the 2008 IEEE Asia-Pacific Services Computing Conference. 2008, IEEE Computer Society.

[2] Y. Liu, A.H.H. Ngu, and L. Zeng. QoS computation and policing in dynamic web service selection. 2004. New York, NY, United states: Association for Computing Machinery.

[3] H. Janicke and M. Solanki. Policy-driven service discovery. in 2nd European Young Researchers Workshop on Service Oriented Computing. 2007.

[4] H.Q. Yu and S. Reiff-Marganiec, Non-functional Property based service selection: A survey and classification of approaches. 2008, Sun SITE Central Europe.

[5] D. Fensel, M. Kerrigan, and M. Zaremba, Implementing Semantic Web Services. 2008, Berlin: Springer.

[6] U.S. Manikrao and T.V. Prabhakar. Dynamic selection of web services with recommendation system. 2005. Seoul, Korea, Republic of: Inst. of Elec. and Elec. Eng. Computer Society.

[7] M. Klusch and P. Kapahne. Semantic Web Service Selection with SAWSDL-MX. in Second International Workshop on Service Matchmaking and Resource Retrieval in the Semantic Web. 2008. Germany.

[8] M. Zuo, S. Wang, and B. Wu. Research on web services selection model based on AHP. 2008. Beijing, China: Inst. of Elec. and Elec. Eng. Computer Society.

[9] Y.-J. Seo, H.-Y. Jeong, and Y.-J. Song. A study on web services selection method based on the negotiation through quality broker: A MAUT-based approach. 2005. Hangzhou, China: Springer Verlag.

[10] E. Al-Masri and Q.H. Mahmoud. Discovering the best web service: A neural network-based solution. 2009. San Antonio, TX, United states: Institute of Electrical and Electronics Engineers Inc.

[11] Y. Liu, A. H. H. Ngu, and L. Zeng. Qos computation and policing in dynamic web service selection. In WWW (Alt. Track Papers & Posters), pages 66–73, 2004.

[12] L. Zeng, B. Benatallah, M. Dumas, J. Kalagnanam, and Q. Z. Sheng. Quality driven web servicescomposition. In WWW, pages 411–421, 2003.

[13] E.M. Maximilien and M.P. Singh, A framework and ontology for dynamic Web services selection. IEEE Internet Computing, 2004. 8(5): p. 84-93.

[14] D. Ardagna and B. Pernici. Adaptive service composition in flexible processes. IEEE Trans. Software Eng., 33(6):369–384, 2007.

[15] Maros. Computational Techniques of the Simplex Method. Springer, 2003.

[16] T. Yu, Y. Zhang, and K.-J. Lin. Efficient algorithms for web services selection with end-to-end qos constraints. ACM Trans. on the Web, 1(1), 2007.

[17] F. L´ecu´e. Optimizing qos-aware semantic web service composition. In ISWC, pages 375–391, 2009.

[18] S. B¨orzs¨onyi, D. Kossmann, and K. Stocker. The skyline operator. In ICDE, pages 421–430, 2001.

[19] Mohammad Alrifai , Thomass Risse, Article “Combining Global Optimization with Local Selection for Efficient QoS-aware Service Composition » . International World Wide Web Conference Committee (IW3C2) 2010

[20] Marios Kokkodis. Implementation Of Skyline Query Algorithms 2003

[21] K. . P. Yoon and C.-L. Hwang. Multiple Attribute Decision Making: An Introduction (Quantitative Applications in the Social Sciences). Sage Publications, 1995.

[22] J. Cardoso, A. P. Sheth, J. A. Miller, J. Arnold, and K. Kochut. Quality of service for workflows and web service processes. J. Web Sem., 1(3):281–308, 2004.

[23] GUTTMAN, A. 1984. R-trees: A dynamic index structure for spatial searching. In Proceedings of the ACM Conference on the Management of Data (SIGMOD; Boston, MA, June 18–21). 47– 57.

[24] D. Fensel, M. Kerrigan, and M. Zaremba, Implementing Semantic Web Services. 2008, Berlin: Springer.

[25] Y. Liu, A.H.H. Ngu, and L. Zeng. QoS computation and policing in dynamic web service selection. 2004. New York, NY, United states: Association for Computing Machinery.

[26] Y. Wang and J. Vassileva. A review on trust and reputation for web service selection. 2007. Toronto, ON, Canada: Institute of Electrical and Electronics Engineers Inc.

[27] S. Lamparter, et al. Preference-based selection of highly configurable web services. 2007. Banff, AB, Canada: Association for Computing Machinery.

[28] C. Schropfer, et al. Introducing preferences over NFPs into service selection in SOA. 2009. Vienna, Austria: Springer Verlag.

[29] Guofeng Chang QoS-Based Web Service Selection Approach. Software Engineering and Knowledge Engineering: Theory and Practice Advances in Intelligent and Soft Computing Volume 115, 2012, pp 887-892

The Appliance Pervasive of Internet of Things in Healthcare Systems

Mir Sajjad Hussain Talpur
School of Information Science & Engineering,
Central South University (CSU), 410083 - Changsha, China.

Abstract

In fact, information systems are the foundation of new productivity sources, medical organizational forms, and erection of a global economy. IoT based healthcare systems play a significant role in ICT and have contribution in growth of medical information systems, which are underpinning of recent medical and economic development strategies. However, to take advantages of IoT, it is essential that medical enterprises and community should trust the IoT systems in terms of performance, security, privacy, reliability and return-on-investment, which are open challenges of current IoT systems. For heightening of healthcare system; tracking, tracing and monitoring of patients and medical objects are more essential. But due to the inadequate healthcare situation, medical environment, medical technologies and the unique requirements of some healthcare applications, the obtainable tools cannot meet them accurately. The tracking, tracing and monitoring of patients and healthcare actors activities in healthcare system are challenging research directions for IoT researchers. State-of-the-art IoT based healthcare system should be developed which ensure the safety of patients and other healthcare activities. With this manuscript, we elaborate the essential role of IoT in healthcare systems; immense prospects of Internet of things in healthcare systems; extensive aspect of the use of IoT is dissimilar among different healthcare components and finally the participation of IoT between the useful research and present realistic applications. IoT and few other modern technologies are still in underpinning stage; mainly in the healthcare system.

Keywords: *Internet of Things (IoT); HealthCare Systems (HCS); Aspects; Quality; Regulations; Principles; Traceability; Tagging; Sensing.*

1. Introduction

In healthcare industry, Internet of Things (IoT) provides an opportunity of discovering healthcare information about a tagged patient or medical object by browsing an Internet address or database entry that corresponds to a particular Radio-Frequency Identification (RFID) tag. But now, it is extended to the general idea of medical things, especially healthcare everyday objects, those are readable, recognizable, locatable, addressable, and/or controllable. These objects may be equipped with devices such as sensors, actuators, and RFID tags, in order to allow patients, doctors, equipments and other healthcare actors to be connected anytime and anywhere with anything and anyone.

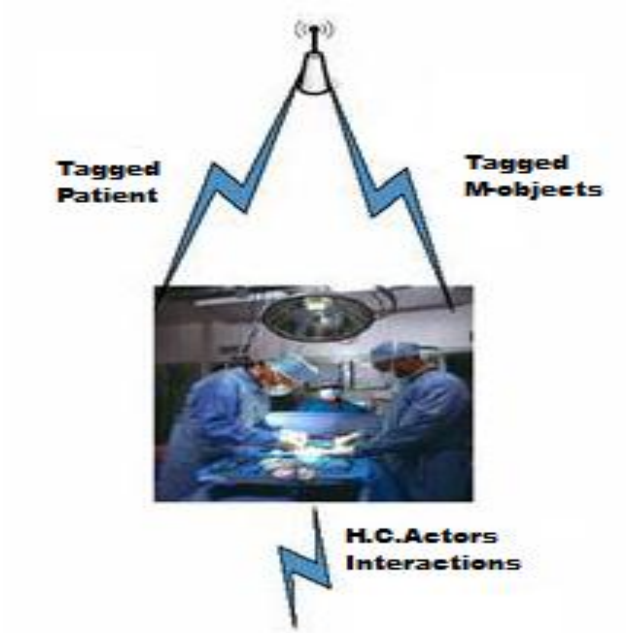


Fig1. IoT modeled Healthcare Actors Interactions

Through the medical object-to-person and object-to-object communications, IoT enables a wide range of smart applications and services to cope with many of the challenges that individuals and organizations face in their everyday lives, such as smart healthcare, smart home, smart earth, smart kitchen, smart transportation and smart office, etc. IoT is not a simple extension of the Internet or an aggregation of Internet systems, and it covers a wide range of technologies including tagging, sensing, networking, computing, storage, and control, which together build feasible complex cybernetic physical and social systems to support these smart applications [1], [2]. Internet of things has dynamic capabilities to connect D2M (Device-to-Machine), O2O (Object-to-Object), P2D (Patient-to-Doctor), P2M (Patient-to-Machine), D2M (Doctor-to-Machine), S2M (Sensor-to-Mobile), M2H (Mobile-to-Human), T2R(Tag-to-Reader), intelligently connects humans, machines, smart devices, and dynamic systems which ensure the effective healthcare system, health monitoring system, medical assets monitoring and medical waste management system[3].

Medical has ever remained one of the major applications of internet. The collaboration of internet and medical formed a sub-field e-health. An application of Internet and other related technologies in healthcare industry to improve the access, efficiency, effectiveness, quality of clinical and business processes utilized by healthcare organizations, practitioners, patients in an effort to improve the health status of patients [4]. Along with many other services, online appointment services in particular are the most common e-health services [3]. E-health is one of small component of IoT based healthcare management system, in which an online interaction of a patient to a doctor is made possible and easy along with easy access to online healthcare record checking for the patients. Whereas IoT based healthcare system consists all this plus identification and tracking of patients and doctors locations, tracking of patient's health records and tracking locations of hospital equipment etc. IoT have also enabled intelligent behavior of some equipment which alarm automatically when near to expiry or auto informing a relevant doctor if concerned patient is brought to Intensive Care Unit. *"Patients and medical objects tracking, tracing and monitoring are demanded by world wide healthcare institutions, governments and other organizations, the universal expectation of all healthcare organizations wants state-of-the-art patients tracking system, has put endorse superior requirements"* [5]. In order to enhance the protection of patients, improve the healthcare system and make them competitiveness of patient care, must develop an effective patient traceability system, traceability system of patients care should have dynamic and efficient management of the healthcare system [6]. Therefore, it has immense challenge for researcher or developer how to use information and communication technology; so the implementation of healthcare management system has become one of key areas. The IoT is a key technology that is quickly gaining ground in the development of modern secure wireless communications.

Tracking/ Tracing	***Sensing***	***Identification and authentication***	***Real time data collection***
Patients tracking and tracing at hospitals and outside to monitor the patient flow	Intelligent medication monitoring (pregnant women or elderly) at home or hospital	Protecting patient privacy	Automatic data collection and transfer is mostly aimed at reducing form processing time, process automation.
Tracking of patient location	home	Patient identification to reduce harmful incidents	Automated care and procedure auditing, and medical information management.
Tracking of drugs, supplies and procedures performed on patient	Sensing will be able to do real time monitoring of patients. Parameters such as blood pressure, glucose levels, heart and breathing rate	Eliminate wrong patient/wrong surgery	Relates to integrating IOT, RFID technology with other health information and clinical application technologies.
Accounting patient time in emergency department	Sensor devices enable function centered on patients, and in particular on diagnosing patient conditions, providing real-time information on patient health indicators	Patient identification to avoid wrong drug, dose, time, procedure	Integrating state-of-the-art physiological parameters monitoring time interval in order to determine actual time period.

Table.1. Potential of IoT in Healthcare System

The main idea of this thought is the pervasive existence around us of a mixture of stuff or substance; such as

Radio-Frequency Identification (RFID) tags, EPC technology, tiny sensors, actuators, smart phones, and so on. These things are capable to interact with each other and collaborate with their neighbors through the unique addressing schemes, in order to achieve their goals [7, 8]. In healthcare industry, IoT is applied for patient tracking by offering intelligent jackets, wristbands containing RFID tags. The tags interact with hospital information system for automating administrative tasks like patient admissions, patient transfers and discharges. *"The U.S Food and Drug Administration (FDA) have recently approved a tag called "veri-chip" in humans. These tags facilitate disoriented, elderly patients more safe by storing a detailed health information record. With this FDA approval in the headlines, suppliers have begun to offer patient wristbands containing tags"* [3]. Even though IoT`s application in patients tracking is less talked about, serious situations could be avoided implementing RFID tags to specific patients atleast, think the severity of this incident that occurred in phoenix recently to emphasize the importance in use of RFID tags, when a patient with dementia wandered from his / her room was found in the storage area, after 3 consecutive days.

2. The Structure of IoT

Through the Internet of things, anything in the healthcare system can be identified tracked and monitored on demand anytime anywhere [1]. Internet of things is considered as remarkable revolution after the blooming of Internet with ICT based industry. Internet of things has three basic components, namely RFID systems, middleware systems and Internet systems Savant. RFID system is one of the major components of IOT and it enables data to be transmitted by a portable device, called *"a tag"*, which is read by an RFID reader and processed according to the needs of a particular application [9]. The data transmitted by the tag may provide identification or location information, or specifics about the patient tagged, such as (e.g. patient ID, age, sex, blood pressure, glucose level) therefore the RFID systems can be used to monitor healthcare objects in real-time, without the need of being in line-of-sight. This allows mapping of real world healthcare system into the virtual world system. Middleware savant system is software that bridges RFID hardware and healthcare applications [10].

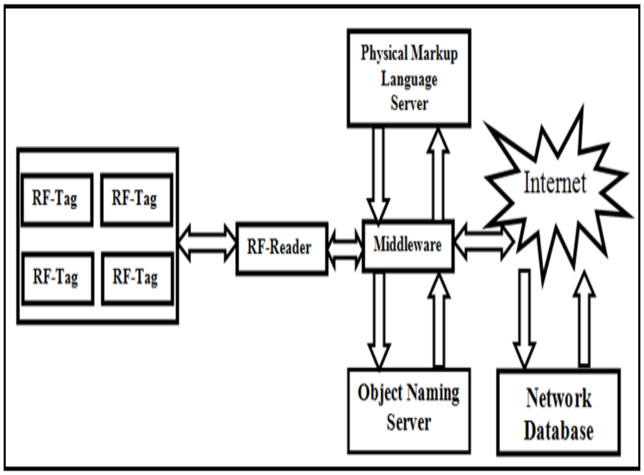


Fig 2. Basic composition diagram of IoT

Indisputably, the primary means of medical data gathering for any Radio frequency identification deployment and it consists of savant server, Object naming service servers, Physical Markup Language server and the corresponding medical data server software [11]. Internet system consists of state-of-the-art computer systems and secure network servers as shown figure2.Vision of healthcare technology mainly rely on accurate patient recognition in reducing healthcare complications, errors and harmful drug effects. The emboldens of IoT technologies in healthcare will ensure the healthcare safety, exact patient, accurate drug, proper dose, right way and exact time, by complying with the principles and regulations of HIPAA (Health Insurance Portability and Accountability Act of 1996) which mentions the standards of data exchange with protection and confidentiality of patient information, JCAHO (Joint Commission on Accreditation of Healthcare Organizations that emphasizes positive patient identification) and AHA (American Hospital Association) stressing guidelines for tamper proof non-transferable wristband minimizing the risk of losing transferred data. The Internet of Things(IoT) system in patient tracking provides non-transferable accurate patient identification which will ensure the safety of patient and reduced the harmful incidents in healthcare system, The IoT will improve efficiency of healthcare system, introduces affordable low cost RFID tags and tiny sensors which are associated to medical devices to reduce service costs, improve the quality of service, control the medical defects, improve identification and authentication of patients, automatic data collection and sensing system [12].IoT has vision to swagger the healthcare based smart communication technologies in order to connect the healthcare actors anywhere and anytime. That is why; Internet of Things played an exigent role in healthcare system.

3. Functioning Philosophy of IoT System

Internet of Things is a technological revolution that represents the future of computing and communications, and its development depends on dynamic technical innovation in a number of important fields, from wireless sensors to nanotechnology [1]. The essential functioning principles of Internet of things based on Radio Frequency Identification which is known as the soul of IoT, EPC technology used global unified products coding and wireless communications technology in order to tracing of healthcare objects, swank integrity of healthcare system [13,14]. Medical products, labeled with EPC code stored electronic tags. Furthermore, in the whole life cycle of the medical product; the EPC code easily recognizes a medical product, as EPC codes for the index in real-time query and modify medical objects information from the healthcare network, but also use it as signs, all movement in the healthcare system to find the medical product tracking. Entire healthcare actors are connected with RFID tags, when a RFID reader read range in its tags to monitor the existence, the label contained in the EPC and its linked data transfer to the savant middleware [10]. Initially the electronic medical product code data is key; in the local Object naming server (ONS) contains the information of medical product for EPC information server's network address, and savant query. According to the EPC information server address, access to medical commodity specific information, the essential cure, to transmit the information back-end healthcare applications to do a deeper level of computing ,at the same time, local EPC information server and source of this information server to record the reader to read and modify the corresponding information.

4. Networking Structure of Healthcare System

A. Healthcare Activities Analysis on the basis of IoT

Healthcare system is the organization of citizens, medical institutions, and healthcare resources to deliver healthcare services to meet the health needs of citizens [8].The healthcare system has been overwhelmed by problems such as patient diagnoses being written illegitimately on paper, doctors not being able to easily access patient information, and limitations on time, space, and personnel for monitoring patients. With advancements in technology, opportunities exist to improve the current state of healthcare to minimize some of these problems and provide more personalized service. Because characteristics of healthcare system, information flow is necessary to pass a higher, in time, need to be quicker, in space, and requires more strict storage conditions. Therefore, it made healthcare traceability an advanced technical requirement, while the development and stability of the healthcare system made a greater requirement [11].All nodes in healthcare system should be a integrated, harmonious division of medical staff and cooperation, and layout of the traditional healthcare system is comparatively complex, healthcare actors activities relatively unstable and so the delay caused by the healthcare organizations and asymmetric information, splits the chain between healthcare actors and enterprise [13]. It can be seen that the traditional healthcare system are often in motion or loose state of the information, timeliness, accuracy, and sharing, based on EPC technology, the application of medical things; a good solution to the above problem. An EPC have tag read and write data function, easy compactness and diversification of the shape, reusable, good penetration and data capacity and other characteristics, can adapt to frequent changes in the healthcare information system, communicate data, acquisition in a timely manner and system commands, widely used in medical products warehouse management, hospital transportation management, medical production management. Healthcare actors tracking, identification and significantly reduces the health complications [13]. Therefore, IoT application for healthcare, the use of RFID and secure databases will integrate all types of medical production, movement and the effective information quality and protection [15]. Medical information collection, storage, transmission, the healthcare system and safety of healthcare actors combined through the entire system, and established healthcare actors traceable system based on IoT. Which archive a summary of the medical data and information in the entire healthcare system to make regulatory and patient information track in the platform, and guarantee transparency throughout the healthcare system process [16].

B. Brunt of Internet of things on Healthcare System

The brunt of patient identification and medical objects identification processes in healthcare system for instance: patient`s identification to reduce harmful incidents to patients (e.g. wrong drug/dose/time etc). Relation to staff, identification and authentication are most frequently used to grant access and to improve employee morale by addressing patient safety issues. In relation to assets, identification and authentication is predominantly used to meet the requirements of security procedures, to avoid thefts or losses of important instruments and products, so in tracking and identification of patients, healthcare staff identification and medical assets identification. In the entire system, the main objective of IoT is to manage the identification process of healthcare actors [8]. RFID tags and tiny sensors are attached to patients and patients also wear bracelet and it has unique identification number of RFID tag, which process and record all relevant information collected on the birth, at the same time, applied RFID chip, tiny sensor in order to manage patients' medical information, assets and other healthcare actors' information.

5. Intensification Necessitate in Health Monitoring System for Quality of Service

Internet of things provides an effective way of real time remote monitoring system of healthcare actors through RFID tags, sensors, and actuators. The RFID tags in healthcare may be applied to patients, assets, medical staff and other objects, allowing the readers on gate frames, hospital wards and other treatment areas of hospital to detect and record interactions. IoT is applied for patient tracking by offering wristbands containing RFID tags. The tags interact with healthcare information system for automating managerial everyday tasks like patients' admissions, transfers and discharges [17], [18]. IoT applications in medical administration to streamline the processes and reduce healthcare harmful incidents the above arguments prove that Internet of things ensure the safety of patients and quality of service.

6. Conclusion

The name, Internet of Things, is syntactically composed of two terms, Internet and Things. As a result, it's usually considered that there are two versions of IoT, "Internet oriented" or "Things oriented". The "Things oriented" version focus on the technology developed to improve object visibility, such as awareness of its status, current location. This is undoubtedly a key component of the path to the full deployment of the IoT vision but it's not only one [19]. In past decades, the Internet has connected numerous devices. Why don't we take advantage of the existing Internet technology and connect smart objects around the world? [13]. We discussed the significance of IoT especially in healthcare system; immense prospects of Internet of things in HCS; extensive aspect of the use of IoT is dissimilar among different healthcare components and finally the participation of IoT between the useful research and present realistic applications. IoT and few other advance technologies are still in underpinning stage; mainly in the healthcare system. This article tries to emphasize a healthcare system not only to realize the illustration and traceability of healthcare actors guarantee the quality but also effectively controls healthcare actors.

Acknowledgment

This work is supported by National Natural Science Foundation (NSF) of Peoples Republic of China under grant numbers 61073037 and 61103035, Ministry of Education fund for doctoral disciplines in Higher Education under the grant number of 20110162110043. The authors are thankful to the Ministry of Education for the financial support.

References

[1] Internet of Things in 2020: Roadmap for the future. (May, 2008)Retrievedfrom:http://ec.europa.eu/information_society/policy/rfid/documents/iotprague2009.pdf.

[2] D. Giusto, A. Iera, G. Morabito, L. Atzori (Eds.), The Internet of Things, Springer, 2010. ISBN: 978-1-4419-1673-0. Italy.

[3] CERP-IoT – Cluster of European Research Projects Vision and Challenges for Realising the Internet of Things March 2010.

[4] IERC(2010) Vision and challenges for realiazing the Internet of things. IERC cluster book. March 2010.

[5] IoT-A(2010) FP7 IP Internet of Things Architecture, online:http://www.iot-a.eu.

[6] Zhang Yingfu, The Technology and Application of The Internet of Things [J].Communication & Information Technology, 20 10(01)5 1-53.

[7] He Ke, The Key Technologies of lOT with Development & Applications [J]. Radio Frequency Ubiquitous Journal, 20 10(0 1)32-35.

[8] Internet of Things P7_TA (2010)0207 European Parliament resolution of 15 June 2010 on the Internet of Things (2009/2224(INI)) (2011/C 236 E/04).

[9] National Intelligence Council. Disruptive Civil Technologies-Six technologies with potential impacts on US Interests out to 2025. Conference Report CR 2008-07. Apr. 2008.

[10] M. Botterman, for the European Commission Information Society & Media Directorate General, Networked Enterprise & RFID Unit, D4, Internet of Things: An Early Reality of the Future Internet, Report of the Internet of Things Workshop, Prague, Czech Republic, May 2009.

[11] M. Evered, S. Bogeholz, A Case Study in Access Control Requirements for a Health Information System, Australasian Information Security Workshop, 2004.

[12] Lu Wenjun, "IoT makes the City Smarter," Science and Culture, Vol 10, pp. 12-13, October 2010.

[13] Yang Zhen. "The development of the Internet of Things," Journal of Nanjing University of Posts and Telecommunications(Social Science). Vol. 12, No. 2, pp. 8-9, June 2010.

[14] Rolf H. Weber, "Privacy Regulations in the Internet of Things", Law and Technology Centre University of Hong Kong, January 24, 2011.

[15] J. Marconi, "E-Health: Navigating the Internet for Health information Healthcare", Advocacy White Paper. Healthcare Information and Management Systems Society, 2003 May 5. pp. 1-6.

[16] O. Rienhoff, Integrated Circuit Health Data Cards (Smart Cards): A Primer for Health Professionals. Washington, DC: PAHO, 2003. pp. 560-565.USA.

[17] Alsinet, T. et al.: Automated monitoring of medical protocols: a secure and distributed architecture, Artificial Intelligence in Medicine, Volume: 27, pp. 367-392. (2003).

[18] Cortes, Ulises et al.: Intelligent Healthcare Managing: An assistive Technology Approach, IWANN 2007, LNCS, pp. 1045-1051 (2007).Canada.

First Author: Mir Sajjad Hussain Talpur received master`s degree in Information Technology from Shah Abdul Latif University Khairpur mirs, Pakistan in 2003.He is currently a Doctoral Degree Candidate at the Trusted Computing Institute, School of Information Science and Engineering, Central South University, Changsha, Hunan China. He is also a Lecturer at the Information Technology Institute, Sindh Agriculture University, Sindh Pakistan. His current research interests include Internet of Things in Healthcare System.

Experimental study on bending constitutive relation of steel box-concrete combined member

Wenjuan Yao [1], Wu Yang[1], Xiaoyu Liu [2]

[1] Department of Civil Engineering, Shanghai University
Shanghai 200072, China
[2]School of Civil Engineering and Architecture, Chongqing Jiaotong University
Chongqing 400074, China

Abstract

According to the direct homogenization theory of inhomogeneous material, a single medium homogenization constitutive model of the confined concrete in the steel box-concrete combined member has been established, and model parameters have been determined by the measured data of the bending test. By analyzing the theoretical ultimate strength and overall process curve, the results obtained is consistent with the curve measured by the test, which proves the bending constitutive relation of steel box-concrete combined member is correct.

***Keywords*: Experimenter, steel box-concrete combined member , bending constitutive relation**

1. INTRODUCTION

When the restrained concrete is acted upon by lateral pressure, its ultimate compressive stress and ultimate compressive strain are improved significantly. Since 1903 when Considere first proposed that using spiral hoop can constrain axial-compression column effectively, people has studied the constitutive model of the restrained concrete for one hundred years. Meantime, many stress and strain constitutive models have been put forward, such as Chan's model, Sargin's model, Kent – Park's model, Zhang Xiuqin's model, Saatcioglu's and Vallenas' model, Sheikh's model, Mander's model, Fafitis' model, Lin Tongyan's model, Xing Qiushun's model, and so on1. Zhang Xiuqin's model2 is based on the research of the work of plain concrete, and studied the stress - strain curves of the restrained concrete under different steel ratios by experiment, besides, and obtained the corresponding equation. In Sheikh's model3 , the strength and ductility of four axial-compression restrained concrete members with different composite reinforcement section form, reinforcement ratio , stirrup spacing has been studied, and stress - strain skeleton curves of the restrained concrete have been put forward. Kent – Park's model4 is also based on the research of ordinary reinforced concrete, without the consideration of the effect of stirrup layout styles on the mechanics properties of concrete. Saatcioglu's 5 is based on the experiment of the memembers with the constraints of circular hoop, simple square hoop, composite reinforcement, or rectangle hoop. However, the ascent stage form adopted is not appropriate, when the strain is approaching to the zero, stiffness will be close to infinity. The expression of Mander's model 6 is concise, with a clear mechanics concept, which can well reflect the phenomenon that concrete ultimate strength and peak strain increases and descending branch changes gently with the increase of effect of restraint .

The steel box-concrete combined member is a new type of stucture, the model mentioned above can not be applied completely to its constitutive relation .Based on the similarity between concrete with the restraint of stiffening rib, steel box and stirrup and concrete with the the restraint of stirrup , this paper adopted the direct homogenization theory of inhomogeneous material 7 and used constitutive model of restrained concrete under a single axle load put forward by Mander et al for reference. The constitutive relationship of restrained concrete under the restraint of stiffening rib, steel box and stirrup has been put forward combined with the characteristics of constraint mechanism of concrete under the restraint of stiffening rib, steel box and stirrup, which has been verified by experiment.

2. EXPERIMENT DESIGN

2.1. Manufacture of Experiment Specimen:

The experimental material is Q235 ordinary hot rolled steel plate, with the depth of 2.5mm, the measured value of tensile strength is 290Mpa. The maximum grain size of the microconcrete is 10mm， and the measured value of concrete cube fc is 49.3Mpa. Four first grade steel bars have been allocated in the concrete, including two steel bars with the diameter of 8mm above and two steel bars with the diameter of 4mm below. PBH rigid connection was used between concrete and steel box. The specific dimensions are shown in Table 1.

Table 1. Parameter Table of Experiment Specimen

Number	Length (mm)	Width (mm)	Height (mm)	Thickness of the steel plate (mm)	Concrete width (mm)	Eccentricity (mm)	Concrete strength (MPa)	Strength of the steel plate (MPa)	Annotation
M-1	1300	120	180	2.5	60	-	49.3	330	Pure bending beam
M-2	1300	120	180	2.5	60	500	49.3	330	Bending beam
M-3	1300	120	180	2.5	60	220	49.3	330	Bending beam

2.2. Loading Project of Experiment

The experiment of bending beams including pure bending beam and bending beam. In order to get the pure bending section, four points loading method was used while distribution beams were allocated under the lifting jack for loading. As shown in Figure 1.

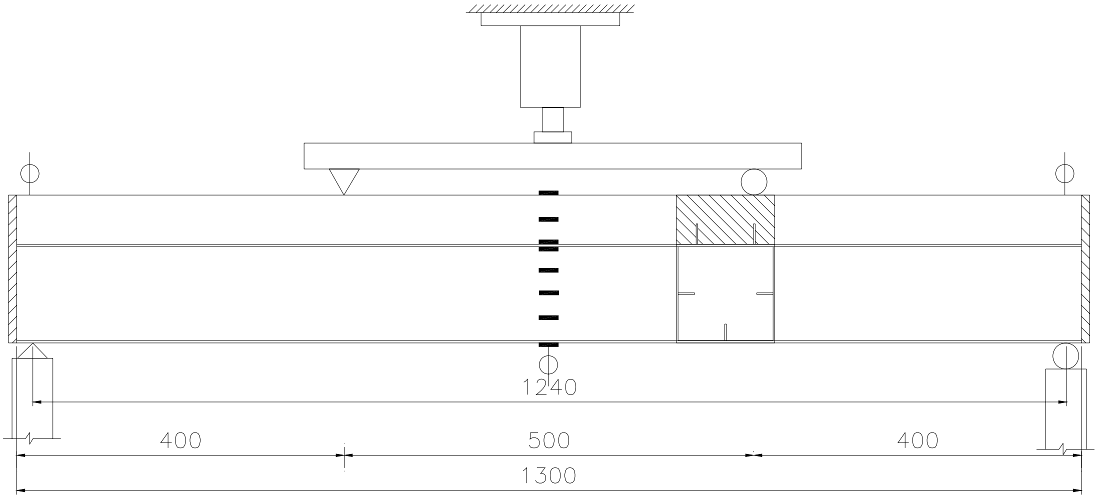
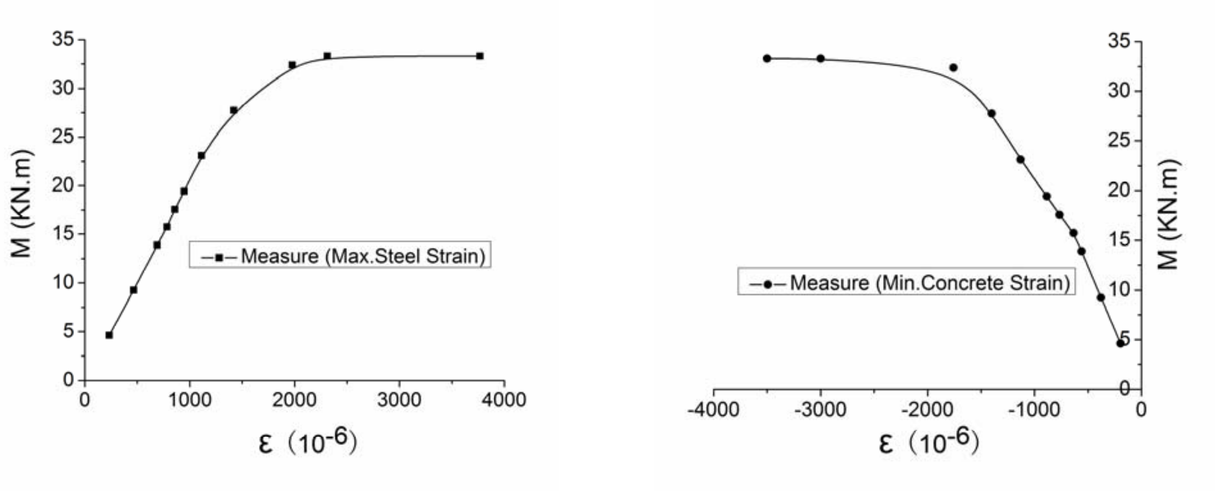


Fig. 1. Loading modes of bending beam in the experiment

The applied force of load , deflection and strain data have been tested during the whole process of the experiment. The concentrated force F was obtained by the force cell. The deflection has been obtained by the dialgage, which has been installed at the supports at the two ends and midspan. The strain of the concrete has been read by the foil gauge with the gauge length of 50mm, and the strain of the steel plate has been read by the foil gauge with the gauge length of 2 mm.

2.3. Experimental results

The maximum measured values of the strain of steel box and concrete have been shown in Figure 2.



(a)

(b)

Figure 2. Measured values of the strain. (a) Measured values of the max steel strain (b) Measured values of the min concrete strain

The strain of every section under load is shown in Figure 3. As shown in Figure 3: When the load increases, the strain value increases gradually along the depth of the section, but the profile of the strain is basically touching angle, which consists with section assumption, and there is no obvious strain change at the boundary. Besides, there is a trend that neutral axis ascends a little.

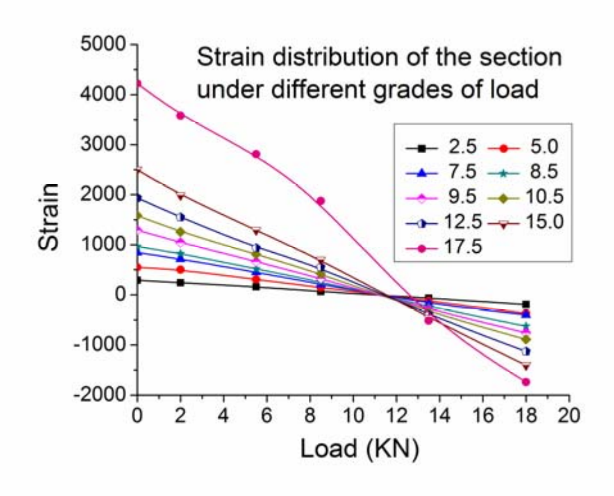


Figure 3 Profile of the strain for each section

3. DETERMINATION OF CONSTITUTIVE MODEL PARAMETER

Mander model is shown in Figure 4, the formulation of each curve is as follows:

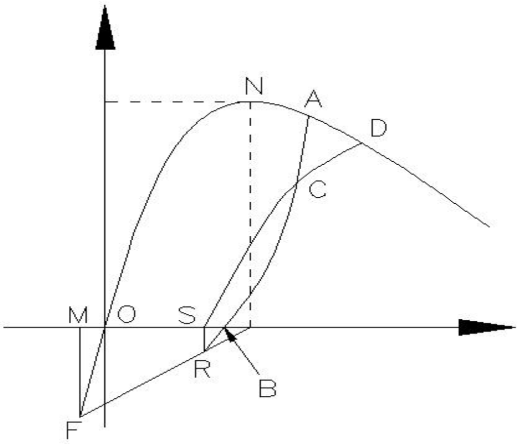


Figure 4. Mander's Model

3.1. The formula of skeleton curve ONAD is as follows:

$$\begin{cases} f_c = \dfrac{f_{cc} x r}{r-1+x'} \\ x = \varepsilon / \varepsilon_{cc} \\ \varepsilon_{cc} = \varepsilon_{c0}\left[1+\eta\left(\dfrac{f_{cc}}{f_{c0}}-1\right)\right] \end{cases}$$

3.2. The formula of unloading curve AB is:

$$\sigma = \sigma_u - \frac{\sigma_u x_2 r_2}{r_2^2 - 1 + x_2^{r_2}}$$

Thereinto, $r_2 = \dfrac{E_u}{E_u - E_{\sec 2}}$;

$$E_{\sec 2} = \frac{\sigma_u}{\varepsilon_u - \varepsilon_{pl}} ;$$

$$x_2 = \frac{\varepsilon - \varepsilon_u}{\varepsilon_{pl} - \varepsilon_u} .$$

Eu is the initial tangent modulus of unloading curve.

3.3. The reloading curve is composed of two parts, and SC is straight line, CD is cubic parabola.

The formula of straight line is :

$$\sigma = \sigma_r + E_r(\varepsilon - \varepsilon_r)$$

The formula of curve is:

$$\sigma = \sigma_{re} + E_{re} x_3 + A x_3^2$$

Thereinto, : $E_r = \dfrac{\sigma_r - \sigma_{new}}{\varepsilon_r - \varepsilon_u}$; $x_3 = \varepsilon - \varepsilon_{re}$;

$$A = \frac{E_r - E_{re}}{-4[(\sigma_{new} - \sigma_{re}) - E_r(\varepsilon_u - \varepsilon_{re})]}$$

fc, ε are the axial stress and strain of core concrete respectively, and f_{c0}, ε_{c0} are axial compression strength and peak strain of unconstrained concrete respectively. f_{cc}, ε_{cc} are axial compression strength and peak strain of confined concrete

respectively. η is corrected parameter of peak strain. Ere is the tangent modulus of point D on the skeleton curve in Figure 6-9. σ_{new} is the corresponding stress of point C.

By means of pilot calculation , homogenization material constitutive relation of the restrained reinforced beam in the steel box-concrete combined bending member has been obtained after many calculations. The key parameters are as follows: fc0=1.20fck、εc0=0.002； fcc=1.81fck、εcc=0.007056；η=0.5.

4. VERIFICATION OF CONSTITUTIVE RELATION

Overall process analysis have been made for the bending moment- deflection curve and bending moment-strain curve, and the curves are compared with the actual measurement overall process curve (Figure 5 and Figure 6).As shown in Figure 5 and Figure 6, The overall process curves are in good agreement .
The bending moment- deflection curve is as follows.

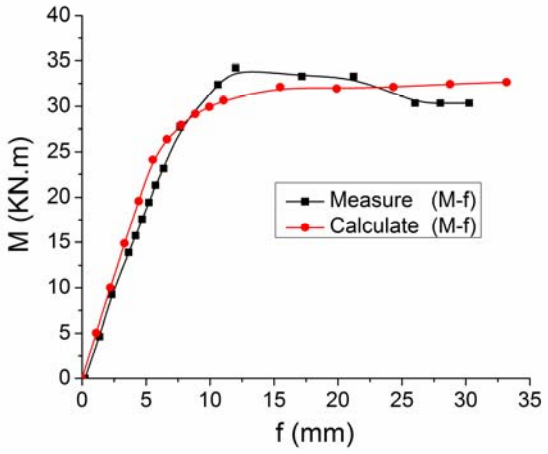
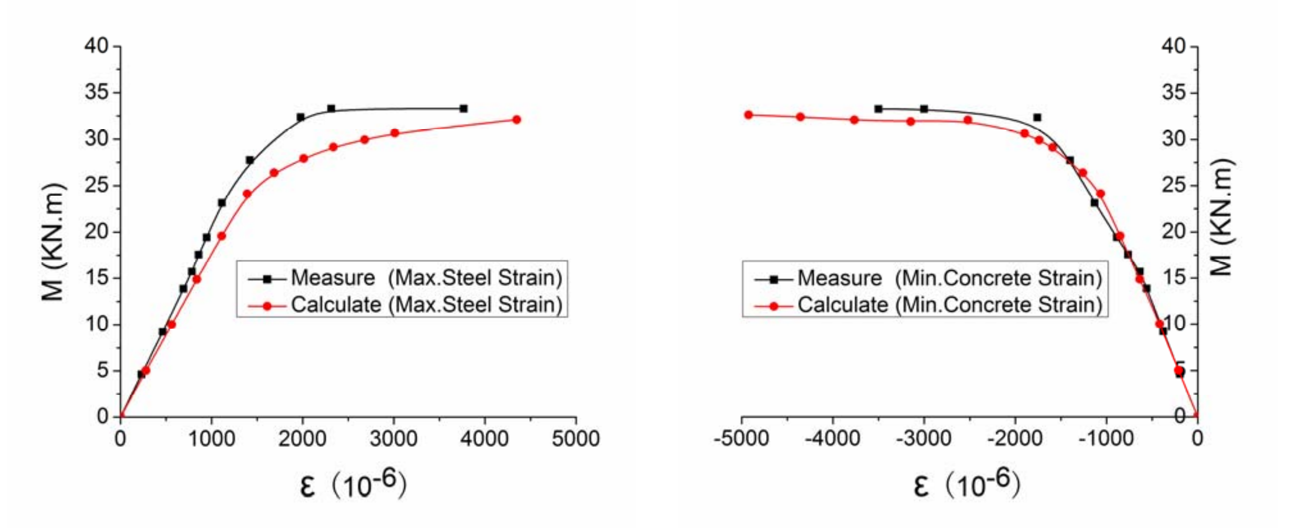


Figure 5: Comparison between experiment value and theoretical value of M-f curve

The bending moment-strain curve is as follows.



(a)

(b)

Figure 6: Comparison between theoretically calculated curve and test curve. (a).Comparison between experiment value and theoretical value of the maximum tension strain for M-steel box. (b). Comparison between experiment value and theoretical value of the maximum compression strain for M-concrete

5. CONCLUSION

1) As shown in Figure 5 and Figure 6, the calculated curve is very similar with the test curve in shape, and there is little difference. The calculated strain for the steel box is a bit larger than the test value. This is partly because that the measurement of strain is not the maximum value absolutely, but the average value along the length of foil gauge. The test value of the maximum compression strain for the concrete is in good agreement with the calculated value .

2) The over process curve and test curve of the ultimate strength are in good agreement, which indicates that the constitutive model in the article is suitable for the structural analysis of steel box-concrete combined bending member.

Acknowledgments

We thank the support of the National Natural Science Foundation of China (50578168) and the western transportation construction project in Ministry of Chinese Communications (200631881448).

References and Notes

[1] W.F.Zhou, Z.M.Huang, S.L. Bai. Introduction and Comparison of Several Representative Confinement Models for Concrete[J] Journal of Chongqing Architecture University，2003，25，(4):121-12

[2] X.Q.Zhang, Z.H.Guo, C.Z.W. The Equation for Complete Stress-Strain Curve of Concrete with Transverse Reinforcement under Cyclic Loading. [J] Journal of Building Structures, 1982,(9):16-20

[3] S.A.Sheikh and S.M.Uzumeri. Analytical Model for Concrete Confinement in Tied Columns [J]. ASCE,1982, (12):2703-2722

[4] R.Park, M.J.N.Priestley, W.D.Gill. Ductility of Square-confined Concrete Columns[J].ASCE,1982, (4):929-951

[5] S.M.Saatcioglu, S.R.Razvi. Strength and Ductility of Confined Concrete [J].ASCE,1992, (6):1590-1607

[6] J.B. Mander, M.J.N. Priestley, R.Park. Theoretical tress-strain Model for Confined Concrete [J].ASCE, 1988, (8):1804-1826

[7] X.M.MAO, Analysis of Interface Behavior on Steel-Concrete Composite Beams and Experimental Study of Load Distribution Width on Stiffened Steel Plate and Concrete Composite Plates.[D]Southwest Jiaotong University,2006.

[8]Mehdi; Adeli, Ali; masoumi, Azra; sargolzae, Mehdi A Comparative Study on ANFIS and Fuzzy Expert System Models for Concrete Mix DesignNeshat, International Journal of Computer Science Issues, v 8, n 3 3-2：196-210

[9] Monfort, Valérie ; Cherif, Sihem，Experimenting a service based connectivity between Adaptable Android, WComp and OpenORB，International Journal of Computer Science Issues, v 8, n 4 4-2：p 1-12

First Author：Wenjuan Yao (1957-), female, PhD, Professor, doctoral tutor, mainly engaged in engineering mechanics, soil structure interaction, model material mechanics theory and numerical method research.

Second Author: Wu Yang (1982-), male, PhD, mainly engaged in bridge structure calculation method research.

Third Author：Xiaoyu Liu (1955-), female, Professor, Master Instructor, mainly engaged in structure design and calculation method research.

Challenges of Online Exam, Performances and problems for Online University Exam

Mohammad A Sarrayrih[1], Mohammed Ilyas[2]

[1] Information System and Technology Department, Sur University College,

[2] Information Systems and Technology Department, Sur University College,

Abstract

In this paper, we propose a system that provides security to improve on-line examination by utilizing technologies such as biometric authentication, internet-firewall, cryptography, network protocol and object oriented paradigms. Furthermore, we propose a framework for conducting on-line exams through insecure internet backbone. However, the proposed system will provide a secure communication based cryptography and group communications. In our research paper, we discuss the performance of student's online course exam with respect to security and main challenges faced by online course exams within the university. We conclude that by improving the security system using biometrics face recognition that can be incorporated into the proposed system to fulfill the challenge of online exam.

Keywords: Biometrics, Course, Online Exam, Security, University Course, Camera, Fingerprint Scanner

Introduction:

Online exam has expanded rapidly [1], [2]. Even so, the off-line exam is usually chosen as evaluation method for both on-line and off-line exams.

Online course examinations are useful to evaluate the student's knowledge using modern computer technology without any effects on the traditional university course exam that uses Pens, Papers and invigilators.

Online exam can improve the standards of student's examination whereas the traditional examination system using the pen and paper requires more effort on the part of students and invigilators.

Online examinations are considered an important source for university exam, and the development of network technology polices has given the possibility to conduct the exams online. Thus, the university students can benefit from these services.

University course exams , using the multiple choice questions and allowing the students to choose only one answer from alternative answers or the true/false questions, are traditionally using the paper and pens and they have always been a heavy load for both students and lecturers. Computer new technology has been generally useful to the fields of education. In attitude and tools, the new computer technology gives the lecturer the advantage of an effective assessment.

The traditional way of identifying the students is checking the student card, driving license, resident card or Passport.

The online process and security of the online exam system helps with eliminating cheating. This paper proposes the usage of biometrics which supports the security control, authentication and integrity of online exam process. E-monitoring of students uses finger prints and cameras for preventing cheating and substitution of the original student.

This paper targets the online exam for Basic computer in university courses with students at particular locations, at a fixed time and same questions for all examinees at the restricted physical location of the examinees.

Literature Review:

Most modern online education uses Web-based commercial courses management software [3] such as Web CT [4], blackboard [5], or software developed in-house. This software is not used widely for online exams, due to security vulnerabilities, and the system must rely on students' honesty or their having an honor code [6].

Online course exam nowadays becomes more efficient than before; online course exam need for enhancing the security. Jung, I.Y proposes an enhanced secure online exam management environment mediated by group cryptography using remote monitoring and control of ports and input [7].

Holding the Online course exam for any substance requires more preparations, whether the teacher or through the support of university students. University on the duties assigned to it to provide the necessary environment entrusted to them. Everyone is there to serve the student and we have to encourage students and train them psychologically for a computerized exam, note that many universities in the world of the complexity of computerized tests on its campus. More of recent research shows the advantage and disadvantage of using online course exam on the university campus such as Al-Mashaqbeh, I.F. and Al Hamad, A. in the Dept. of Computer. Educ., Al al-Bayt Univ., Mafraq, Jordan reached to good results showed that there was a positive perception towards the adopting of online exam. They measured students' perceptions toward the use of online exam as an assessment tool on university campus within a Decision Support System Course at Al al Bayt University [8].

A study has been conducted on online exam and traditional exam which indicates that an online exam has better results than traditional exams. [9]

Considerable discussion has taken place on group protocols and group-mediated communications to ensure secure communications among group members [10], [11]. This discussion has included the consideration of secure group composition, secure intergroup communication using a public key, and secure intragroup communication using the symmetric key through the Diffie-Hellman key exchange [12]. This paper adopts two groups for secure communication between distributed entities in the online exam system. The intergroup communication is protected through public key infrastructure (PKI), while intragroup communication uses several symmetric Diffie-Hellman

Keys. The “group” in this paper is a concept for entities with similar roles. [12].

In this research, we try to bring out the challenges and some best solutions that may solve the problems. This paper considers the **Challenge of personal identity** and unauthorized invention of other users in the **network using other clients**

Solutions for the above challenge

1) Challenge of personal identity:

The special cameras of 360^o and finger print recognition device will be incorporated for identifying the identity. The camera and the finger print device will be placed at one location in each lab. The biometric scan devices (finger print scanner and camera 360^0) will check the students from the data base which is collected and stored in the registration department. The 360^0 camera is used for dual purpose of identifying and controlling of examination hall activities. Thus, we are utilizing the same resource for identifying the students.

2) Unauthorized interference of other users in the network using other clients

To solve this challenge of students entering from different IPs into the domain and attempting the exam for their fellow students, we propose a system, where we create a domain with the set of students user id’s allocated by the university domain and each instructor will add all the students user id’s of his course; then he will give them the specific permissions like read and write for the specific time of that particular course exam.

The students who enter from the different IP’s cannot use the allocated domain and thus the system is secure.

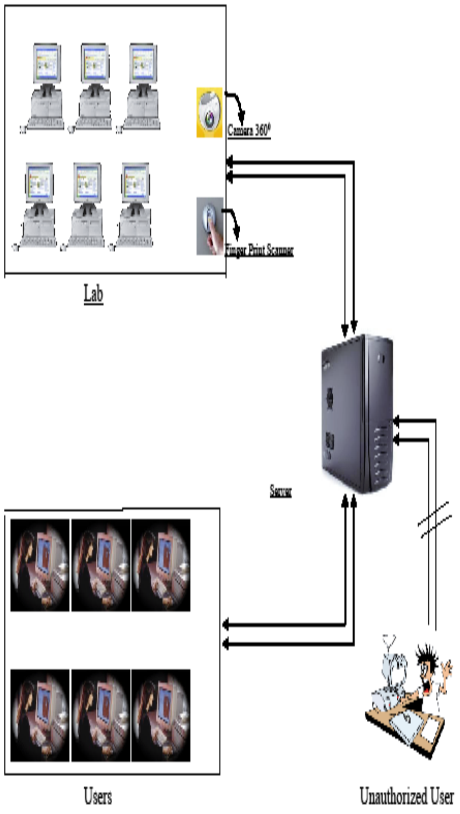


Figure1: System architecture

Figure1: The systems are connected using the star topology. The camera and finger print scanner inside the lab are connected to the security server; once the security server authenticates the biometrics of user, then the users are allowed to write the exam at the specific terminal provided to them. When an unauthorized user attempts to access the system from different location he is not allowed.

The proposed system

The special exam group is created by grouping the hostnames / IP of clients for a specific location (Computer Lab) and time.

To avoid the malpractice in the exams we use different types of biometrics as a means to log into the exam.

We use the camera and finger print scanner to identify the students as shown in the **figure1** above.

The user after identified login into the system uses the user-id and password provided by the university, which are authenticated by the server.

This gives him/ her permission to open the exam from the server otherwise the students cannot login into the system.

The unauthorized users attempting to log into the system from remote computers are blocked by the proposed system

Once the session begins the timer is on, the student completes his exam within the allocated time and once the time is up the system send an alert and logs the user off.

Figure 2: The **figure 2** shows the flow chart of the secured online exam system proposed. It shows the series of steps of online exam starting with the secured login using biometrics and system login through server till the end of exam results.

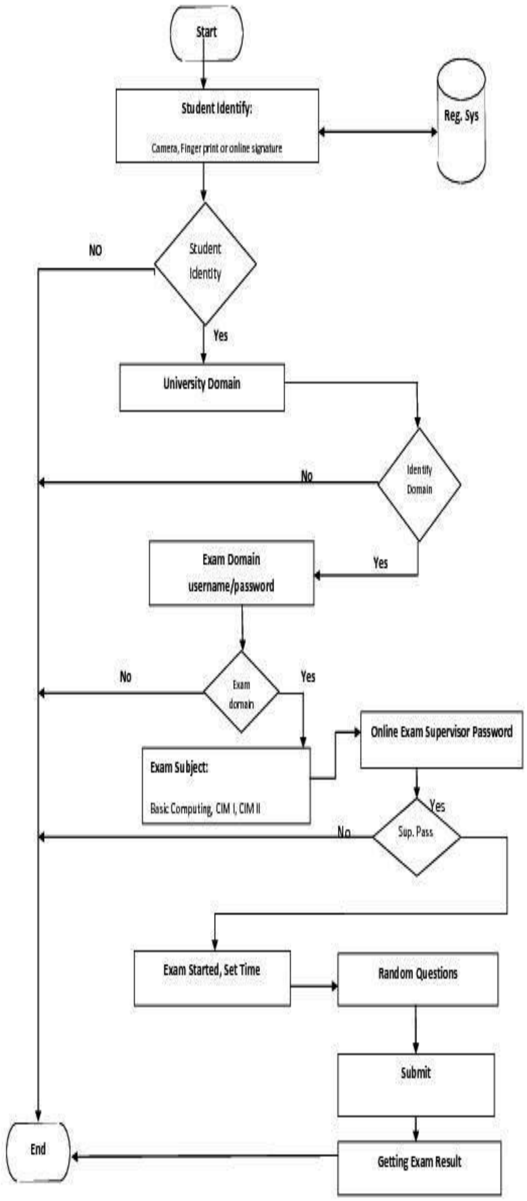


Figure2: Flowchart of secure online exam proposed

Algorithm:

Step 1: Student Identification: The system will check the identity of the student by using biometrics which will take the picture and the fingerprint before entering the exam. This will also check whether the student is eligible for that particular exam.

Step 2: **University Domain Login:** The student will log into the domain of the university with the user name and password provided by the university domain login (Ex: username: SUC, Password: suc).

Step 3: **Special login into exam domain:** The system asks the user to write the user name and password. If the user name and password are correct, then the user will be able to log into the exam.

Step 4: Access the Exam: The user will complete the exam file that is located in the domain desktop window (Online Exam)

Step 5: Online Exam Supervisor Password: The supervisor password is given to the students who are successfully logged into the exam domain. This gives them access to the exam and the exam session begins for that specific exam.

Step 6: Random questions and Results: The random questions are given to the students, who submit the answers to the server; when the session is completed, the system generates the result of the exam.

Step 7: End.

Conclusion:

We believe the online format is considerably superior to paper-and-pencil exams for our courses.
We have come to the conclusion that the above mentioned challenges can be solved by introducing the following security systems. Using biometrics we overcome the traditional way of checking the ID cards of the students after they start the exam. Biometrics will identify the student as he enters the exam hall.
The IP address check allows as follows:

1-Using online signature or displaying student photo

2-Using fingerprint

3-We can provide more security to identify the students by using online cameras which are more useful than the traditional method of checking the ID cards. Since we check the identity before the start of the exam, there are some more security problems regarding the questions and answers which are for a further research.

This type of online exam system reduces the examination work.

The future scope of this research can be the security of online remote exam systems.

References:

[1] IC3 “Online”. Available: http://www.ucertify.com/certifications/Certiport/ic3.html

[2] The WebCT, SIMON FRASER UNIVERSITY “Online” available: https://webct.sfu.ca/webct/entryPageIns.dowebct

[3] J. C. Adams and A. A. Armstrong, “A Web-based testing: A study in insecurity,” World Wide Web, vol. 1, no. 4, pp. 193–208, 1998.

[4] C. Rogers, “Faculty perceptions about e-cheating during online testing,” J. Comput. Sci. Colleges, vol.22, no. 2, pp. 206-212, 2006.

[5] The Blackboard Northern Illinois Univ. [Online]. Available: http://www.blackboard.niu.edu/blackboard/

[6] J. C Adams and A. A. Armstrong, “A Web-based testing: A study in interesting,” World Wide Web, vol. 1, no. 4, pp. 193-208, 1998.

[7] Jung, I.Y “Enhanced Security for Online Exams Using Group Cryptography” IEEE vol52, issue: 3 Page(s): 340 – 349 Aug 2009.

[8] Al-Mashaqbeh, I.F. Al Hamad, A. “Student's Perception of an Online Exam within the Decision Support System Course at Al al Bayt University” Conference publication Pages: 131 – 135 7-10 May 2010.

[9] Eros Desouza, Matthew Fleming, “A Comparison of In-Class and Online Quizzes on Student Exam Performance”, Journal of Computing in Higher Education, Vol. 14(2), pp. 121-134, spring 2003.

[10] D. Agarwal, O. Chevassut, M. R. Thompson, and G. Tsudik, “An integrated solution for secure group communication in wide-area networks,” in Proc. IEEE Symp. Comput. Commun., 2001, pp. 22–28.

[11] K. Berket, D. A. Agarwal, P. M. Melliar-Smith, and L. E. M. Ernest, “Overview of the intergroup protocols,” Lecture Notes in Comput. Sci., vol. 2073, pp. 316–325, 2001

[12] E. Bresson, O. Chevassut, and D. Pointcheval, “Provably-secure authenticated group Diffie-Hellman key exchange,” ACMTrans. Inf. Syst. Security J., vol. 10, no. 3, 2007, Article 10.

A Brief Author’s Biography

Mohammad A Sarrayrih - graduated from Mutah University in 1999-Jordan, major in Computer Science. Upon completing his Master's degree from Al-Neelain University 2004 Amman Branch, he started working as a teacher of Computer Science, his experience of date being of more than 12 years. Mr. Al-Sarrayrih has also a two-year experience as Programmer in the Banking Systems and Online Systems (Phone banking and ATM's). Currently, Mr. Al-Sarrayrih is acting as Deputy Chair of Information Systems and Technology at Sur University College as well as a full time instructor.

Mohammed Ilyas: Mr. Mohammed Ilyas has a bachelor degree in Computer Science and Engineering; India, Higher Diploma in Software Engineering India, Master degree in Master in Computer Application, Computer Science, Worked as Software Engineer in Seer Software, Hyderabad, India. Worked as Instructor in Computer Science and Engineering, College, Hyderabad, worked as a Lecturer in the Department of Computer Science of Sultan Qaboos University (SQU), Oman. Currently, he is a lecturer of computer Sciences at Sur University College (SUC), Oman.
He published paper in conferences, Paper presented in International Conference on Leading Beyond the horizon: Engaging Future, Annamali University Paper Title: Knowledge Management in educational processes – A qualitative approach

Fast Algorithm for In situ transcription of musical signals : Case of lute music

Lhoussine Bahatti[1], Omar Bouattane[1], Mimoun Zazoui[2] and Ahmed Rebbani[1]
[1] Ecole Normale Superieure d'Enseignement Technique , Université Hassan II Mohammedia – Casablanca, Morocco
[2] Faculté des sciences et techniques , Université Hassan II Mohammedia – Casablanca, Morocco

Abstract

In this paper, we propose a fast and accurate transcription method of a musical signal. It consists of extracting the musical information from the temporal evolution of the generated signal by the instrument. Each note (primary) will mainly be represented by a finite set of basic attributes (pitch, partial, energy, duration,). To do so, we begin by extracting each note by selecting the beginning of its appearance (onset detection), then proceeding by segmenting the signal, in order to delimit each note, which is to be identified later by determining its remaining features.

The proposed method is an extension of the well known spectral based method. It is specially designed for oriental music which is characterized by its richness in tone that can be extended to ¼ tone. It aims to detect and isolate notes from a real audio signal recorded from an Oriental lute in an ordinary environment, then exploits the constraints of the lute's sound to improve the performance of the proposed transcriber. This method also includes, preprocessing and post processing based mainly on the surrounding noise, and echo. Subsequently, we present an interpretation of the results and rigorous assessment of the method through modeling the lute string motion.

Keywords: *musical note, oriental music; onset, pitch, lute, tone, RAST, string motion*

1. Introduction

Music is a field where individual sounds are combined to compose melodies, rhythms and songs. It is an art in which information is transmitted and distributed by audio signals; it can be represented either by a symbolic level or a signal one.

In the Symbolic level case, the musical content is described in terms of structures according to the music theory (partition). While the signal Level representation corresponds to a time evolution of an analog signal from which the information can be extracted by signal processing tools.

Automatic transcription refers to the analysis and the automatic extraction of settings from a musical signal in an efficient manner to describe it.

Despite attempts dating back to 1970 [1] [2] and recent processes [3] [4] [5], Oriental music represents a large area of signals rich in term of information that requiring a deep exploration. This prompted us to focus our study on the signal of a string instrument as a lute. The choice of this instrument is made because its special physical structure allows us to split the octave by a fine quarter tone units. Unlike the most of standard musics that are played on semitone units.

In this sense, the proposed transcript procedure consists, initially, on the detection of the starting point of the note (onset) in the signal as in [6] and subsequently, the identification of the selected note using the pitch estimation procedure [7] [8].

This paper is organized as follows: in Section 2, we present a brief overview of an oriental music back ground. Theoretical modeling of the motion of a string lute is studied in section 3. Section 4 is devoted to the proposed transcriber and its various stages for signal analysis and its implementation, as well as the obtained results, and their discussions. A conclusion of this work, remarks and future perspectives are presented in the last section.

2. Oriental music back ground

Oriental music differs from occidental one by the existence of the quarter tone range that results in a rich melody. The intervals used in this music are closely represented by a ratio of successive integer numbers (n+1)/n as in [9] as follow: the tone (9/8), the diatonic semitone (20/19), the quarter-tone (37/36) and the complementary to the quarter-tone called the three quarters of tone (12/11).

The main range of that music which has the tonic note C is called the RAST range, and its symbolic representation is illustrated in Figure 1; it is the analogue of C major for the range of occidental music.

As mentioned above, we present the symbolic level of a music basing on the genetic code. Thus the genetic code of RAST range is: $1\frac{3}{4}\frac{3}{4}11\frac{3}{4}\frac{3}{4}$ so, the range RAST is:

C-D-Eƀ-F-G-A-Bƀ-C, which can be combined with its genetic

code and becomes: C 1 D ¾ E ¾ F 1 G 1 A ¾ B ¾ C.
For some partitions of the oriental music, the notes are called as in [9]: Rast(C); Doukah(D); Bousselik(E); Djahaka(F); Naoua(G); Housseini(A); Mahour(B).

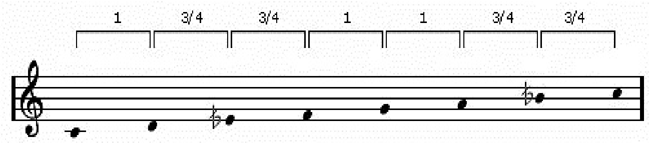


Figure 1: RAST range and its genetic code
ƀ: half-flat:

The signal level of the same RAST range is illustrated by the fundamental frequency of each note. Thus in the first octave, the notes and their frequencies are summarized as in table 1 below.

Table 1: Frequency notes of RAST range, in the first octave

Note	C	D	Eƀ	F	G	A	Bƀ	C
Freq (Hz)	65	73	79	87	98	110	120	131

Code	1	3/4	3/4	1	1	3/4	3/4
Gap (Hz)	8	6	8	11	12	10	11

According to the table 1, the developed transcription system must be able to extract the notes whose frequency difference can be about 6 Hz. In the octave 0, the transcriber resolution must be 3Hz.

3. String lute modeling

In this section, we represent a theoretical model of the string motion of lute when the string is excited at any time t. we assume also that the string length at rest is L, and it is attached at both ends (0, 0) and (0, L). We represent the vibrations of the string motion by a curve y= f (t, x) in a plane (x0y). The string in query is considered uniform and vibrates transversely,

Assumptions

1. Vibrations are only transverse, so the algebraic moving f(t, x), is lateral. (Figure 2)
2. Disturbances are small,: $|f(t,x)| \ll L$; and $\left|\frac{\partial f(t,x)}{\partial x}\right| \ll 1$
3. The gravity efforts are negligible
4. *Once the string is excited, it is leaved free, and without damping within a fixed slot time.*

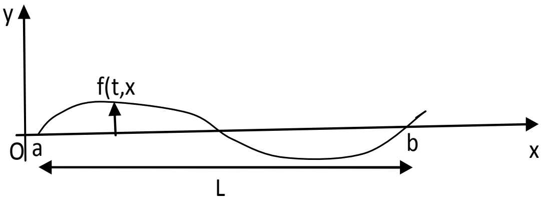


Figure 2: Modeling of string motion

Physical properties of the string:

- A linear density μ.
-A tension T, which is adjusted when tuning the lute by the artist,
- A length L: For the Arabic lute, Turkish lute and Iranian one, the total length is between 58cm and 65cm. Practically L is the distance between the bridge and the finger position on the handle of the lute.
Taking into account the different hypotheses, the string motion is governed by the following fundamental equation of dynamics:

$$\frac{\partial^2 f(t,x)}{\partial t^2} - C^2\,\frac{\partial^2 f(t,x)}{\partial x^2} = 0 \qquad (1)$$

Where $C = \sqrt{\frac{T}{\mu}}$: Wave velocity propagated by the vibrating string.

Initial and boundary conditions:

- f(0,t)=f(L,t)=0 ; ∀ t ≥ 0
- f(x,0)=f_0(x) ;and $\frac{\partial f}{\partial t}$(x,0)=$f_1$(x) ; for 0 ≤ x ≤ L
- f_0(x) and f_1(x) are the functions satisfying the following conditions: $f_0(0)=f_0(L)=0$, and $f_1(0)= f_1(L)=0$
Solving this equation leads to the expression 2 as in [10]

$$f(x,t) = \sum_{n=1}^{\infty} \frac{2\alpha L^2}{(n\pi)^2 c}.sin(k_n x)\,.\,sin(2\pi f_n t) \qquad (2)$$

Where: $k_n = \frac{n\pi}{L}$; and $f_n = \frac{nc}{2L}$:
f_n: is the nth harmonic frequency.

The sound of the lute can be expressed by the effect exerted by the string on the bridge. The latter is connected to the harmonics table from which the acoustic signal representing the note is emitted. The sound s(t) emitted by the lute, over the time, is defined by*:*

$$s(t) = \left.\frac{\partial f(t,x)}{\partial x}\right|_{x=L} = \sum_{n=1}^{\infty} d_n.\,sin(2\pi f_n t) \qquad (3)$$

The resulted note has theoretically an infinite number of harmonics whose amplitude decreases as n increases. However in the practical domain, we restrict our study to a harmonic number P (partial) and we can write:

$$s(t) = \sum_{n=1}^{p} d_n . sin\,(2\pi f_n t) \qquad (4)$$

Where d_n represents the n^{th} partial amplitude.
The fundamental frequency expression of the musical note is then:

$$f_0 = \frac{K}{2L} \qquad (5)$$

Once the parameter **K** is adjusted by the artiste through the tension T, f_0 is inversely proportional to the length L. In practical case, the first string of the lute is generally assigned for the note C. thereafter the other musical notes can be determined according to the following table.

Table 2: Position of notes on the handle of a lute

Notes		Position (length from the bridge)
Octave 0	* (C)	L
Octave 1	C	L/2
	D	8L/9=0,88L
	E	64L/81=0,79L
	F	3L/4=0,75L
	G	2L/3=0,66L
	A	16L/27=0,59L
	B	128L/243= 0,52L
	Dƀ (≈C#)	2048L/2187=0,93L
	F#[≈Gƀ]	6144L/8748=0,7L
	Eƀ	2368L/2916=0,81L
	F#	108L/148=0,72L

** Base note of the octave 0: It is freely adjusted through the string tension*

4. Transcription system

The proposed music lute transcriber has three main building blocks, designed Onset detection, Signal segmentation and Attributes extraction. It is indicated by the flow sheet in Figure 3.

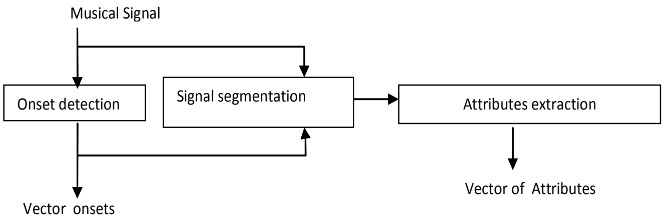


Figure 3: Transcriber scheme

The different stages of our proposed system are described as follow:

4.1. Onset detection and segmentation

4.1.1. Approach principle

Any musical note has a temporal allure showing some principal features:

- a start (onset: start time of the note)
- a transitional state: during which the spectral content is rapidly variable
- a steady state

The amplitude variation of the note is enclosed in a shape called the ADSR shape (an acronym of the words Attack, Decay, Sustain, Release), and has a behavioral pattern as in figure 4

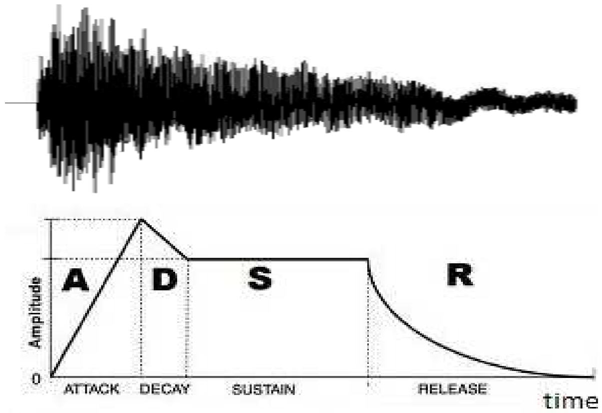


Figure 4: Allure of a note and its ADSR envelope

The four phases of this shape are:

- Attack: time during which the energy increases and the amplitude rapidly reaches its highest value.

- Decay: after the attack, a part of the initial energy is lost and the amplitude decreases.
- Sustain: amplitude maintains an almost constant level during this slot time.
- Release: the amplitude progressively decreases until it becomes negligible. .

Generally in the sustain phase, the amplitude of each high frequency decreases rapidly while low frequencies stay for long time.

In most cases, the detection of onset is based on the well known algorithm as in Figure 5a.

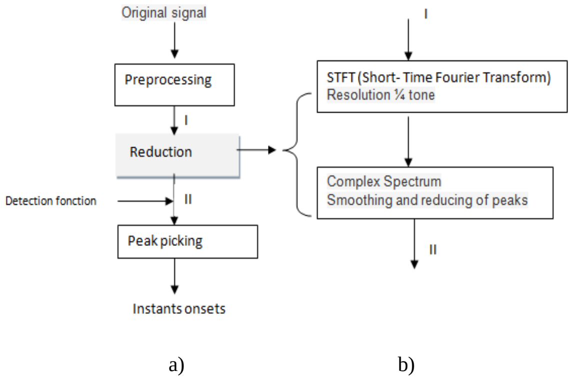


Figure 5: Diagram of the onset detection algorithm

The proposed onset detection algorithm is composed of the following steps:

-The Preprocessing stage: It consists: first, removing the present silences at the beginning and at the end of the signal. Secondly, split the signal into slices of width 4s each to facilitate and standardize the processing, because, this time is large and sufficient to extract the different features. Finally eliminate the DC component as Eq 6 and Eq 7, and normalize the signal as Eq 8.

$$s(n) = s(n) - \overline{s(n)} \qquad (6)$$

Where

$$\overline{s(n)} = \frac{1}{L}\sum_{n=0}^{L-1} s(n) \qquad (7)$$

And: $$s(n) = \frac{s(n)}{max_n|s(n)|} \qquad (8)$$

-The detection function: The ideal detection function is the one corresponding to a Dirac pulse train whose abscissas coincide with the moments of onsets. In reality, this function exhibits the peaks at each onset. In the literature, Bello et al **[6]** studied a set of detection functions based on: the temporal characteristics, energy, frequency and probabilistic behavior of the signal. The detection function can also be established on the basis of a non-negative factorization of the amplitude spectrum **[11]**. Also the onset detection can be achieved by a sequential algorithm based on computing a statistical distance measure between two autoregressive models **[12].** However the performance of each method is closely related to the signal (depending on the instrument) in query. Our method exploits both the temporal characteristics (string instrument) and frequency one (range of oriental music) of a lute music signal by calculating the Short-term Fourier transform (STFT) for any time n, and a frequency k of a signal x by:

$$X_k(n) = \sum_{m=-N/2}^{\frac{N}{2}-1} x(nh+m)w(m)e^{-2j\pi mk/N} \qquad (9)$$

h: is the hope size : space between two slice analysis centers.

w(m) : is the weighting window of length N. That length is directly related to the resolution of the Short Term Fourier Transform.

According to Table No. 1, this resolution must be 6Hz for the first octave. In order to take into account of the non-stationarity of the musical signal and the distribution of frequencies in a quarter tone, the best manner is to choose a Hann window of variable size [13]. The variation of this size depends on the Constant Q Transform (CQT) approach, applied to the Rast range of oriental music **[13].** Its expression is:

$$N_k = 37.\frac{F_s}{k} \qquad (10)$$

Where: F_s is the sample rate.

The amplitude spectrum of $\mathbf{X_k(n)}$ is considered it as an N-dimensional vector, so the approach based on changes in this spectrum is to formulate the detection function as a “distance” between two successive vectors as in Eq 11 below.

$$SD(n) = \sum_{k=-N/2}^{\frac{N}{2}-1} \{H(|X_k(n)| - |X_k(n-1)|)\}^2 \qquad (11)$$

Where: $H(x) = \frac{x+|x|}{2}$ to take into account just the increasing state of the energy.

The phase spectrum $\varphi_k(n)$ is evaluated from the instantaneous frequency of the spectrum $X_k(n)$ by:

$$f_k(n) = (\varphi_k(n) - \varphi_k(n-1))/2\pi h.$$

And

$$f_k(n-1) = (\varphi_k(n-1) - \varphi_k(n-2))/2\pi h \qquad (12)$$

So for a stationary sinusoid:$f_k(n) = f_k(n-1)$. Therefore

$$\varphi_k(n) - \varphi_k(n-1) = \varphi_k(n-1) - \varphi_k(n-2) \qquad (13).$$

We define the phase deviation, characterizing the breakdown of stationarity of a signal for the frequency bin k, by the following equation:

$$\Delta\varphi_k(n) = \varphi_k(n) - 2\varphi_k(n-1) + \varphi_k(n-2) \qquad (14).$$

The aggregate measure of stationarity is then defined by a mean of the absolutes deviations as Eq 15.

$$\eta_p(n) = \frac{1}{N}\sum_{k=1}^{N} |\Delta\varphi_k(n)| \qquad (15)$$

In our case, we combined information about the deviation of the phase and the amplitude difference spectrum using the complex spectrum.

In fact: at time n, the estimated spectrum $\hat{X}_k(n)$ from $X_k(n-1)$ (in the case of Stationarity) is:

$$\hat{X}_k(n) = |X_k(n-1)| e^{j\Delta\varphi_k(n)} \qquad (16)$$

The Stationarity of the signal is measured by calculating the Euclidean distance between the estimated spectrum and the observed one. This distance is given by the following equation:

$$\Gamma_k(n) = \left\{|\hat{X}_k(n)|^2 + |X_k(n)|^2 - 2|\hat{X}_k(n)||X_k(n)| \cos(\Delta\varphi_k(n))|\right\}^{\frac{1}{2}} \qquad (17)$$

These distances are then summed across the frequency-domain to generate the onset detection function

$$\eta(n) = \sum_{k=1}^{N} \Gamma_k(n) \qquad (18)$$

The obtained result is shown in figure 6 below

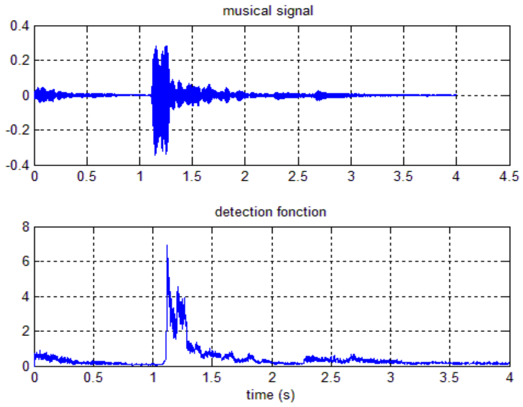


Figure 6: musical Signal and detection function

4.1.2. Smoothing and peaks selection

The physical structure of an oriental lute is particular by the fact that it can generate onsets with peaks of the detection function having low and high levels. Some peaks of high level do not necessarily mean that they are an instant of onset. Therefore, the detection function has significant peaks and insignificant ones that must be discarded. To do so, we proceed by:

• Smoothing the detection function using a median filter to reduce noise at the edges

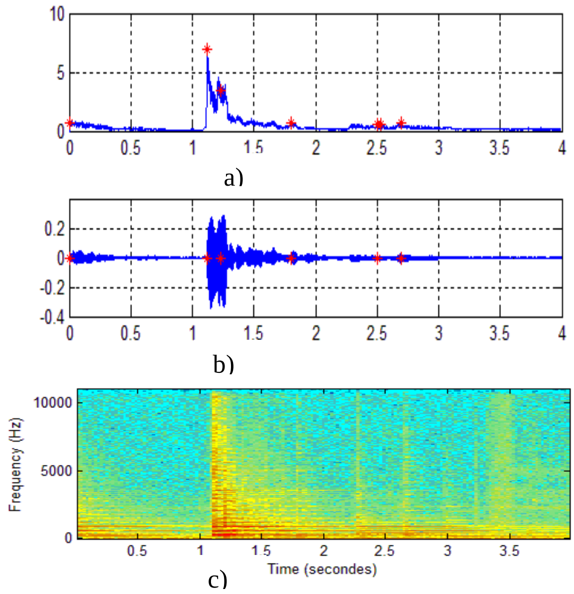
$$d(n) = median\big(d(n-1) : d(n+1)\big) \qquad (19)$$

• Reduction and selection of peaks.

Despite of the smoothing stage, the number of peaks in the detection function is often greater than the number of real onsets of the signal. In order to eliminate the faulty peaks, three operations are executed:

- Keep the peaks that are greater than adaptive threshold which depends on the instantaneous energy of the signal, the threshold is calculated as [6] as follows:
  $$\tilde{\delta} = \delta + \lambda\,\mathrm{median}\{|d(n-M)|, \dots, |d(n+M)|\} \qquad (20)$$
  In our case, δ: static threshold set at 0.22, λ is set at 2. M: Number of samples of the detection function of a window size of 50ms. The size 50ms is considered as the optimal width where the signal features do not have significant changes. This size allow us to have a good handle on the number of detected peaks
- Among the peaks that are greater than the adaptive threshold, we calculate the variation between two successive peaks: Δ (i) = peak (i)-peak (i-1) and we consider a peak(i) as significant and therefore retained, if Δ(i) is greater than 25 % of peak(i-1).

The result of our onset detection technique, applied to a real signal of a lute, is represented by the following figure



a)

b)

c)

Figure 7: Results of onsets detection for a real signal lute
a) Detection function and significant peaks
b) Source signal and times of onsets
c) The music signal spectrogram

4.1.3. Segmentation

The task of the segmentation operation is to extract and isolate the musical notes from the signal in order to process them. The result of the segmentation procedure, according to the determined onset times, is shown in Figure 8 below

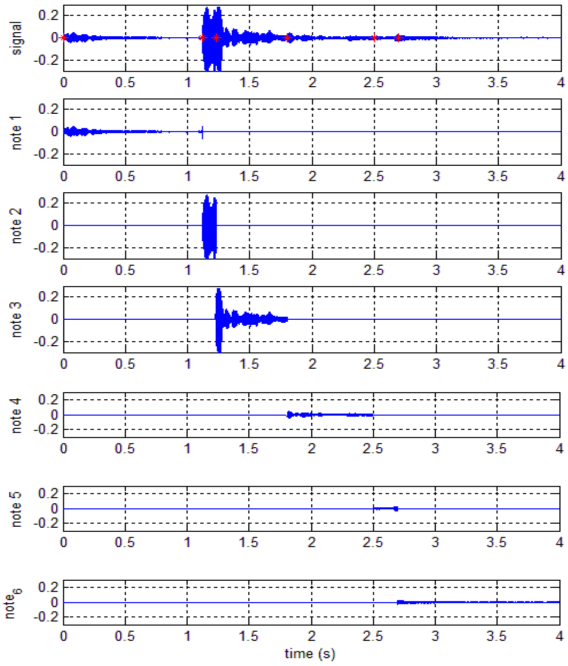


Figure 8: Switching signal according to onset

4.2. Attributes extractor

Musical notes are considered as carry meaning musical entities. Our task is to extract information allowing the passage of raw data to a more compact representation. Thus, each isolated note must be purified of noise (especially the echo) before starting the transcription process

4.2.1. Echo removing

The problem of echo can be described by a simple model of the form:

$$y(n) = x(n) + a.x(n - \Delta) \qquad (21).$$

Where:

- x(n): the original (echo free) signal;
- y(n): signal with echo.
- **a** and Δ are the amplitude and delay of the echo respectively.

The first task of this part is to estimate the parameters **a** and Δ for each note, using the autocorrelation technique:

$$C_{yy}(\tau) = E[y(n)y(n-\tau)]$$

$$= (1+a^2)C_{xx}(\tau) + a.C_{xx}(\Delta - \tau) + a.C_{xx}(\Delta + \tau) \qquad (22)$$

The autocorrelation function $C_{xx}(u)$ is maximum for u = 0, so $C_{yy}(\tau)$ exhibits peaks at the instants 0, Δ and −Δ, then

$$\Delta = Arg(max\,(C_{yy}(\tau))\,; with\ \Delta \neq 0. \qquad (23)$$

For the parameter a: we consider the ratio **r** defined by:

$$\boldsymbol{r} = \frac{\boldsymbol{C_{yy}(0)}}{\boldsymbol{C_{yy}(\Delta)}} = \frac{\boldsymbol{1+a^2}}{\boldsymbol{a}} \qquad (24)$$

Therefore **a** is the real solution of the following equation:

$$\boldsymbol{a^2} - a.r + 1 = 0 \qquad (25)$$

Having the parameters a and Δ, the echo phenomenon eliminator can be modeled by a filter transmittance,

$$\boldsymbol{H(z)} = \frac{\boldsymbol{Y(z)}}{\boldsymbol{X(z)}} = \frac{\boldsymbol{1}}{\boldsymbol{1+a.z^{-\Delta}}} \qquad (26)$$

So to remove the echo from y(n) and extract x(n) , we use a FIR filter transmittance F(z) with:

$$\boldsymbol{F(z)} = \frac{\boldsymbol{Xs(z)}}{\boldsymbol{Y(z)}} = \boldsymbol{1 + a.z^{-\Delta}} \qquad (27)$$

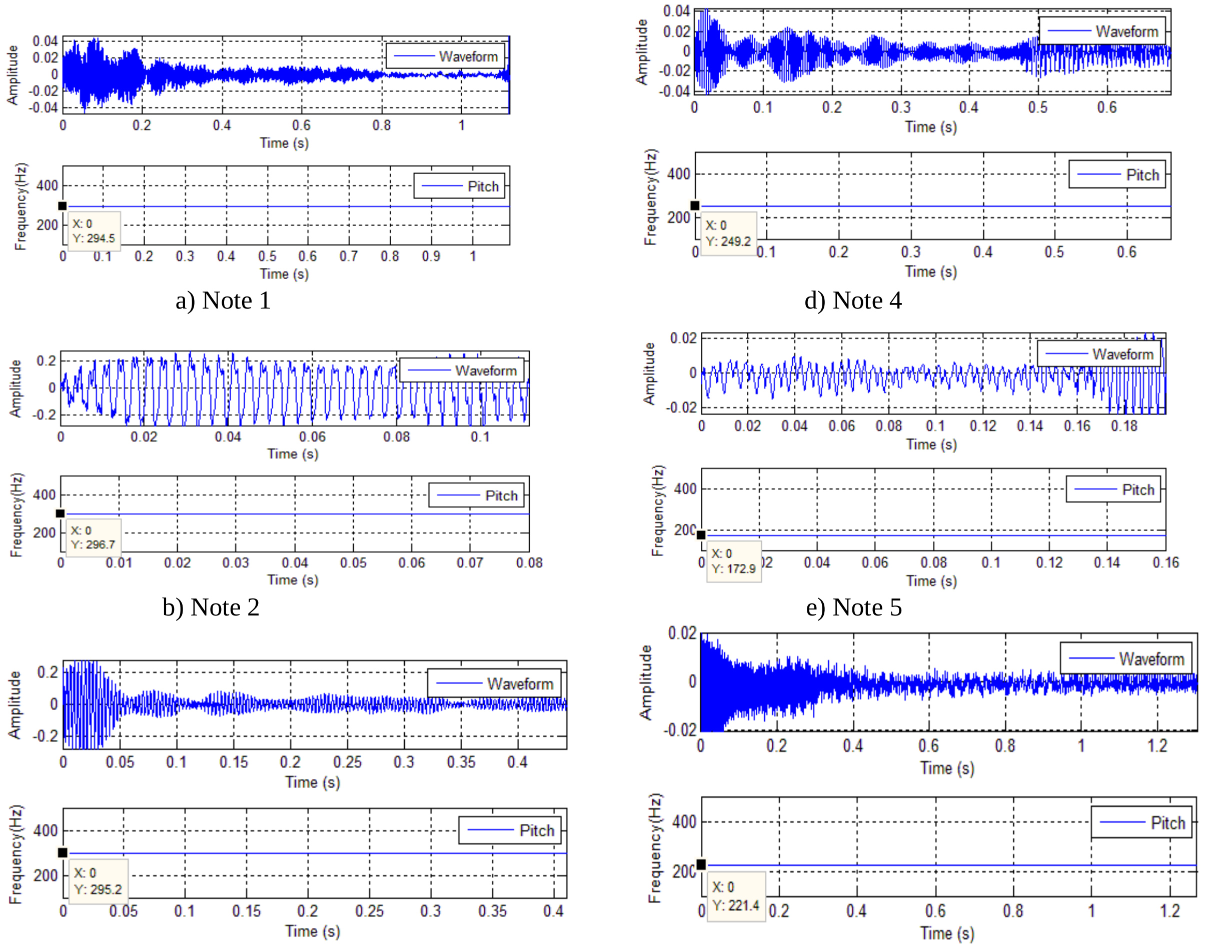
Xs(z) denote the original signal without echo

4.2.2. Pitch estimation

The pitch is the basic perceptual attribute used to characterize sound events and it is closely related to the fundamental frequency. The estimation of pitch is usually based on spectrum, autocorrelation, or cepstrum, or a mixture of these strategies **[7] [8]**. The main method presented in this paper to determine the pitch is the autocorrelation process .This is a most robust method that is based on the periodic characteristic of the music signal, and independent on its amplitude. The procedure for estimating the pitch of a note using the autocorrelation method is as follows:

1. Divide the signal expressing the note into frames of 5ms
2. Determinate the pitch of each frame.
3. Determinate the average pitch

The results of the signal segmentation according to the instants onset, followed by the determination of pitch for each note are illustrated in figure 9 bellow:



a) Note 1

d) Note 4

b) Note 2

e) Note 5

c) Note 3

f) Note 6

Figure 9: Alluring single notes and their pitches

The following table summarizes the results related to the transcript of each isolated note:

Table 3: Transcription results

Index note	Duration (s)	Measured Frequency Hz)	Normalized Frequency (Hz)	Absolute frequency error (Hz)	note Identification	Energy
1	1,1	294,5	294	0,5	D4	1.877
2	0,1	296,7	294	2,7	D4	62.291
3	0,45	295,2	294	1,2	D4	37.786
4	0,7	249,2	247	2,2	B3	1.380
5	0,16	172,9	174	1,1	F3	0.247
6	1,3	221,4	220	1,4	A3	0.625

4.3. Obtained result analysis

The studied signal is recorded in normal conditions. Some retrieved fundamental frequencies do not correspond exactly to the normalized values. The error calculated as follow: see table 3.

$$\varepsilon = abs\,(F_{mesured} - F_{normalized})$$

That error is maximum for the musical note 2 (ε =2,7Hz), and it corresponds to the shortest musical note present in the signal. Generally, the errors are due to the fact that the position of the musician's finger may not exactly match the theoretical location. For an Arabic lute, the notes locations on the handle require a good selectivity to properly play the desired notes as mentioned above in Table 2.

To assess our procedure, we synthesized the musical signal, with the same notes, the same durations and the same energy but fundamental frequencies are normalized as mentioned in table 3. This signal is then analyzed by our transcriber. The obtained results are summarized in the following table 4:

Table 4. Measured Frequency of synthesized signal

Note index	Normalized Frequency (Hz)	Duration (s)	Measured Frequency (Hz)
1	294	1,1	293,8
2	294	0,1	292,5
3	294	0,45	294,8
4	247	0,7	248
5	174	0,16	173
6	220	1,3	220,3

Comparing the results of table 3 and table 4, we can see that the detection Pitch error increases as well as the note duration decreases.

According to the obtained result in figure 9, we can say that onsets moments depend on the previous note. For example, note 4 in figure 9.d is detected after a delay time of 0,25 sec, while the note 5 in figure 9.e is detected after 0,03 sec. However, the notified delays do not have significant impact on the result of transcription

5. Conclusion and future works

The proposed method is designed to detect and isolate notes from a recorded audio signal issued from an Oriental lute. The amplitudes of some tested notes were very low, and the beginning of a given note does not necessarily correspond to the end of the previous one. Our identification system is then more reliable, and more accurate for the long duration notes. In order to have accurate results, it is recommended to make pitch detection on the Sustain phase of the note, where the amplitude of the musical note stays unchanged. In fact, in this phase, the note is steady and indicates its periodicity. In the real context, this periodicity is used by the human ear to correctly distinguish the pitch of the note.

As the perspectives of this work, we propose to take into account the overlapping between successive notes (real context), and extend the method to more robust descriptors.

Applying these descriptors to several samples of real music, we will try to address a new process of extraction of the fingerprint and signature of musician.

For our mathematical modeling, this approach reflects just the steady state of the string motion when the oscillations are forced. In order to take into account the transitional regime and benefit from the signal and physical modeling techniques, a model that is based on the state space representation is required. This is another modeling problem that requires a deep analysis and this make our future work.

References

[1] J.A. Moorer, “On the segmentation and analysis of continuous musical sound by digital computer”. PhD thesis, CCRMA, Stanford University, 1975.

[2] M. Piszczalski and B. A. Galler, “Predicting musical pitch from component frequency ratios,” J. Acoust. Soc. Am., 66 (3), 710–720, 1979.

[3] Klapuri, A, “Signal processing methods for the automatic transcription of music,” Ph.D. dissertation, Tampere University of Technology, 2004

[4] Klapuri, A, "Automatic music transcription as we know it today," Journal of New Music Research, Vol. 33, No. 3, pp. 269-282, Sep. 2004.

[5] Klapuri, A.”Sound Onset Detection by Applying Psychoacoustic Knowledge”, Proceedings IEEE Int. Conf. Acoustics Speech and Sig.Proc. (ICASSP), pp. 3089–3092, Phoenix AR,USA March 1999.

[6] J. P. Bello, L. Daudet, S. Abadía, C. Duxbury, M. Davies and M. B. Sandler, “A tutorial on onset detection in music signals”, IEEE Transactions on Speech and Audio Processing. September,2005.

[7] Chunghsin YEH, “Multiple fundamental frequency estimation of polyphonic recordings”, PhD thesis, université Paris VI - Pierre et Marie curie, 2008

[8] G. Peeters, “Music pitch representation by periodicity measures based on combined temporal and spectral representations”, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2006

[9] B. Marzouki. “Application de l’arithmétique et des groupes cycliques à la musique”. Département de Mathématiques et Informatique Faculté des Sciences Oujda. Maroc

[10] D.EUVRARD : « Résolution numérique des équations aux dérivées partielles». Edition MASSON chapitre3: Equation des cordes vibrantes

[11] WenwuWang, Yuhui Luo,Jonathon A. Chambers, and Saeid Sanei , “Note Onset Detection via Nonnegative Factorization of Magnitude Spectrum” , Hindawi Publishing Corporation EURASIP Journal on Advances in Signal Processing Volume 2008, Article ID 231367.

[12] Jehan, T: “Musical Signal Parameter Estimation”. Msc. Thesis, CNMAT. Berkeley, 1997

[13] L.Bahatti; M.Zazoui; O.Bouattane; A.Rebbani. “Short-term sinusoidal modeling of an oriental music signal by using CQT transform”. Journal of Signal and Information Processing”; accepted for publication: to appear: Article ID3400246

Lhoussine BAHATTI was born in 1968 in MIDELT, Morocco. He is now a teacher of control, automatic regulation and signal processing, and researcher at the University Hassan II Mohammedia, ENSET Institute. His research is focused on Music information retrieval. He Received the B.S. degree in Electronics in 1987 and the Aggregation Electrical Engineering degree in 1995 from the ENS CACHAN France. He received the DEA diploma in information processing in 1997 from the Ben M'sik University of Casablanca MOROCCO

Omar BOUATTANE was born in 1962 in FIGUIG, south of Morocco. He has his Ph.D. degree in 2001 in Parallel Image Processing on Reconfigurable Computing Mesh from the Faculty of Science Ain Chock, CASABLANCA. He has published more than 30 research publications and brevets in various National, International conference proceedings and Journals. His research interests include Massively Parallel Architectures, cluster analysis, pattern recognition, image processing and fuzzy logic.

Mimoun ZAZOUI is a professor at the Faculty of Science and Technology of Mohammedia (FSTM). Chair of the Moroccan Society of Renewable Energy. He has coordinated various research projects and training at national and international level and within the framework of cooperation. He has also gained considerable experience in the area of academic capacity; he is Vice Dean and Director of Graduate Studies at the FSTM. He is also responsible for unit training and research in Thin Film Materials and Systems for Energy Conversion. He is also Director of the Laboratory of Physics of Condensed Matter at the FSTM.
.

Ahmed REBBANI was born in 1967, in TAZA, Morocco. He is now a teacher of computer networks, and researcher at the University Hassan II Mohammedia, ENSET Institute. His research is focused Source separation. He Received the B.S. degree in Electronics in 1988 the M.S. degree in Applied Electronics in 1992 from the ENSET Institute, Mohammedia, Morocco. He received the DEA diploma in information processing in 1997 from the Ben M'sik University of Casablanca MOROCCO

Computer Simulation to Detect the Blind Spots in Automobiles

Hazem (Moh'd Said) Hatamleh[1], Ahmed A.M Sharadqeh[2], As'ad Mahmoud Alnaser[3], Omar Alheyasat[4] and Ashraf Abdel-Karim Abu-Ein[5]
[1,3] Department of Computer Science Ajloun University College Al-Balqa' Applied University
Ajloun, Jordan
[2,4,5] Computer Engineering and Computer Technology Department , Al-Balqa' Applied University
Amman,Jordan

Abstract

During driving Changing lanes can be very hazardous on a busy highway. There is region called “blind spot” which is a problem for every car driver since it's not covered by the driver's mirrors. Relying solely on the mirrors while changing lane can lead to a collision with another vehicle. This paper focuses on this situation by ensuring that the blind spots of the vehicle are clear prior to the driver attempt to change lanes. This computer simulation incorporates the need for detection and warning of objects present within the blind spot on either side of the vehicle to the driver along with distance measurement of the object relative to the vehicle, incase the driver decides to change lanes. This simulation is constructed using the theory of embedded systems and will alert the driver if there is another car on the blind area.

***Keywords: blind spot, computer simulation, embedded system, automobile*.**

1. Introduction

The blind spot is the place behind your vehicle that the driver cannot see in the rear or side view mirrors or even by turning your neck out the driver's side window. Generally speaking , All vehicles have a blind spots & the larger the vehicle, the larger the blind spot, [6]. Blind spots for shorter drivers tend to be significantly larger as well. In addition, the elevation of the driver's seat, the shape of a vehicle's windows and mirrors, and the slope of a driveway can affect the size of the blind spot behind a vehicle. Blind spots are areas in adjacent lanes of traffic that are blocked by various structures in the automobile. The physical constraints in eye movement and head and body rotation make certain areas invisible to the driver. The blind spots in cars depend on their construction, figure 1. The direct blind spots are:

- Area covered by the A-pillar, between the front door and windshield
- Area covered by the B-pillar, behind the front door
- Area covered by the C-pillar, ahead of the rear windshield

Apart from these, there are certain indirect blind spots like the region between the driver's peripheral vision on the sides and the area that is covered by the rear view mirror. The unseen areas are immense for drivers of medium and heavy trucks as compared to drivers of passenger vehicles. Areas directly to the right of the cab extending past the trailer, directly behind the trailer, to the immediate left of the cab and directly in front of the cab are blind spots for truck drivers.

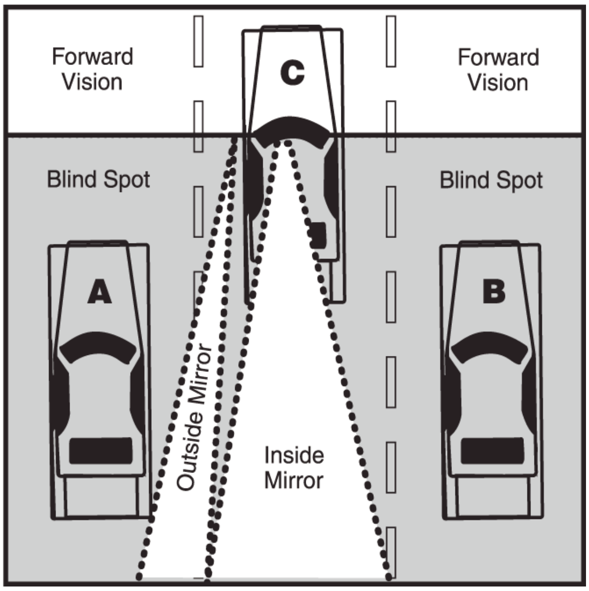


Fig. 1: blind spots[6]

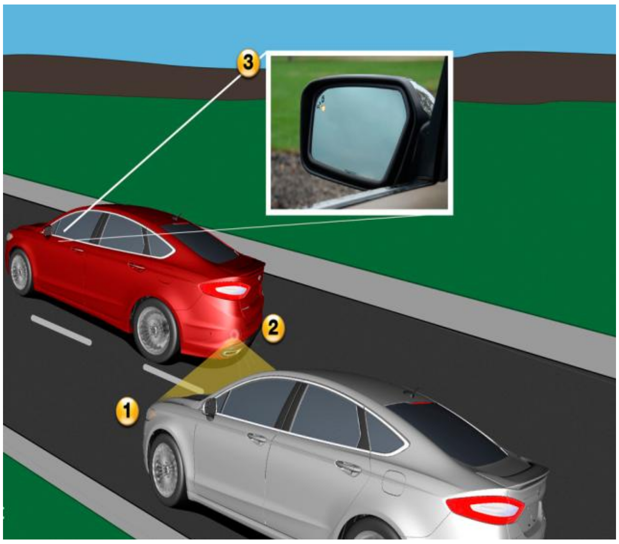


Fig.2: blind spot seen by two drivers

There are little literature discussed such issue, R. Andrew, et al. 2005, designed a passenger side mirror for an automobile that does not have a blind spot and that does not distort the image. The model consists of a coupled pair of partial differential equations that do not have a common solution. Using a best mean-square-error functional, they find approximate solutions using nonlinear optimization. In one case a local minimum provides a mirror that solves the problem, but it does not reverse the image, [10]. Christoph M., et al. 2000, discussed their findings of transit buses driving through very cluttered surroundings and being involved in many different types of accidents where currently available CWS do not work effectively. One of the focuses of their work is pedestrians around the bus and their detection,[11].

2- Results and discussion

The working principle of the our system simulation depends on :

Inputs ⟶ processing & ⟶ comparison Outputs

The PIC receives inputs from the turn signal & the ultrasonic sensor . then processes this input data & compares them with the data stored in its memory, depending on the comparison process , the PIC will activate or deactivate the warning systems . There are 3 cases explain the simulation working in details :

Case 1 : No received signal from the turn signal to the PIC

In this case the turn signal is " off " . therefore , the system can't work because no received signal to the PIC (the switch is open) . therefore , the PIC sends order to the LCD to display : " No signal ": this case is shown in figure 3.

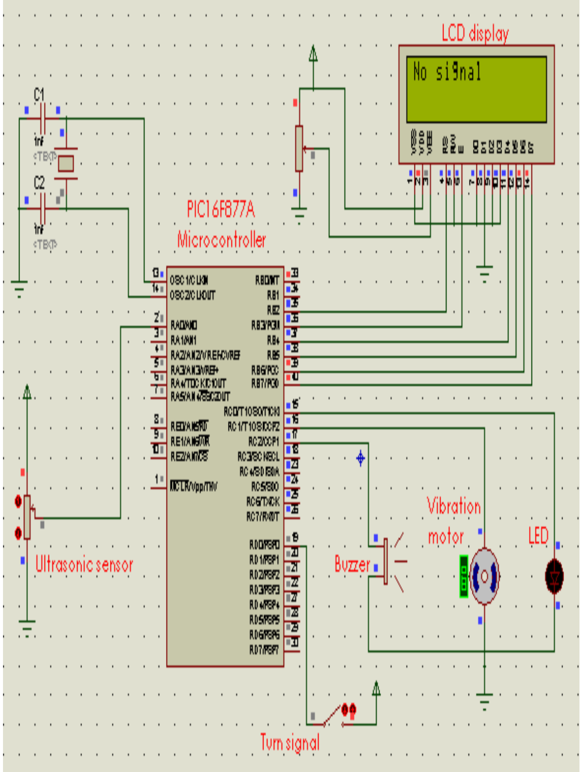


Fig.3: Case 1: No received signal from the turn signal to the PIC.

Case 2 : the turn signal is " on " & the measured distance is less than 300 cm

In this case the turn signal is " on " (by pressing on the switch to close) . therefore , the PIC receives input from the turn signal , then the system starts to run then the PIC receives input from the ultrasonic sensor which measures the distance of object in the blind spot . we will input the distance manually into the PIC by varying position of the wiper of the potentiometer to down . for example , the input distance is 154 cm . the PIC will compare this distance with 300 cm . the result of comparison shows that the measured distance is less than 300 cm . therefore , the PIC will display on the LCD that there is object in the blind spot & display its distance

(154 cm) , as well as , activate the warning systems which represented by buzzer , vibration motor & LED . in details , the buzzer will emit an intermittent sound , & the vibration motor will vibrate the steering wheel & the LED will emit a red light, this case is shown in figure 4.

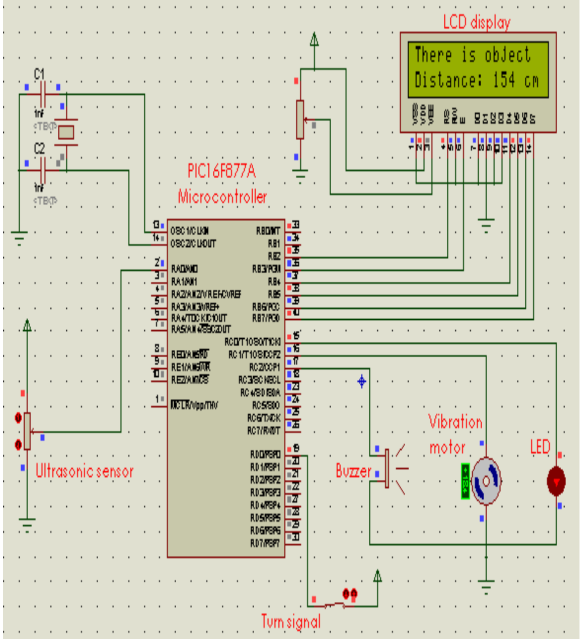


Fig.4 : Case 2 the turn signal is " on " & the measured distance is less than 300 cm.

Case 3 : the turn signal is " on " & the measured distance is more than 300 cm

In this case also the turn signal is " on " (by pressing on the switch to close) . therefore , the PIC receives input from the turn signal , then the system starts to run then the PIC receives input from the ultrasonic sensor which measures the distance of object in the blind spot . we will input the distance manually into the PIC by varying position of the wiper of the potentiometer to up . for example , the input distance is 350 cm . the PIC will compare this distance with 300 cm . the result of comparison shows that the measured distance is more than 300 cm . therefore , the PIC will display on the LCD that there is " No object " in the blind spot & display " The road is safe " , as well as , deactivate the warning systems which represented by buzzer , vibration motor & LED . in details , the buzzer will not emit an intermittent sound , & the vibration motor will not vibrate the steering wheel & the LED will not emit a red light, this case is shown in figure 5.

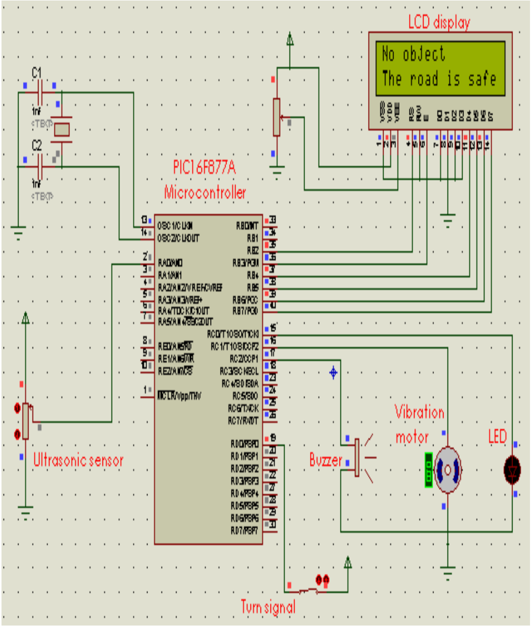


Fig.5: Case 3 the turn signal is " on " & the measured distance is more than 300 cm.

3. Conclusion

From results of the simulation the following conclusions can be estimated:

- The system detects presence of objects in the blind spot .
- The system measures the distance of objects by the sensor.
- The system displays the measured distance on an LCD if an object is present else it displays that there is no object.

The system gives out audible warning, vibration warning and lighting warning if an object is present in the range and the turn-signal is high.

Reference

[1] Milan Verle , 2008 , PIC Microcontrollers , 1st edition , mikroElektronika ISBN-13: 978-86-84417-15-4

[2] John Lovine , 2009, Pic Microcontroller, McGraw_Hill .

[3] Shibuya-ku , piezoelectric sound components application manual , International Division , Tokyo 150-0002, Japan

[4] Nishi –Ikebukuro , LED Light Emitting Diodes , Sanken Electric Co.,Ltd. , Tokyo , PHONE: 03-3986-6164

[5] Moreno, Ivan 2008 , Modeling the radiation pattern of LEDs, McGraw_Hill .

[6] Shaun Milano , 2007 , Fully Integrated Hall Effect Motor Driver for Brushless DC Vibration Motor Applications , Allegro MicroSystems, Inc.

[7] Mateja, J., 1995, Warning System Sheds Light on Blind Spot, Traffic Safety , May/June 1995.

[8] Navet and Françoise Simonot-Lion , 2009 , Automotive Embedded Systems Handbook , Taylor & Francis Group, LLC.

[9] Dogan Ibrahim , 2000 , Microcontroller Projects in C , Newnes.

[10] R. Andrew Hicks and Ronald K. Perline, 2005, Blind-spot problem for motor vehicles, 1 July 2005 _ Vol. 44, No. 19 _ APPLIED OPTICS, pp:3893-3897.

[11] Christoph Mertz1, Sue McNeil2, and Charles Thorpe1 2000, Side Collision Warning Systems for Transit Buses, Proceedings of the 2000 Intelligent Vehicles Conference, The Ritz-Carlton Hotel, Dearborn, MI, USA, October 4-5, 2000

Dr.hazem (Moh'd said) Hatamleh was Born in Irbid Jordan in 1973,Doctor of Philosophy (Ph.D Engineering Science) “Computers, Computing System and Networks//National Technical University of Ukraine 2007.Assistant Professor in Al-Balqa' Applied University, His current research Are computer networks, operating system and Image processing

Dr. AHAMD SHARADQH was born in Irbid, Jordan, in 1978. He received the M.Sc. and PH.D degrees in 2007 from the National Technical University of Ukraine". He is currently Assistant professor of computer Engineering and Computer Technology Department of Al-Balqa' Applied University. His current research interests are computer networks, operating systems, Microprocessors, programming and digital logic design.

Dr.Asad Mahmoud Asad Alnaser .I was born in Irbid / Jordan, 1974. I completed my Ph.D in 2004 in National Technical University of Ukraine “Kyiv Polytechnic Institute” I specialized in Computer Systems and Networks. Now, I'm working as an Assistant Professor in Al-Balqa ' Applied University, Ajlun University College, Department of Computer Science. My Research area concentrates on Computer Networks, Neural Networks, Information Security, and Image Processing

Dr.Omar AlHeyasat is currently Assistant professor of computer Engineering and Computer Technology Department of Al-Balqa' Applied University. His current research interests are computer networks, operating systems, Microprocessors, programming and digital logic design

Dr. Ashraf Abdel-Karim Abu-Ein was born in Irbid, Jordan, in 1979. He received the M.Sc. and PH.D degrees in 2007 from the National Technical University of Ukraine". He is currently Assistant professor of computer Engineering and Computer Technology Department of Al-Balqa' Applied University. His current research interests are computer networks, operating systems, Microprocessors, programming and digital logic design.

User Contribution Measurement in Online Forum with Fraud Immunity

Guo-Ying WANG[1] and Shen-Ming QU[2]

**[1] Information Engineering College, Zhejiang A&F University
Hangzhou, 311300, China**

**[2] Computing Center，Henan University
Kaifeng, 475004, China**

Abstract

It's very important to reward the contributive users of online forums, for that almost all contents are provided by users in such forums. There should be some rewards for contributive users, and rewards should be proportional to the contributions. So the determination and measurement of user contributions are needed in online forums. At the same time, some users may do some fake contributions to obtain more rewards. In this paper, we analyzed possible frauds in online forum, examined features of each kind of fraud, and proposed some fraud-tolerant parameters according the features of frauds. Results of our experiment show that almost 81% users in the examined online forum have fraudulent activities and pure advertising users can be discovered according to the fraud-tolerant parameters we considered. On the other hand, the experiment results also show that the biggest count of fraud type detected is with the parameter minimum intervals of posts from the same users, and followed by the parameter minimum length of posts. While minimum average rate value of post after specified rates is the parameter that was used for least times. Based on the idea of this paper, frauds of user contributions could be discriminated well, and user contributions can be measured quantitatively and fraud-tolerantly, which provides a basis for online forums to reward users in various ways.

Keywords: *contribution measurement, fraud immunity, online forum.*

1. Introduction

Success of an online forum depends on two key aspects: the forum infrastructure and the contents. The contents are all provided by forum users. To encourage users to provide more valuable contributions, the forums usually give some rewards to their contributive users.

A good contribution-reward mechanism can motivate users to provide more valuable contents. Obviously, the rewards should be proportional to the contributions. But currently most forums do not possess a good approach to measure the contributions of their users. On the other hand, some users do some fake contributions to cheat more rewards. How to discriminate the fraud from contributions is another problem.

The purpose of this paper is to model user contributions with some features of user actions, measure user contributions in a quantitative way, and provide a basis for online forums to award their users in various ways.

The rest of the paper proceeds as follows. Section 2 describes some existing technologies of measuring and motivating user contributions in different environments. Section 3 lists the possible frauds in online forums. Section 4 gives the fraud-tolerant parameters according to each feature of frauds, and then describes the fraud-tolerant user contributions measurement method. Section 5 describes a user contribution measuring experiment and its results. Section 6 concludes the paper.

2. Related Works

Some researches have been done about user contribution motivation and measurement.

Cheng [1] introduced hierarchical membership levels (gold, silver and bronze) to motivate user contributions in P2P communities, and five contribution relevant factors are used to measure user contributions.

Abtoy [2] proposed a model that supports monitoring the quality of content according to its life on the Web. The model emphasized on the prioritizing information and users both.

One simple way of measuring user contribution is allowing users to rate the quality of contents provided by other users. A similar method is a member-controlled reward mechanism, that is, users who posted questions could rate the quality of other users' answers [3].

Klamma [4] evaluated user contributions with 5 factors such as postings, replies to posts, post ratings, replies to their posts and post ratings received. Chai [5] proposed a model of user contributions measurement with 16 factors, and user contribution score are calculated by summing up the values of 16 factors with different weights individually. This method didn't consider the possible frauds, and a user with many junk posts or advertisement posts may be considered more contributive.

Shi [6] studied the patterns of user participation behavior, and the feature factors that influence such behavior on different forum datasets, found that users' community joining behaviors display some strong regularities, built social selection models, Bipartite Markov Random Field (BiMRF), quantitatively evaluated the prediction performance of those feature factors and their relationships, and showed that we show that some features carry supplementary information, and the effectiveness of different features vary in different forums.

Wikipedia is a website in which all contents are provided and edited by users from the whole world voluntarily. Adler [7] considered the problem of measuring user contributions to versioned, collaborative bodies of information, such as wikis, considered and compared various alternative criteria that take into account the quality of a contribution, in addition to the quantity, and analyze how the criteria differ in the way they rank authors according to their contributions, proposed to adopt total edit longevity as a measure of author contribution. Edit longevity is resistant to simple attacks, since edits are counted towards an author’s contribution only if other authors accept the contribution.

3. Possible Frauds in Forums

In some online forum, many particular actions of users, such as postings and replies to posts, are considered as contribution. In order to be rewarded as contributive ones or for other purposes, some users may posts as many topics as possible, even useless topics. Table 1 lists some fraud types of fraud usually occurred.

3.1 Junk posts

Junk posts are useless and have no contribution to a forum, which is interesting to few users. In worst case, junk posts may makes users leave the forum if where is full of junk posts. Junk posts have some common features as follows.

Table 1: Fraud Types in Online Forums

Code	*Fraud Types*
T1	Junk posts
T2	Advertisement posts
T3	Advertisement messages
T4	Dummy replies
T5	Dummy rates

Short length: Most junk posts are created by users who just pursued the counts he posted, are always very short and even zero length.

Short interval: Junk posts by the same user are usually posted with short intervals so as to achieve a large number of posts in short time.

Few views, few replies: Because of the useless of junk posts, they can't incur others users' interest and obtain few view.

Low rated or Judging replies: As to such junk posts, common users surely rated them low or give judging replies such as "junk", "useless", and so on.

3.2 Advertisement posts

Some users join forums in order to post advertisements. Such posts are similar to junk posts: they have show interval, few views and replies and are low rated.

Short interval in different subforums: To achieve as many viewers as possible, the advertiser users usually post their advertisements in several even all subforums. The advertisements are prepared well, and just copied and pasted when posting, and the post interval is short in many cases.

Not empty and the same contents: Advertisement posts are usually copied and pasted using a prepared template, so they are not empty and have the same contents.

Few views, few replies: Few users are fascinated by advertisement posts and give replies.

Low rated or Judging replies: Advertisement posts usually are low rated and even receive some replies that give some advertising judgments.

3.3 Advertisement messages

Users can send personal messages to other users in most forums. Sending messages means user's active level and is also considered as a feature of user contribution. Some advertisement users do not post advertisements in forum, but send advertisements as personal message to other users. Such messages can be featured as following.

Sent to many users: Advertisement messages are usually sent to as many users as possible to achieve more viewers.

Short interval: The advertiser users tend to send advertisement messages with short intervals.

Not empty and the same contents: A advertisement sent to many users is not empty and with the same content.

3.4 Dummy replies

To achieve more user contribution score or pretend not to be a junk or advertisement post, a user may forge some dummy replies to a topic he/she posted. These dummy replies may be produced by the same account, other accounts of the same user or accounts of the user's friends. Dummy replies can not be distinguished easily.

Dummy replies can be featured as many replies from a few users (may including topic owner). When a user want to forge replies, only few dummy replies is useless to forge contribution. So the amount of dummy replies of a topic is usually not a small number. But the involved user accounts are from few users.

3.5 Dummy rates

To influence the rate result, a user may also forge dummy rates, while which can be identified as a very few rates deviate from dominate rates if there are many rates.

4. User Contribution Measurement

In this section we will describe our method of fraud-tolerant user contribution measurement.

4.1 Common features

Here we use the features listed in Chai's paper [5], which are shown in table 2. Using these features, we can only measure user contribution without the consideration of frauds in the contributions.

4.2 Fraud-tolerant parameter

Now we introduce fraud-tolerant capability into our user contribution measurement method, and some parameters are introduced. As is shown in table 3, in which the denotation of each fraud is the same to that in section 3.

Table 2: Common Features

Code	*Name*	*Weight*
F1	# of posts created by a user	1
F2	# of voting polls created by a user	4
F3	# of votes cast by a user	1
F4	# of questions asked by a user	1.5
F5	# of questions answered by a user	2
F6	# of topics created by a user	1.5
F7	# of sticky topics created by a user	4
F8	# of topics that the user has provided the first reply	2
F9	# of responses received user topics	1.5
F10	# of views received for user topics	0.1
F11	# of personal messages sent	0.1
F12	# of personal messages received	0.2
F13	# of topic update notifications	0.1
F14	# of board update notifications	0.1
F15	# of quality posts created	3
F16	Frequency of user posts	3

Threshold of post's minimum length (P1): The parameter is used to resolve the "short length" problem of junk posts. It is not used when a user do a post, but used when calculated a user's contribution, that is, subtracting the counts of posts whose length is shorter than this threshold from the total posts by the user.

Threshold of post's minimum interval (P2): This is according to "shot interval" feature of junk and advertisement posts. A post is considered as junk or advertisement posts and is omitted when its post time and the former post's time by the same user is too close to the threshold.

Threshold of post's minimum views in its first *X* hours (P3): This is used to check posts with "few views" to discriminate junk posts or advertisement posts.

Threshold of post's minimum replies per *Y* views (P4): This is used to discriminate junk posts or advertisement posts with "few replies", based on we considered that normal topics may be replied at least once every Y views.

Threshold of post's minimum average rate value after *Z* rates (P5): This is according to the "low rated" feature of junk posts or advertisement posts. Here we considered that first Z rates can't represent the real average rates for the small samples count.

Table 3: Fraud-Tolerant Parameters

Code	*Fraud-Tolerant Parameter*	*Frauds type*	*Fraud feature*
P1	Minimum length of post	T1	Short length
P2	Minimum interval of post	T1, T2	Short interval
P3	Minimum views of post in the first X hours	T1, T2	Few views
P4	Minimum replies of post per Y views	T1, T2	Few replies
P5	Minimum average rate value of post after Z rates	T1, T2	Low rated
P6	Maximum proportion of filter-matched replies	T1, T2	Judging replies
P7	Maximum users be sent messages by a user continually without responses	T3	Sent to many users
P8	Maximum messages sent by a user per minute	T3	Short interval
P9	Maximum message sent by a user with same contents	T3	The same contents
P10	Maximum replies per user	T4	Many replies from a few users
P11	Minimum proportion of a rate value	T5	Few rates deviate from dominate rates

Threshold of post's maximum proportion of filter-matched replies (P6): When too many replies contain specified words such as "junk", "useless", "advertisement", this post is considered as a junk or advertisement post and should not be counted in.

Threshold of maximum users be sent messages by a user continually without responses (P7): This is used to discriminate advertisement messages on the basis of "sent to many users" feature. When a user sends too many messages continually without any response, the user is considered as sending advertisement.

Threshold of maximum messages sent by a user per second (P8): This is used to discriminate advertisement messages on the basis of "short interval" feature.

Threshold of maximum message sent by a user with same contents (P9): This is used for the "the same contents" feature of advertisement messages.

Threshold of post's maximum replies per user (P10): This is used to distinguish dummy replies.

Threshold of post's minimum proportion of a rate value (P11): This is used to find dummy rates, which is on the basis of that the proper rate of post should not be given by only a very small proportion of replies.

4.3 Fraud-tolerant user contribution measurement

In Chai's paper [4], the user contribution score (*UCS*) is calculated by summing up all features with different weight, as is shown in equation 1, in which *u* means a user, *m* means the count of features, f_{iu} means the i^{th} feature value of user *u*, w_i means weight of the i^{th} feature.

$$UCS_u = \sum_{i=1}^{m} W_i f_{iu} \tag{1}$$

According the fraud-tolerant parameters, we describe the fraud-tolerant user contribution score (*FUCS*) as equation 2, in which f'_{iu} means the result of f_{iu} subtracts count of posts or messages that were consider as frauds using corresponding threshold of fraud-tolerant parameters in table 2.

$$FUCS_u = \sum_{i=1}^{m} W_i f'_{iu} \tag{2}$$

5. Experimental Evaluation

We examined the contributions of 200 random users in an online forum in one month. In this forum, contributive users are rewarded with virtual coins. Coins can be used to ask for help in the forum, and users are granted different titles representing different levels according to the count of coins. So some users would like to post more topics or give meaningless replies to gain more coins.

Using the features and weights in table 2, we computed the *UCS*s of these users according to formula (1). Then according to the fraud-tolerant parameters in table 3 and values of each parameter in table 4, the *FUCS*s were computed using formula (2). Results are shown in figure 1.

Table 4: Experimental Values of Fraud-Tolerant Parameters

Parameters code	*Values*
P1	50 chars
P2	60 seconds
P3	20 views (*X*=1)
P4	1 replies (*Y*=10)
P5	C level (*Z*=10)
P6	40%
P7	20
P8	20
P9	5
P10	5
P11	5%

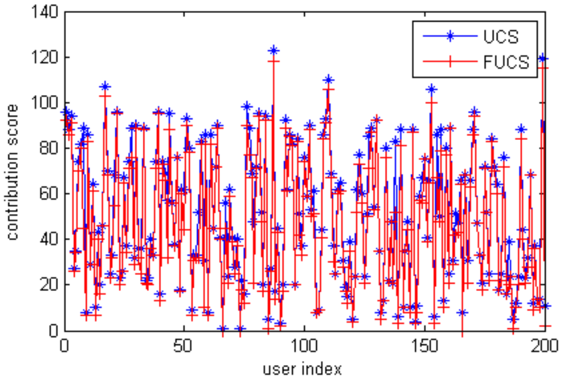


Fig. 1 *UCS*s and *FUCS*s of 200 randomly chosen users in an online forum during one month.

From the results of experiment we found that almost 81% users got a bigger *UCS* than *FUCS*, as is shown in figure 1, the reason of which is that users may do some frauds unconsciously for the allurement of rewards. A few users got very low *UCS* and even zero *FUCS*, which may be considered as a discriminative result of pure advertising users.

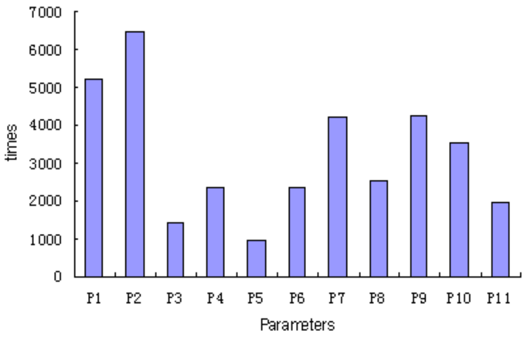


Fig. 2 Times of frauds detected using each parameter in an online forum during one month

Figure 2 shows the result of frauds detected with all parameters. As can be seen that the biggest count of frauds is detected using the parameter P2, minimum intervals of posts from the same users, and followed by the parameter P1, minimum length of posts, while P5 is the parameter that is used for least times.

6. Conclusions

In this paper, several possible frauds of user contribution in online forum are examined, and according to which, corresponding parameters are considered in the measurement of user contribution. These fraud-tolerant parameters can discriminate frauds from real user contribution. The result of this paper can be used in all kinds of online forum to build fair and effective contribution-reward mechanisms.

Acknowledgments

This work is supported by Zhejiang Provincial Natural Science Foundation of China (LY12F02016), and Henan provincial Natural Science Foundation of China (112300410009; 112300410125).

References

[1] R. Cheng and J. Vassileva, "User Motivation and Persuasion Strategy for Peer-to-peer Communities", in Proc. of the 38th Annual Hawaii International Conference on System Sciences (HICSS’05), 2005, pp. 193a.

[2] A. Abtoy, N. Aknin, B. Sbihi, A. El Moussaoui and K.E. El Kadiri, “Content validation as a tool for new pertinent Web 2.0 Blogs”, International Journal of Computer Science Issues, Vol. 9, Issue 3, No 3, May 2012, pp. 146-151.

[3] C. Wiertz and K. Ruyter, "Beyond the Call of Duty: Why Customers Contribute to Firm-hosted Commercial Online Communities", Journal of Organization Studies, vol. 28, no. 3, 2007, pp. 347-376.

[4] R. Klamma, M. A. Chatti, E. Duval, H. Hummel, E. T. Hvannberg, M. Kravcik, E. Law, A. Naeve, and P. Scott, "Social Software for Lifelong Learning", Journal of Educational Technology \& Society, vol.10, no. 3, 2007, pp. 72-83.

[5] K. Chai, V. Potdar and E. Chang. "User Contribution Measurement Model for Web-based Discussion Forums", In Proc. of 3rd IEEE International Conference on Digital Ecosystems and Technologies, June 2009, pp. 347-352.

[6] X. Shi, J. Zhu, R. Cai, and L. Zhang. "User Grouping Behavior in Online Forums", in Proc. of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2009), 2009, pp.777-786.

[7] B. Thomas Adler, L. Alfaro, I. Pye, V. Raman, "Measuring Author Contributions to the Wikipedia", Technical Report UCSC-SOE-08-08, School of Engineering, University of California, Santa Cruz, CA, USA. May 2008.

First Author Guo-Ying WANG received the B.S. degree in computer science from Beijing JiaoTong University, Beijing, China, in 1999 and M.S. degree in computer science from Guangxi University, Nanning, China, in 2004. In the same year, He joined the faculty of the Computer Science Department, Information Engineering College, Zhejiang A&F University, where he is currently a lecturer. His research interests include computer networks, peer-to-peer networks, wireless ad-hoc and sensor networks, and mobile networks.

Second Author Shen-Ming QU received the B.Eng. degree from Hebei University and the M.S. degree from Henan University. Currently, he is a lecturer in Computing Center, Henan University. His research interests include Computer Network, information retrieval and computer vision.

Detection of Pulsing DoS Attacks at Their Source Networks

Ming Yu[1], Xiong-wei Li[2]

**[1]School of Information and Communication Engineering, Dalian University of Technology
Dalian, 116024, China**

**[2]Department of Computer Engineering, Ordnance Engineering College,
Shijiazhuang, 050081, China**

Abstract

Pulsing Denial of Service (PDoS) is a type of DoS attack. Its attacking behavior is intermittent rather than constant, which helps it avoid being detected. In this paper, an adaptive detection method is proposed for source-end detection of PDoS attacks. It has three distinctive features: (i) its detection statistic is based on the discrepancy in the aggregated outbound and inbound packets; (ii) a self-adaptive detection threshold adapts it quickly to the variations of network traffic and the latest detection result; (iii) random abnormalities in the normal network traffic can be filtered by consecutive accumulation of threshold violations. Experimental results show the minimum attack traffic that can be detected is less than 35% of the background traffic, under the requirements that probability of false alarms is less than 10^{-6}, probability of a miss during an attack is less than 10^{-2} and detection delay is within 7 sampling periods.

Keywords: *Pulsing DoS, Attack Detection, Adaptive Detection, Source-end Defense, Network Security.*

1. Introduction

At SIGCOMM 2003, Kuzmanovic and Knightly proposed a new generation of DoS attacks, which could decrease the throughput of normal TCP traffic by periodically sending high-volume traffic in a short period. They named it “shrew attack”[1]. By further and deep study on shrew attacks, X.Luo *et.al*. proposed a generic definition of PDoS (Pulsing Denial of Service)[2]. That is, a DoS attack can be called a PDoS attack only if its attack traffic is sent in a intermittent way. By this definition, a shrew attack is considered as a kind of PDoS attacks. Different from traditional DoS attacks, PDoS traffic is sent periodically and lasts for a short time within each attacking period. Thus, it is more difficult to detect PDoS attacks.

According to the different deployment locations, an autonomous DoS defense systems can be classified into source-end defense, victim-end defense and intermediate-network defense[3]. Among them, source-end refers to those networks that unwittingly host attacking machines; victim-end refers to the target network or the network that hosts the target machines; intermediate-network means the infrastructure between the attacking machines and the target. In recent years, source-end defense against DoS attacks has been a hotspot in network security. Several methods have been proposed for anomaly detection of the source-end traffic. Among them, the one used in the D-WARD system[4,5] is widely accepted. It adopts a set of legitimate traffic models to identify legitimate traffic and detect or constrain malicious traffic. Unfortunately, these models need to be updated periodically and therefore cannot adapt to the frequent changes in network traffic. This paper expatiates on our latest study on source-end defense against PDoS attacks.

Rest of this paper is organized as follows. Section 2 discusses the previous detection algorithms proposed for PDoS attacks. Section 3 presents the design of an adaptive method for source-end detection of PDoS attacks. Section 4 gives a performance analysis of the proposed method. Section 5 explains the parameter configuration in the proposed method. Section 6 presents the experiments and the detection results. Section 7 concludes this paper.

2. Related Work

Luo and Chang proposed a two-stage detection system to detect PDoS attacks on the receiver side[2]. Their method is based on the presence of two types of traffic anomalies induced by PDoS attacks: periodic fluctuations in the inbound TCP data traffic and a decline in the trend of the outbound TCP acknowledgement (ACK) traffic. In the first stage, the detection system monitors the inbound data and outbound ACK traffic using discrete wavelet transform. In the second stage, a nonparametric CUSUM algorithm is employed to detect the anomalies. Experiment results show the system is effective in detecting PDoS attacks with constant attack periods. However, it is ineffective in detecting flooding-based DoS attacks because such attacks will not cause periodic fluctuations in TCP traffic.

Hussain *et al*. proposed to differentiate between single-source and multi-source DoS attacks[6] by analyzing spectrum of the network traffic. Chen *et al*. found the power spectrum density of a traffic stream containing shrew attacks has much higher energy in low-frequency band as compared with legitimate traffic. Based on this observation, they proposed a spectral template matching method to detect shrew attacks[7,8]. YU *et al*. proposed a similar method to detect SYN flooding attacks[9]. However, all these spectrum-based methods are ineffective in detecting PDoS attacks with different attacking frequencies and intervals.

Sun *et al*. proposed to detect shrew attacks using a dynamic time warping method which is divided into two stages[10]. In the first stage, autocorrelation is used to extract the periodic patterns in the inbound network traffic and eliminate the problem of time shifting. In the second stage, a slightly modified dynamic time warping algorithm is used to detect the signature of a shrew attack based on its autocorrelation coefficient. However, performance of this method is unsatisfactory when used in detecting PDoS attacks which are not separated by a constant interval. Moreover, such methods are ineffective in detecting flooding-based DoS attacks because the assumed square-wave patterns in such methods do not exhibit in the traffic under attack.

The D-WARD system is designed and implemented for source-end defense of DoS attacks. It adopts a useful metric that computes ratio of the inbound TCP traffic to the outbound TCP ACK traffic in detecting DDoS attacks [4]. Such a metric is also adopted in the Vanguard DoS detection system[11,12]. In both systems, however, a fixed ratio of the inbound TCP traffic to the outbound TCP ACK traffic is used to distinguish an attack flow from legitimate ones, which cannot adapt to the frequent changes in network traffic. Therefore, it is of crucial importance to design an “intelligent” detection method which can automatically adjust its detection parameters to adapt to the changing network conditions.

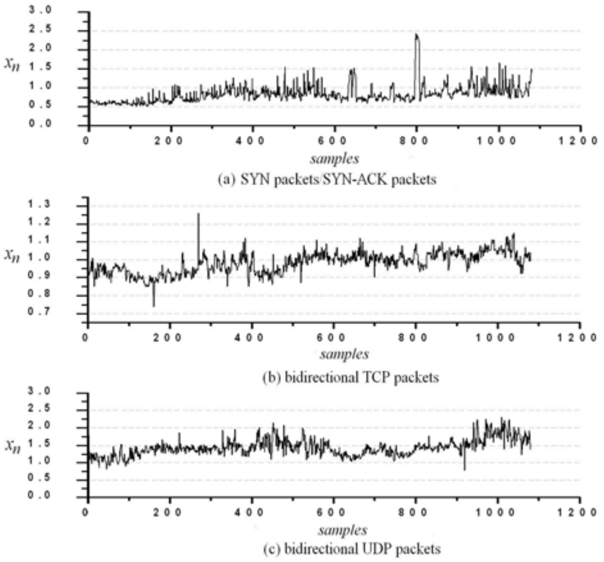
This paper expatiates on our latest study of source-end defense against PDoS attacks. Our main contribution is to propose an adaptive detection method for source-end detection of PDoS attacks.

3. Design of an Adaptive Method for Source-end Detection of PDoS Attacks

3.1 Problem Formulation

Let us begin by giving the problem formulation of PDoS detection before we go deep into the design of the proposed adaptive detection method.

Suppose $X=\{x_n, n=1,2,\cdots\}$ is a sequence of independent random variables observed sequentially, and $x_n=(O_n+1)/(I_n+1)$. Respectively, O_n and I_n denote the number of outbound requests and inbound replies collected within the n^{th} observation period. For legitimate traffic, O_n is approximately equal to I_n, thus we have $x_n\approx 1$. Normally, the mean of X (denoted by μ_X) is stable and close to 1. This conclusion has been referred by Mirkovic[4]. It is also supported by our analysis on some real traffic datasets collected at Dalian University of Technology. Fig. 1 gives the result of our analysis on one of those datasets.



(a) SYN packets/SYN-ACK packets

(b) bidirectional TCP packets

(c) bidirectional UDP packets

Fig.1 Analysis on x_n based on one of the real traffic dataset.

At a certain moment (random and unknown), an anomalous event occurs and μ_X is increased. When the anomaly ends, the mean of X is decreased to normal. Fig. 2 illustrates this process. However, no prior knowledge is known about the probability distribution function of X. The aim of a PDoS detection method is to accurately detect the start and the end of the PDoS attack as soon as possible.

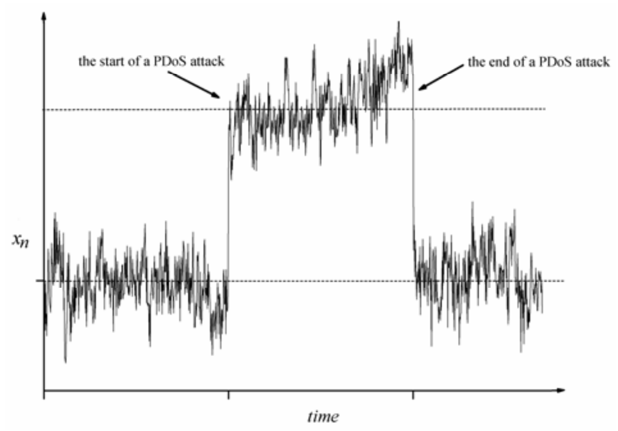


Fig.2 Illustration of PDoS detection.

3.2 Design of the Method

Essentially, the proposed method belongs to the area of sequential change-point detection[13,14]. It monitors network traffic at a regular interval and analyzes it to determine if any abnormalities are in the traffic. As we know, sequential change-point detection has to employ smaller amounts of data in order to keep such detection simple and efficient. At the same time, the data can not be so few that meaningful statistical characters can hardly be drawn. In the proposed method, data smoothing algorithms are considered to tradeoff these considerations. In addition, sequential change-point detection may also require that any alarms on attacks or anomalies be raised with a delay as short as possible. This is statistically guaranteed in the proposed method.

Three measures in design of the proposed method distinguish it from others. Respectively, they are as follows.

(i) Adopt the simple moving average algorithm in processing $X=\{x_n, n=1,2,\cdots\}$. Consequently, a new series, $U=\{u_n, n=1,2,\cdots\}$, is obtained. Here, u_n is used as the detection statistic and N is the sliding window size. Moreover, $u_n=\frac{1}{N}\sum_{i=n-N+1}^{n} x_i$.

(ii) Adopt the exponentially-weighted moving average (EWMA) algorithm in estimation of the mean of U after the n^{th} sampling period (denoted by $\hat{\bar{u}}_n$) when there are no alarms. We get $\hat{\bar{u}}_n = p\hat{\bar{u}}_{n-1} + (1-p)u_n$, where p is the EWMA factor. Once an alarm is raised, update of $\hat{\bar{u}}_n$ is suspended and the current $\hat{\bar{u}}_n$ is referred until the alarm is canceled.

(iii) Make consecutive estimations of the standard deviation of U after the nth sampling period (denoted by $\hat{\sigma}_n$) when there are no alarms. We get

$$\hat{\sigma}_n = \sqrt{2\sum_{i=N}^{n}(u_i - \overline{u}_{i-1})^2 / (n-1)}$$

Once an alarm is raised, update of $\hat{\sigma}_n$ is suspended and the current $\hat{\sigma}_n$ is referred until the alarm is canceled.

To reduce disturbance of random abnormalities in the normal network traffic, two variables are set. Respectively, they are AI for accumulation of threshold violations, and d_n for alarm decisions. The decision rules for threshold violations are as follows. Here, $\eta(\eta > 0)$ is a parameter indicating PDoS attacks in the network traffic.

```
------------------------------------------------------
 IF $u_n \geq \hat{\bar{u}}_{n-1} + \eta\hat{\sigma}_n$
   $AI$=$AI$+1;
   IF $AI$ equals $K$
     $AI$=$K$-1;
 ELSE
   IF $AI$>0
     $AI$=$AI$-1;
 END
------------------------------------------------------
```

The decision rules for raising alarms are as follows. Here, ‘0’ is for no alarms and ‘1’ is for raising alarms.

$$d_n = \begin{cases} 0, & \text{if } AI < K \\ 1, & \text{if } AI \geq K \end{cases}$$

To reduce miss of alarms during an attack, η is set as a function of d_n, that is,

$$\eta = \begin{cases} \eta_H, & \text{if } d_n \text{ equals } 0 \\ \eta_L, & \text{if } d_n \text{ equals } 1 \end{cases}$$

In summary, the proposed method is described in Fig.3.

4. Performance Analysis

As is mentioned in section 3, no prior knowledge is known about the probability distribution function of X. However, it is generally accepted that the discrepancy between the numbers of outbound packets and inbound packets is due to some transmission failures and the subsequent retransmissions. And usually, transmission failures are caused by various random network anomalies, such as network congestion, routing loops, link failures and server failures. To date, there is little evidence indicating these anomalies are closely correlated. Thus, it is a reasonable assumption that U is a stochastic series following Gaussian distribution $p_0 = N(\mu_0, \sigma_0)$ under normal conditions and $p_1 = N(\mu_1, \sigma_1)$ during a PDoS attack. Due to the data smoothing measure, we assume $\hat{\bar{u}}_n = u_0$ and $\hat{\sigma}_n = \sigma_0$ under normal conditions.

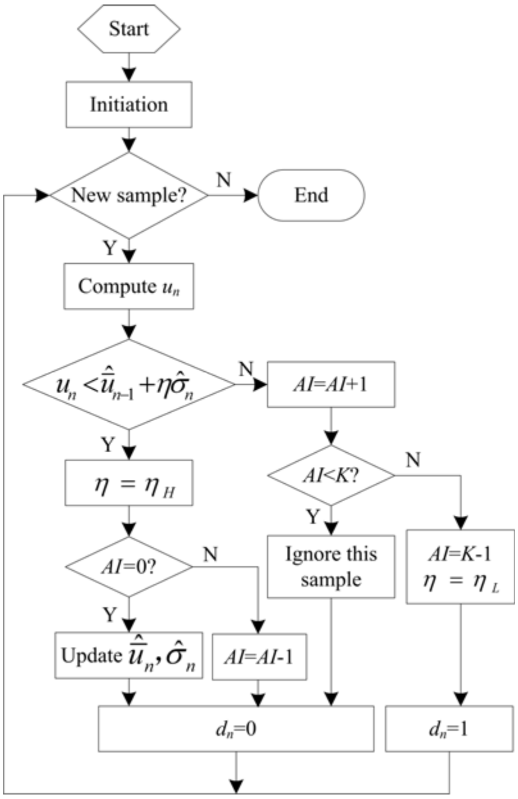


Fig.3 Flowchart of the proposed method.

For convenience of subsequent analysis, we define the following variables.

(i) O_N: the averaged number of outbound packets under normal conditions in one sampling period.

(ii) O_M: the averaged number of outbound packets sent by attacking machines in one sampling period during a PDoS attack.

(iii) I_N: the averaged number of inbound packets under normal conditions in one sampling period.

(iv) λ(λ>0): the proportion of O_M to O_N , and $\lambda = O_M / O_N$.

And we get

$$\mu_1 = E[\frac{O_M + O_N + 1}{I_N + 1}] \approx (1+\lambda)E[\frac{O_N}{I_N}] = (1+\lambda)\mu_0 \quad (1)$$

$$\sigma_1^2 = D[\frac{O_M + O_N + 1}{I_N + 1}] \approx D[(1+\lambda)\frac{O_N}{I_N}] = (1+\lambda)^2\sigma_0^2 \quad (2)$$

$$\sigma_1 = (1+\lambda)\sigma_0 \quad (3)$$

Based on Eq. (1)~Eq.(3), following results on performance of the proposed method can be obtained.

(i) Probability of false alarms (P_f)

According to the proposed method, false alarms are raised mainly by random network anomalies which last K sampling periods at least. Thus, we get

$$P_f = [\int_{\mu_0+\eta_H\sigma_0}^{\infty} p_0(x)dx]^K = Q^K(\eta_H) \quad (4)$$

Here, $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^{\infty} e^{-u^2/2}du$.

(ii) Probability of a miss during an attack (P_m)

$$P_m = \int_0^{\mu_0+\eta_L\sigma_0} p_1(x)dx \approx Q(\frac{\mu_1 - \mu_0 - \eta_L\sigma_0}{\sigma_1}) \quad (5)$$

(iii) Probability of detection (P_d) and detection delay (τ)

In PDoS detection, it is meaningless to talk about probability of detection without referring to the corresponding detection delay. In this paper, PDoS detection delay is defined as the time between the beginning of an attack and the first alarm of it. The unit of τ is the observation period t_s. If an attack is launched when 0<AI<K, $\tau \geq 0$. In the design of the proposed method, we consider the worst case that an attack is launched when AI is zero. Then, τ can be expressed as $\{\tau = K + 2m, m=0,1,2,3,\cdots\}$. However, it is practical to focus on the probability of detection with $\tau \leq K+2$. Thus, we get

$$P_d(\tau = K) = [\int_{\mu_0+\eta_H\sigma_0}^{\infty} p_1(x)dx]^K = [1 - Q(\frac{\mu_1 - \eta_H\sigma_0 - \mu_0}{\sigma_1})]^K \quad (6)$$

Let $P_d(\tau = K) = P_{d(K)}$, then

$$P_d(\tau = K+2) = (K-1)P_{d(K)} \mathrm{P}_{d(K)}^{1/K}(1 - \mathrm{P}_{d(K)}^{1/K}) \quad (7)$$

$$P_d(\tau \leq K+2) = P_{d(K)} + (K-1)P_{d(K)} \mathrm{P}_{d(K)}^{1/K}(1 - \mathrm{P}_{d(K)}^{1/K}) \quad (8)$$

5. Parameter Specification

Suppose: the specified requirements for PDoS detection is $P_{\mathrm{f}} \le \alpha$, $P_{\mathrm{m}} \le \beta$, $P_{\mathrm{d}}(\tau \le K+2) \ge \gamma$ and $\lambda_{\min} = \rho$. By Eq. (4) and Eq. (5), we get

$$Q^K(\eta_{\mathrm{H}}) \le \alpha \Rightarrow \eta_{\mathrm{H}} \ge Q^{-1}(\sqrt[K]{\alpha}) \tag{9}$$

$$Q(\frac{\mu_1 - \mu_0 - \eta_{\mathrm{L}}\sigma_0}{\sigma_1}) \le \beta \Rightarrow \frac{\mu_1 - \mu_0 - \eta_{\mathrm{L}}\sigma_0}{\sigma_1} \ge Q^{-1}(\beta) \tag{10}$$

Here, $Q^{-1}(x)$ is the inverse function of $Q(x)$.

Based on Eq. (8), we get

$$P_{\mathrm{d}(K)} + (K-1)P_{\mathrm{d}(K)}\, P_{\mathrm{d}(K)}^{1/K}(1 - P_{\mathrm{d}(K)}^{1/K}) \ge \gamma \tag{11}$$

Suppose $\xi(K,\gamma)$ is the minimum value which satisfies inequation $x^K + (K-1)x^{K+1}(1-x) \ge \gamma$ and $0 < \xi(K,\gamma) < 1$, then $P_{\mathrm{d}(K)}^{1/K} \ge \xi(K,\gamma)$. By Eq. (6), we get

$$\frac{\mu_1 - \eta_{\mathrm{H}}\sigma_0 - \mu_0}{\sigma_1} \ge Q^{-1}(1 - \zeta(K,\gamma)) \tag{12}$$

Based on Eq. (3)、Eq. (9)、Eq. (10) and Eq. (12), we get

$$\eta_{\mathrm{H}} \ge Q^{-1}(\sqrt[K]{\alpha}) \tag{13}$$

$$\eta_{\mathrm{L}} \le \frac{\mu_1 - \mu_0}{\sigma_0} - (1+\lambda)Q^{-1}(\beta) \tag{14}$$

$$\frac{\mu_1 - \mu_0}{\sigma_0} \ge Q^{-1}(1 - \zeta(K,\gamma)) + \frac{\eta_{\mathrm{H}}}{1+\lambda} \tag{15}$$

To fulfill the minimum value of the requirements, key parameters in the method can be set as follows.

$$\eta_{\mathrm{H}} = Q^{-1}(\sqrt[K]{\alpha}) \tag{16}$$

$$\eta_{\mathrm{L}} = Q^{-1}(1 - \zeta(K,\gamma)) + Q^{-1}(\sqrt[K]{\alpha}) - Q^{-1}(\beta) \tag{17}$$

6. Experiments

Five real traffic traces are used to validate the proposed method. These traces were all collected by a Endace® DAG card at Dalian university of technology with an OC-48c PoS link connected to CERNET. For each trace, two types of PDoS traffic were included. Respectively, they are SYN flooding traffic and UDP flooding traffic. All the attacks are launched every 10 minutes, and the bursting time of each attacking machines is 5 minutes. In order to reflect the advantages of the proposed method in detecting low intensity attacking traffic over those with a fixed detection threshold, a comparison is made on the detection results between the proposed method and a direct detection with a fixed threshold.

In order to make an impartial comparison, all experiments were done with the same requirements. Respectively, they are $P_{\mathrm{f}} < 10^{-6}$, $P_{\mathrm{m}} < 10^{-2}$ and $P_{\mathrm{d}}(\tau \le 7) \ge 0.7$. Based on these requirements, we get $\eta_{\mathrm{H}} = 1.6$, $\eta_{\mathrm{L}} = 0.7$ and K=5 according to Eq. (8)、Eq. (16) and Eq. (17). Other parameters that are related to the proposed method are set as follows. The sliding window size N is set to 3. The sampling interval t_s is set to 20 seconds. The EWMA factor p is set to 0.1. Initiation of $\hat{\bar{u}}_n$ and $\hat{\sigma}_n$ is set as $\hat{\bar{u}}_2 = 2$, $\hat{\sigma}_2 = 0.2$. Configuration of these parameters is fit for various requirements on the proposed method, and they seldom change in our experiments. In the fixed threshold detection, the detection threshold is set to 1.2, 1.6 and 2 respectively.

The experiments were carried out with the intensity of attacking traffic varied from 10% to 10 times of the normal background traffic. Table 1~Table 8 give the detection results on pulsing SYN flooding traffic and pulsing UDP flooding traffic by the proposed method and fixed threshold detection respectively. These results are obtained after 300 experiments on each trace.

Table 1 Detection of pulsing SYN flooding traffic by the proposed method

	Trace-1	Trace-2	Trace-3	Trace-4	Trace-5
P_{f}	1.9×10^{-3}	0.0	0.5×10^{-3}	0.0	0.0
P_{d}	99.4%	96.9%	98.8%	100.0%	100.0%
$\tau(t_{\mathrm{s}})$	6.7	6.8	6.6	6.7	6.9
$\lambda_{\min}$	0.59	0.20	0.38	0.34	0.29

Table 2 Detection of pulsing SYN flooding traffic by fixed threshold detection with the threshold set to 1.2

	Trace-1	Trace-2	Trace-3	Trace-4	Trace-5
P_{f}	3.0×10^{-2}	—	12.8×10^{-2}	12.1×10^{-2}	6.5×10^{-3}
P_{d}	99.7%	—	99.5%	99.2	100.0%
$\tau(t_{\mathrm{s}})$	0.1	—	0.1	0.1	0.1
$\lambda_{\min}$	0.66	—	0.29	0.17	0.23

Table 3 Detection of pulsing SYN flooding traffic by fixed threshold detection with the threshold set to 1.6

	Trace-1	Trace-2	Trace-3	Trace-4	Trace-5
P_f	1.9×10^{-3}	10.2×10^{-2}	1.3×10^{-2}	2.2×10^{-3}	0.0
P_d	100.0%	88.2%	100.0%	100.0%	100.0%
$\tau(t_s)$	0.1	0.0	0.1	0.1	0.1
λ_{min}	1.15	0.13	0.84	0.71	0.78

Table 4 Detection of pulsing SYN flooding traffic by fixed threshold detection with the threshold set to 2

	Trace-1	Trace-2	Trace-3	Trace-4	Trace-5
P_f	0.9×10^{-3}	0.0	3.4×10^{-3}	0.0	0.0
P_d	100.0%	100.0%	100.0%	100.0%	100.0%
$\tau(t_s)$	0.1	0.1	0.1	0.1	0.1
λ_{min}	1.64	0.38	1.42	1.25	1.33

Table 5 Detection of pulsing UDP flooding traffic by the proposed method

	Trace-1	Trace-2	Trace-3	Trace-4	Trace-5
P_f	4.6×10^{-3}	0.0	2.0×10^{-3}	0.0	0.0
P_d	91.7%	100.0%	81.7%	96.6%	92.0%
$\tau(t_s)$	6.3	7.0	6.4	6.6	6.4
λ_{min}	0.31	0.69	0.32	0.65	0.74

Table 6 Detection of pulsing UDP flooding traffic by fixed threshold detection with the threshold set to 1.2

	Trace-1	Trace-2	Trace-3	Trace-4	Trace-5
P_f	6.0×10^{-2}	0.0	10.2×10^{-2}	4.4×10^{-2}	10.4×10^{-2}
P_d	31.5%	100.0%	81.7%	99.6%	100.0%
$\tau(t_s)$	0.0	0.2	0.1	0.2	0.3
λ_{min}	0.12	0.74	0.26	0.92	0.80

Table 7 Detection of pulsing UDP flooding traffic by fixed threshold detection with the threshold set to 1.6

	Trace-1	Trace-2	Trace-3	Trace-4	Trace-5
P_f	5.7×10^{-2}	0.0	1.1×10^{-2}	2.2×10^{-3}	7.4×10^{-3}
P_d	98.1%	100.0%	96.9%	100.0%	100.0%
$\tau(t_s)$	0.1	0.2	0.2	0.2	0.4
λ_{min}	0.36	1.32	0.63	1.56	1.42

Table 8 Detection of pulsing UDP flooding traffic by fixed threshold detection with the threshold set to 2

	Trace-1	Trace-2	Trace-3	Trace-4	Trace-5
P_f	1.6×10^{-2}	0.0	3.0×10^{-3}	0.0	0.0
P_d	100.0%	100.0%	98.6%	100.0%	100.0%
$\tau(t_s)$	0.0	0.3	0.2	0.3	0.4
λ_{min}	0.70	1.93	1.05	2.22	2.06

In these tables, "—" stands for invalid detections which occur when the detection threshold is too low to avoid false alarms. λ_{min} denotes the lowest intensity of the attacks that can be detected. Our explanation on all of the detection results are give as follows.

(i) As we can see, part of the results on P_f in Table 1 and Table 5 are between 0.5×10^{-3} and 4.6×10^{-3}. This is related to insufficiency of trace data in the corresponding traces. Analysis shows there are 1080 sample data (the traffic data was collected in 6 hours) in trace-1 and 2030 sample data (the traffic data was collected in 6 hours about 11 hours) in trace-3. Even if one or two false alarms occur, P_f will rise to 10^{-3}. In practice, this is acceptable. In our opinion, P_f will decrease if more traffic is collected. This opinion can be supported by trace-2、trace-4 and trace-5 which contain more traffic data than that in trace-1 and trace-3.

(ii) The detection results on P_d, τ and λ_{min} obtained by using the proposed method fulfill the requirements on the detection. Since the key parameters η_H 、 η_L and K configured in the proposed method are derived from Eq. (8)、Eq. (15) and Eq. (16), accuracy of the performance analysis in section 4 is proved.

(iii) Fixed threshold detection is not ideal for PDoS detection. A proper threshold is hard to determine in the detection. If the detection threshold is lower, P_f will increase to 10^{-2}. This can not be improved by gathering more traffic data. If the detection threshold is higher, attacks may be missed. This is reflected by $\lambda_{\min}$ which is higher in fixed threshold detection than in the proposed method.

7. Conclusions

Adaptive detection of PDoS attacks is a newly proposed research area. In this paper, an adaptive method is proposed based on the assumption of normal distribution of the detection statistic which is a ratio between the outbound packets and the inbound packets in the source-end networks of the attacking machines. Distinct characters in design of the proposed method include: (i) its detection statistic is based on the discrepancy in the aggregated outbound and inbound packets; (ii) a self-adaptive detection threshold adapts it quickly to the variations of network traffic and the latest detection result; (iii) random abnormalities in the normal network traffic can be filtered by consecutive accumulation of threshold violations. Performance analysis of the proposed method is made in terms of probability of false alarms, probability of a miss during an attack, probability of detection, and detection delay. Experiments on real traffic traces validate the accuracy of our performance analysis on the proposed method and show the efficacy of the proposed method in source-end detection of pulsing SYN flooding and pulsing UDP flooding.

Acknowledgments

This work is supported by (1) National Natural Science Foundation of China (Grant No.61172059); (2) the Scientific Research Foundation for Ph.Ds of Liaoning Province, China (Grant No.20111022).

References

[1] A.Kuzmanovic, and E.W.Knightly, "Low-rate TCP-targeted Denial of Service Attacks: the Shrew vs. the Mice and Elephants ", in ACM SIGCOMM 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, 2003, Vol.1, pp.75-86.

[2] X.Luo, and R.Chang, "On a New Class of Pulsing Denial-of-Service Attacks and the Defense", in Network and Distributed System Security Symposium, 2005, pp.67-85.

[3] Yu Ming, "A Nonparametric Adaptive CUSUM Method and Its Application in Source-End Defense against SYN Flooding Attacks", WuHan University Journal of Natural Science, Vol. 16, No. 5, 2011, pp.414-418.

[4] Jelena Mirkovic, and Peter Reiher, "D-WARD: A Source-End Defense Against Flooding Denial-of-Service Attacks", IEEE Transactions on Dependable and Secure Computing, Vol. 2, No. 3, 2005, pp. 216-232.

[5] Bekravi Masoud, Jamali Shahram, Shaker Gholam, "Defense against SYN-flooding Denial of Service Attacks Based on Learning Automata", International Journal of Computer Science Issues, Vol. 9, No. 3, 2012, pp.514-520.

[6] A. Hussain, J. Heidemann, and C. Papadopoulos, "A Framework for Classifying Denial of Service Attacks", in ACM SIGCOMM 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, 2003, Vol.1, pp.99–110.

[7] Y. Chen, and K. Hwang, "Collaborative Detection and Filtering of Shrew DDoS Attacks Using Spectral Analysis", Journal of Parallel and Distributed Computing, Vol. 66, No. 9, 2006, pp. 1137–1151.

[8] Y. Chen and K. Hwang, "Spectral Analysis of TCP Flows for Defense against Reduction-of-Quality Attacks", in IEEE International Conference on Communications, 2007, Vol.1, pp.1203-1210.

[9] Ming Yu, and Xi-yuan Zhou, "An Adaptive Method for Anomaly Detection in Symmetric Network Traffic", Computers & Security, Vol.26, No.6, 2007, pp.427-433.

[10] H. Sun, J. C. S. Lu, and D. K. Y. Yau, "Defending against Low-rate TCP Attacks: Dynamic Detection and Protection", in the 12th IEEE International Conference on Network Protocols, 2004, Vol.1, pp. 196–205.

[11] Xiapu Luo, Edmond W. W. Chan, and Rocky K.C.Chang, "Detecting Pulsing Denial-of-Service Attacks with Nondeterministic Attack Intervals", EURASIP Journal on Advances in Signal Processing, Vol.2009, 2009, pp.1-13.

[12] C. W. Zhang, Z. P. Cai, W. F. Chen, et.al., "Flow Level Detection and Filtering of Low-rate DDoS", Computer Networks, Vol. 56, No.15, 2012, pp.3417-3431.

[13] M.Basseville, and I.V.Nikiforov, Detection of Abrupt Changes: Theory and Applications, New Jersey: Prentice Hall, 1993

[14] Ullah Fasee, Tariq Waqas, Arshad Muhammad, *et.al.*, "Analysis of Security Techniques for Detecting Suspicious Activities and Intrusion Detection in Network Traffic", International Journal of Computer Science Issues, Vol. 9, No. 2, 2012, pp.259-265.

Ming Yu received the BS degree in electronics engineering in 1998 from Shandong University, China. He received the MS degree and Ph.D degree in information and telecommunication system in 2004 and 2008 from Xidian University, China. He is currently an associate professor in Dalian University of Technology, China. He is also a member of IEEE Computer Society. So far, he has 15 papers published in international journals. His research interests include network security, cloud computing and DoS defense.

Xiong-wei Li received the BS degree in electronics engineering in 1998 from Wuhan Air Force Radar Academy, China. He received the MS degree and Ph.D degree in information and telecommunication system in 2004 and 2008 from Ordnance Engineering College, China. He is currently an associate professor in Ordnance Engineering College, China. He is also a member of IEEE Computer Society. So far, he has 10 papers published in international journals and conferences. His research interests include network security, and DoS defense.

Software Development Process Improvement Framework (SDPIF) for Small Software Development Firms (SSDFs)

Mejhem Yousef Al-Tarawneh[1], Mohd Syazwan Abdullah[2], and Jasem Alostad[3]

[1,2] College of Arts and Sciences, School of Computing, Universiti Utara Malaysia
06010 UUM Sintok, Kedah, Malaysia

[3]The Public Authority for Applied Education and Training (PAAET), College of Business
P.O. Box.23167, Safat 13092, Kuwait

Abstract

Most of the software development organizations all over the world are Small Software Development Firms (SSDFs). These firms have realized that it is necessary to organize and improve their software development and management activities. Traditional software process improvement (SPI) models and standards are generally not possible to be implemented directly by SSDFs, as these firms are not capable of investing the cost of implementing these programs due to limited resources and strict deadlines to complete the projects. In addition, the existing regional SPI models which were developed for SSDFs are not suitable to be implemented by SSDFs all over the world. Furthermore, SSDFs also have ignored the software development practices to explain "how to do the improvement"; as they only focus on "what to do for improvement". This paper presents a new software development process improvement framework (SDPIF) for SSDFs based on eXtreme programming (XP) as the software development method and Capability Maturity Model Integration version 1.2 for Development (CMMI-Dev1.2) as the SPI model.

Keywords: *CMMI-Dev1.2, XP Method, Software Development Process Improvement Framework.*

1. Introduction

Technological advancements affect our life in many ways, and control our way of living in all sectors. In the software development field, we can see the high spread of SSDFs all over the world [1]. These firms play a crucial role in the economy of many countries, where they develop a large portion of the required software applications, offer many job opportunities, and exploit new technologies [2]. Unfortunately, these firms are suffering from problems related to developing their software products as they are unaware of the basic software development best practices. The main reason to this is that most of them are using ad-hoc manner for the software development [3].

In addition, theses firms have lack of understanding of the success factors of SPI and do not have enough people to perform all the SPI activities. Therefore, they find themselves to be very far from implementing formal SPI traditional models and standards, such as ISO 9001 Series, ISO/IEC 15504 (SPICE), ISO/IEC 12207 and BOOTSTRAP, where these models and standards were developed for large and very large firms, very complicated and too expensive to be implemented by SSDFs [4].

Even though the SPI traditional models and standards are difficult to be implemented directly by SSDFs, SPI in these firms is still possible through simplification of these models and standards [1] [5]. There are some regional initiatives of SPI which were developed for SSDFs, such as: OWPL in Belgium; ASPE-MSC in Brazil; PRISMS in Britain; iFLAP in Sweden; MESOPYME in Spain, MoProSoft in Mexico; and MPS in Brazil [6][7]. However, these initiatives are not suitable for SSDFs all over the world, as they were developed based on the characteristics, environments, and infrastructures of firms in these specific countries where the models originated. In addition, the development of these initiatives were based on simplifying the SPI traditional models or standards by selecting the suitable Key Process Areas (KPAs) of SPI traditional models or standards which are suitable for SSDFs in the specific country, without identifying the suitable software development practices that would achieve global quality level [6][8].

Pikkarainen [9], and Lina and Dan [10] indicated the need for a suitable SDPIF for SSDFs. This improvement framework should specify how to carry out the tasks of improving the software processes [4]. In addition, Richardson [11] stressed the need of these firms to have SDPIF, which focuses on the software processes activities, and providing faster return on investment, flexible, and easy-to-use. The framework will allow these firms in knowing "what to do for improvement" by the SPI model and "how to do the improvement" by software development best practices [9].

This paper highlights the reasons of selecting XP as a software development method and CMMI-Dev1.2 as a SPI model in developing the new SDPIF for SSDFs as discusses in Section 2. Section 3 details out the establishment of the SDPIF and the findings of the stages used in developing the framework. The new framework and how it can be used is discussed in section 4. Section 5 concludes the paper and discusses future direction in this area.

2. Why XP Method & CMMI-Dev1.2 Model?

Lightweight software development methods are more suitable for SSDFs such as agile methods that are more applicable for these firms [12]. Agile development methods [13] have been designed to solve the problem of delivering high-quality software on time under constantly and rapidly changing requirements. The XP method is considered as the most popular and effective method compared to other agile development methods for software development processes in SSDFs. Due to the flexibility and agility of XP method, this method is reflect to as extreme, as it take good aspects in developing the software and applies these aspects extremely.

XP method has been used in this study as a baseline development method in developing the SDPIF for SSDFs these several reasons such as: (1) XP is more applicable for small, medium-scale and less complex projects and it is the most widely used agile methods as well as one of the more prominent approaches that adheres to agile principles; (2) XP is easy to use; (3) XP could be easily adapted with changing requirements; (4) XP achieves SPI better than agile methods; it conforms to level two in CMMI-Dev1.2; and (5) XP practices work tightly together by carefully applying different practice at a time that will eventually lead to SPI [9][14][15].

As for CMMI, this model is the comprehensive software improvement model of the Software Engineering Institute (SEI) based on some emerging CMM models which are Capability Maturity Model for Software (SW-CMM) v2.0 draft, Systems Engineering Capability Model (SECM), and Integrated Product Development Capability Maturity Model (IPD-CMM) v0.98. Furthermore, the CMMI model was represented in eight versions which are: CMMI for Development v1.02 (2000), CMMI for Development v1.1 (2002), CMMI for Development v1.2 (2006), CMMI for Acquisition v1.2 (2007) and CMMI for Services v1.2 (2007), CMMI for Development v1.3 (2010), CMMI for Acquisition v1.3 (2010) and CMMI for Services v1.3 (2007) [16].

In this study, CMMI-Dev1.2 was chosen as a SPI model in developing the SDPIF for SSDFs for several reasons such as: (1) This model was written specifically for the software industry to guide the software development improvement processes [17] and also to improve upon the best practices of other improvement models in many important ways [18]; (2) CMMI-Dev1.2 provides a comprehensive integrated solution for development and maintenance activities applied to products and services [4]; (3) CMMI-Dev1.2 is a widely-used beneficial approach for identifying the key weaknesses of a software development process which need immediate attention and improvement especially with agile development methods [9]; (4) CMMI-Dev1.2 can aid SSDFs in achieving their quality goals when used as guides for SPI [19]; and (5) Even though CMMI-Dev1.3 is the newest version of the CMMI generations, CMMI-Dev1.2 has been broadly used for assessing and improving the organizational maturity and process capability of most software development firms in the world, as this model presents extensive descriptions of how the various good practices fit together [4].

In addition, CMMI-Dev1.2 model and XP practices could be used as a combined approach to integrate the best abilities of both together, where these aspects not only can co-exist, but they even support each other [10]. Furthermore, CMMI-Dev1.2 is considered a suitable way to improve the software process of XP method [15], where high levels of CMMI-Dev1.2 would be possible to be achieved by extending XP method [20]. However, there is no extension work carried out in this respect, but it rather focuses on mapping XP method to CMMs KPAs. Accordingly, XP method is used with CMMI-Dev1.2 model in this study to develop the SDPIF for SSDFs.

3. Establishment of the SDPIF for SSDFs

Ramsin [21] used stages strategy to develop software modeling analysis methodology, where each stage has goals and tasks to achieve these goals. Stages strategy is useful and suitable to be used in this study to establish the SDPIF for SSDFs. In this study, four stages are required to be followed sequentially to establish the desired, which are [8]:

- **Stage One:** Aligning XP method to CMMI-Dev1.2.
- **Stage Two:** Developing the proposed SDPIF.
- **Stage Three:** Verifying the the proposed SDPIF.
- **Stage Four:** Validating the verified SDPIF.

Sections 3.1 to 3.4 explain these stages and discuss the results of each stage in detail.

3.1 Stage One: Aligning XP practices to the specific goals of CMMI-Dev1.2 KPAs.

This stage aimed to identify the coverage ratio of XP practices to the specific goals of CMMI-Dev1.2 KPAs. This alignment was based on the specific goals, because all the generic goals are repetitive throughout the specific goals [22]. In this stage, three scales had been used to identify the coverage ration of XP practices to the KPAs of CMMI-Dev1.2 which are: (1) Largely supported: XP practices largely support the specific goals of the KPA; (2) Partially supported: XP practices partially support the specific goals of the KPA; and (3) Not-supported: XP practices do not support or not applicable for the specific goals of the KPA.

As a result of this alignment, the CMMI-Dev1.2 KPAs are classified into there groups as follows [8]:

- **Largely supported (++):** This group consists of twelve KPAs, which are: project planning, project monitoring and control, configuration management, technical solution, product integration, verification, validation, integrated project management + IPPD, risk management, decision analysis and resolution, and quantitative project management, causal analysis and resolution.
- **Partially supported (+):** This group consists of eight KPAs, which are: requirements management, measurement and analysis, process and product quality assurance, requirements development, organizational process definition + IPPD, organizational training, organizational process performance, and organizational innovation and deployment.
- **Not-supported (-):** This group consists of two KPAs, which are: supplier agreement management and organizational process focus.

At the end of this stage, the coverage and missing specific goals of each KPA are known and used as inputs in stage two.

3.2 Stage Two: Developing the Proposed SDPIF

This stage starts by extending XP method to fulfill the partially and not-supported KPAs of CMMI-Dev1.2. In this regard, the Extension-Based Approach (EBA) of the Situational Method Engineering (SME) theory was suitable to be adapted in this study to extend the XP method, as this approach was developed for extending the existing methods to achieve specific issues [23]. Figure 1 shows the processes of the adapted EBA.

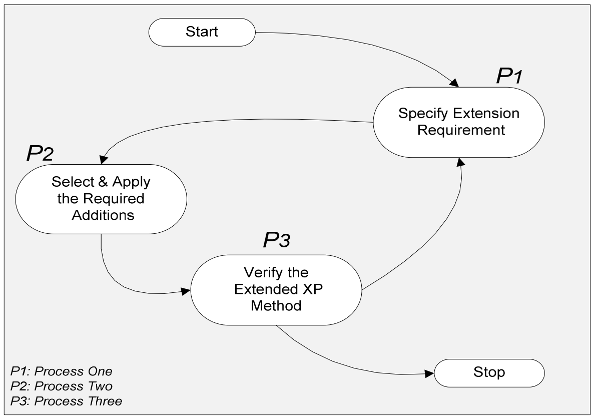

Start
P1
Specify Extension Requirement
P2
Select & Apply the Required Additions
P3
Verify the Extended XP Method
Stop
P1: Process One
P2: Process Two
P3: Process Three

Fig. 1 EBA in to Extend XP Method (adapted from [23]).

As shown in Figure 1, three main processes (P1, P2, and P3) are used in extending XP method, which are [8]:

- **P1 (Specify extension requirements):** this process aims to extract the required software development, management, and improvement addition that are needed to cover the partially and not-supported KPAs of CMMI-Dev1.2. In this regard, the required additions were extracted to fulfill the missing KPAs of CMMI-Dev1.2.
- **P2 (Select & apply the required additions):** this process aims to extract the new phases of the proposed Extended-XP method and harmonize these phases to be comprehensive for all the popular software development methodologies. In addition, it distribute the required software development, management,

and improvement additions into the new phases of the proposed Extended-XP method based on the need for these additions during the software development lifecycle.

In this regard, the popular software development models (Waterfall, Spiral, incremental, and prototyping) and the required software development, management, and improvement additions were used to extract the new phases of the Extended-XP method to be harmonized and homogeneous compared to generic activities of the popular development models.

- **P3 (Verify the Extended-XP method):** this process aims to verify the commitment of the proposed Extended-XP method to the principles of XP method. This process is very important to ensure that the Extended-XP method is still applicable as an agile method. In this respect, XP values that reflect the XP principles were used as a main question during the verification process in stage three.

In general, the SPI framework consists of four generic elements, which are [24]: (1) Software Process: a set of tools, practices, and methods to produce software products according to specific plan; (2) Assessment: this element is used to assess the current state of the software process and this can be done by implementing the suitable assessment methods; (3) Capability Determination: this element is used to know the capability level of the software process and what motivates an organization to do process improvement by identifying the capability and risks of a process; and (4) Improvement Strategy: based on the capability determination results, the improvement strategy will identify the changes which should be made to the process.

In this study, the generic elements of the SPI framework were used as a baseline in developing the desired framework. Thus, these elements had been re-arranged to be suitable for software development and improvement issues by integrating the CMMI-Dev1.2 as assessment model and the proposed Extended-XP method as a development method into the generic elements of SPI framework. Figure 2 shows the foundation of the elements in the proposed framework.

As shown in Figure 2, there are three generic processes in the new proposed software development improvement framework which are: (1) assess the current software development processes; (2) adopt the Extended-XP method; and (3) identify the best practices of the current project firms. Section 4 explains in detail each stage of the framework after the verification process.

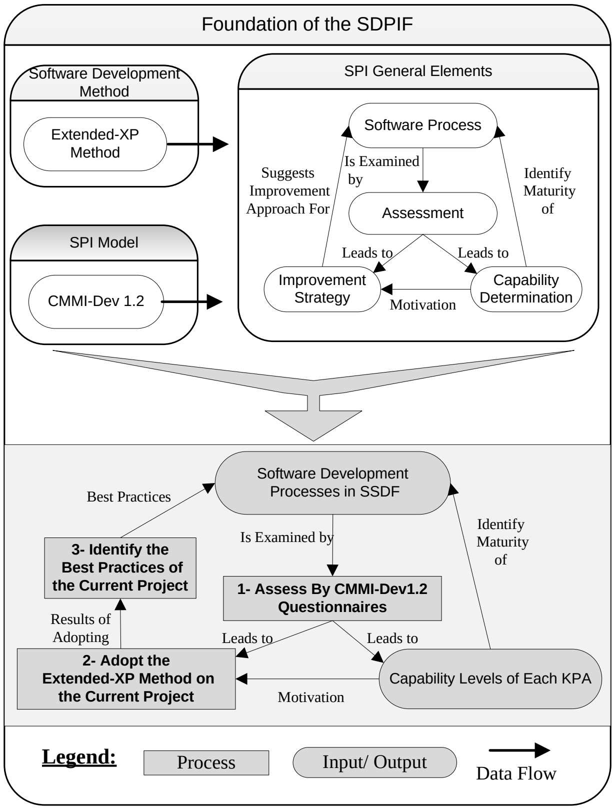


Fig. 2 Generic elements of the new framework.

At the end of this stage, the newly proposed framework goes to the third stage for the verification process.

3.3 Stage Three: Verifying the Proposed SDPIF

In this stage, the focus group method coupled with Delphi technique had been used to: (1) verify the compatibility of the proposed SDPIF to the specific goals of CMMI-Dev1.2 KPAs; (2) verify the commitment of the proposed Extended-XP method to XP values; (3) verify the suitability of the proposed framework and proposed Extended-XP roles for their related practices; and (4) verify the suitability of the proposed framework and the proposed Extended-XP structures for the software development and improvement issues.

In this regard, seven professional developers and managers with three expert researchers have participated in verifying the proposed SDPIF, where three rounds were performed in conducting this verification process. Table 1 shows these rounds.

As a result of the verification process, several modifications had been made to the proposed framework. The major modification was to remove the related activities of the organizational innovation and deployment process area from the proposed framework, as this area is not suitable to be implemented by SSDFs. Section 4 presents the verified framework (including the verified Extended-XP method.

Table 1: Focus group rounds.

Round	*Session*	*Activities*
R.1	S.1	- Researcher introducing himself. - Thanking the focus group members for the participation. - Presenting the research problem. - Presenting the purpose of the research.
	S.2	- Explaining the verification questions. - Answering the verification question individually. - Explaining if there are any inquiries about the verification questions.
	S.3	- Discussing the answers and suggestions of each focus group member by all the members.
R.2	S.1	- Modifying the proposed framework as suitable suggestions of focus group members.
R.3	S.1	- Viewing the verified framework to the members. - Asking if there is need for more new modifications.

At the end of this stage, the proposed SDPIF has been verified as suggested by focus group members [8]. To ensure that the verified framework is suitable SSDFs, there is a need to validate the suitability of this framework with more professional managers and developers who are working in these firms. Section 3.4 explains the two approaches used in validating the verified framework.

3.4 Stage Four: Validating the Verified SDPIF

Two validation approaches were used to validate the verified SDPIF, which are: a quantitative research method that involved a survey to validate the suitability of this framework for SSDFs, and qualitative research method (descriptive statistic) that involved two case studies to validate the applicability and effectiveness of this verified SDPIF for SSDFs. Section 3.4.1 & 3.4.2 explain the results of the two validation approaches.

3.4.1 Validating Suitability of the Verified SDPIF for SSDFs

A formal validation for suitability of the verified framework by SSDFs has been undertaken by using CMMI-Dev1.2 questionnaires as the main items in this validation. The questionnaire format consists of two parts: the first part is to know the general demographic information about the respondents; while the second part is to include all the specific goals of the suitable CMMI-Dev1.2 KPAs.

In this respect, the verified framework should be clearly read and understood by the professional developers and managers in these firms to evaluate it according to the characteristics of their firms and the requirements of the specific goals of each CMMI-Dev1.2 KPAs. Therefore, a hard copy which includes the detailed description of the verified framework (included the verified Extended-XP method) and the descriptions of CMMI-Dev1.2 KPAs were attached with these questionnaires.

These questionnaires were given to 80 professional developers and managers who are working in different SSDFs in Jordan. A total of 80 questionnaires distributed, and only 37 questionnaires were returned and 7 of them returned with missing data of more than (30%) for each questionnaire. Therefore, only 30 questionnaires have been used for the validation process. Sections 3.4.1.1 & 3.4.1.2 present the results of the two parts answers.

3.4.1.1 Part One: Respondents' Profile

This part consists of four questions: current position; current work; size of firm; and software experience. As a result of the answers of this part, the majority of the respondents were members of software engineering process group (40%), while the rest are distributed: managers (26.66%), technical members constituted (20%), and project or team leaders were (13.33%). Additionally, with regard to the current work activities the highest ratio was for code and unit testing (26.66%), where software design (16.66%), software quality assurance (17%), each of configuration management and software requirement (13.33%), and finally software process improvement (6.66%). In term of CMMI training, (90%) of respondents did not receive any CMMI training, while (10%) have received training. With regard to the software experience term, (66.66%) of respondents had 6-10 years, where the other two periods (less than 5 years & 11 years and above) got the same ratio (16.66%). Concerning the firm's size, (46.66%) of the respondents had working in firms that had 20-31 employees while (20%) of respondents had working in firms that had 41-50 employees. The firms that had 10 – 20 employees and 31 - 40 employees got the same ratio (16.66%).

3.4.1.2 Part Two: Validating the Suitability of the Verified SDPIF for SSDFs

In this part, the respondents were asked to rate the level of the suitability of the verified framework components for SSDFs. The questions of this part consisted of scaled-response from 1 to 5, 1= Strongly Unsuitable and 5= Strongly Suitable. Based on the interval scale, the appropriate interval is calculated as: appropriate interval = (number of scales -1) / (number of scales) [25]; the appropriate interval for this study is (4/5) = 0.8. Therefore, Table 2 shows the definitions of the intervals scales that explain the level used for each interval scales.

Table 2: Intervals scale definition of the suitability.

Interval	*Degree of Suitability*
From 1 To 1.80	Strongly Unsuitable
From 1.81 To 2.61	Unsuitable
From 2.62 To 3.42	Average
From 3.43 To 4.23	Suitable
From 4.24 To 5	Strongly Suitable

From the answers of the questions in this part, is can be concluded that [8]:

- **CMMI-Dev1.2 Level Two:** at this level, four KPAs achieved at the level of strongly suitable as follows: requirement management (4.66); project planning (4.50); project monitoring and control (4.56); and measurement and analysis (3.46), while the remaining three KPAs only achieved the level of suitable as a follows: supplier agreement management (3.50); process and product quality assurance (3.60); and configuration management (4.16).
- **CMMI-Dev1.2 Level Three:** at this level, just three KPAs achieved the level of strongly suitable as follows: technical solution (4.50); verification (4.30); and validation (4.36), while the remaining eight KPAs achieved the level of suitable as a follows: requirements development (3.83); product integration (4.13); organizational process focus (3.93); organizational process definition +IPPD (3.56); organizational training (3.93); integrated project management +IPPD (3.50); risk management (3.63); and decision analysis and resolution (3.60)
- **CMMI-Dev1.2 Level Four:** the two KPAs of this level achieved the level of suitable as a follows: organizational process performance (3.46); and quantitative project management (3.66).
- **CMMI-Dev1.2 Level Five:** causal analysis and resolution (3.63) is the only one KPA at this level, and achieved the level of suitable.

Therefore, it can be concluded that the verified SDPIF is suitable for SSDFs as all the related activities in the verified framework that aimed to achieve the requirements of the specific goals of the suitable CMMI-Dev1.2 KPAs are rated strongly suitable or suitable for these firms.

3.4.2 Validating the Applicability and Effectiveness of the Verified SDPIF for SSDFs

In order to validate the applicability and effectiveness of the verified framework for SSDFs, two Jordanian SSDFs implemented this framework to improve their software development processes in developing their software projects. The first firms used the verified framework in developing a computer skills online examination system, while the second firm used the verified framework in developing a brokerage online system [8].

At the end of implementing the verified SDPIF by the two firms, three evaluation criteria had been used to ensure that the verified framework is effective for these firms, which are: gain satisfaction [26], interface satisfaction [26], and task support satisfaction [26][27].

For this, the interview method has been used as a data collection method to evaluate the modified framework. The primary purpose of the interview method is to understand the meanings that the interviewees attach to issues and situations in context that are not structured in advance by the researcher's assumptions [28]. Therefore, it was suitable in this study to conduct interview with the projects team members of the two SSDFs to evaluate the effectiveness of the modified framework for SSDFs.

The results of evaluating this framework in terms of gain satisfaction, interface satisfaction, and task support satisfaction that are concluded from the answers of evaluating the verified framework by the two case studies can be summarized as follows:

- Gain Satisfaction Criteria

- **Perceived usefulness:** the verified framework enable the project team to execute their roles correctly and with high efficiency, because the practices of each role are clear. Therefore, the productivity of each member of the project team was good compared to ad-hoc manner which had been used in the firm before implementing the verified framework. Furthermore, the distributions of the roles in the first stage of the verified framework are very similar the current role of each member which led them to execute their roles easily.

- **Decision support satisfaction:** project managers were responsible for decision making, where this is the main role of the managers. Furthermore, the continuous communications between team members assist in making better decisions.
- **Comparison with other guidance:** the verified framework stages were very clear for the project teams, were each member has specific roles to perform and as such, there is no overlap between their roles.
- **Cost (effectiveness):** the verified framework was cost-effective because the coach enabled the project to keep on the right path and kept the team working on the current features for the actual iteration. Also, the tracker was careful not to interrupt the project too many times. Furthermore, the verified framework guidance helped the team in ensuring that they are implementing their roles in the right ways.
- **Clarity (clear and illuminate the process):** the verified framework was very clearly for the project team by the training that was conducted in the first stage of this framework. Moreover, the roles of coach and tracker helped the project team for any inquiry during the development processes.
- **Appropriateness for task:** the verified framework stages were comprehensive, where stage one helped to motivate the firm to implement the verified Extended-XP method (stage two) as the development method and stage three helped to identify the best practices of using this method.

- Interface Satisfaction Criteria

- **Perceived ease of use:** during the training process, the verified framework was easily understood.
- **Internally consistent:** the roles of each member in the team were very clear, where these roles ensure consistent development process.
- **Organization (well organized):** the verified framework was well organized and structured, where the sequence of verified framework stages and verified Extended-XP phases were useful to understand the activities easily.
- **Appropriate for audience:** based on developing the systems by the verified framework, the audiences were satisfied with product releases that helped them to add more features on the required products, as the verified Extended-XP method is incremental and iterative software development method.
- **Presentation (readable and useful format):** framework-SEPG members indicated that the verified framework is readable and is in the appropriate format. The project team also highlighted that the phases of the verified Extended-XP method were very clear, as the training process helped them to understand all things about this method.

- Task Support Satisfaction Criteria

- **Ability to produce expected results:** as a result of stage three for each case study, project managers and framework-SEPG members indicated that the implementation of the verified framework returned high capability levels compared to the levels before implementing this framework.
- **Ability to produce usable results:** the managers and framework-SEPG members indicted that the completed systems were usable for the end users because the customers participated in developing the systems (On-Site Customer), so the products were highly usable for the systems owners.
- **Completeness (adequate or sufficient):** managers and framework-SEPG members indicated that the verified framework was comprehensive for improving the software development and management processes in SSDFs. However, they argued it would be more sufficient when all KPAs of CMMI-Dev1.2 level five were included.
- **Ease of implementation:** framework-SEPG members indicated that the verified framework was very easy to implement, where the descriptions of each phases were clear. Therefore, it was easy to know the roles of each member in the developing process. The project teams also argued that the verified Extended-XP method was easy for implementation.
- **Understandability (simple to understand):** framework-SEPG members indicated that the framework was understandable. Project teams also argued the activities of the verified Extended-XP

method were easy to understand especially after the training processes.

4. The Verified SDPIF for SSDFs

As a result of verification process, the proposed framework had been verified. This section explains the stages of the verified SDPIF for SSDFs. In addition, section 4.1 presents the verified Extended-XP method. Figure 3 shows the verified SDPIF. The verified framework consists of three stages as a following:

- **Stage One: Assessing the Current Software Development Processes**
  Prior to implementing the verified framework, the framework-SEPG members are responsible for determining a suitable simple repository to be used during the implementation of the framework. To achieve this, the Microsoft Office is suggested as a tool for data storing issues. In this stage, framework-SEPG members start to assess the current software development processes by using the questionnaires of CMMI-Dev1.2 KPAs to determine the capability levels of these processes. Three scales can be used in this assessment. These are: (1) largely supported: the current software development processes achieve the majority of the specific goals of the KPA; (2) partially supported: the current software development processes achieve some of the specific goals of the KPA; and (3) not supported: the current software development processes can not achieve the specific goals of the KPA. As a result of this self-assessment, the firm will know the weaknesses of the current software development processes.

  Then, the Framework-SEPG members are responsible for rearranging the current software development processes to be suitable with the required roles of the verified framework based on the assessment results. This can be done by distributing the new roles of the verified framework to the project team members as to commensurate with their experiences. At the end of this stage, the new roles will be known for each employee in the firm.

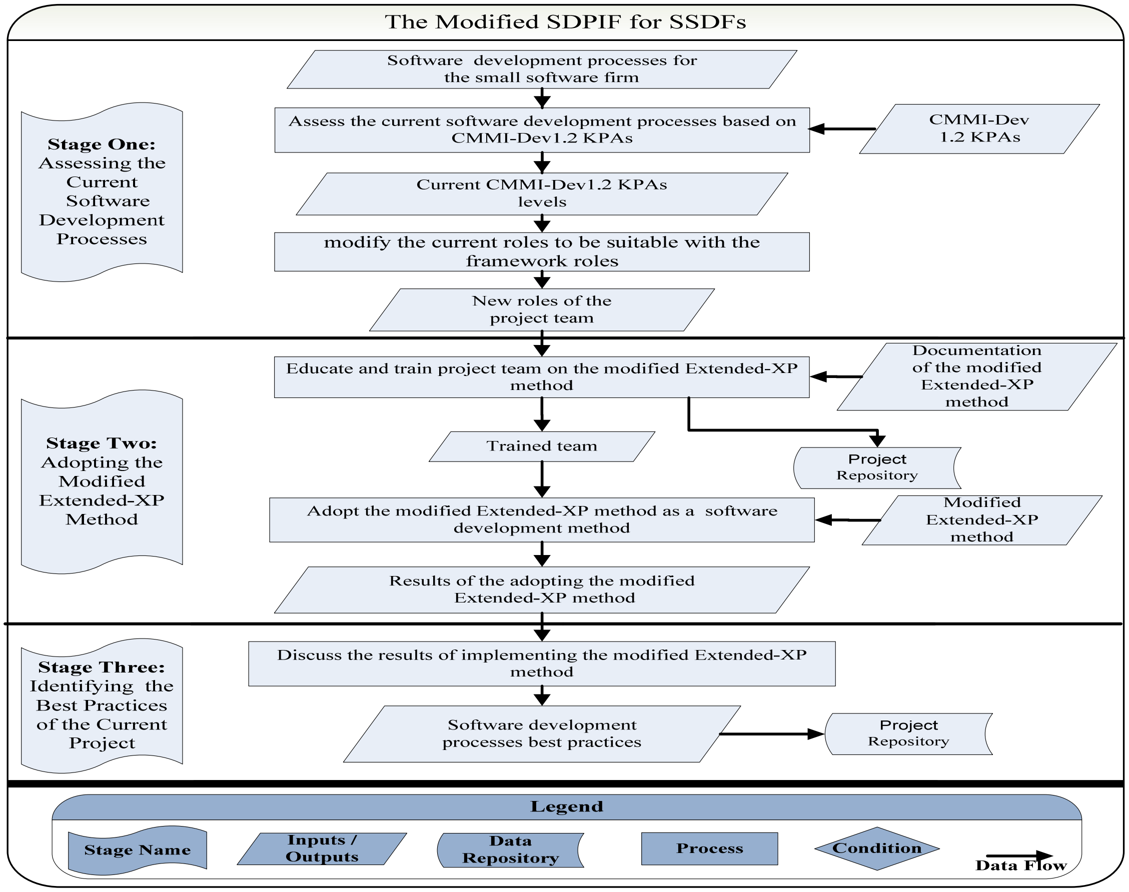

The Modified SDPIF for SSDFs
Stage One: Assessing the Current Software Development Processes
Software development processes for the small software firm
Assess the current software development processes based on CMMI-Dev1.2 KPAs
CMMI-Dev 1.2 KPAs
Current CMMI-Dev1.2 KPAs levels
modify the current roles to be suitable with the framework roles
New roles of the project team
Stage Two: Adopting the Modified Extended-XP Method
Educate and train project team on the modified Extended-XP method
Documentation of the modified Extended-XP method
Trained team
Project Repository
Adopt the modified Extended-XP method as a software development method
Modified Extended-XP method
Results of the adopting the modified Extended-XP method
Stage Three: Identifying the Best Practices of the Current Project
Discuss the results of implementing the modified Extended-XP method
Software development processes best practices
Project Repository
Legend
Stage Name
Inputs / Outputs
Data Repository
Process
Condition
Data Flow

Fig. 3 SDPIF for SSDFs.

- **Stage Two: Adopting the Verified Extended-XP Method**
  In order to implement the verified Extended-XP method in the right way; all the team members involved in the software development processes must have a good knowledge in their roles and must be trained. The best way to learn the verified Extended-XP method is through training courses. Furthermore, there is a need to support the team with the required XP books and with the documentation of the verified Extended-XP method during the training and the development lifecycle. Here, the framework-SEPG members are responsible for carrying out the training process prior the implementation of the verified Extended-XP method by the firm, while the members are responsible for: establish plans for training the developers, estimate the time required of training, determine if there is a need for outsourcing professional team in the training process, training the developers on how they can implement the activities of the verified Extended-XP, assessing the project teams efficiency to know if they are ready to implement the verified Extended-XP method or there is need for more training, and recording the training efficiencies in the project repository.

  As a results of the training process (through the assessment of the team's efficiency), it will be known if there are needs for more training or not. Accordingly, the teams will be ready to adopt the Extended-XP method in the right way. Section 4.1 discusses the phases of the verified Extended-XP Method.

- **Stage Three: Identifying the Best Practices of the Current Project**
  Referring to the results of the second stage, framework-SEPG members are responsible for meeting the project team to discuss the best practices of implementing the verified framework by using the specific practices of CMMI-Dev1.2 KPAs as the main items in this discussion. In this questionnaire; three choices have been used to answers these questions which are; "Yes" when the practice is well established and consistently performed; "Don't Know" when there are uncertainty about how to answer the question; "Does Not Apply" when the required knowledge about the project or firm and the question asked, but the question does not apply to the project; and "No" when the practice is not well established or is inconsistently performed. Through this, the best practices of implementing the verified framework for the current project can be determined. Then, the framework-SEPG members are responsible for documenting these best practices in the project repository to be taken into account for the coming projects.

As for the roles in the SDPIF, this framework has the same roles of XP method [13] with several additional practices to programmers, coach, and tracker. Furthermore, there are two new roles that have been added to this framework compared to XP roles which are framework-SEPG members and Extended-XP-SEPG members. The roles in the SDPIF are as follows:

- **Programmers, Customer, Tester, Coach, Tracker, Consultant, and Big Boss:** these roles have the same practices of the XP method roles [13]. In this framework, these roles are used during the implementation of the modified Extended-XP method.
- **Coach and Tracker:** together coach and tracker are responsible for implementing the required metrics to achieve the objective of process and product quality assurance and the process performance at the third phase of the proposed Extended-XP method which are: (1) calculating the difference between estimates and actual time spent on user stories or tasks; (2) calculating the velocities of the projects and the length of pair programming sessions and store these in the project repository; and (3) calculating the number of failed acceptance tests, and number of severity defects after release.
- **Programmers and Extended-XP-SEPG Members:** these members are responsible for implementing the supplying process at the first phase of Extended-XP method, where programmers are responsible for extracting the required unavailable development tools or services; while Extended-XP-SEPG members are responsible for executing the supplying process with the external suppliers.
- **Framework-SEPG Members:** these members are responsible for: (1) specifying the suitable simple project repository in the first stage of the verified framework to keep the important date during the implementing of this framework; (2) assessing the current software development processes in the first stage of the framework; (3) modifying the roles of the current software development processes to be suitable with the verified framework roles in the first stage of the verified framework; (4) arranging the required organizational training before starting to adopt the verified Extended-XP method in the

second stage of the verified framework, which are: establishing planning for training the programmers, estimating the time required of training, determining if there is need for out sources professional team in training process, training the project team on how they can implement the activities of the verified Extended-XP, recording the training results and assessing the training efficiencies in the project repository, and meeting the project team in the third stage of the verified framework to extract the best practices of the current project and storing these practices to help incoming projects in the same firm.

4.1 The Verified Extended-XP Method

The verified Extended-XP method consists of four generic phases which are:

- **Phase One: Requirement Management**
This phase consists of the same contents of the three generic processes which are: (1) exploring the customer requirements process: this process consists of the same activities of the exploration phases in XP method [13]; (2) Supplying the required development tools and services process: this process is used to support the project with the unavailable required development tools or services, where programmers are responsible to identify the unavailable required development tools and services. Then, the Extended-XP SEPG members are responsible for determining the type of acquisition that will be used for the products to be acquired, selecting suppliers, establishing and maintaining agreements with suppliers, executing the supplier agreement, monitoring selected supplier processes, evaluating selected supplier work products, accepting delivery of acquired products, and transitioning acquired products to the project; and (3) Planning process: this process consists of the same activities of the planning phases in XP method [13].
- **Phase Two: Development**
This phase has the same activities of iteration to release phase in the XP method [13].
- **Phase Three: Product Delivery and Product & Process Efficiency**
This phase consists of the same activities of productionizing phase in XP method [13] and other additions such as: (1) several metrics that could be appropriate for objectively verifying the products and the process which are: release plan adherence, percentage of test cases that are running successfully, percentage of acceptance tests that are running successfully, length of pair programming sessions, and project velocity; (2) convey the metrics through defined channels to the affected parties and senior management. At the end of this phase, there is a need to keep the metrics results to help in the measurement of the same user requirements for the coming projects.
- **Phase Four: Maintenance & Death**
This phase consists of the same activities of maintenance and death phases of XP method [13].

5. Conclusion and Future Work

SSDFs represent a high proportion of software firms around the world. Unfortunately, most of these firms are adopting add-hoc manner in developing their software products and are unaware of the basic software best practices and the existence of SPI traditional models and standards, where all of these models and standards were developed for large firms. In addition, the regional SPI initiatives are not suitable to be implemented by SSDFs all over the world, as they were developed based on the environments of firms in these specific countries where the models originated, and they neglect the software development practices to do the suitable improvement.

This study aimed to help SSDFs in developing a suitable SDPIF to enable them for managing and improving the software development activities in a systematic way to keep these firms alive and make it more effective. In this respect XP method and CMMI-Dev1.2 model have been selected in this study to develop a new framework for several reasons as discusses in Section 2.

This paper presented the results of the research stages used in developing the new framework (aligning XP method to CMMI-Dev1.2 KPAs, developing the proposed SDPIF, verifying the proposed framework, and validating the verified framework)

Referring to the results of the first approach of validating the modified framework, all the software, development, and improvement practices that are used in developing the modified framework to achieve the twenty one KPAs of CMMI-Dev1.2 were suitable for SSDFs. Furthermore, based on the responses of the team members of the two case studies on the evolution criteria questions at the second validation approach; it can be concluded that the modified framework is useful, useable, satisfy user needs and valid for use by SSDFs.

There are several potential directions for extracting or complementing this research, such as: fulfilling the missing KPAs and specific practices of several KPAs; using more agile practices; and applying the framework to case studies of larger scope.

References

[1] A. Baruah, “Contribution of Software Process Improvement Approaches For Small and Medium Scale Enterprises”, International Journal of Computing and Corporate Research, Vol. 2. No.2, 2012, pp.1-10.

[2] P. Savolainen, H. M. Sihvonen, and J. Ahonen, "SPI with lightweight software process modeling in a small software company”, Lecture Notes in Computer Science, 4764, 2007, pp. 71-81.

[3] G. Valdes, M. Visconti, and H. Astudillo, “The Tutelkan Reference Process: A Reusable Process Model for Enabling SPI in Small Settings”, in Systems, Software and Service Process Improvement: 18th European conference (EuroSPI 2011), 2011, pp. 179-190, Roskilde, Denmark, 2011.

[4] I. Garcia, C. Pacheco, and J. Calvo-Manzano, "Using a web-based tool to define and implement software process improvement initiatives in a small industrial setting", Software, IET, vol. 4, NO 4, 2010, pp. 237-251.

[5] H. Altarawneh, and S. Amro, “Software Process Improvement In Small Jordanian Software Development Firms”, in the 7th International Conference on Perspectives in Business Informatics Research (BIR’2008), 2008, pp. 175-189.

[6] D. Mishra, and A. Mishra, “Software process improvement in SMEs: A comparative view”, Computer Science and Information Systems, Vol. 6, No. 1, 2009, pp. 111-140.

[7] A. B. M. Isawi, “Software Development Process Improvement for Small Palestinian Software Development”, M.S. thesis, Faculty of Graduate Studies, An-Najah National University, Nablus, Palestine, 2011.

[8] M. Y. Al-tarawneh,, “Harmonizing CMMI-Dev1.2 and XP Method to Improve the Software Development Process in Small Sofware Development Firms”, Ph.D. thesis, School of Computing, Universiti Utara Malaysia, Sintok, Malaysia, 2012.

[9] M. Pikkarainen, and V. T. tutkimuskeskus, “Towards a framework for improving software development process mediated with CMMI goals and agile practices, Academic Dissertation, Faculty of Science, Department of Information Processing Science, University of Oulu, Finland, 2008.

[10] Z. Lina, and S. Dan, “Research on Combining Scrum with CMMI in Small and Medium Organizations”, in the International Conference on Computer Science and Electronics Engineering (ICCSEE), 2012, pp. 554-557.

[11] I. Richardson, "SPI models: what characteristics are required for small software development companies?", Software Quality, Vol. 10, No. 2, 2002, pp. 100-113.

[12] D. Fernandeas, “Study on the correlation between CMMI and agile practices and their application in SMEs”, M.S. Thesis, Computer Faculty, university Polytechnic of Madrid, Spain, 2009.

[13] K. Beck, Extreme programming explained:embrace change: 3th End.Reading, Mass, addition-Wesley. Boston, 2000.

[14] J. A. H. Alegra and M. C. Bastarrica, "Implementing CMMI using a Combination of Agile Methods", CLEI ELECTRONIC JOURNAL, Vol. 9, No 1, 2006, pp. 1-15.

[15] M. Fritzsche, and P. Keil, “Agile Methods and CMMI: Compatibility or Conflict?,” e-Informatica Software Engineering Journal, Vol. 1, No. 1, 2007, pp. 9-26.

[16] C. P. Team, “CMMI for Development (CMMI-DEV): Version 1.3”, Technical Report, CMU/SEI-2010-TR-033, Software Engineering Institute, 2010.

[17] T. Galinac, “Analysis of Quality Management in Modern European Software Development”, Electronic form only: NE Eng. Rev, Vol. 28, No. 2, 2008, pp. 65-76.

[18] D. Goldenson, and D. Gibson, “Demonstrating the impact and benefits of CMMI: an update and preliminary results”, CMU/SEI-2003-SR-009, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, 2003.

[19] G. V. Boas, A. R. C. da Rocha, and M. Pecegueiro do Amaral, "An Approach to Implement Software Process Improvement in Small and Mid Sized Organizations", in the Seventh International Conference on the Quality of Information and Communications Technology, 2010, pp. 447-452.

[20] H. Mehrfard, H. Pirzadeh, and A. Hamou-Lhadj, “Investigating the Capability of Agile Processes to Support Life-Science Regulations: The Case of XP and FDA Regulations with a Focus on Human Factor Requirements”, Software Engineering Research, Management and Applications, Vol. 296, 2010, 241-255.

[21] R. Ramsin, “The engineering of an object-oriented software development methodology”, Ph.D Thesis, Department of Computer Science, university of York, UK, 2006.

[22] D. Vasiljevic, and S. Skoog, “A Software Process Improvement Framework for Small Organizations”, M.S. thesis. Department of Software Engineering and Computer Science Blekinge Institute of Technology, Sweden, 2003.

[23] J. Ralyté, R. Deneckère, and C. Rolland, “Towards a Generic Method for Situational Method Engineering”, in the 15th International Conference Advanced on Information Systems Engineering (CAiSE2003), 2003, pp. 95-110.

[24] T. Rout, (project manager), SPICE: Software Process Assessment-Part 1: Concepts and Introductory Guides, 2002. Downloadable from http://www.sqi.gu.edu.au/ spice/suite/download.html.

[25] S. Birisci, M. Mentin, and M. "Karakas, Prospective Elementary Teacher’s Attitudes Toward computer and internet use: A Sample from Turkey", World Applied Sciences Journal, Vol.6, No. 10, 2009, pp. 1433-1440.

[26] B. A. Kitchenham, "Evaluating software engineering methods and tool part 1: The evaluation context and evaluation methods", ACM SIGSOFT Software Engineering Notes, Vol. 23, No. 5, 1998, pp. 11-14.

[27] E. J. Garrity and G. L. Sanders, "Dimensions of information systems success," Information systems success measurement, Idea Group Publishing (IGP), Hershey, pp. 13-45, 1998.

[28] M. Easterby-Smith, R. Thorpe, and A. lowe, “Management Research: An Introduction,” Sage Publications Ltd, London, 1991.

Empirical Studies on Methods of Crawling Directed Networks

Junjie Tong, Haihong E , Meina Song and Junde Song
School of Computer, Beijing University of Posts and Telecommunications
Beijing, 100876, P. R. China

Abstract

Online Social Network has attracted lots of academies and industries to look into its characteristics, models and applications. There are many methods for crawling or sampling in networks, especially for the undirected networks. We focus on sampling the directed networks and intend to compare the efficiency, the accuracy and the stability between them. We consider the sampled nodes and links as a whole and separated from the original one. We evaluate experiments by deploying the snow ball method, the random walk method, DMHRW and MUSDSG with different sampling ratios on the datasets. The snow ball method and random walk method both have bias towards low outdegree nodes while the snow ball method tends to sample more hub nodes. DMHRW and MUSDSG can sample the network parallel but more complex than the snow ball and the random walk under the same sampling ratio. DMHRW will be the best choice of all while the computation capability and time are sufficient.

Keywords: *Sampling Method, Directed Networks, Measurements, Graph Sampling.*

1. Introduction

In recent years, the population of Online Social Networks (OSNs) has experienced an explosive increase. Twitter for example, has attracted more than 600 million individuals by August 2012 [1] counted by Twopcharts. The world-wide spreading of OSNs has motivated a large number of academies and researchers do studies and researches on the analysis and model on the structures and characteristics of OSNs. However, the complete dataset is typically not available for privacy and economic considerations at some extent. Therefore, a relative small but representative subset of the whole is desirable in order to study properties, characteristics and algorithms of these OSNs. How to get the relative small but representative subset of the whole accurately, efficiently and stably becomes an important problem.

Various graph sampling algorithms have been proposed for producing a representative subset of OSNs users. Currently, the algorithms for crawling OSNs can be roughly divided into two main categories: graph traversal based and random walk based. For the graph traversal based methods, each node in the connected component is visited only once, if we let the process run until completion. For the random walk based methods, they allow node re-visiting. BFS, in particular, is a basic technique that has been used extensively for sampling OSNs in the past studies [2, 3, 4]. And the comparison between the graph traversal based methods and the random walk based methods can be shown in Table 1. And we denote the graph traversal based methods as T and the random walk based methods as W in the method type column.

Table 1. Comparisons Between Main Methods

Method Name	*Method Type(T/W)*	*For Directed/Undirected Networks*	*Biased/Unbiased*
BFS	T	Both	Towards high degree nodes [5], underestimate the level of symmetry [6]
Snow-ball [7]	T	Both	Underestimate the power-law coefficient [5]
Random Walk [8]	W	Both	Towards high degree nodes
Metro-Hastings RW [9]	W	Undirected	Unbiased
Re-Weighted RW [8]	W	Undirected	Unbiased

There are many other random walk based methods in [8]. Although the random walk bases methods may be biased but the bias can be analyzed using classic results from Markov Chains and corrected by re-weighting the estimators [10].

Currently, there are a lot of works on new unbiased sampling method and bias analysis on existing sampling methods. The comparisons between BFS, Metro-Hastings RW and Re-Weighted RW in crawling undirected Facebook can be seen in [11, 12]. USDSG has been proposed for unbiased sampling in directed ONSs [13]. And the most widely used baseline sampling method is UNI which is usually called ground truth. This simple method is a textbook technique known as rejection sampling [14] and in general it allows to sample from any distribution of interest with some limitations [11].

For the efficient crawling of OSNs, sometimes we have to adopt parallel processing which not only benefits the efficient but also overcomes the limitations such as capabilities on computation and storage and API requesting times etc. For example, the snow ball sampling and random walk sampling often choose several initiative nodes to start. For implementing MHRW and RWRW, we have to consider the convergence more seriously. And the Geweke diagnostic [15] and the Gelman-Rubin diagnostic [16] are widely used.

The focus of our work is how to crawling the directed networks accurately, efficiently and stably. We describe and implement several crawling methods on crawling directed networks not only limited to online social networks under practical experiment. And compare the stability, efficiency and bias between them. Our main contributions are the followings but not limited:

- Modifying and implementing the crawling algorithms in directed networks. Most of the current methods especially the unbiased methods are for crawling on undirected networks. And we implement the algorithms in various directed networks not only to the directed OSNs.
- Crawling without knowing the whole set. Most implementations of current sampling methods treat sampled nodes separately and focus mainly on the degree distribution but ignore the topology characteristics such as clustering coefficient and diameter.
- Comparisons between several crawling algorithms in various aspects which include efficiency, accuracy and stability.

2. Crawling Methodology

2.1 Scope and Assumptions

The directed networks can be modeled as a directed graph G=(V,E) , where V is a set of nodes (users) and E is a set of edges (denote relationships of some type). Let k_v^{in} and k_v^{out} be the indegree and outdegree of node v. Let k^{in} denotes the average indegree and r denote the sample ratio which is set before starting the crawling. N denotes the node number of the whole set and M denotes the edge number of the whole set. In this paper:

- We begin our crawling at some initial nodes without knowing the whole set so when we analyze the characteristics of the sampled networks we just consider the crawled nodes and corresponding connections between them.
- The initial nodes are selected in the largest SCC (strongly connect components) of each dataset for fast crawling and feasible experiments.
- We do not consider the missing links and implicit links. The datasets which we used are considered fully collected. We just consider the static snapshot of each dataset and do not consider its dynamics.

2.2 Sampling Methods

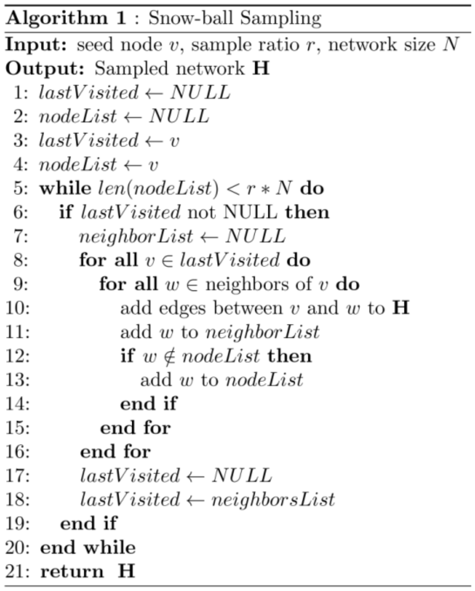
The crawling of the directed graph starts with one or several initial nodes and proceeds iteratively even in parallel. In every operation, we visit a node and discover all its neighbors. There are many ways, depending on the particular sampling method, in which we can proceed. In this section, we describe the crawling methods we implemented and compared in this paper.

1) Snow ball Sampling: Snow ball sampling typically selects one node as the seed node to start and processes iteratively. At each new iteration the neighbors of sampled nodes but not-yet-visited are explored. As BFS is widely used in graph traversal, snow ball sampling is an incomplete BFS to cover only some specific region of the graph. Sometimes, several seed nodes are selected for efficient crawling. The snow ball sampling method we adopted can be depicted as follows in Figure 1. And the above algorithm can be easily extended in choosing more than one seed node and processing parallel. The selection of the seed node will be described later in this section.
2) Random Walk: In the classic random walk [8], the next-hop node is chosen uniformly at random among the neighbors of the current node. And the classic random walk sampling is biased towards high degree nodes. The random walk smapling method we implemented as follows as in Figure 2.
3) Directed Metro-Hastings Random Walk (DMHRW): We modify the existing MHRW which is usually used in sampling undirected networks for crawling the directed networks. As the choice of the next-hop node depends on the ratio of indegree and outdegree of currentNode as depicted in Figure 3, it bias towards nodes with high indegree and low outdegree.

And While implementing this method, we use multiple parallel walks and we have to consider the covergence of the sampling process. We use the Geweke diagnostic to detect the convergence of the sampling process. The details of the Geweke diagnostic will be described later in this section.

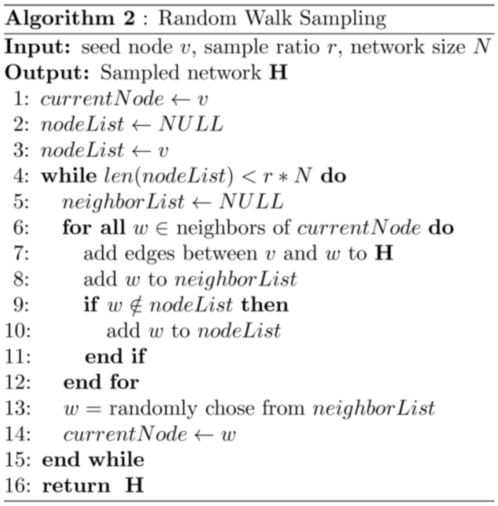
And as shown between line 8 and 9 in algorithm 3, the algorithm will exit while current node numbers exceed some point. As in the experiment, we find out that sometimes it will not converge at last while the node number reaches the defined one before starting sampling.

The algorithm can be divided into two parts as algorithm 3 and algorithm 4. And algorithm 4 in Figure 4 calls algorithm 3 and use the Geweke diagnositc after each iteration.



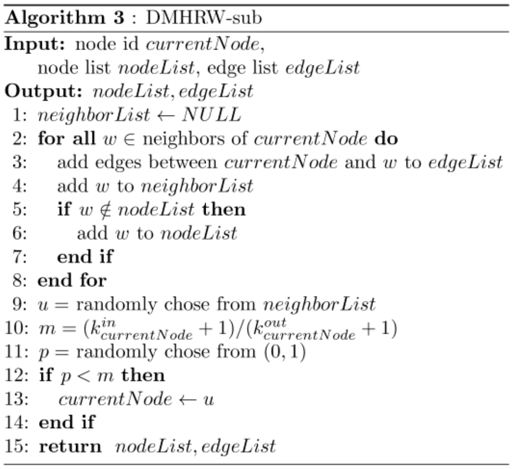
```
Algorithm 1 : Snow-ball Sampling
Input: seed node $v$, sample ratio $r$, network size $N$
Output: Sampled network $\mathbf{H}$
$lastVisited \leftarrow NULL$
$nodeList \leftarrow NULL$
$lastVisited \leftarrow v$
$nodeList \leftarrow v$
while $len(nodeList) < r * N$ do
  if $lastVisited$ not NULL then
    $neighborList \leftarrow NULL$
    for all $v \in lastVisited$ do
      for all $w \in$ neighbors of $v$ do
        add edges between $v$ and $w$ to $\mathbf{H}$
        add $w$ to $neighborList$
        if $w \notin nodeList$ then
          add $w$ to $nodeList$
        end if
      end for
    end for
    $lastVisited \leftarrow NULL$
    $lastVisited \leftarrow neighborsList$
  end if
end while
return $\mathbf{H}$
```

Fig. 1 Algorithm 1.



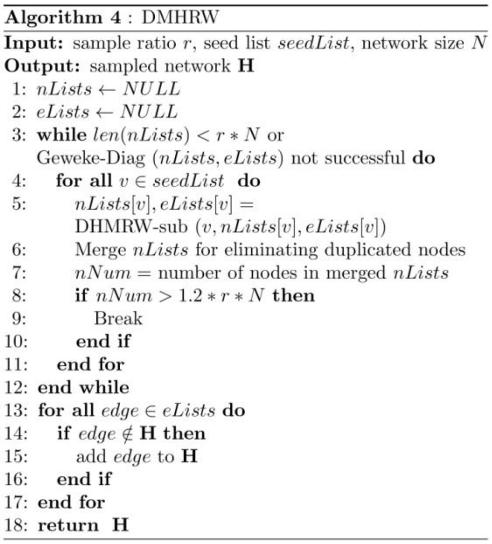
```
Algorithm 2 : Random Walk Sampling
Input: seed node $v$, sample ratio $r$, network size $N$
Output: Sampled network $\mathbf{H}$
$currentNode \leftarrow v$
$nodeList \leftarrow NULL$
$nodeList \leftarrow v$
while $len(nodeList) < r * N$ do
  $neighborList \leftarrow NULL$
  for all $w \in$ neighbors of $currentNode$ do
    add edges between $v$ and $w$ to $\mathbf{H}$
    add $w$ to $neighborList$
    if $w \notin nodeList$ then
      add $w$ to $nodeList$
    end if
  end for
  $w =$ randomly chose from $neighborList$
  $currentNode \leftarrow w$
end while
return $\mathbf{H}$
```

Fig. 2 Algorithm 2.



```
Algorithm 3 : DMHRW-sub
Input: node id $currentNode$,
    node list $nodeList$, edge list $edgeList$
Output: $nodeList, edgeList$
$neighborList \leftarrow NULL$
for all $w \in$ neighbors of $currentNode$ do
  add edges between $currentNode$ and $w$ to $edgeList$
  add $w$ to $neighborList$
  if $w \notin nodeList$ then
    add $w$ to $nodeList$
  end if
end for
$u =$ randomly chose from $neighborList$
$m = (k^{in}_{currentNode} + 1)/(k^{out}_{currentNode} + 1)$
$p =$ randomly chose from $(0, 1)$
if $p < m$ then
  $currentNode \leftarrow u$
end if
return $nodeList, edgeList$
```

Fig. 3 Algorithm 3.



```
Algorithm 4 : DMHRW
Input: sample ratio $r$, seed list $seedList$, network size $N$
Output: sampled network $\mathbf{H}$
$nLists \leftarrow NULL$
$eLists \leftarrow NULL$
while $len(nLists) < r * N$ or
Geweke-Diag $(nLists, eLists)$ not successful do
  for all $v \in seedList$ do
    $nLists[v], eLists[v] =$
    DHMRW-sub $(v, nLists[v], eLists[v])$
    Merge $nLists$ for eliminating duplicated nodes
    $nNum =$ number of nodes in merged $nLists$
    if $nNum > 1.2 * r * N$ then
      Break
    end if
  end for
end while
for all $edge \in eLists$ do
  if $edge \notin \mathbf{H}$ then
    add $edge$ to $\mathbf{H}$
  end if
end for
return $\mathbf{H}$
```

Fig. 4 Algorithm 4.

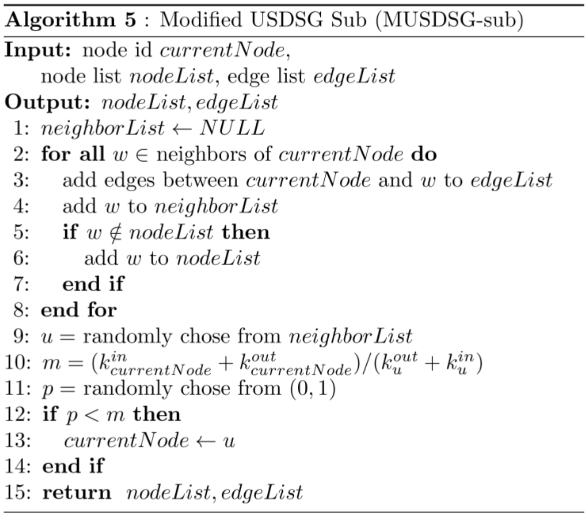
4) Modified Unbiased Sampling for Directed Social Graphs (MUSDSG): We have modified USDSG as the followings: using multiple parallel walks and the Geweke diagnostic to detect the sampling process, adding the neighbors of currentNode to sampled network and setting the upper bound of the sampled node number while the iteration processing is not converged. As we consider the topology of sampled network, we can compare topology in sampled network not just the degree of node and degree distribution which are already compared in [13]. The algorithm is depicted in Figure 5.

5) Modified Uniform Sampling (MUNI): The UNI [12] method allows us to obtain uniformly random users' ID by generating random IDs in certain space. This

algorithm has limitations. First, the ID space must not be sparse for this operation to be efficient. Secondly, the operation which enables us to verify the user and retrieve the users whom he connects to can be easily implemented. As we have obtained the whole set and the IDs are integers without interval numbers, the ID space is not sparse and we can easily obtain the connections between any two users. We modify UNI as MUNI and add the connections between sampled nodes to form the sampled network. And the algorithm we implemented can be shown as follows in Figure 6.

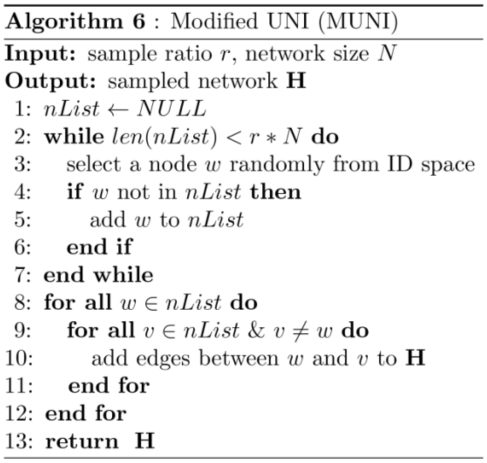
6) The Selection of the Initial or Seed Nodes: All the above sampling or crawling methods have to select one or more than one seed nodes. While we implement snow ball sampling and random walk sampling, we select one seed node randomly from the largest SCC separately. And while we implement DMHRW and MUSDSG, we select several seed nodes randomly from the largest SCC without multiple ones. Although the seed nodes in largest SCC will facilitate the crawling process, the diameter will be underestimated. As the small-world effect [17] leads to the small diameter, this will not affect much.

7) Detecting Convergence with Geweke Diagnostic: While we implement DMHRW and MUSDSG using multiple parallel walks, we have to detect convergence. The Geweke diagnostic detects the convergence of a single Markov chain. Let X be a single sequence of samples of our metric of interest. Geweke considers two subsequences of X, its beginning X_a (typically the first 10%), and its end X_b (typically the last 50%). Based on X_a and X_b, we compute the z-statistic: $z = (E(X_a) - E(X_b))/\sqrt{Var(X_a) + Var(X_b)}$. With increasing number of iterations, X_a and X_b move futher apart, which limits the correlation between them. And we declare rigid convergence when all the values of sequences fall in the [-1,1] interval.



```
Algorithm 5 : Modified USDSG Sub (MUSDSG-sub)
Input: node id $currentNode$,
    node list $nodeList$, edge list $edgeList$
Output: $nodeList, edgeList$
$neighborList \leftarrow NULL$
for all $w \in$ neighbors of $currentNode$ do
    add edges between $currentNode$ and $w$ to $edgeList$
    add $w$ to $neighborList$
    if $w \notin nodeList$ then
        add $w$ to $nodeList$
    end if
end for
$u$ = randomly chose from $neighborList$
$m = (k^{in}_{currentNode} + k^{out}_{currentNode})/(k^{out}_u + k^{in}_u)$
$p$ = randomly chose from $(0, 1)$
if $p < m$ then
    $currentNode \leftarrow u$
end if
return $nodeList, edgeList$
```

Fig. 5 Algorithm 5.



```
Algorithm 6 : Modified UNI (MUNI)
Input: sample ratio $r$, network size $N$
Output: sampled network $\mathbf{H}$
$nList \leftarrow NULL$
while $len(nList) < r * N$ do
    select a node $w$ randomly from ID space
    if $w$ not in $nList$ then
        add $w$ to $nList$
    end if
end while
for all $w \in nList$ do
    for all $v \in nList$ & $v \neq w$ do
        add edges between $w$ and $v$ to $\mathbf{H}$
    end for
end for
return $\mathbf{H}$
```

Fig. 6 Algorithm 6.

3. Datasets and Evaluation Methodology

This section contains two main parts. First, we describe our datasets and measure their topology characteristics. Second, we give details on how to compare or evaluate the different crawling methods.

3.1 The Datasets

We download four datasets from [18] including: Gnutella peer-to-peer network of August 2002 (gpn08 [19]), EU email communication network (eue [19]), Slashdot social network of November 2008 (slashdot01 [20]) and Slashdot social network of February 21 2009 (slashdot02 [20]).

1) Explanations on Measured Characteristics: We have measured many topological characteristics of the networks in the datasets including the diameter, the correlation coefficient between indegree and outdegree r_0 [21], the coefficient of link reciprocity r_1[22, 23], the assortative coefficient r_2 [24], The average clustering coefficent in directed networks c. And r_2 falls in [-1,1]. An interesting observation is that essentially all social networks measured appear to be assortative, but other types of networks (information networks, technological networks, biological networks) appear to be disassortative [25]. And more discussions about clustering coefficent in directed networks are in [26].

2) Measurements on the Datasets: The characteristics are measured with NetworkX [27]. We use it to calculate the diameter and store the network. The characteristics of the datasets are shown in Table 2.

Table 2. Characteristics of the Networks in the Datasets

Network	gpn08	eue	slashdot01	slashdot02
N	6301	265214	77360	82168
M	20777	420045	905468	948464
k^{in}	3.297	1.5838	11.7046	11.543
d	9	14	12	13
effective diameter	6	5	5	5
c	0.015	0.4913	0.0228	0.0164
r_0	1.375	76.076	10.553	10.698
r_1	-0.00052	0.131	0.4391	0.427
r_2	0.194	-0.071	0.0072	0.0018

3.2 Evaluation Methodology

We compare the crawling methods described in section II in three aspects including efficiency, accuracy and stability. Before the description, we first define some parameters.

Each algorithm runs for Num times under certain sampling ratios R. To reach certain sampling ratio $r_i \in R$, each algorithm has to operate literately in t_j times in the jth ($j \leq Num$) sampling. And in num sampling times under certain sampling ratio r_i, the crawling process fails $fail_i$ times by using DMHRW and MUSDSG. While the sampled number exceeds certain upper bound without convergence, we consider it fails. And $c_j, d_j\ and\ k_j^{in}$ denote the average clustering coefficient, the diameter and the average indegree of the jth sampled network.

1) *Efficiency:* to evaluate the efficiency of the crawling method, we propose two parameters. One is average sampling time under certain sampling ratio defined as $t_{r_i} = \sum_{j=1}^{Num} t_j/Num$. And the other one is the successful ratio under certain sampling ratio which defined as $s_{r_i} = (Num - fail_i)/Num$. And the second one is only for DMHRW and MUSDSG.

2) *Accuracy:* to consider the accuracy of the crawling algorithm, we compare the node indegree and outdegree distribution, the average clustering coefficient, the diameter and the average node indegree of sampled networks by implementing the above algorithms and MUNI separately under certain sampling ratio.

3) *Stability:* we have to consider the expectation and standard deviation of the characteristics under certain sampling ratio during the separately Num sampling times by implementing the same algorithm. And we also consider the degree distribution seriously.

4. The Experiment and Analysis

In this section, we will compare snow ball sampling, random walk, DMHRW and MUSDSG in efficiency, accuracy and stability. And while we compare the characteristics of the sampled network, we consider MUNI as the ground truth. We evaluate the experiments of the different sampling methods on the datasets.

4.1 The Ground Truth: MUNI

First, we look into MUNI which is usually considered as the ground truth while comparing different crawling methods. We compare the characteristics under various sampling ratio of the four different directed networks to the whole set in two aspects: its stability and accuracy.

We run MUNI for 10 times under different sampling ratio for the four different datasets or directed networks. And Table 3 shows the expectation *E* and the stand deviation *D* of the characteristics under different sampling ratio for the four directed networks.

1) *Accuracy:* As shown in Table 3, we can easily compare the characteristics between the sampled networks and the whole network.
For the diameter, while the sample ratio becomes larger and larger, it gets closer and closer to the value of the whole network. And the different between the two is small. And it’s also the same as the average indegree.
But for the clustering coefficient, for the two online social networks (slashdot01 and slashdot02), the value of the

sampled networks is nearly 200 times over the value of the whole network. It turns out that for the density directed network, MUNI is more likely towards the nodes with links to the connected nodes. And it is somehow like the proximity bias of link growth in Flickr [28] and FriendFeed [29].

As shown in Figure 7, the cumulative distribution of indegree and outdegree of the sampled network are totally fit for corresponding distributions of the whole network for slashdot02.

2) Stability: We run MUNI for 10 times under different sampling ratio for the four different dataset. And we can easily figure out from Table III that the standard deviations are small under different sampling ratio. And the standard deviations are often ten percent as the expectations.

The larger the sample ratio is, the closer the expectation is to the real value. But while the ratio is small, the MUNI sampling method may cause significant biases in characteristics such as the clustering coefficient of slashdot01 and slashdot02.

4.2 The Efficiency Comparison

The average sampling time under various sampling ratio are given in Figure 12. From that figure, we can easily figure out that under the same sample ratio, the snow ball method is the fastest but with extra store. And the random walk method is the slowest method. Both the snow ball method and the random walk method cannot be deployed parallel.

For the parallel methods: DMHRW and MUSDSG, DMHRW is usually faster than MUSDSG. But their successful ratios are different. We run DMHRW and MUSDSG for sampling gpn08 three times independently and show the results in Table 4. The successful ratio is low for both DMHRW and MUSDSG and the two methods cost much as parallel deployment. In Figure 8, we plot the convergence with Geweke Diagnostic in the only one successful time while we sampling gpn08 with the sample ratio is 0.4.

4.3 The Accuracy Comparison

From Figure 9, we can easily figure out that the cumulative indegree distributions are similar while using random walk method and snow ball sampling. And they are all close to the distribution of the whole network. But for the cumulative distributions of outdegree, significant bias towards the low outdegree nodes exists by using the random walk method. And from Figure 9, the bias reduces as the increasing of the sample ratio. And it gets closer and closer to the cumulative distribution of the whole network as shown in Figure 10.

After implementing DMHRW and MUSDSG, only one time successful in each one. And Figure 13 shows the CDD of indegree and outdegree of the sampled network in the successful deployment of DMHRW and MUSDSG. Significant bias towards low outdegree nodes exists in both of them.

From Table 5, the data are expectation and standard deviation about the diameter and clustering coefficient under different sample ratios in different datasets. The networks have smaller diameter and larger indegree by using snow ball method other than random walk method no matter to the sample ratio and the network. As the random walk method is likely towards the nodes with low indegree and less connections with other nodes, the diameter and the clustering coefficient is much smaller.

And from Table 4, although the successful ratio is low for both DMHRW and MUSDSG, the diameter and the clustering coefficient are closer to the whole network than snow ball method and random walk method.

4.3 The Stability Comparison

From Table 5, the standard deviation is only about ten percent of the corresponding expectation. For random walk method, the standard deviations decrease as the increasing of the sample ratio. And the deviations are smallest by using DMHRW.

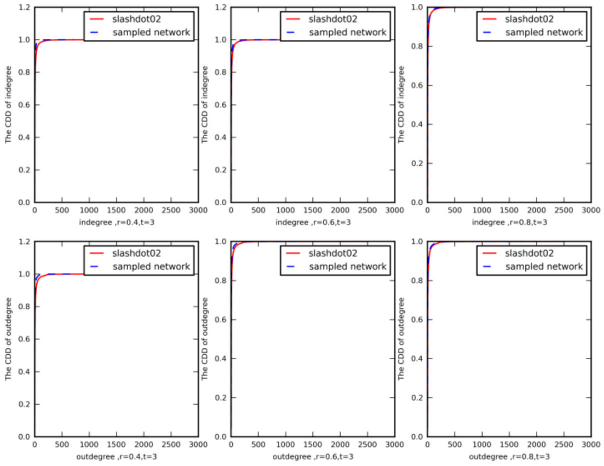


Fig. 7 The CDD of indegree and outdegree of slashdot02 by using MUNI

Table 3. Expectation and Standard Deviation of the Characteristics (MUNI)

Network	r	$E(d)$	$D(d)$	$E(c)$	$D(c)$	$E(k^{in})$	$D(k^{in})$
gpn08	0.4	13.9	0.539	0.00845	0.00176	1.316	0.0449
	0.6	11	0.775	0.009	0.00105	1.979	0.0423
	0.8	10.1	0.7	0.0102	0.000374	2.62	0.0571
	1.0	9	-----	0.015	-----	3.297	-----
slashdot01	0.4	11.7	0.781	2.455	0.0172	5.265	0.157
	0.6	12.4	0.8	2.106	0.0146	7.437	0.132
	0.8	11.6	0.49	1.846	0.0108	9.595	0.130
	1.0	12	-----	0.0228	-----	11.7046	-----
slashdot02	0.4	12.2	0.748	2.302	0.01	5.21	0.110
	0.6	12.2	0.6	1.99	0.0135	7.25	0.135
	0.8	12.2	0.748	1.74	0.00567	9.43	0.0667
	1.0	13	-----	0.0164	-----	11.543	-----
eue	0.4	11.7	0.64	0.215	0.0113	0.616	0.0396
	0.6	13.7	2.002	0.308	0.0114	0.945	0.0241
	0.8	13.75	1.199	0.40	0.00654	1.25	0.0154
	1.0	14	-----	0.491	-----	1.5838	-----

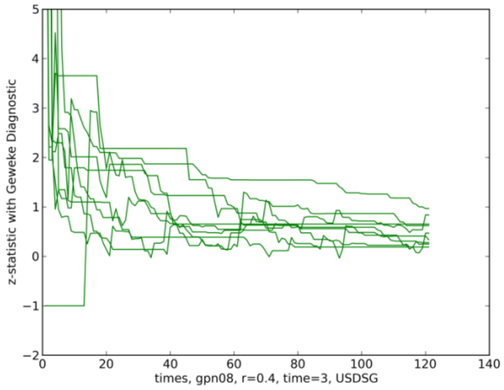


Fig. 8 The convergence with Geweke Diagnostic (MUSDSG)

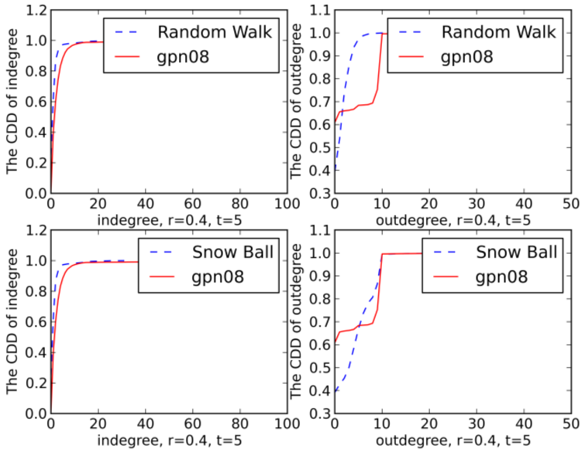


Fig. 9 The CDD of indegree and outdegree(r=0.4, t=5)

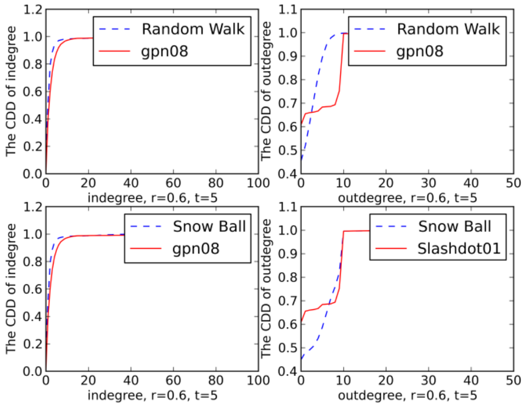


Fig. 10. The CDD of indegree and outdegree(r=0.6, t=5)

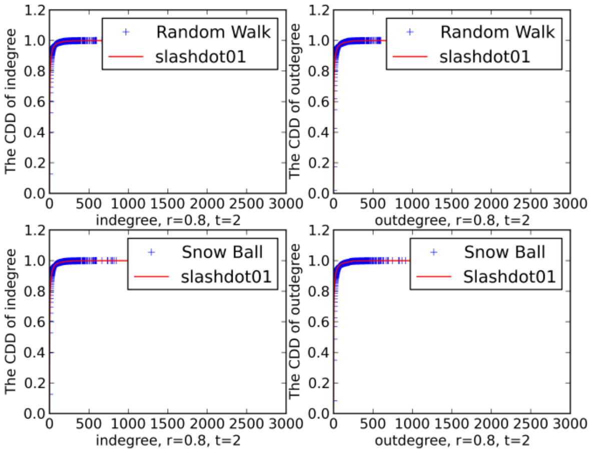


Fig. 11. The CDD of indegree and outdegree for slashdot02

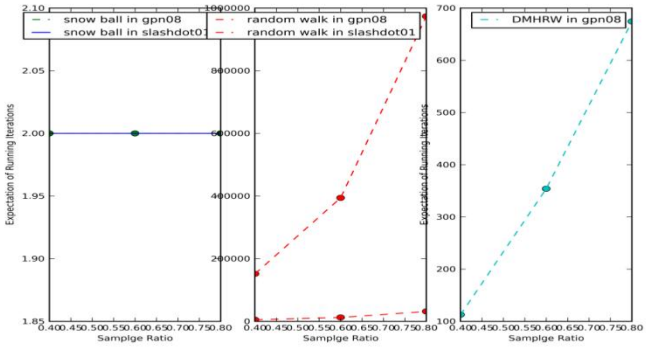


Fig 12. The sampling time under various sampling ratio

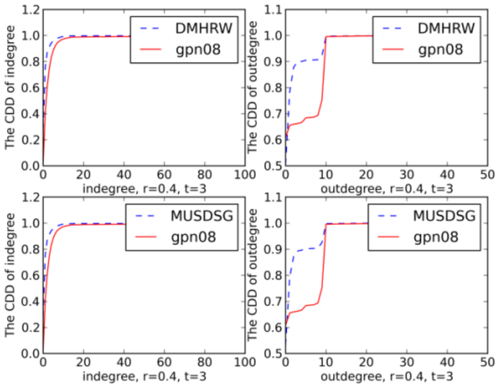


Fig. 13. The CDD of inderee and outdegree(DMHRW and MUSDSG)

Table 4. The successful ratio for DMHRW and MUSDSG in gpn08

r	**Sample Method**	S_r
0.4	DMHRW	33.3%
	MUSDSG	0
0.6	DMHRW	33.3%
	MUSDSG	0

5. Conclusions

In this paper, we have to explore into the sampling or crawling methods on directed networks. And we aim to give some insights into the methods in efficiency, accuracy and stability.

Despite the size of the original network and the difference of the sampling ratio, the MUNI performs well in accuracy and stability compared to different sampling methods. But it requires some conditions, such as getting the whole information of the user, the whole network is not sparse and etc.

The random walk method and snow ball method both have bias towards low outdegree nodes and increasing the sampling ratio can reduce the bias. As the increase of the sampling ratio, the sampling time of random walk increase much faster than the snow ball method while the snow ball method just needs several iterations.

Table 5. Expectation and Standard Deviation of the Characteristics under Different Sampling Ratio

Network	**r**	**Sample Method**	**E(d)**	**D(d)**	**E(c)**	**D(c)**	**E(k^{in})**	**D(k^{in})**
gpn08	0.4	Snow Ball	7	0	0.0224	0.0019	4.049	0.1824
		Random Walk	16	0.67	0.00466	0.00135	1.589	0.0187
		DMHRW	10	0	0.01	0.0019	1.59	0.046
		MUSDSG	9	0	0.0134	0.004	1.745	0.18
	0.6	Snow Ball	7	0	0.0209	0.0139	4.109	0.0278
		Random Walk	12	0	0.007	0.00125	2.138	0.0325
		DMHRW	9	0	0.01	0.001	2.37	0.0048
		MUSDSG	9	0	0.0132	0.0018	2.487	0.047
	0.8	Snow Ball	8	0	0.0134	0.00024	3.76	0.0113
		Random Walk	9	0	0.01	0.0006	2.923	0.0287
	1.0	-----	9	-----	0.015	-----	3.297	-----

MHRW and USDSG can be both deployed parallel and we modified both MHRW and USDSG for investigating into the unbiased sampling method in directed networks. Both the DMHRW and MUSDSG are more complex for computation and we have to detect the convergence while deploying the latter two. Although the successful ratio of DMHRW and MUSDSG is low, the characteristics of the sampled network are closer to the whole network than the random walk method and snow ball method.

For the lack of computation capability and the limits of the time, we suggest use the snow ball and the random walk methods. And sample the network with high sample ratio is much better for inducing the biases. If the computation capability and the time are enough, the DMHRW will be the better choice than MUSDSG with the characteristics closer to the whole network.

While comparing the sampling methods, we assume we know all the nodes and the links explicitly and the network is static. The network may have implicit links and nodes, and it is often dynamic [30]. Our future work will focus on completing the network using its growth mechanisms and the previous characteristics.

Appendix

Proposition I: For a given directed network $G=(V,E)$ with no loops and no self-loops, and convert it to an undirected network G'as the following way: if there is an edge between u and v and the edge (u,v) not in G', then add the edge (u,v) to G'. Then at last it forms an undirected network without loops and $d_{G'} \geq d_G$.

Proof 1: Suppose $d_{G'} < d_G$ and the length of the shortest path between u and v is equal to $d_{G'}$ inG'. As it has supposed that $d_{G'} < d_G$, then in G there is much shorter path ℓ between u and v. While converting G to G', the path is obviously added to G. And that means $d_{G'} < d_{G'}$, it is obviously not right. So our suppose is wrong and $d_{G'} \geq d_G$.

Acknowledgments

This work is supported by the National Kay project of Scientific and Technical Supporting Programs of China (Grant No. 2009BAH39B03); the National Natural Science Foundation of China (Grant No. 61072060); the National High Technology Research and Development Program of China Grant No. 2011AA100706); the Research Fund for the Doctoral Program of Higher Education (Grant No. 20110005120007); the Fundamental Research Funds for the Central Universities (2012RC0205); the Co-construction Program with Beijing Municipal Commission of Education; Engineering Research Center of Information Networks, Ministry of Education..

References

[1] http://twopcharts.com/twitter500million.php 2012-8-19.

[2] A. Mislove, M. Marcon, P. K. Gummadi and P. Druschel et al., Measurement and Analysis of Online Social Networks, IMC, 2007, pp. 29-42.

[3] Y. Ahn, S. Han, H. Kwak and S. Moon et al., Analysis of Topological Characteristics of Huge Online Social Networking Seervices, WWW, 2007, pp. 835-844.

[4] C. Wilson, B. Boe, A. Sala and K. P. Puttaswamy, User Interactions in Social Networks and Their Implications, EUROSYS, 2009, pp. 205-218.

[5] S. H. Lee, P. J. Kim, H. Jeong, Statistical Propertities of Sampled Netowrks, PHYS REV E, 2006, vol. 73, no. 1.

[6] L. Becchetti, C. Castillo, D. Donato and A. Fazzone, A Comparison of Sampling Techniques for Web Graph Characterization, LinkKDD, 2006.

[7] S. K. Thomson, Sampling, John Wiley & Sons, Inc., New York, 2002.

[8] L. Lovasz, Random Walks on Graphs: A Survey, Journal of Combinatorica, 1996.

[9] N. Metropolis, A. Rosenbluth, M. Rosenbluth and A. Teller et al., Equation of State Calculations by Fast Computing Machines, J. Chemical Physics, vol. 21, 2004, pp. 1087-1092.

[10] A. H. Rasti, M. Torkjazi, R. Rejaie and N. G. Duffield, Respondent-Driven Sampling for Characterizing Unstructured Overlays, INFOCOM, 2009, pp. 2701-2705.

[11] M. Gjoka, M. Kurant, C. Butts and A. P. Markopoulou, Walking in Facebook: A Case Study of Unbiased Sampling of OSNs, INFOCOM, 2010, pp. 2498-2506.

[12] M. Gjoka, M. Kurant, C. Butts and A. Markopoulou, Practical Recommendations on Crawling Online Social Networks, Journal on Selected Areas in Communications, vol. 29, no. 9, 2011, pp. 1872-1892.

[13] T. Y. Wang, Y. Chen, Z. B. Zhang and P. Sun et al., Unbiased Sampling in Directed Social Graph, ACM SIGCOMM, 2010, pp. 401-402.

[14] A. Leo-Garcia, Probability and Random Processes for Electrical Enginerring, 2nd ed, MA: Addison-Wesley Publishing Company, Inc., 1994.

[15] J. Geweke, Evaluating the Accuracy of Sampling-based Approaches to Calculating Posterior Moments, in Baysesian Statist. 4, 1992.

[16] A. Gelman and D. Rubin, Inference From Iterative Simulation Using Multiple Sequences, Statist. Sci., vol. 7, 1992.

[17] C. Karte, S. Milgram, Acquaintance Linking between Whie and Negro Populations: Application of the Samll World Problem, Journal of Personality and Social Psychology, vol. 15, no. 2, 1970, pp. 101-108.

[18] http://snap.stanford.edu/data/index.html, 2012-8-19

[19] J. Leskovec, J. Kleinberg and C. Faloutsos, Graph Evolution: Densification and Shrinking Diameters, ACM TKDD, vol. 1, no. 1, 2007, Articale 2.

[20] J. Leskovec, K. J. Lang, A. Dasgupta and M. W. Mahoney, Community Structure in Large Networks: Natural Cluster Size and the Absence of Large Well-Defined Clusters, Internet Mathematics, vol. 6, no. 1, 2009, pp. 29-123.

[21] R. Albert, A. L. Barabasi, Statistical Mechanics of Complex Networks, Reviews of Modern Physics, vol. 74, no. 1, 2002, pp. 47-97.

[22] D. Garlaschelli, M. I. Loffredo, Patterns of Link Reciprocity in Directed Networks, Physical Review Letters, vol. 93, no. 26, 2004, 268701.

[23] F. Zhao, T. Zhou, L. Zhang and M. H. Ma et al., Research Progress on Wikipedia, J. Univ. Electron. Sci. Technol., vol. 39, no. 3, 2010, pp. 321-334.

[24] M. E. J. Newman, Mixing Patterns in Networks, Phys. Rev. E, vol. 67, no. 2, 2003, 026126.

[25] M. E. J. Newman, The Structure and Function of Complex Network, SIAM Review, vol. 45, no. 2, 2003, pp. 167-256.

[26] G. Fagiolo, P. Martiri, Clustering in Complex Directed Networks, Physical Review E, vol. 76, no. 2, 2007, 026107.

[27] http://netowrkx.lanl.gov/#.

[28] A. Mislove, M. Marcon, P. K. Gummadi and P. Druschel et al., Growth of the Flickr Social Network, IMC, 2007, pp. 29-42.

[29] S. Garg, T. Gupta, N. Carlsson and Anirban Mahanti, Evolution of an Online Social Aggregation Network: An Empirical Study, IMC, 2009, pp. 315-321.

[30] J. Kunegis, D. Fay and C. Bauckhage, Netowrk Growh and the Spectral Evolution Model, CIKM, 2010, pp. 739-748.

Junjie Tong received his bachelor degree in computer science from China University of Mining and Technology in Beijing, and now is a Ph.D. candidate at the Department of Computer Science and Technology of Beijing University of Posts and Telecommunication. His research interests include CDN and complex networks.

Haihong E is a Ph.D., Lecturer at the Department of Computer Science and Technology of Beijing University of Posts and Telecommunication. Her research interests include service sciences and engineering, service network, and trusted service.

Meina Song is a Ph.D. Associate Professor at the Department of Computer Science and Technology of Beijing University of Posts and Telecommunication. Her research interests include service methodology, service system architecture, service sciences and engineering.

Junde Song is a Ph.D. Professor, Doctoral Advisor at the Department of Computer Science and Technology of Beijing University of Post & Telecom. His research interests include parallel computing, service sciences and engineering.

Energy efficient routing in mobile ad-hoc networks for Healthcare Environments

Sohail Abid[1], Imran Shafi[2] and Shahid Abid[3]

**[1] Department of Computing and Technology
IQRA University, Islamabad, Pakistan**

**[2] Department of Computing and Technology
Abasyn University Islamabad campus, Pakistan.**

**[3] Foundation University Institute of Engineering
and Management Sciences, Pakistan**

Abstract

The modern and innovative medical applications based on wireless network are being developed in the commercial sectors as well as in research. The emerging wireless networks are rapidly becoming a fundamental part of medical solutions due to increasing accessibility for healthcare professionals/patients reducing healthcare costs. Discovering the routes among hosts that are energy efficient without compromise on smooth communication is desirable. This work investigates energy efficiency of some selected proactive and reactive routing protocols in wireless network for healthcare environments. After simulation and analysis we found that DSR is best energy efficient routing protocol among DSR, DSDV and AODV, because DSR has maximum remaining energy.

Keywords: *Energy Efficient Routing in healthcare environment, Energy Awareness, energy-efficiency, medical application, Simulation of Energy Efficient Routing Protocol in NS2.*

1. Introduction

Today Mobile Ad-hoc Network (MANET) is a rapidly growing technology, due to its unique nature of distributed resources and dynamic topology. The routing protocols in MANET have standards which controls the number of nodes that harmonies the way to route packets between all the mobile nodes in their networks. There are lots of challenges to MANET like security, network stability, performance, efficiency and energy efficiency etc. Now a day wireless MANET's are becoming very popular and many routing protocols have been suggested by researchers. We give brief introduction on applications of wireless networks in the medical field and discuss the issues and challenges regarding to performance and efficient use of energy. We are concerning with energy efficiency and select some well known energy efficient routing protocols and simulate these protocols in NS2 and analyze energy efficiency in different cases.

Most of the healthcare equipments which are equipped with Wireless technology need energy efficient routing. It is vital in healthcare environments because The Energy efficient routing protocols were introduced years ago. Today Energy efficient routing protocols are used in wireless MANET and WSN. A Wireless Mobile Ad-hoc Network (MANET) is a set of mobile nodes that are randomly and dynamically placed in such a way that the interconnections between hosts are proficiently changing on a regular basis. The energy efficient routing protocol is used to discover routes between hosts to smooth the progress of communication within the network and try to utilize minimum energy consumption. The primary objective of such a MANET energy efficient routing protocol is best, correct and efficient route establishment between a pair of nodes so that messages may be delivered in a timely manner and save energy. A minimum overhead and bandwidth consumption should be done in the creation of route and maintenance of route [1].

Now a day mobile ad hoc networks have focused much more attention to the convenience of building mobile wireless networks without any need for a pre-existing infrastructure. A mobile ad hoc network is a group of wireless mobile nodes which are capable and agree to establish relations, using without any centralize supervision and infrastructure [2]. Mobile ad-hoc networks provide an environment, in which each node acts as a router for example receives packets and forwards to the nearest node or next hop, in order to reach final destination through various hops.

Wireless technology is ideal for medical applications and equipment due to its everywhere accessibility and mobility. Most of the heavy and expensive machines have limited mobility. These machines are not being fully used for a long time. Wireless technologies provide new interfaces to these heavy machines and make them interrelate with any new machines and use it everywhere. Some wireless technologies which are used in medical field are CodeBlue, MobiHealth and project connect etc. [3], [4]. Different medical applications for hospital have been developed like hospital management, equipment management and patient management etc. due to use of these applications efficiency of hospitals is increasing. Wireless technology for healthcare applications and equipment is rapidly growing. Some of these wireless medical equipments which are being used in hospitals are ECG anywhere, Life Source Products, Health Trax, LifeSync Wireless ECG System etc. [4].

The wireless networking has a bright future in the field of medical applications. Now day access and cost reductions are two hottest issues in the field of medical or healthcare. Both of these ends are successfully achieved by wireless networking. Around the world medical care organizations are rapidly getting complex, especially in the United States. Nearly 98000 patients die every year due to preventable medical errors. Wireless Network provides tools that can help reduce such medical errors. In wireless applications which are used in medical field, one of the major issue is availability of power. It is being guaranteed that the routing protocol is energy efficient and show full performance in less energy environment.

1.1 Mobile Ad-Hoc Network Challenges

Mobile ad-hoc network uses a broad range of applications. In real world for the impact of these applications, we require professional, reliable and more efficient algorithms and protocols. The MANET faces lot of challenges, which understand carefully and unmistakably [5]. Some of these challenges are discussed below:

Quality of Service: It is also important factor that the packet of data which is sent by source reached at destination timely and reliably. The protocol quality of service (QoS) is very critical in some applications like audio, video streaming.

Scalability: In ad-hoc network it is most important when a network is extended or expanded the protocol handle it smoothly and reliably. The protocol is flexible enough to respond and operate with such large number of hosts.

Ad-hoc Deployment: In a particular area ad-hoc network deployment is different according to the application. In ad-hoc network hosts are randomly deployed in the region without any knowledge of topology and prior infrastructure. In this situation the distribution of nodes and identification of connectivity between nodes are depend upon nodes.

Fault-Tolerance: In unfriendly environment, a host may fail due to certain problems or lack of power or energy. If a host fails, it is the responsibility of the protocols to accommodate these changes in the network.

Physical Resource Constraints: Limited battery power is most important and challenging constraint forced on MANET network host. The power supply is determined by MANET host directly. The energy consumption is the main issue in MANET.

In this paper, we focus on the MANET energy-efficient routing techniques regarding the network protocols that have been developed in recent years.

2. Background

2.1Ad-Hoc Routing Protocol

Three different types of ad-hoc routing protocols describe in Fig. 1, details are as under.

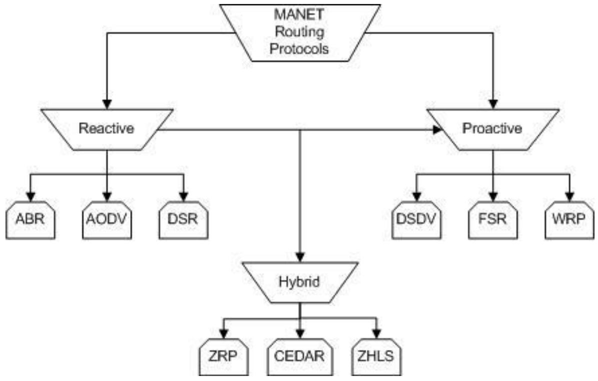


Fig. 1: Mobile ad-hoc routing protocols

7. Proactive Routing Protocol

In proactive routing protocols, first routing table is created and then data has to be sent from one host to another. These protocols regularly sends request to their neighbor nodes to organize their network topology and as well as make the routing table.

7. Reactive Routing Protocol

In this type of protocols when a host sends data to another host, it asks their neighbor nodes for a route. If neighbor nodes have no route, they broadcast the request to their neighbor nodes and so on.

7. Hybrid Routing Protocol

These protocols are the combination of reactive and proactive protocols. The important achievement in hybrid protocol is to minimize latency and broadcast. In hybrid protocols the network is divided into many small parts and each part has a gateway, inner part use reactive routing and inter-part use proactive routing.
In this paper DSDV is simulated and compared with reactive routing protocols AODV and DSR. The main goal of this task is to examine how to find different routing techniques impact the energy consumption in MANET. The summary of the three routing protocols are given below.

AODV

The process of rout discovery in AODV (Ad-hoc On-demand Distant Vector Routing) is that it initiate route discovery and initially the routing information of the table is empty. First of all a host or a node send / broadcast RREQ (route request) or acknowledgement (Ack) packet to all its neighbor nodes. The RREQ packet is a collection of broadcast ID, source address, source sequence no., destination address, destination sequence no. and total no. of hop count. AODV algorithm is appropriate for a dynamic user-starting network like when user wants to utilize ad-hoc network. AODV gives free of loop routing environment still when recovering broken links.

DSDV

DSDV (Destination-sequenced Distance Vector) was developed by C. Paerkins in 1994. Bellman-Ford algorithm is used in DSDV. This algorithm was developed for graph search applications and this is also used for routing purpose. Like every table-driven protocol, when the data has to be sent from source to destination, the DSDV minimize the latency by having a route. In the network each host maintain a routing table, the routing table count hops and tell how many hops arrived from source to destination. The routing table in DSDV protocol consists of Seq. no, hop count and route. The updated routing table is sent to each and every host of network in order to update table regularly.

DSR

DSR (Dynamic Source Routing) is working on the concept of source routing and on-demand routing protocol [6]. Each host is mandatory to maintain route caches which restrain the source routes that are acknowledged to it. The route caches are constantly updated by the host as it continues to find out new routes.

It has two major operations, first route discovery and second route maintenance. If a packet is sent to a destination node, the node will check with its own route cache for its obtainable route to the destination which is not expired. If a route present, the packet will be sent through the existing route. But, if a route does not present, it will send RReq (Route Request) to all host in the network. This is a point where route discovery phase start. The Route Request (RREQ) packet contain of unique identification number, source and destination address. When route request reached to a host, the host compare the new route to its own routing table, if does not match, it will add in routing table and forward the packet to it neighbor. This process will continue until all nodes in the network have a route to the destination.

3. RELATED WORK

Different routing protocols have been produced by the researchers with the help of simulation software. Some of them have also been used to minimize the energy consumption. L. M. Feeney presented in his paper a comparison of energy consumption for DSR, AODV in NS2 [7]. The analysis considers the cost for sending and receiving traffic, for dropped packets, and for routing overhead packets. Fr´ed´eric Giroire and his team present a link which connects the two routers. The two network interfaces join via this link [8]. Their goal is to find new routes that reduce links between source and destination while completing all requirements. Li Layuan, Li Chunlin, and Yuan Peiyan presents energy level based routing protocol "ELBRP" and compare with two other protocol RDRP and AODV [9]. Saoucene Mahfoudh and Pascale Minet enhanced OLSR to EOLSR by replacing multipoint relays (MPRs) with energy-aware multipoint relays (EMPRs) [10]. In this review paper Neeraj Tantubay, Dinesh Ratan Gautam and Mukesh Kumar Dhariwal present a summary of different energy control techniques and various powers saving methods have been proposed in different research articles [11]. Dr. S. P. Setty and B. Prasad (The author) compares QOS in energy consumption for proactive and reactive routing protocols with the impact of network size [12]. Ved Prakash, Brajesh Kumar and A. K. Srivastava analyze and compare energy efficiency of topology based and location based routing protocols [13]. Feeney L. M. divides the methods which are used in energy efficient awareness routing protocols in ad-hoc networks [14]. In first method when a host transmitting packets, the routing protocol minimized the total energy consumed during transmitting [15], [16], [17]. In second method load balance between hosts to increase the life time of whole network, instead of managing energy consumption for individual packet [18], [19], [20].

Nicolas Chevrollier and Nada Golmie investigate the impact of Bluetooth and wireless standard IEEE 802.15.4 in medical environment. Moreover, they find the importance of both technologies with respect to scalability issues [21].

4. SIMULATIONS

Case I: Methodology Used in Our Simulation-I

The Reference Point Group Mobility model (RPGM) has been used with node speed between 0.5 to 5.0 m/s, simulation time is 900 seconds, transport protocols is UDP and traffic generator source is CBR. The node density and simulation area varies from 20 to 80 nodes and 500mx500m to 2000mx2000m respectively. The initial energy of each node is 1000 joules and two-ray ground is used as propagation model. The other network parameter used in our simulation is described in table 1.

Table 1

Simulation I Parameters	
Parameters	Values
MAC Type	IEEE 802.11
Antenna	Omni directional
Simulation Time	900 sec
Transmission range	500 x 500 – 2000 x 2000
Node speed	0.5m/s to 5.0 m/s
Traffic Type	CBR
Data payload	512 bytes/ packet
Packet rate	8 packet/sec
Node Pause Time	0
Mobility Model	RPGM
Interface Queue Type	Drop Tail/Priori Queue
Interface Queue Length	50
No. of Nodes	20 to 80

In table 1 some parameters are constant and some are variable. These parameters varying during simulation to test and verify the results (simulation area, node pause time, mobility model and number of node).

Energy Consumption Model

There are four states of energy consumption of mobile devices which are given in table 2.

Table 2

Energy Consumption Parameters	
ei:	Energy Consumption during Idle mode
es:	Energy Consumption during Sleep mode
et:	Energy Consumed during Transmitting mode
er:	Energy Consumed during Receiving mode

The fifth parameter Energy consumed during forwarding mode is not used directly because during forwarding mode the host first received packets and send to the next hop. The two parameters er: and et: involved.

Case II: Methodology Used in Our Simulation-II

The energy model has presented by Santashil Pal Chaudhuri and David B. Johnson in [22], and Dr. S. P. Setty and B. Prasad used this energy model in their paper. The Random waypoint Mobility model has been used with node speed 1 to 10 m/s, simulation time is 300 seconds, transport protocols is UDP and traffic generator source is CBR. The node density varies from 5 to 25 nodes and simulation area 600x600. The initial energy of each node is 1000 joules and two-ray ground is used as propagation model. The other network parameter used in our simulation is described in table 3.

Table 3

Simulation II Parameters	
Parameters	Values
MAC Type	IEEE 802.11
Antenna	Omni directional
Simulation Time	300 sec
Transmission range	600 x 600 m
Node speed	1 m/s to 10 m/s
Traffic Type	CBR
Data payload	512 bytes/ packet
Packet rate	8 packet/sec
Node Pause Time	0
Mobility Model	Random Waypoint
Interface Queue Type	Drop Tail/Priori Queue
Interface Queue Length	50
No. of Nodes	5, 10, 15, 20, 25

We analyze the performance indexes and consumed energy depending on the following operations.

1. Consumed Energy in Rx Mode
2. Consumed Energy in Tx Mode
3. Consumed Energy in Idle Mode
4. Average Remaining Energy
5. Routing Overhead (RO)
6. Packet delivery Ratio / Function (PDR)

7. Average Throughput

Energy Consumption Model: The energy consumption model [22], [23], [24] describe total host energy spent in the following modes: (1) TX Mode (2) RX Mode (3) Idle Mode and (4) Overhearing Mode. These modes are describe as under

1. TX Mode

When a node send packet to other nodes, it is in TX mode. The energy required during transmit packet is called TX Energy [24], [11] of a node. TX Energy depends on packet size (in bits). TX energy can be described as follows.

$$TX = (\text{Pkt-size} \times 330) / 2 \times 10^6$$

And

$$P_{TX} = TX / T_{TX}$$

Where P_{TX} is transmitting power, TX is transmitting energy and T_{TX} is time take during packet transmit and Pkt-size is the size of packet in bits.

5. RX Mode

When a node receives packet from other nodes it is said to be in RX mode. The energy required during receiving packet is called RX energy [25], [26]. The RX energy can be formulated as

$$RX = (\text{Pkt-size} \times 230) / 2 \times 10^6$$

And

$$P_{RX} = RX / T_{RX}$$

Where P_{RX} is receiving power, RX is receiving energy and T_{RX} is time take during receiving a packet and Pkt-size is the size of packet in bits.

5. Idle/ Listening Mode

According to idle mode, the node does not send or receive any data packet. But in this mode energy consumed because the node continuously listening the wireless channel and ready to receive packet. When a packet is arrived and the node is converted from idle mode to RX mode. The power consumed in idle mode is as under.

$$P_{Idle} = P_{RX}$$

Where P_{RX} is power consumed in receiving mode and P_{Idle} is power consumed in idle mode.

5. Drop / Overhearing Mode

When a packet is receive by a node which is not design for this node it is called overhearing mode. The power consumed in overhearing mode is describe as under.

$$P_O = P_{RX}$$

Where P_O is power consumed in overhearing mode and P_{RX} is power consumed in receiving power.

5. Results

5.1 Simulation-I

The energy consumption in DSR, DSDV and AODV protocols are evaluated in term of average energy consumed. The node density varies from 20 nodes to 80 nodes.

1. Energy Consumption in Idle Mode:

According to the Fig 2, it is proved that energy consumption is maximum in DSR protocol, AODV protocol consumes medium energy and DSDV protocol consumes minimum energy in idle mode.

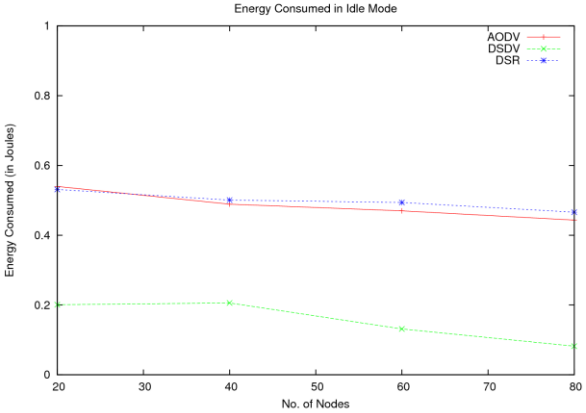


Fig 2: Comparison of average energy consumed in DSDV, DSR and AODV protocols in idle mode.

2. Energy Consumption in TX Mode:

According to the Fig 3, it is proved that energy consumption in AODV protocol is maximum DSR protocol consumes medium energy and DSDV protocol consumes minimum energy in TX mode.

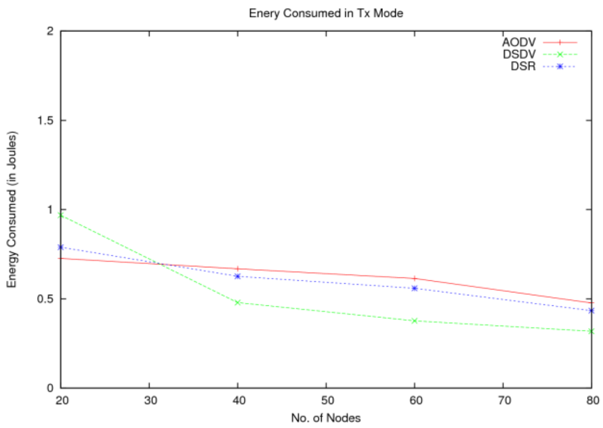


Fig 3: Comparison of average energy consumed in DSDV, DSR and AODV protocols in TX mode.

7. Energy Consumption in RX Mode:

According to the Fig 4, it is interesting that energy consumption of DSR in RX mode is less among all the three routing protocols. DSDV protocol consumes minimum energy in idle and TX mode and consumes maximum energy in RX mode. Because DSDV is a proactive protocol so it update table periodically.

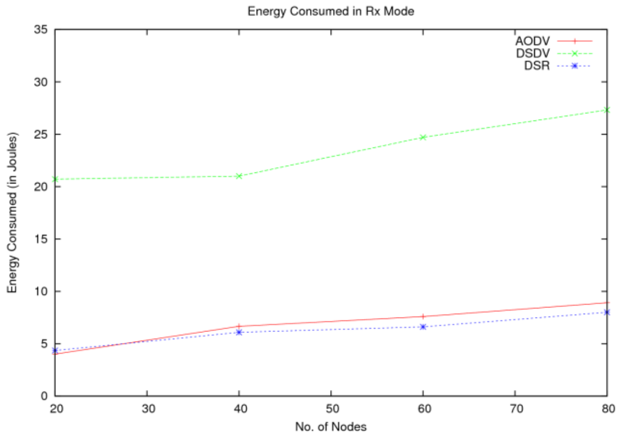


Fig 4: Comparison of average energy consumed in DSDV, DSR and AODV protocols in RX mode.

5. Average Remaining Energy:

According to the Fig 5, it is experimental prove that average remaining energy in DSDV protocol is minimum and medium in AODV and maximum in DSR protocol. It means that the performance of DSR is best in this scenario.

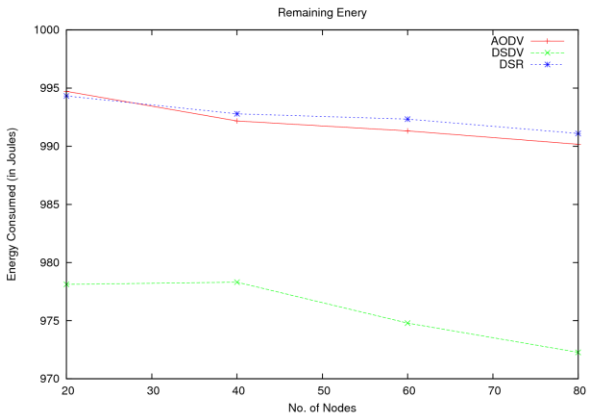


Fig 5: Average remaining energy.

5.2 Simulation-II

The energy consumption in DSR, DSDV and AODV protocols are evaluated in term of average energy consumed. The node density varies from 5 nodes to 25 nodes.

1. Energy Consumption in Idle Mode:

According to the Fig 6, it is experimental prove that energy consumption in DSDV protocol is maximum, AODV protocol consumes medium energy and DSR protocol consumes minimum energy in idle mode.

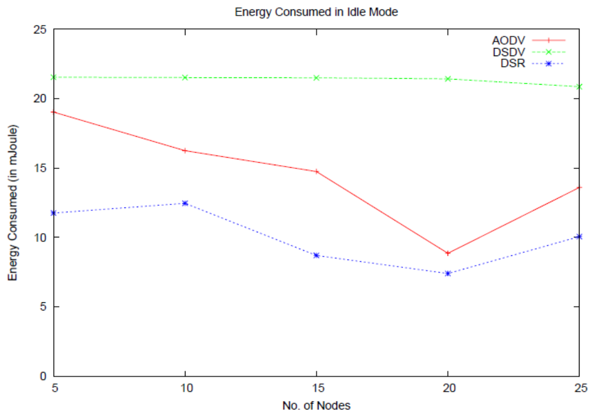


Fig 6: Comparison of average energy consumed in DSDV, DSR and AODV protocols in idle mode.

2. Energy Consumption in TX Mode:

According to the Fig 7, it is proved that energy consumption in AODV protocol is maximum DSR protocol consumes medium energy and DSDV protocol consumes minimum energy in TX mode.

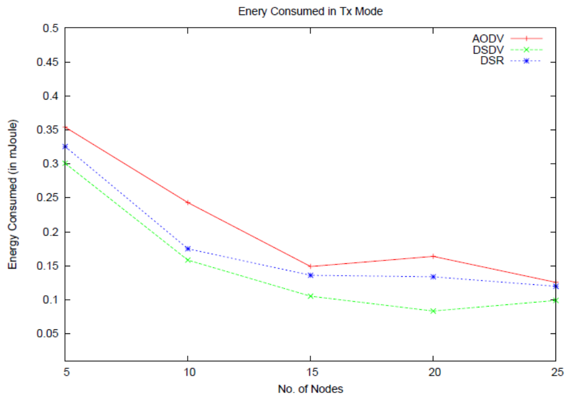


Fig 7: Comparison of average energy consumed in DSDV, DSR and AODV protocols in TX mode.

7. Energy Consumption in RX Mode:

The energy consumption of DSR in RX mode is less among all the three routing protocols in Fig 8. DSDV protocol consumes medium energy and AODV consumes maximum energy.

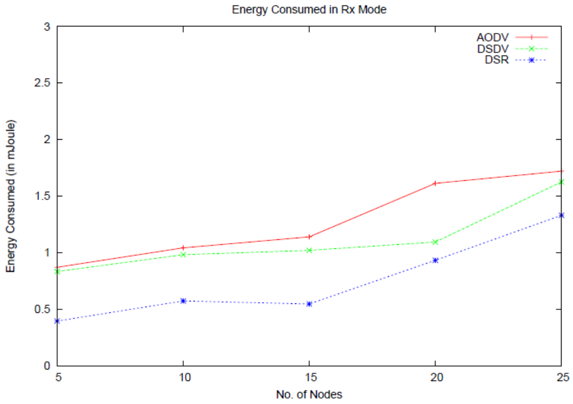


Fig 8: Comparison of average energy consumed in DSDV, DSR and AODV protocols in RX mode.

4. Average Remaining Energy:

According to the Fig 9, it is proved that average remaining energy in DSDV protocol is minimum and medium in AODV and maximum in DSR protocol. It means that the performance of DSR is best in this scenario.

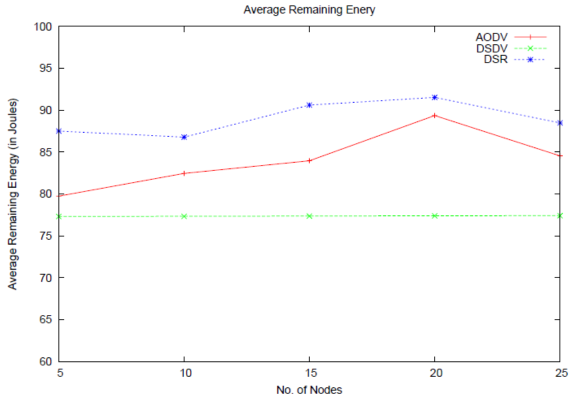


Fig 9: Average remaining energy.

5. Routing Overhead (RO):

According to the fig 10, it is analyzed that DSDV has maximum routing overhead AODV has medium but very close to DSR and DSR has minimum RO.

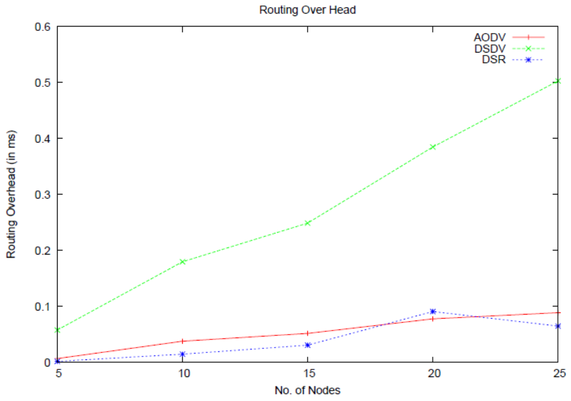


Fig 10: Routing Overhead.

6. Packet Delivery Ratio:

According to the fig 11, it is analyzed that DSDV has less packet delivery ratio AODV has medium but very close to DSR and DSR has maximum PDR.

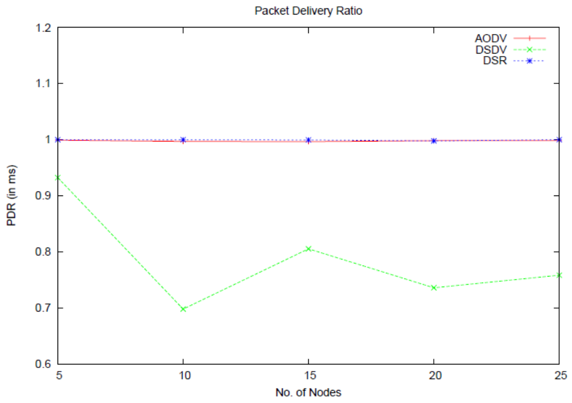


Fig 11: Packet Delivery Ratio.

7. Average Throughput:

According to the fig 12, it is analyzed that DSDV has less throughput AODV has medium but close to DSR and DSR has maximum throughput.

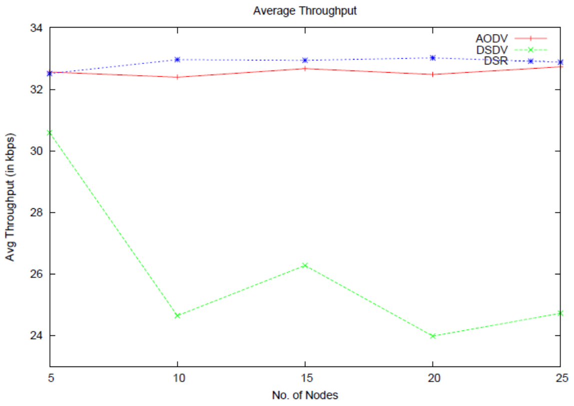


Fig 12: Average Throughput.

CONCLUSION

In this paper we simulate and compare three routing protocols to investigate the energy consumption in healthcare environment and find in both cases during Idle and TX mode performance of DSR is best when node density is less than and equal to 20 and when node density is greater than 20 the performance of DSDV is best as an energy efficient routing protocol. It is according to average remaining energy graph DSR is maximum remaining energy protocol in both cases On the other hand the performance of DSR is also good on the basis of RO, PDR and throughput.

Acknowledgments

I greatly acknowledged the kind supervision of Dr. Imran Shafi, who taught me the subject Mathematical Methods in Communication, ANN and encouraged me for this work and to write this paper.

References

[1] N. A. Pantazis, S. A. Nikolidakis, D. D. Vergados, " Energy-efficient routing protocols in wireless wensor networks for health communication systems", PETRA 09, Corfu, Greece, ACM ISBN 978-1-60558-409-6, 2009.

[2] P. Bergamo, Alessandra, " Distributed power control for energy efficient routing in ad hoc networks ", Wireless Networks 10, pp. 29–42, 2004.

[3] V. Shnayder, B. Chen, K. Lorincz, T. R. F. FulfordJones, M. Welsh, "Sensor networks for medical care", this document is a technical report. It should be cited as: Technical Report TR-08-05, Division of Engineering and Applied Sciences, Harvard University, 2005.

[4] I. Noorzaie, "Survey Paper: Medical applications of wireless networks", last modified: April 12, 2006. Note: This paper is available on-line at http://www.cse.wustl.edu/~jain/cse574-06/ftp/medical_wireless/index.html.

[5] I. F. Akyildiz, I. H. Kasimoglui, "Wireless sensor and actor networks:research challenges", (Elsevier) Journal, pp. 351-367, 2004.

[6] D. B. Johnson, D. A. Maltz, J. Broch. "DSR: The dynamic source routing protocol for multi-hop wireless ad-hoc networks", Ad-hoc Networking, chapter 5, pp. 139-172, 2001.

[7] L. M. Feeney, "An energy consumption model for performance analysis of routing protocols for mobile ad hoc networks", Mobile Networks and Applications 6, pp. 239–249, 2001.

[8] F. Giroire∗ , D. Mazauric∗ , J. Moulierac∗ , B. Onfroy∗ , "Minimizing routing energy consumption: from theoretical to practical result", IEEE / ACM International Conference on Green Computing and Communications, 2010.

[9] L. Layuan, L. Chunlin, Y. Peiyan, "An energy level based routing protocol in ad-hoc networks", proceedings of the IEEE / WIC / ACM International Conference on Intelligent Agent Technology (IAT), 2006.

[10] S. Mahfoudh, P. Minet, "Energy-aware routing in wireless ad hoc and sensor networks", IWCMC, Caen, France, ACM 978-1-4503-0062-9/10/06, 2010.

[11] N. Tantubay, D. R. Gautam, M. K. Dhariwal , "A review of power conservation in wireless mobile ad-hoc network (MANET)", IJCSI International Journal of Computer Science Issues, ISSN (Online), Vol. 8, Issue 4, No 1, 2011.

[12] S. P. Setty, B. Prasad, "Comparative study of energy aware QoS for proactive and reactive routing protocols for mobile ad-hoc networks", International Journal of Computer Applications, Volume 31, No.5, 2011.

[13] V. Prakash, B. Kumar, A. K. Srivastava, "Energy efficiency comparison of some topology-based and location-based mobile ad-hoc routing protocols", ICCCS 11, Copyright © 2011 ACM 978-1-4503-0464-1/11/02, Rourkela, Odisha, India, 2011.

[14] Feeney, "Energy efficient communication in ad hoc wireless networks",

http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.9379.

[15] Bergamo, Giovanardi, Travasonia, "Distributed power control for energy efficient routing in Ad-hoc a networks", published in journal of Wireless Networks, Volume 10 Issue 1, pp. 29–42, 2004.

[16] Gomez, Campbell, Naghshineh, "Conserving transmission power in wireless ad-hoc networks ", [A]. In Proc of IEEE Conference on Network Protocols (ICNP. 01), 2001.

[17] Y. Wei, L. Jangwon, "DSR-based energy-aware routing protocols in ad hoc networks", in Proc. of the International Conference on Wireless Networks, 2002.

[18] W. Zhao, K. Ramamritham, "Virtual time CSMA protocols for hard real–time communications", IEEE Transactions on Software Engineering, Vol. 13, No. 8, 1987.

[19] W. Zhao, J. Stankovic, K. Ramamritham, "A window protocol for transmission of time constrained messages", IEEE Transactions on Computers, Vol. 39, No. 9, 1990.

[20] Rahman, Gburzynski, "On constructing minimum-energy path-preserving graphs for ad-hoc wireless networks", ICC 2005: IEEE international conference on communications, Vols.1, pp. 3083-3087, 2005.

[21] N. Chevrollier, N. Golmie, "On the use of wireless network technologies in healthcare environments", White Paper (U.S Department of Commerce) published in 2005. (Note: This paper is available on-line at http://w3.antd.nist.gov/pubs/aswn05.pdf)

[22] S. P. Chaudhuri, D. B. Johnson, "Power mode scheduling for ad-hoc networks", IEEE International Conference on Network Protocols, 2002.

[23] J. H. Chang, L. Tassiulas, "Energy conserving routing in wireless ad-hoc networks", Proc. IEEE INFOCOM, Tel Aviv, Israel, pp. 22-31, 2000.

[24] M. Fotino, "Evaluating energy consumption of proactive and reactive routing protocols in a MANET", Proc. 1st Int. Conf. On Wireless Sensor and Actor Networks, pp. 119-130, 2007.

[25] R. Zheng, R. Kravets, "On-demand power management for ad-hoc networks", published in IEEE INFOCOM, 2003.

[26] T. H. Tie, C. E. Tan, S. P. Lau, "Alternate link maximum energy level ad-hoc distance vector scheme for energy efficient ad-hoc networks routing", In Proceedings of International Conference on Computer and Communication Engineering (ICCCE 2010), Kuala Lumpur, Malaysia, 2010.

AUTHOR'S PROFILES

Sohail Abid: (Mobile No: +92-321-5248497)
Sohail Abid Student of MS (TN) at IQRA University Islamabad and working as System Administrator at Foundation University Institute of Engineering and Management Sciences.

Dr. Imran Shafi: (Mobile No: +92-334-5323402)
Dr. Imran Shafi is working as Assistant Professor at Abasyn University Islamabad campus, Pakistan.

Shahid Abid: (Mobile No: +92-333-5656413)
Shahid Abid having Master in Computer Science and working as Assistant System Administrator at Foundation University Institute of Engineering and Management Sciences.

Various Approaches for Enhancing The Performance of Wireless Sensor Networks

Hasan Al Shalabi[1] & Ibrahiem M. M. El Emary[2]

**[1]Computer Engineering Department, Al Hussein Bin Talal University
Ma'an, Jordan**

**[2]Information Technology Deanship, King Abdulaziz University
Jeddah, Saudi Arabia**

Abstract

In the current time and next decades, Wireless Sensor Networks (WSNs) represents a new category of ad hoc networks consisting of small nodes with three functions: sensing, computation, and wireless communications capabilities. Many routing, power management, and data dissemination protocols have been designed for WSNs where energy awareness is an essential design issue to improve the overall performance of WSN. There are many approaches and techniques explored for the optimization of energy usage in wireless sensor networks. Routing represents one of these areas in which attempts for efficient utilization of energy have been made. In this paper, we report on the current state of the research on optimizing the performance of WSN using various advanced approaches. There are various directions to enhance and optimize the performance as: avoiding congestion and keep it within certain controlled value, selecting the optimum routing approach, reducing the level of power consumption to increase the life time of the sensor node and others. So, the major objective of this paper is to investigate the various techniques used in improving and enhancing the performance of WSN to let it be more reliable in various applications like: health care and biomedical treatment, environment monitoring, military survival lance , target tracking, greenhouse monitoring,…etc .

Keywords: *WSN, ad hoc, DAQ, TDMA, TRAMA, RETSINA, ACCP, Dijkstra algorithm*

1. Introduction

The wireless sensor network is some type of an ad-hoc network. Mainly it consists of small light weighted wireless nodes called sensor nodes, deployed in physical or environmental condition. It measure the physical parameters such as sound, pressure, temperature, and humidity. These sensor nodes deployed in large or thousand numbers and collaborate to form an ad hoc network capable of reporting to data collection sink (base station).

Wireless sensor network have various applications like habitat monitoring, building monitoring, health monitoring, military survival lance and target tracking. However wireless sensor network is a resource constraint if we talk about energy, computation, memory and limited communication capabilities.

A typical wireless sensor network is comprised of tens, hundreds, or even thousands of sensor nodes. Typically each sensor node is composed of a microcontroller, a radio transceiver, one or more micro sensors, power source and other components. The microcontroller samples the micro sensors, send the data, either with or without processing, through radio links to the locations where the information is needed. Due to the limited radio range and the relatively larger target areas, in many cases a multiple hop ad hoc wireless network is formed for the information transmission. The devices that gather the information from the wireless sensor networks are defined as base stations. There may be one or more base stations for a wireless sensor network. The base stations may be static or mobile. However, for many applications, the sensor nodes themselves are not moving, either due to the scenario requirements, or technical or economical hindrance.

All sensor nodes in the wireless sensor network are interacting with each other or by intermediate sensor nodes [1]. A sensor nodes that generates data, based on its sensing mechanisms observation and transmit sensed data packet to the base station (sink). This process basically direct transmission since the base station may locate very far away from sensor nodes needs (see Fig.1). More energy to transmit data over long distances so that a better technique is to have fewer nodes sends data to the base station. These nodes called aggregator nodes and processes called data aggregation in wireless sensor network.

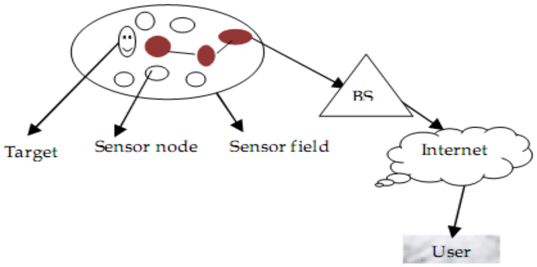


Fig. 1 Architecture of the Sensor network

In a wireless sensor network, sensing nodes with limited power, computation, communication and storage resources cooperate to fulfill monitoring and tracking functionalities. Compared with conventional sensors, wireless sensor networks have the following advantages: Since the relevant technologies have become technologically and economically feasible, people want to gather much more information from more places in the physical world, which was either impossible due to technological difficulties or formidable due to high cost, in terms of money and human power. Decreasing form factors and costs of micro sensors make deployment of hundreds even thousands of sensors much more feasible than conventional sensors that are too cumbersome and expensive [2].

Many conventional sensors send data to data acquisition (DAQ) modules in personal computers or workstations, and typically the sensors, personal computers and workstations are interconnected in a wired manner for data collection, aggregation and processing. When the number of sensors deployed in a certain area increases exponentially, this approach obsoletes. A more appropriate approach is to connect the sensors with low cost microcontrollers and to send and receive information through wireless links. This approach is made feasible by decrease in costs, form factors and power consumption of microcontrollers and radio transceivers. Based on the sensing ranges of quite a few commonly used types of sensors, a deployment density of one node per one hundred square meters is typical [2].

One of the design optimization strategies applied in WSN is to deterministically place the sensor nodes in order to meet the desired performance goals. In such case, the coverage of the monitored region can be ensured through careful planning of node densities and fields of view and thus the network topology can be established at setup time. However, in many WSNs applications sensors deployment is random and little control can be exerted in order to ensure coverage and yield uniform node density while achieving strongly connected network topology. Therefore, controlled placement is often pursued for only a selected subset of the employed nodes with the goal of structuring the network topology in a way that achieves the desired application requirements. In addition to coverage, the nodes' positions affect numerous network performance metrics such as energy consumption, delay and throughput. For example, large distances between nodes weaken the communication links, lower the throughput and increase energy consumption.

2. Literature Review

This section review the prior work on improving the congestion control over wireless sensor networks as one approach for enhancing the performance of WSN. In [3], a cross-layer TDMA-based protocol that guarantees collision-free communication by scheduling slots for each node and results in significant energy savings was presented. This technique has the capability of determining the collision-free slots that are to be assigned to wireless nodes in a multiple-hop network. In [4], another approach was proposed in which TRAMA that organizes time into frames and uses a distributed election scheme based on traffic information at each node to determine which node can transmit at a particular slot. TRAMA uses a distributed hash function to determine a collision-free slot assignment and builds a scheduling scheme when a node has data to send. This random scheduling scheme increases the queuing delays.

Another technique called Queue based Congestion Control Protocol with priority support, using the queue length as an indication of congestion degree was presented in [5]. In this approach, the rate assignment to each traffic source is based on its priority index as well as its current congestion degree. A node priority-based congestion control protocol for wireless sensor networks was proposed in [6]. In this technique, the node priority index is introduced to reflect the importance of each node and uses packet inter-arrival time along with packet service time to measure a parameter defined as congestion degree and imposes hop-by-hop control based measurement as well as node priority index. In [7], it was proposed an energy efficient congestion control scheme for sensor networks called Enhanced Congestion Detection and Avoidance which comprises of three mechanisms. First, the approach uses buffer and weighted buffer difference for congestion detection. Secondly, proposed a bottleneck-node-based source data sending rate control scheme and finally uses a flexible queue scheduler for packets transfer. A new and more recent approach was proposed in [8] called a cluster head method to allow parallel transmission of data packets to form a schedule by arranging data transfer at each round. The cluster head accepts request for data transfer and assigns a slot for each node wishing to transmit. Each node of data transfer is divided into contention, data transmission and idle period. In WSN the single point of failure is eliminated by providing a decentralized control and nodes that have no data to send waste time slots in the contention period where idle listening and overhearing occurs.

In [9] a suggested approach called an adaptive rate control for congestion avoidance in WBANs was presented. The scheme performs rate control dynamically each node based on a predication model which uses rate function including congestion risk degree and valuation function, without requiring congestion detection and congestion notification steps. There is another advanced approach presented in [10] based on a distributed and scalable algorithm that eliminates congestion within a sensor network, and ensures the fair delivery of packets to a central node or a base station. This routing structures often results in the sensors closer to the base station experiencing congestion, which inevitably cause packets originating from sensors to have a higher probability of being dropped. The problem of single-path upstream congestion control in wireless sensor networks through the traffic control was investigated in [11], where authors of this work proposed a multi-agent system based approach to control the traffic in the upstream congestion. The traffic generated in a wireless sensor node is of two types named, source traffic and transit traffic. The source traffic is generated from each wireless sensor node and the transit traffic is generated from other wireless sensor nodes.

A Reusable Task-based System of Intelligent Networked Agents (RETSINA) is a cooperative multi-agent system that consists of three classes of agents: interface agents, task agents and information agents. RETSINA provides a domain-independent, componentized, and reusable substratum to (a) allow heterogeneous agents to coordinate in a variety of ways and (b) enable a single agent to be part of a multi-agent infrastructure. RETSINA [12] provides facilities for reuse and a combination of different existing low-level infrastructure components, and it also defines and implements higher level agent services and components that are reconfigurable and reusable. Paper [11] proposed an upstream congestion control model by using RETSINA multi-agent named Agent-based Congestion Control Protocol (ACCP). ACCP reduce the packet loss by its intelligent scheduling schemes. Fig.2 illustrates the proposed congestion control model in a wireless sensor node. ACCP consists of four components: Execution Monitor, Communicator, Planner, and Scheduler [13].

The execution monitor identifies the congestion based on the packet arrival time (ta) and packet service time (ts) at the Medium Access Control (MAC) layer. The packet arrival time (ta) is the time interval between two subsequent packets arrived from any source and the packet service time (ts) is the time interval between arrival of packets at the MAC and its successful transmission. These two parameters are monitored at each node by the execution monitor on a packet-by packet manner.

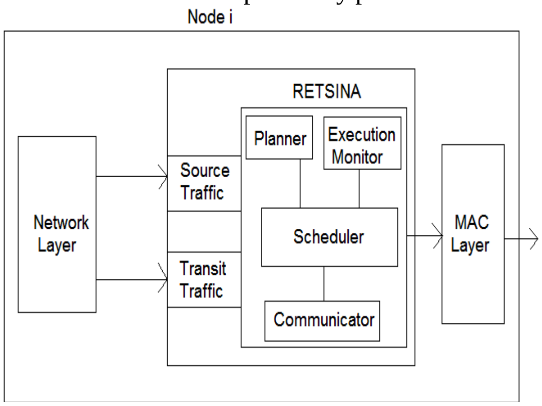


Fig.2 Congestion control model [11]

From this, a congestion index (Cx) is calculated and it is defined as the ratio of average packet service time over average packet arrival time at each wireless sensor node. The congestion index at node i is given [11] by:-

$$Cx(i) = ts / ta \quad (1)$$

The execution monitor also takes the agent's next intended action and prepares, monitors, and completes its execution. The communicator module communicates all the notifications at each wireless sensor node in the packet header to be forwarded. From the congestion index the communicator module computes a global congestion priority index by summating source congestion priority index and the global congestion priority index of the lower level wireless senor nodes. The planner receives goals through communication message packets and finds alternative ways to fulfill them. The planning component is reusable and capable of accepting different planning algorithms in an intelligent way. The scheduler has two queues for the source traffic and the transit traffic. By adjusting the scheduling rate the congestion can be reduced. The scheduling algorithm uses the earliest deadline-first heuristic. A list of all actions is scheduled and the action with the earliest deadline is chosen for execution. When a periodic action is chosen for execution, it is reinstated into the schedule with a deadline equal to the current time plus the action's period.

The four modules of RETSINA multi-agent are implemented for the upstream congestion control as autonomous threads of control to allow concurrent planning and scheduling actions, and execution in an efficient way. Furthermore, all modules are executed as separate threads and are able to execute concurrently. So almost all the packets are forwarded to the next wireless sensor node without any loses.

3. WSN Architecture Parameters and Design Requirements

A typical sensor network operates in five phases: the planning phase, deployment phase, post-deployment phase, operation phase and post-operation phase. In the planning phase, a site survey is conducted to evaluate deployment environment and conditions, and then to select a suitable deployment mechanism. In the deployment phase, sensors are randomly deployed over a target region. In the post deployment phase, the sensor networks operators need to identify or estimate the location of sensors and to access coverage. The operation phase involves the normal operation of monitoring tasks where sensors observe the environment and generate data. The post-operation phase involves shutting down and preserving the sensors by settings the sensors to sleep mode for future operations or destroying the sensor network. In a WSN setup, the nodes may be deployed in an ad-hoc manner with no predefined topology. The nodes automatically setup a network by communicating with one another in a multihop fashion. New nodes can malfunction, be added or removed from the network at any time. Newly added nodes must integrate into the network seamlessly and the network must detect and react quickly when nodes are removed to avoid affecting the reliability of message delivery services. The timely detection, processing, and delivery of information are indispensable requirements in a real-time WSN application. In SPEED there are two types of communication associated with data delivery

- **unicast** (a specific node will receive the packet) area-multicast (where a copy of the packet is send to every node inside the specified area)
- **area-unicast** (copy of the packet is sent to at least one node inside the specified area)

For efficient communication both the route discovery cost and resulting route length are important. Unlike wired networks, where the delay is independent of the route length, in multihop wireless sensor networks, the end-to-end delay depends on not only single hop delay, but also on the distance a packet travels.

Any real time protocol should satisfy three design objectives: stateless nodes, load balanced routes and congestion control mechanism. The architectures of WSNs emerged from the experience gained from devising architectures for self-organizing, mobile, ad hoc networks . The latter show emphasis on the need for decentralized, distributed form of organization and this is a shared characteristic with WSNs. They benefit from the evolutions in real-time computing, peer-to-peer computing, active networks and mobile agents/swarm intelligence. Besides the networking and computing concepts just mentioned, many other factors play a significant role when devising architectures for a WSN.

The critical factors that distinguished between different WSNs architectures are listed as follows:-

- **Fault tolerance:** WSNs are mainly monitoring important phenomena. Therefore, it is essential for a WSN to sustain its functionality without disruptions, even if some nodes malfunction or die. Usually, WSNs are deployed in hostile environments where nodes may be damaged, due to environmental interference, or eventually die due to the impracticality of recharging or replacing their batteries. Nodes in a WSN are prone to failures and this may result in sever situations like partitioning the network. The design of a WSN should guarantee that its functionality and services are never degraded by these failures.
- **Scalability:** Sensor nodes are deployed densely to form a WSN. This huge number of nodes has a direct impact on the design of schemes and protocols at different layers. For example, a MAC protocol (data-link layer) should be able to grant, in a fair fashion, each node access to the medium while minimizing or preventing collisions, which is very difficult given the huge number of available nodes. Also, a routing protocol (network layer) that depends on exchanging routing tables among nodes may not be efficient since there will be excessive control traffic that underutilizes the bandwidth of the medium.

- **Production cost:** The cost of a single sensor node should be minimized since it determines the overall cost of the network under design.
- **Network topology:** The fact that Wireless Sensor Networks are constituted by a huge number of nodes raises the challenge of network topology maintenance and modification. The challenge occurs starting at the early stage of nodes deployment. Sensor nodes can be either thrown in a mass (e.g., from a plane) or manually placed one by one (e.g., by a human or a robot) in the field. Also, after nodes deployment, topology may change due to failures in some nodes, changes in nodes locations, lack of reachability (due to jamming for instance), and huge reductions in power resources at some nodes (which affect their transmission power levels to the limit that they vanish from the vicinity of neighboring nodes). The WSN should be able to adapt to these sudden changes to avoid any degradations in its functionality.
- **Security:** In the environment of deployment, sensor nodes are either deployed very close to the phenomenon of interest or directly inside it. As a result, we can see that WSNs are usually not supervised (especially in remote geographic areas). This means that WSNs may be targeted by intruders to exploit any security vulnerability.
- **QoS support:** Time-sensitive applications (especially in military) require support for real-time communication that provisions guarantees on maximum delay, minimum bandwidth, or other QoS parameters.
- **Power consumption:** this is a primary design factor for any WSN. Power consumption should be made minimal in order to prolong the lifetime of the network. In fact, “power conservation” is a distinguishing factor between designing a WSN and designing other classes of wireless networks. The latter may consider QoS parameters (like, delay, throughput, fairness, etc.) as key design requirements. Based on this observation, research activities target the development of power-aware protocols and algorithms for sensor networks. That is, power-awareness should be incorporated in every stage of designing a WSN. In fact, power-awareness imposes constraints on the size and complexity of a sensor node’s platform. In this context, hardware of sensor nodes should be designed to be power-efficient.

1. 4. Various Methodologies Applied for Enhancing the WSN Performance

In this section, we list and report the important works achieved for improving the WSN performance in view point of different directions given as follows:-

- **Firstly regarding the congestion control problem:** [3] proposed a cross-layer TDMA-based protocol that guarantees collision-free communication by scheduling slots for each node and results in significant energy savings. This has the main challenge to determine the collision-free slots that are to be assigned to wireless nodes in a multiple-hop network. [4] proposed TRAMA that organizes time into frames and uses a distributed election scheme based on traffic information at each node to determine which node can transmit at a particular slot. TRAMA uses a distributed hash function to determine a collision-free slot assignment and builds a scheduling scheme when a node has data to send. This random scheduling scheme increases the queuing delays. [5] proposed a novel upstream congestion control protocol for WSNs named Priority based Congestion Control Protocol, which introduced node priority index to reflect the importance of each sensor node. This utilizes a cross-layer optimization and imposes a hop-by-hop approach to control congestion. [6] presented a new Queue based Congestion Control Protocol with priority support, using the queue length as an indication of congestion degree. In this approach, the rate assignment to each traffic source is based on its priority index as well as its current congestion degree. [5] proposed a node priority-based congestion control protocol for wireless sensor networks. In this, the node priority index is introduced to reflect the importance of each node and uses packet inter-arrival time along with packet service time to measure a parameter defined as congestion degree and imposes hop-by-hop control based measurement as well as node priority index. [9] proposed an energy efficient congestion control scheme for sensor networks called Enhanced Congestion Detection and Avoidance which comprises of three mechanisms. First, the approach uses buffer and weighted buffer difference for congestion detection. Secondly, proposed a bottleneck-node-based source data sending rate control scheme and finally uses a flexible queue scheduler for packets transfer.

- **Secondly regarding the avoidance of routing problems:** The limited energy supply of sensor nodes necessitates energy-awareness at most layers of networking protocol stack including the network layer. In addition, many applications of sensor networks require the deployment of a large number of sensor nodes making it impractical to build a global addressing scheme. Moreover, in contrary to contemporary communication networks almost all applications of sensor networks require the flow of sensed data from multiple sources to a particular sink. These unique characteristics of sensor networks have made efficient routing of sensor data one of the technical challenges in wireless sensor networks [15]. While a number of routing protocols pursued a data centric methodology by naming the data, some considered clustering the sensor nodes in order to decrease the number of transmitted messages to the sink node and have a more scalable setup. Other protocols either adopted a location-based routes setup or strived to achieve energy saving through activation of a limited subset of nodes. In addition, with the increasing interest in the applications that require certain end-to-end performance guarantees, a few routing protocols have been proposed for providing energy-efficient relaying of delay-constrained data [16, 17]. While the goals of most published techniques are increasing network lifetime and on-time delivery of data through clever architecture and management of the network, none of the work considered the possibility of relocating the sink (gateway) node for enhanced network performance.

Gateway positioning has been also investigated in the context of wireless local area network and cellular infrastructure [18, 19]. The gateways, also called base stations, in these systems are stationary in nature and are placed in order to achieve coverage of an area or a building using the minimal number of gateway units. The considered model was using the gateway node as a direct router for a group of mobile nodes that would be otherwise unreachable due to topological reasons such as blockages. The problem addressed was to find the optimal place for the gateway node to best serve the group in terms of latency and throughput. The pursed approach was to move the gateway to the weighted geographic centroid of the group by considering the location and traffic generated by nodes regardless the established routes.

Determining the gateway position was formulated as an optimization problem. The gateway applies the repositioning algorithm at each sampling instant to gradually get closer to the nodes that generate the highest traffic and to respond to changes in the node locations. The work in [14] is path-based and does not just consider nodes. They argue that it is not possible to optimally place the gateway without considering the network topology and inter-node links. They observe that frequent changes to the network topology can impose overhead that can surpass the value of the relocation from a system's point of view.

- **Thirdly regarding the problem of power saving**: Ideally, we would like the sensor network to perform its functionality as long as possible. Optimal routing in energy constrained networks is not practically feasible (because it requires future knowledge). However, we can soften our requirements towards a statistically optimal scheme, which maximizes the network functionality considered over all possible future activity. A scheme is energy efficient (in contrast to 'energy optimal') when it is statistically optimal and causal (i.e. takes only past and present into account).

In most practical surveillance or monitoring applications, we do not want any coverage gaps to develop. We therefore define the lifetime we want to maximize as the worst case time until a node breaks down, instead of the average time over all scenarios. However, taking into account all possible future scenarios is too computationally intensive, even for simulations. It is therefore certainly unworkable as a guideline to base practical schemes on. Many routing and data transfer protocols have been specifically designed for WSNs [21-24]. Most sensor network routing protocols are, however, quite simple and for this reason are sometimes insecure. In what follow, we present a discussion of major attacks against them. Most network layer attacks against sensor networks fall into one of the following categories [25].

1- Sinkhole attacks: to make the compromised node look attractive to surrounding nodes with respect to the routing choice.

2- Spoofed, altered routing attack: to replay routing information, create routing loops, and to extend or shorten source routes.

3- Selective forwarding: refuse to forward certain messages, to simply drop them, and to attract or repel network traffic

4- Sybil attacks: a single node presents multiple identities to other nodes in the network.

5- Wormholes: adversary tunnels messages received in one part of the network over a low latency link and replays them in a different part.

6- HELLO flood attacks: broadcast HELLO packets to announce themselves to their neighbors and define new node.

7- Acknowledgement spoofing: spoof link layer acknowledgments for "overheard packets" addressed to neighboring nodes.

As in [26], a centralized approach which allows using a source routing methodology (see fig.3) was adopted. Before realizing the route discovery, a first phase consists in providing each link in the network a specific weight. This weight depends on the energy of the destination node, in order to relay the information by the nodes having the higher remaining energy, and the distance between nodes, in order to prefer short distance transmissions.

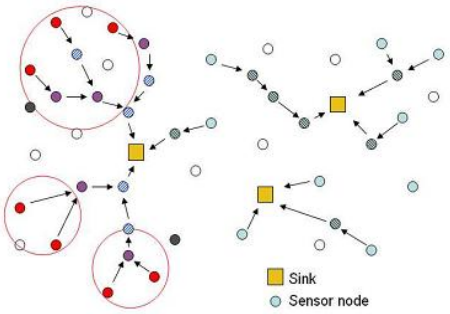


Fig. 3 Routing in a sensor network

Once this weight computed, we obtain a graph on which Dijkstra algorithm can be applied to find the shortest path between a sink and each node in the network. However, in our case the network contains more than one sink. So the aim now is to find the shortest path towards the nearest sink. We will proceed by exploring all the shortest paths towards all the sinks, and we will conserve for routing needs the one leading to the nearest sink. At this point, it is worth to notice that a more efficient searching method can be found, nevertheless our approach has the benefit to easily maintain, for each node, the n nearest sinks and their corresponding paths, which are necessary for the configuration evaluation. The algorithm complexity is acceptable since we achieve p Dijkstra in O(n).

One of the major and probably most important challenges in the design of WSNs is the fact that energy resources are significantly more limited than in wired networks [27,28]. Recharging or replacing the battery of the sensors in the network may be difficult or impossible, causing severe limitations in the communication and processing time between all sensors in the network. Note that failure of regular sensors may not harm the overall functioning of a WSN, since neighboring sensors can take over, provided that their density is high. Therefore, the key parameter to optimize for is network lifetime, or the time until the network gets partitioned [29].

Another issue in WSN design is the connectivity of the network according to the selected communication protocol [27]. The most common protocol follows the cluster-based architecture, where single-hop communication occurs between sensors of a cluster and a selected cluster head sensor that collects all information gathered by the other sensors in its cluster. Usually, connectivity issues include the number of sensors in each cluster, because a cluster head can handle up to a specific number of connected sensors, as well as coverage issues related to the ability of each sensor to reach some cluster head.

Finally, design issues that have been rather neglected in the research literature are those that depend on the particular application of WSNs. Energy and connectivity issues are certainly important in a WSN design, but one must not forget the purpose of the sensor network, which is the collection and possibly management of measured data for some particular application. This collection must meet specific requirements, depending on the type of data that are collected. These requirements are turned into specific design properties of the WSN, which in this work are called ''application specific parameters'' of the network.

Several analyses of energy efficiency of sensor networks have been realized [21–24] and several algorithms that lead to optimal connectivity topologies for power conservation have been proposed [30–34]. However, most of these approaches do not take into account the principles, characteristics and requirements of application-specific WSNs. When these factors are considered, then the problem of optimal design and management of WSNs becomes much more complex.

Fig. 4 shows the percentage of sensors (over the entire grid of 900 sensors) with battery capacities below certain percentage-levels after each measuring cycle, based on the assumption that all sensors had 100% battery capacity at the beginning of the first measuring cycle. It is clear that the percentage of sensors with battery capacity below 40% is kept very low during the 15 measuring cycles, even while at the end of the 15th measuring cycle there is no sensor with battery capacity below 20%. Corresponding results on the analysis of remaining sensors with battery capacities above certain percentage levels also showed high conservation of energy resources.

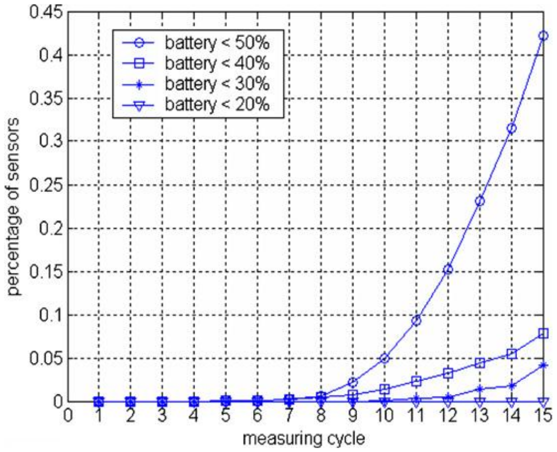


Fig. 4 Percentages of sensors with battery capacities below certain values (as percentages of full battery capacity) at the end of each measuring cycle of the adaptive WSN design.

5. Concluded Remarks

In the current time, there is a new era of ubiquitous computing . One type of such ubiquitous is wireless sensor technologies which is characterized with a great potential in opening a world of sensing applications. This paper provides the different approaches used in enhancing the performance of such WSN with great focus on three important factors: congestion control, optimal routing approaches and reducing the consumption power. However the consumption power was touched in deep since Wireless sensor networks are battery powered, therefore prolonging the network lifetime through a power aware node organization is highly desirable. An efficient method for energy saving is to schedule the sensor node activity such that every sensor alternates between sleep and active state. One solution is to organize the sensor nodes in disjoint covers, such that every cover completely monitors all the targets. These covers are activated in turn, in a round-robin fashion, such that at a specific time only one sensor set is responsible for sensing the targets, while all other sensors are in a low-energy, sleep state.

References

[1] Kiran Maraiya, Kamal Kant, Nitin Gupta, “Wireless Sensor Network: A Review on Data Aggregation “, International Journal of Scientific & Engineering Research Volume 2, Issue 4, April - 2011

[2] Ting Yan, “Analysis Approaches for Predicting Performance of Wireless Sensor Networks” , Ph.D. Dissertation, August 2006

[3] Cheng Tien Ee, and Ruzena Bajcsy, “Congestion Control and Fairness for Many-to-One Routing in Sensor Networks”, Proceeding of the ACM Conference SenSys ’04, Baltimore, Maryland, USA, November 2004.

[4] Chonggang Wang, Kazem Sohraby, Victor Lawrence, Bo Li, and Yueming Hu, “Priority-based Congestion Control in Wireless Sensor Networks”, Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC ’06), IEEE Computer Society, 2006.

[5] Jain Liu, and Suresh Singh, “ATCP : TCP for Mobile Ad-hoc Networks”, IEEE journal on selected areas in Communications, Vol. 19, No. 7, pp 1300-1315, July 2001.

[6] Lazarou G. Y., Li J., and Picone J., “A Cluster-based Power-Efficient MAC Scheme for Event-Driven Sensing Applications”, Proceedings of the ACM Conference on Ad hoc Networks, Vol. 5, No. 7, pp. 1017-1030, Sep. 2007.

[7] Liqiang Tao and Fengqi Yu, “ECODA: Enhanced Congestion Detection and Avoidance for Multiple Class of Traffic in Sensor Networks”, Proceedings of the 15th Asia-Pacific Conference on Communications (APCC 2009), pp. 726-730, 2009.

[8] Mohammad Hossein Yaghmaee and Donald Adjeroh, “A New Priority Based Congestion Control Protocol for Wireless Multimedia Sensor Networks”, Proceeding of the IEEE, 2008.

[9] Paolucci M., D. Kalp, A. Pannu, O. Shehorg, and K. Sycara, “A planning Component for RETSINA Agents”, Internet Draft.

[10] Rajendran V., Obraczka K., and Garcia-Luna-Aceves J. J., “Energy Efficient Collision-Free Medium Access Control for Wireless Sensor Networks”, Proceedings of the First ACM International Conference on Embedded Networked Sensor Systems (SenSys ’03), March 2003.

[11] V. Vijaya Raja, R. Rani Hemamalini, A. Jose Anand, “Multi Agent System Based Upstream Congestion Control in Wireless Sensor Networks”, European Journal of Scientific Research, Vol.59 No.2 (2011), pp.241-248

[12] Sichitiu M. L., “Cross-Layer Scheduling for Power Efficiency in Wireless Sensor Networks”, Proceedings of the IEEE INFOCOM ’04, March 2004.

[13] Young-mi Baek, Byung-hwa Lee, Jilong Li, Qin Shu, Ji-hum Han, and Ki-ju Han, “An adaptive rate control for congestion avoidance in wireless body area networks”, Proceeding of the IEEE, 2009.

[15] K. Akkaya and M. Younis, “A survey on routing protocols for wireless sensor networks,” in Elsevier Journal of Ad Hoc Networks (to appear).

[16] T. He et al., “SPEED: A stateless protocol for real-time communication in sensor networks,” in the Proceedings of International Conference on Distributed Computing Systems, Providence, RI, May 2003.

[17] K. Akkaya and M. Younis, “Energy-aware Routing of Delay-constrained Data in Wirel ess Sensor Networks,” Journal of Communication Systems, special issue on QoS support and service differentiation in wireless networks, (to appear).

[18] R. Rodrigues et al., “Optimal Base Station Placement and Fixed Channel Assignment Applied to Wireless Local Area Network Projects, “ in the Proceedings of IEEE International Conference on Networks, Australia, 1999.

[19] T. Fruhwirth and P. Brisset, “Optimal Placement of Base Stations in Wireless Indoor Telecommunication,” in Lecture Notes in Computer Science, Volume 1520, 1998.

[20] Kemal Akkaya, Mohamed Younis and Meenakshi Bangad, “Sink Repositioning For Enhanced Performance in Wireless Sensor Networks”
www. www.cs.umbc.edu

[21] Jamal N. Al-Karaki, Ahmed E. Kamal, “Routing Techniques In Wireless Sensor Networks: A Survey”, IEEE Wireless Communications December 2004

[22] Chris Karlof and David Wagner, Secure Routing in Wireless Sensor Networks: Attacks and Countermeasures”, Sensor Network Protocols and Applications, 2003. Proceedings of the First IEEE. 2003 IEEE International Workshop on 11 May 2003 Page(s):113 - 127

[23] Brad Karp and H. T. Kung. "GPSR: Greedy Perimeter Stateless Routing for Wireless Networks". In Proceedings of the 6th Annual International Conference on Mobile Computing and Networking (MOBICOM’00), ACM Press, N. Y., 2000.

[24] M. Younis, M. Youssef and K. Arisha, “Energy-aware Routing in Cluster-Based Sensor Networks”, in the Proceedings of the 10th IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, (MASCOTS2002), Fort Worth, Texas, October 2002. no. 5.408.237, April 1995

[25] B. Krishnamachari, F. Ordo ́ nez, Analysis of energy-efficient, fair routing in wireless sensor networks through non-linear optimization, in: Proc. IEEE Vehicular Technology Conference – Fall, Orlando, FL, 2003, pp. 2844–2848.

[26] S. Ghiasi, A. Srivastava, X. Yang, M. Sarrafzadeh, Optimal energy aware clustering in sensor networks, Sensors 2 (2002) 258–269.

[27] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, Wireless sensor networks: a survey, Computer Networks 38 (2002) 393–422.

[28] S. Slijepcevic, M. Potkonjak, Power efficient organization of wireless sensor networks, in: Proc. IEEE Int. Conf. on Communications, Helsinki, Finland, 2001, pp. 472–476.

[29] Konstantinos P. Ferentinos *, Theodore A. Tsiligiridis, “Adaptive design optimization of wireless sensor networks using genetic algorithms”, Computer Networks 51 (2007) 1031–1051

[30] V. Rodoplu, T.H. Meng, Minimum energy mobile wireless networks, IEEE J. Select. Areas Commun. 17 (8) (1999) 1333–1344.

[31] W.R. Heinzelman, A. Chandrakasan, H. Balakrishnan, Energy-efficient communication protocol for wireless microsensor networks, in: Proc. 33rd Hawaii Int. Conf. on System Sciences, Maui, Hawaii, 2000.

[32] J.-H. Chang, L. Tassiulas, Energy conserving routing in wireless ad-hoc networks, in: Proc. IEEE INFOCOM’00, Tel Aviv, Israel, 2000, pp. 22–31.

[33] D.J. Chmielewski, T. Palmer, V. Manousiouthakis, On the theory of optimal sensor placement, AlChE J. 48 (5) (2002) 1001–1012.

[34] C. Zhou, B. Krishnamachari, Localized topology generation mechanisms for wireless sensor networks, in: IEEE GLOBECOM’ 03, San Francisco, CA, December 2003.

First Author Biography:

Hasan Mohammad Al-Shalabi*:* He received his Ph.D. in Computer engineering in 1995. In the same year he joined Al-Isra' Private University then Princess Sumaya University for Technology. Currently he is the Dean of student’s affairs at Al-Hussein Bin Talal University. His research interest includes computer networks, object-oriented programming, data security and e-learning.

Second Author Biography: Ibrahim M. M. El Emary received the Dr. Eng. Degree in 1998 from the Electronic and Communication Department, Faculty of Engineering, Ain shams University, Egypt. From 1998 to 2002, he was Assistant Professor of Computer sciences in different faculties and academic institutions in Egypt. From 2002 to 2010, he worked as visiting Assis. Professor and Assoc. Professor of computer science and engineering in two universities in Jordan. Currently, he is a Professor of computer science and engineering at King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia. His research interests cover: various analytic and discrete event simulation techniques, performance evaluation of communication networks, application of intelligent techniques in managing computer communication network, and performing comparative studies between various policies and strategies of routing, congestion control, subnetting of computer communication networks. He published more than 150 articles in various refereed international journals and conferences covering: Computer Networks, Artificial Intelligent, Expert Systems, Software Agents, Information Retrieval, E-learning, Case Based Reasoning, Image Processing and Pattern Recognition, Wireless sensor networks, cloud computing and Robotic engineering. Also, he participates in publishing seven book chapters in three international books (Published by Springer Verlag, IGI publisher & NOVA Science Publisher) as well as Editor of two books edited by international publishers (LAP Lampert- Germany). He has been included in Marquis Who's Who in the World 2013 edition.

RFID Technology Based Attendance Management System

Sumita Nainan[1], Romin Parekh[2], Tanvi Shah[3]

[1] **Department of Electronics & Telecommunication Engineering, NMIMS University
Mumbai, Maharashtra 400 056. INDIA**

[2] **Department of Computer Engineering, NMIMS University
Mumbai, Maharashtra 400 056. INDIA**

[3] **Department of Computer Engineering, NMIMS University
Mumbai, Maharashtra 400 056. INDIA**

Abstract

RFID is a nascent technology, deeply rooted by its early developments in using radar[1] as a harbinger of adversary planes during World War II. A plethora of industries have leveraged the benefits of RFID technology for enhancements in sectors like military, sports, security, airline, animal farms, healthcare and other areas. Industry specific key applications of this technology include vehicle tracking, automated inventory management, animal monitoring, secure store checkouts, supply chain management, automatic payment, sport timing technologies, etc. This paper introduces the distinctive components of RFID technology and focuses on its core competencies: scalability and security. It will be then supplemented by a detailed synopsis of an investigation conducted to test the feasibility and practicality of RFID technology.

Keywords: *RFID technology, RFID detection, RFID applications, RFID in management, RFID components.*

1. Introduction

RFID, which stands for Radio Frequency Identification, is an automatic identification technology used for retrieving from or storing data on to RFID Tags without any physical contact [1]. An RFID system primarily comprises of RFID Tags, RFID Reader, Middleware and a Backend database. RFID Tags are uniquely and universally identified by an identification sequence, governed by the rubrics of EPCglobal Tag Data Standard[2]. A tag can either be passively activated by an RFID reader or it can actively transmit RF signals to the reader [3]. The RFID reader, through its antenna, reads the information stored on these tags when it's in its vicinity. The reader, whose effective range is based on its operational frequency, is designed to operate at a certain frequency. The operational frequency of the reader ranges from 125 KHz – 2.4 GHz [5]. The Middleware encompasses all those components that are responsible for the transmission of germane information from the reader to the backend management systems [8]. The Middleware can include hardware components like cables and connectivity ports and software components like filters that monitor network performance of the system [2, 9]. The Backend database stores individual tag identifiers to uniquely identify the roles of each tag. The database stores record entries pertaining to individual tags and its role in the system application. The RFID system is interdependent on its core components to achieve maximum efficiency and optimum performance of the application. Due to its high degree of flexibility, the system can be easily adopted for an array of applications ranging from small scale inventory cabinets to multifarious and highly agile supply chain management systems [4, 6]. Although, the cost of incorporating this technology has restricted its outreach, the technology promises to have untapped potential [10, 11].

2. Evolution of RFID

The success of RFID technology primarily centres on the advent of radio technology [12]. The developments in radio technology were a prerequisite to harness the essence of RFID technology. There is significant growth over the past couple of decades in this technology (see figure 1). RFID technology is rife in modern industries that demand data integrity and high efficiency of the system. This technology is used for tracking vehicles and goods, courier services and luggage handling [18]. Other applications include animal tracking, secure toll payments, inventory management systems, access control mechanisms, etc. Figure 1 depicts the evolution of RFID technology.

[1] Radio Detection and Ranging is a communication medium to subliminally detect objects that are miles away, invisible to the naked eye.

[2] It defines the guidelines of how key identifiers must be encoded on the tag to define industry based standardization.

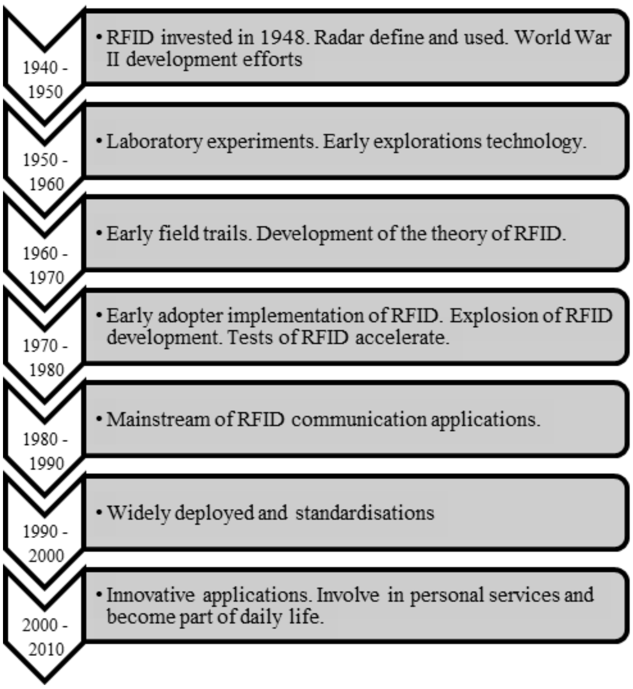

1940 - 1950
• RFID invested in 1948. Radar define and used. World War II development efforts
1950 - 1960
• Laboratory experiments. Early explorations technology.
1960 - 1970
• Early field trails. Development of the theory of RFID.
1970 - 1980
• Early adopter implementation of RFID. Explosion of RFID development. Tests of RFID accelerate.
1980 - 1990
• Mainstream of RFID communication applications.
1990 - 2000
• Widely deployed and standardisations
2000 - 2010
• Innovative applications. Involve in personal services and become part of daily life.

Fig. 1 A depiction of the evolution of RFID technology adapted from [22].

3. Components of an RFID System

An RFID system consists of various components that are connected to one another by a dedicated communication path (see figure 2). The individual components are integrated into the system to implement the benefits of RFID solution [15]. The list of components is as follows:

- Tags – an object that is attached to any product and uses a unique sequence of characters to define it. It comprises of a chip and the antenna.
- Antenna – it is responsible for the transmission of information between the reader and tag using radio waves.
- Reader – a scanning device that uses the antenna to realise the tags that are in its vicinity. It transmits signals at a certain frequencies.
- Middleware – it is a communication interface to interpret and process data being fed by the readers into information. It takes into account all relevant ports of communication and a software application to represent this information.
- Backend database – a repository of information, which is designed specific to the application. The database stores records of data specific to individual tags.

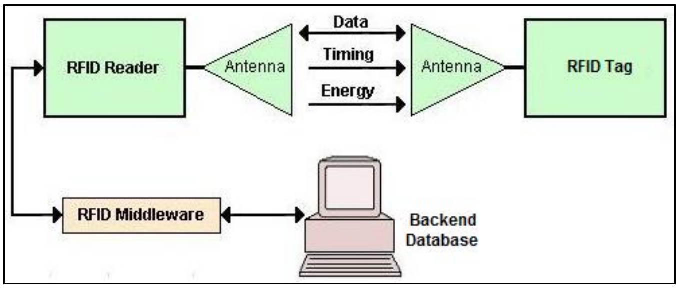


Fig. 2 Components of an RFID system

2.1 Tags

A tag consists of a microchip that stores a unique sequence identifier that is useful in identifying objects individually. The sequence is a numeric serial, which is stored in the RFID memory. The microchip includes minute circuitry and an embedded silicon chip [14, 18]. The tag memory can be permanent or re-writable, which can be re-programmed electronically by the reader multiple times. Tags are designed specific to its applications and environment. For example, paper-thin tags are attached to books in a library management system [12].

Tags are available in various shapes and sizes (see figure 3). Tags that are initiated by the reader are known as Passive tags, whilst those that do not require external initiation are called Active tags. A Semi-Passive tag exists, which has the features of both Active and Passive tags [21]. Each tag type has its distinct characteristics, which are discussed in table 1.

Tags are operable on Microwave (2.4 – 2.5 GHz), Ultra High Frequency (UHF) (860 – 1500 MHz), High Frequency (HF) (13.56 MHz) and Low Frequency (LF) (125 kHz) [22].

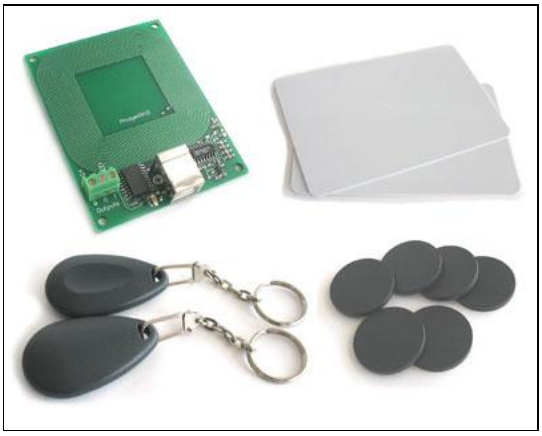


Fig. 3 Types of RFID Tags

Table 1: Features of Types of Tags

Feature	Type of Tag		
	Passive	Active	Semi – Passive
Read Range	Short (up to 10m)	Long (up to 100m)	Long (up to 100m)
Battery	No	Yes	Yes
Lifespan	Up to 20 years	Between 5-10 years	Up to 10 years
Cost	Cheap	Very expensive	Expensive
Availability	Only in field of reader	Continuous	Only in field of reader
Storage	128 bytes read/ write	128 Kbytes read/ write	128 Kbytes read/ write
Application	EZ-Pass toll payment booths	Monitor the condition of fresh produce	Measurement of temperature periodically

2.2 Antenna

The antenna is medium through which the tag and reader communicate with each other. It antenna can activate a passive tag and transfer data by emitting wireless impulses that has electromagnetic properties [20]. The antenna comes in various designs (see figure 4). They come in following types: (1) Stick antennas, (2) Di-pole or multi-pole antennas, (3) Beam-forming or phased-array element antennas, (4) Circular polarized, (5) Gate antennas, (6) Patch antennas, (7) Linear polarized, (8) Adaptive antennas, and (9) Omni directional antennas [19].

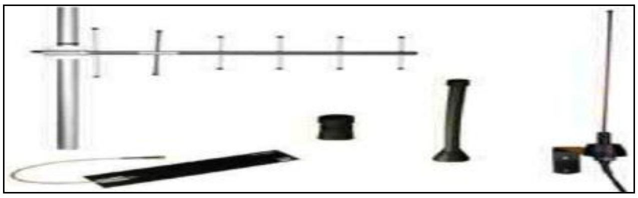


Fig. 4 Types of antenna [7]

2.3 Reader

The reader is the most fundamental part of the RFID system. It reads raw data from the tag and transmits it to the Middleware for further processing [16]. The reader attempts to interrogate the tags at varying frequencies. The reader communicates by transmitting a beam of impulses, which encapsulate commands to the tag and listens for the tag's response [14]. The reader also contains built in anti-collision processes, which allows the reader to read multiple tags simultaneously [15]. The reader is connected to the computer for data processing via a USB cable or over a wireless connection.

2.4 Middleware

The middleware is an interface required to manage the flow of data from the reader and to transmit it efficiently to the backend database management systems [18]. The middleware monitors the number of tags present in the system and extracts relevant information from the readers [12]

2.5 Backend Database

The backend database primarily deals with the storage of relevant information recorded by the reader and communicated by the middleware [16]. For example, the middleware in an automated security control system will store all tag readings taken by the reader in the database. This helps create log entries for the system [19].

4. Research

RFID technology has a widened horizon as it transcends into an era of emerging applications [1]. A detailed research must be conducted to assay the limitations and feasibility of implementing an RFID system [3, 4]. This paper focuses on the development of an attendance management system using RFID technology to monitor the attendance for a group of students [2]. This paper attempts to evaluate the benefits of implementing RFID technology to an existing system. The implementation of RFID in student management will provide additional capabilities like high efficiency and overall ease in management of the system [11]. The objectives of the research should be clearly organised to successfully develop the system.

5. Application Description

The primary aim of the research is to uniquely identify individual students based on their unique tag identifiers [22]. The research should shower light on how scalable and efficient the system is [15]. A systematic and serialised approach is required to solve this conundrum. The key characteristics of the application include:

- Perform automated attendance
- Generate report of attendees for a particular course
- Error free tag identifier detection
- Easy scalability to incorporate more records
- Integrity and security in data storage

This paper concentrates on the principal purpose to overcome the human errors while recording student attendance and the creation of a data centric student attendance database system with an improved overall efficiency. The application graphical user interface (GUI) is designed using Visual Basic 6.0[3] and Microsoft Access is used as the database provider. The Atmel[4] AT89S52 is the heart of the system, which is a low-power high performance CMOS 8-bit microcomputer with 8K bytes of downloadable flash programmable and erasable read only memory [11]. It is operable in two modes namely (1) Idle mode and (2) Power down mode [9, 11]. The microcontroller can be programmed with the 80C51 instruction set along with additional standardised features like:

- 256 bytes of RAM[5]
- 32 Input/ Output data lines
- Three 16bit timers/ counters
- SPI[6] serial interface
- Power off flag

The circuit contains a 16x2 LCD[7] display panel, which is the output device of the system [17, 19]. It displays the user's information when the stored tag is read by the reader. The serial interface allows connectivity to a local database for data storage and retrieval [20]. The input to the system is the unique tag identifier stored in the RF tag, which is sensed by the reader [21]. The components are mounted on the printed circuit board for interconnectivity between them.

[3] Visual Basic is a high level programming language developed by Microsoft.

[4] Atmel Corporation is a worldwide leader in the design and manufacture of microcontrollers, capacitive touch solutions, advanced logic, mixed-signal, non-volatile memory and radio frequency (RF) components [9].

[5] RAM is an acronym for Random Access Memory, which is a volatile type of memory required by the computer at runtime.

[6] SPI is an acronym for Serial Peripheral Interface, which is 4-wire serial communications interface to provide stable rate of data transferring [6].

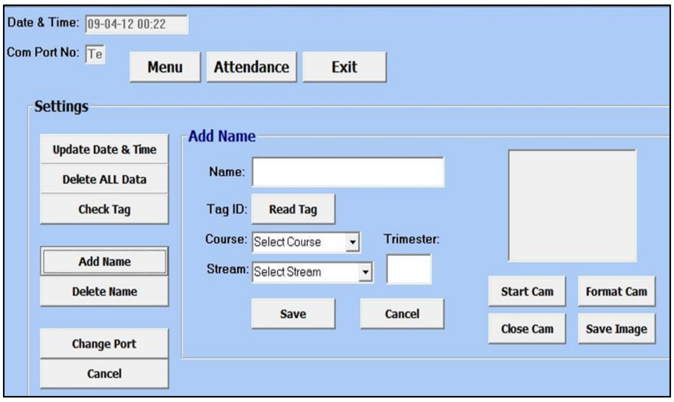
[7] LCD stands for Liquid Crystal Display, which paints a picture on the screen by correcting the orientation of the liquid crystals by applying alternating currents [9].

The software module of the middleware processes the raw data fed in by the hardware circuit. The raw data fed into the middleware are:

- Unique tag sequence number
- Timestamp of data entry

The middleware obtains the unique identifier from the reader and compares it with the list of stored tags. If the identifier sequence is present, then the details are fetched and displayed on the LCD display and the GUI (see figure 5). If the identifier is not present then a new record is created with the corresponding timestamp and it is stored in the database. The student will be prompted to fill in the following details:

- Name
- Course details
  - → Course
  - → Stream
  - → Trimester



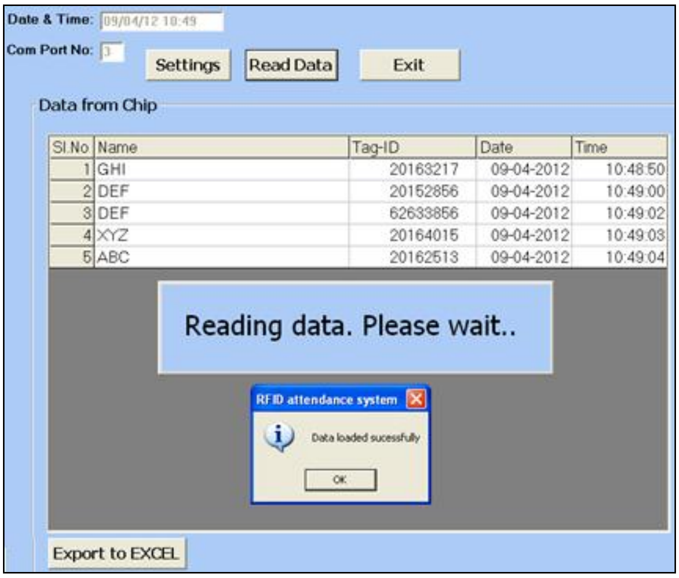
Date & Time: 09-04-12 00:22

Com Port No: Te

Fig. 5 GUI form to enter new student details

Figure 5 shows the new student registration page drafted using Visual Basic 6.0. The added functionality of capturing an image of the student provides visual authenticity whilst recording attendance. Data once stored in the database can only be modified by the system admin.

The RFID reader used in this research operates at a frequency of 125 KHz with an effective read range of 10cm only [13]. A short read range is preferred so as to maintain the authenticity and security of the attendance being recorded. Figure 6 depicts the display on the GUI as the system in the process of recording attendance. Data being recorded can be easily exported to a Microsoft Excel file for report generation. The database can be easily scaled to incorporate more details about the student.



Date & Time: 09/04/12 10:49

Com Port No: 3

Data from Chip

Sl.No	Name	Tag-ID	Date	Time
1	GHI	20163217	09-04-2012	10:48:50
2	DEF	20152856	09-04-2012	10:49:00
3	DEF	62633856	09-04-2012	10:49:02
4	XYZ	20164015	09-04-2012	10:49:03
5	ABC	20162513	09-04-2012	10:49:04

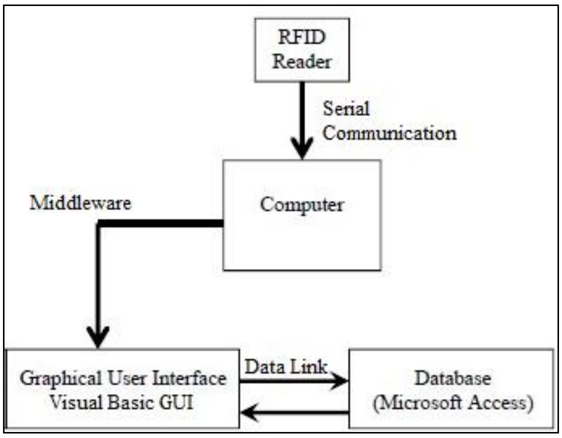
Reading data. Please wait..

RFID attendance system

Data loaded sucessfully

Fig 6 Recording student attendance

The overall system design is holistically depicted in figure 7, which is a block representation of the system. The figure shows the interconnection of two modules which are RFID module and Visual Basic 6.0 module. On the contrary, figure 8 displays the actual experimental setup of the research along with its individual components. The implementation of RFID technology in the system must be evaluated in a holistic to quantify its success.



RFID Reader
Serial Communication
Middleware
Computer
Graphical User Interface Visual Basic GUI
Data Link
Database (Microsoft Access)

Fig. 6 Block representation of RFID system

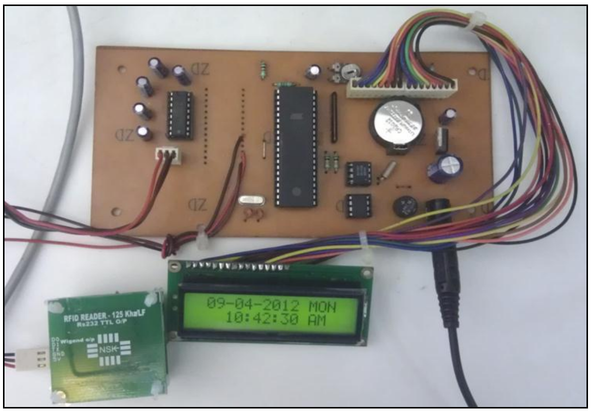


Fig. 7 Experimental setup

6. Results

The research was conducted on a sample of 60 students, enrolled in a particular course. The implementation of RFID technology has definitely quickened the entire of process of recording attendance. The traditional method of recording attendance involves individual manual entry; an arduous and a time consuming process. On average, based on experiment, the total time taken to record the attendance of a class of 60 students by manual entry method took approximately 10 minutes. This implies that approximately 10 seconds per student was required to record their attendance. This time duration includes visual and written authentication, after which the teacher records the attendance. In comparison (see figure 8), the total time taken for recording the attendance of 60 students using barcode and RFID technology is 120 seconds and 12 seconds respectively (see table 2). Based on the relationship obtained, a projection for a batch of 100 students is also forecasted.

Table 2: Results of the Study

Method	Total Number of Students			
	1	10	60	100
Manual Entry	10 seconds	100 seconds	600 seconds	1000 seconds
Bar Code	2 seconds	20 seconds	120 seconds	200 seconds
RFID technology	0.2 seconds	2 seconds	12 seconds	20 seconds

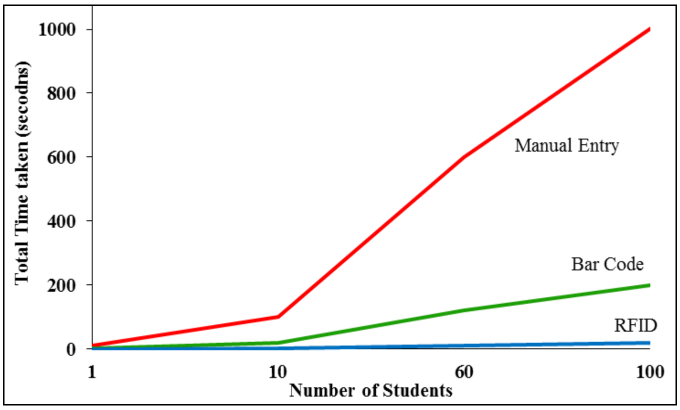


Fig 8 A line graph showing the comparison of total time taken to record the attendance of students.

As shown in table 2, compared with the time consumption in data entry for different technologies, RFID technology saves considerable amount of time and greatly improves the operation efficiency. Also with the adoption of this technology the process and product quality can be improved due to reduction in entry errors by manual human operations. Therefore, labour cost is reduced to perform the value added functions.

7. Conclusion

The study has identified and explained the key benefits of RFID technology. RFID will open doors to a pool of applications from a plethora of industries [8]. Although the focal challenge to thwart the adoption is its investment cost, RFID technology provides an ocean of lucrative business opportunities that could convince several firms adopt it [14]. The first part of the paper explains the evolution of RFID technology and the role of its individual components within the system. The second part of the paper discusses the feasibility of employing RFID technology and how it is benefactor of improved efficiency at lowered costs. The last part of the paper highlights one of the numerous practical implementations of RFID technology.

RFID technology definitely promises an increased effectiveness and improved efficiency for business processes [22]. In the long run, with reducing unit tag and reader costs, several businesses will be able to leverage the benefits of RFID technology.

References

[1] L. Sandip, “RFID Sourcebook”, IBM Press, USA, (2005) ISBN: 0-13-185137-3.

[2] E. Zeisel & R. Sabella, “RFID+”, Exam Cram, (2006), ISBN: 0-7897-3504-0.

[3] US. Department of Homeland Security, “Additional Guidence and Security Controls are needed over Systems using RFID and DHS”, Department of Homeland Security (Office of Inspector General), (2006), OIG-06-53.

[4] US. Department of Homeland Security, “Enhanced Security Controls needed for US-Visit’s System using RFID Technology”, Department of Homeland Security (Office of Inspector General), (2006), OIG-06-39.

[5] US. Government Accountability Office, “Information Security: Radio Frequency Identification Technology in the Federal Government”, (2005), Report to Congressional Requesters, GAO-05-551.

[6] K. Ahsan, H. Shah, P. Kingston, “Role of Enterprise Architecture in healthcare IT”, Proceeding ITNG2009, (2009), IEEE.

[7] Intermec, “ABCs of RFID: Understanding and using radio frequency identification”, White Paper, (2009).

[8] Juels, A., Weis, S.A. (2005). Authenticating Pervasive Devices with Human Protocols. Advances in Cryptology – Crypto ’05. Lecture Notes in Computer Science. Volume 3621. pp 293-308.

[9] S. Shepard, (2005), “RFID Radio Frequency Identification”, (2005), USA, ISBN:0-07-144299-5.

[10] Roy, W. (2006). An Introduction to RFID Technology. Pervasive Computing and Communications. IEEE Press. pp 25-33.

[11] International Organization for Standardization (ISO). (2003). Identification cards -- Contactless integrated circuit(s) cards -- Vicinity cards. ISO/IEC 14443.

[12] R. Want, “Enabling Ubiquitous Sensing with RFID,” *Computer*, vol. 37, no. 4, 2004, pp. 84–86.

[13] Ibid. (2004). RFID for Item Management. ISO/IEC 18000.

[14] RFID Journal. (2003). Gillette Confirms RFID Purchase. RFID Journal. Available at: http://www.rfidjournal.com/article/articleview/258/1/1/.

[15] Krane, J. (2003). Benetton clothing to carry tiny tracking transmitters. Associated Press.

[16] Albrecht, K., and McIntyre, L. (2005). Spychips : How Major Corporations and Government Plan to Track Your Every Move with RFID. Nelson Current Publishing.

[17] Rieback, M.R., Crispo, B., Tanenbaum, A.S. (2006). Is your cat infected with a computer virus? Pervasive Computing and Communications. IEEE Press. pp 169-179.

[18] Stockman, H. (1948). Communication by Means of Reflected Power. Proceedings of the Institute of Radio Engineers. October. pp 1196-1204.

[19] J. Schwieren1, G. Vossen, “A Design and Development Methodology for Mobile RFID Applications based on the ID-Services Middleware Architecture”, IEEE Computer Society, (2009), Tenth International Conference on Mobile Data Management: Systems, Service and Middleware.

[20] J. Bohn, “Prototypical implementation of location-aware services based on a middleware architecture for super-distributed RFID tag infrastructures”, Pers Ubiquit omputing, (2008) Journal 12:155-166.

[21] Application Notes, “Introduction to RFID Technology” CAENRFID: The Art of Identification (2008).

[22] L. Srivastava, RFID: Technology, Applications and Policy Implications, Presentation, International Telecommunication Union, Kenya, (2005).

Research on Reliability and Cost Integrated Optimization Algorithm of Construction Project Logistics System

Xiaoping Bai[1], Xiaomin GU[2]

**[1] School of Management, Xi'an University of Architecture and Technology,
Xi'an,710055, China**

**[2] School of Management, Xi'an University of Architecture and Technology,
Xi'an,710055, China**

Abstract

The structure of the construction project logistics system is decomposed in detail; some uncertain factors affecting system reliability are analyzed. This paper applies the probability-influence coordinate graph to screen out the logistics subsystem that has great influence on system reliability under the conditions of occurring failure, and establishes an allocation model of the reliability index in construction project logistics system based on the restriction of cost, makes use of the presented model and algorithm to calculate the cost-based reliability, contrasts it with the index reliability assigned by the scoring method to the optimal distribution value of reliability index and the optimal cost. The presented detailed methods and steps can offer the meaningful reference for reliability optimization management of the logistics system in the construction project.

Keywords: Reliability, Optimization, Algorithm, Costs, Logistics, Construction, Project management

1. Introduction

In the process of engineering construction, there are purchasing, transportation, safekeeping, inventory and other activities of a lot of special mechanical equipment, raw materials, and preliminary products. Considering some characteristics existed in construction projects, such as high investment, large scale, complex technology, long construction cycle, having an important influence on the development of national economy, and etc, studying the reliability of logistics system in construction project is very necessary. However, the high reliability of the construction project logistics system often means increasing the investment cost. The purpose of research is to make the operation of logistics system of construction to meet the requirements of the corresponding reliability index, at the same time to realize the optimal investment cost and to guarantee to transport the correct amount of materials and equipment to the right place in the right time.

2. Detailed analyzing some related references

Until now, there have been many references studying construction logistics, and many concepts about it have been set up. In reference [1], a best-in-class solution to the supplier selection problem has been presented by means of an intelligent evaluation engine to rank suppliers via a hybrid fuzzy mechanism. The proposed mechanism has been carefully implemented and verified via a real world case study in a large building and construction corporation [1]. The reference [2] considers the applicability of logistics management in construction and facilitates a better understanding of construction supply chains by studying the logistical functions of builders' merchants [2]. The reference [3] aims to identify the possible savings in time and cost due to different amounts of buffer stock on site, by introducing an activity-based simulation model, details and data of a residential project involving substantial amounts of pre-cast components are collected, the project participants are asked to unveil the constraints on site and throughout the material delivery and storage processes [3]. The reference [4] reports a research that employs logistic regression (LR) to predict the probable relationship between negotiator tactics and negotiation outcomes, to achieve this, three main stages of work were involved, and Negotiator tactics and negotiation outcomes were identified from the literature [4]. The reference [5] presents the analysis of three chosen variants of supply the construction in building materials, the costs connected with the supply the construction in building materials and the benefit as an effect of deduction in price in materials are discussed [5]. In reference [6], based on the principle of cyberspace for a workshop on meta-synthetic engineering, real-time dispatching command system for cement and fly-ash in Three Gorges Project was developed [6]. The reference [7] determines the optimal carrier selection based on a multi-commodity reliability criterion for a logistics network subject to budget, a genetic algorithm integrating minimal paths and Recursive Sum of Disjoint Products is developed to identify an optimal carrier selection strategy [7]. In reference [8], new procurement strategies have been developed by both public and private sectors to focus on the R, M and S characteristics inherent to the design of a system. One such strategy known as Performance Based Logistics (PBL) has gained popularity due to its success in improving the operational effectiveness of the system [8]. The reference [9] discussed the characteristics, heterosexual, timeliness and one-time, about a construction project, and points out the uncertainty factors which is due to its characteristics, in the logistics system of construction projects, but did not do any further analysis to put these uncertainty factors on the influence of construction project logistics system reliability [9]. The reference [10] constructs the structure frame model of the construction project logistics system, the key link include owner, design business, the contractor and supplier; discussed that the establishment of a specialized logistics management department and the study of their operation process and responsibility is necessary, it has a great help to the division of the construction project logistics system's logic structure, the division is the foundation for the research of reliability allocation problems [10]. The reference [11] proposes a dynamic stocking policy that adaptively replenishes the inventory to meet the time-varying parts demand; the study provides theoretical insights into the performance-driven service operation in the context of changing system fleet size due to new installations [11], and etc [12-14].

By detailed analyzing these references, we find that they general studies the optimization problems of the construction project logistics system from two independent sides including reliability and cost, and many meaningful outcomes has been gotten, but the special studies aiming to reliability and cost integrated optimization research of this system are scarce. However, the high reliability of the construction project logistics system often means increasing the investment cost. This paper considers these integrated factors, and studies how making the operation of logistics system of construction to meet the requirements of the corresponding reliability index, at the same time to realize the optimal investment cost.

3. Construction project logistics system

3.1 Characteristics of construction project logistics

The essence of engineering project construction is the consumption of materials. The final purpose of the construction project logistics system is transporting the right quantity equipment, raw materials to the right place at the right time to meet the requirements of the project progress and quality. The different characteristics distinguished construction project logistics system from general logistic system include: 1) disposable, just exist for a construction project; 2) uncertainty; 3) supply chain end when the project completion, 4) high risk, the occurrence of risk always lead to serious financial loss; 5) system reliability is complex, and controllability is weak .

3.2 The Structure of construction project logistics

The construction project logistics system is a complex system consisted of project owner, design unit, the general contractor, professional subcontractors, material suppliers, mechanical equipment suppliers by certain contract relations. The owner is an eventual member of the logistics chain, which plays a leading role.

Aim at a certain construction project, the structure of the construction project logistics system is shown as Fig. 1

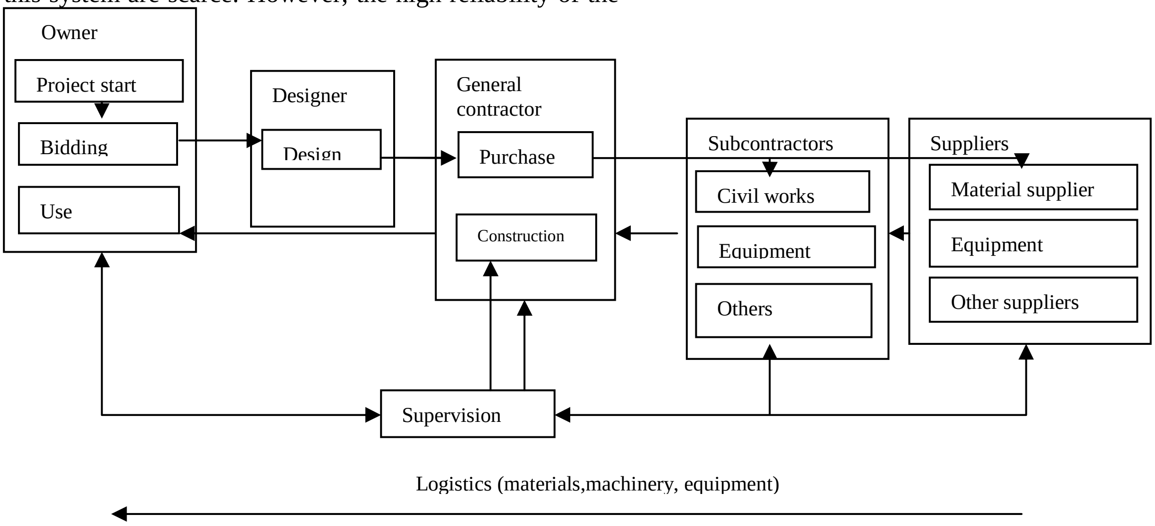


Fig. 1 the structure diagram of construction project logistics system

4. The uncertainty factors affecting the reliability

4.1 The definition of reliability of construction project logistics system

Reliability refers to the ability to complete required function within the prescribed time and regulations conditions. In the point of this paper, the reliability for construction project logistics system refers to the ability that, through the organization and coordination, construction project deals in completing the material supply, storage, fabrication processing, human resource supply, site layout, equipment layout, site logistics management and all exchanging of information related logistics and service flow in the provision of time, quantity and quality to guarantee delivering for use on schedule smoothly.

4.2 Analyzing uncertainty factors

1) Uncertainties caused by project participants
The operation of the project is influenced greatly by the owner, designer, supervisors and the general contractor. During the construction, some unforeseen and changeable cases frequently occur, every detail changes need timely logistics guarantee. Such as the change of civil sub-project schedule will cause changes of concrete products on time or quantity in demand, it will extend to the quantity and time of raw material supply, eventually can cause adjustment of the whole storage, transportation, loading and unloading handling system of construction project. Such schedule delay, material supply delay, and etc., will inevitably lead to the changes of the local material supply scheme. But the logistics system management can't obtain the material supply quantity and exact time information.
2) Uncertainties caused by logistics operation link
The equipment or materials transportation and distribution are greatly influenced by the climate conditions and geographical environment. For arriving to the construction site on time, flexible time should be fully considered in logistics solutions. The contractor usually contracts the logistics transportation business to the third party, which is logistics enterprise. Such logistics operation is not in the unified management environment, which increases uncertainties. Therefore, the selection of the logistics enterprise, transportation capacity and credibility of the enterprise should be taken into consideration.
3) The uncertain factors caused by management of the construction project logistics system
Another uncertainty affecting the reliability of logistics system is the organization management. The lack of enforceable reliability norms causes the uncertainty of managing. Project management of our country uses matrix functional organization, although there is a special material management department in construction projects, but stick division and department division are serious, the rights of materials department are weakened, it takes difficult for system organization, and logistics system management.

5. The fault condition impact assessment of construction project logistics system

According to the influence of fault state on the reliability parameters, sorts the system by the influence degree, which will intuitively show which events have a high failure rate, and which events have high impact resistance. The analysis result can be expressed in the probability-influence coordinate chart, shown as Fig. 2.

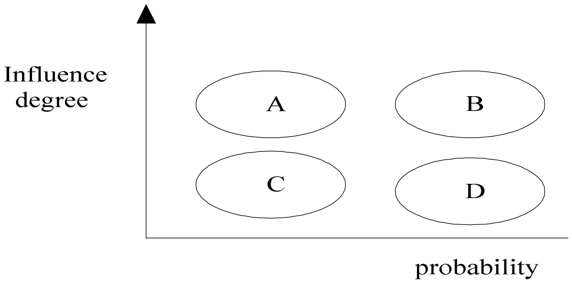


Fig. 2. Probability - impact coordinate chart

By state analyzing in Figure 2, the failure happed in B area is high probability and high influence events, once it occurs, which will take serious harm to the system reliability. So the event in the B area must be taken a key consideration in distribution of reliability index. Large operation equipments are often used in construction project. There are usually problems during the transportation and installation of these devices, because of no spare solution. If any failure occurs during the logistics process, it will be affect other activities in the logistics system. Such fault is origin fault; it is also the weak links of the system. Marking the location of origin faults in the total logistics system will help the project management staff to make out corresponding strategies.

6. The distribution of the reliability index

Before the project started, management staff of subproject logistics must clear about the reliability index that should be achieved in logistics link. The reliability index distribution can be used to look for the situation about index implementation of construction project logistics subsystems from aspects of human, time, resources, and find a weak link.

The high reliability often means the high cost; therefore, it is needed to find out a break-even point between the reliability requirement and the optimal cost. The minimum cost and the highest reliability indexes are an ideal state in the optimization. In the actual project, we can only be close to this state area and make a relatively reasonable choice of acceptable cost and reliability index.

6.1 Establishes the cost and reliability relationship of logistics subsystem

It is difficult to establish the function of the cost and reliability in the subsystem of construction project because of two reasons, such as 1) lack of statistical data; 2) too many factors influence the cost and the reliability relationship of subsystem unit, such as the environment, technology level and resources, and etc. Strictly speaking, the relationship between cost and reliability is not one-to-one. The characteristics of function $C_i(R_i)$ have 1) low reliability accompanying low cost, high reliability accompanying high cost; 2) cost is monotone increasing function of reliability, cost to reliability derivation is monotone increasing function.

According to the above experience, the relationship between reliability and cost of subsystem is

$$R_i = 1 - e^{-\alpha_1(C_i - \beta_i)} \quad \text{For} \quad i = 1,2,\dots n; \qquad (1)$$

6.2 Establishes the ideal cost reliability index distribution model

According to the structure of the construction project logistics system and time scale network planning arrangement, the back closely activity cannot begin until the front closely activity construction have been finished. The conclusion is drawn that this system is a complicated series-parallel mixed system. That partition logistics system conforms to the Time-Scaled Network Diagram is reasonable. Around the critical path, combines the single engineering that in parallel into a logistics unit. Finally, the system is divided into n series logistics subsystems in the critical path.

We select the minimizing cost of logistics system as the target and select the maximum reliability as a constraint, use Lagrange multiplier method for the reliability index distribution, C^* is the ideal cost, R_s^* is the reliability index. The model is as follows.

$$\begin{cases} \min \sum_{i=1}^{n} C_i \le C^* \\ \prod_{i=1}^{n} R_i \ge R^* \end{cases} \quad \text{For} \quad i = 1,2,\dots n \qquad (2)$$

Introducing Lagrange multiplier λ,

$$H = \sum_{i=1}^{n} C_i + \lambda(R_S{}^* - \prod_{i=1}^{n} R_i) \qquad (3)$$

Make $\frac{\partial H}{\partial C_i} = 0$, then $\frac{\partial R_i}{\partial C_i} = \alpha_i(1 - R_i)$, so:

$$\begin{cases} \frac{R_1}{\alpha_1(1-R_1)} = \frac{R_i}{\alpha_i(1-R_i)} \\ \prod_{i=1}^{n} R_i = R_s{}^* \end{cases} \qquad (4)$$

Resolves the equations, then get the subsystem reliability R_i, within the limits of the ideal cost, take it into the type (1).

$C_i = \beta_i - \frac{\ln(1-R_i)}{\alpha_i}$, The total cost of the system is: $C = \sum_{i=1}^{n} C_i$;

Where α_i, β_i for backlog experience parameters, β_i for the cost when the reliability is 0 in the logistics system, α_i decided the curve trend, α_i is smaller, the forepart of $C_i(R_i)$ curve slope is flat, the posterior segment is steeper, it shows that if R is small, improve R ,the cost is small, if want to improve it in the case of R is bigger, it needs high cost.

On the basis of above analysis, the cost is the only factor in the distribution of the reliability indexes, that is not scientific, the origin fault events in the construction project logistics system demand high reliability. It can't reduce the reliability index just because of the high cost, which will affect the success of the construction project because 1) The uncertainties that affect the reliability is more; 2) Lacking data on predictive each subsystem reliability; 3) The failure of logistics subsystems are not independent. So chooses scoring method to redistribute R_s^*, then gets R_i^P, contrasts with R_i^C, screening R_i^C, which do not satisfy R_i^P, in accordance with the scoring method distribution, for further optimization, then uses with the scoring method for reliability index of construction project logistics system.

The factors that the traditional evaluation method considering is not comprehensive, in view of the construction project logistics subsystems, some factors such as {important degree, complexity, the technical level of members, operation time and operation environment condition, investment cost, expressed by j said, (j = 1, 2 ... 6)} should be considered.

Calculates the score coefficient: ω_i --score coefficient of i subsystems

$$\omega_i = \frac{\prod_{j=1}^{6} r_{ij}}{\sum_{i=1}^{n}\prod_{j=1}^{6} r_{ij}} \tag{5}$$

r_{ij} : Score evaluation of the j factor in i subsystems

The required reliability index of Logistics system, R_s^*

Puts R_s^* in accordance with equal principle distributes subsystem reliability index, then subsystem reliability index is $\overline{R_s^*}$, so

$$\overline{R_s^*} = 1 - \omega_i (1 - R_i^P) \tag{6}$$

The distribution value of the i logistics system reliability is

$$R_i^P = 1 - \frac{1 - \overline{R_s^*}}{\omega_i} \text{ ; (i=1, 2, ……n)} \tag{7}$$

According to this method, allocates a reliability index to each subsystem unit step by step. Refines the reliability index, makes scientific quantitative index to the basis of construction project logistics system management. Similarly, if subsystem unit is parallel structure, then

$$R_i^P = 1 - \frac{(1 - R_s)^{\frac{1}{n}}}{\omega_i} \text{ , } i = 1, 2 \ldots n \tag{8}$$

Contrasting the two distribution methods:

1) $R_i^C = R_i^P$, Keep R_i^C , cost and reliability index are optimal;

2) $R_i^C \ge R_i^P$, The cost is ideal, and a reliability index is high;

3) $R_i^C \le R_i^P$, Screening this part of logistics subsystems, make further analysis.

Takes this situation, $R_i^C \le R_i^P$ for analysis. First, find out the coordinate position of the subsystems in fault state probability/influence. Judging on whether it belongs to the area that the high failure rate and high impact, if so, improve R_i^C , make $R_i^C = R_i^P$; if not, then making judgment according to the actual situation.

Finally, finds a set of new relative optimal reliability index $R_i^*, (i = 1, 2, \ldots n)$, then the system reliability $R_S > R_s^*$, the cost C also realized the optimization, take R_i^* in to the type (1), concluded C. Although the ideal cost is improved, but it makes sure the logistics system highly reliability operating, so the distribution of reliability index is more scientific, achieve the optimization target.

7. Conclusions

The risk of construction project extends to all the logistics activity, the reliability of its logistics system is related to the whole project economic value, and even decides the success of the construction project. Because the management of engineering construction logistics is all most extensive in our country, the reliability of logistics system operation is far more than the allowed range of variation, so the cost is out of control.

The paper puts forward a structural division of the construction logistics system, and finds out the weak links of the logistics system, makes a preliminary distribution of the system reliability index from cost angle, and then integrates more influential factors to distribute reliability index by scoring method. Contrasting the two indexes, combining with the failure probability/influence coordinates, screening out the subsystem which need to improve the reliability index, and ultimately distributes a new scientific reliability index set R_i^* , calculates the optimal cost C. The presented detailed methods and steps can offer the meaningful reference for reliability optimization management of the logistics system in the construction project.

Acknowledgments

This work was supported in part by NSFC (59874019), Shanxi Province Education Department Research Project (12JK0803) and Shanxi Province Natural Science Basic Foundation (2010JM7004).

References

[1] Soroor, Javad ; Tarokh, Mohammad J.; Abedzadeh, Mostafa, “Automated bid ranking for decentralized coordination of construction logistics”, Automation in Construction, v 24, July 2012, pp. 111-119.

[2] Vidalakis, Christos ; Tookey, John E.; Sommerville, James, “The logistics of construction supply chains: the builders' merchant perspective, Engineering”, Construction and Architectural Management, v 18, n 1, 2011, pp. 66-81.

[3] Ng, S. Thomas ; Shi, Jonathan; Fang, Yuan , “Enhancing the logistics of construction materials through activity-based simulation approach, Engineering”, Construction and Architectural Management, v 16, n 3, May 1, 2009, pp. 224-237

[4] Yiu, Tak Wing; Cheung, Sai On; Chow, Pui Ting, “Logistic regression modeling of construction negotiation outcomes”, IEEE Transactions on Engineering Management, v 55, n 3, 2008, pp. 468-478.

[5] Czobot, Pawel, “The analysis of the logistic costs of operating the construction project, as an example of individual house Prace”, Naukowe Instytutu Budownictwa Politechniki Wroclawskiej, n 91, 2008, p p.37-44

[6] Fei, Qi ; Chen, Xue-Guang; Wang, Hong-Wei; Liu, Zhen-Yuan, “Application of cyberspace for workshop of meta-synthetic engineering in logistics of large scale construction projects - Real-time dispatching command system for cement and fly-ash in Three Gorges Project” , System Engineering Theory and Practice, v 31, n SUPPL. 1, October 2011, p p.171-180

[7] Lin, Yi-Kuei; Yeh, Cheng-Ta, “Carrier selection optimization based on multi-commodity reliability criterion for a stochastic logistics network under a budget constraint”, International Journal of Innovative Computing, Information and Control, v 8, n 8, August 2012, pp. 5439-5453

[8] Kumar, U. Dinesh; Nowicki, David; Verma, Dinesh, “A goal programming model for optimizing reliability, maintainability and supportability under performance based logistics”, International Journal of Reliability, Quality and Safety Engineering, v 14, n 3, June 2007, pp. 251-261

[9] Tai xin Chen; “General Contract Project Logistics Management Development Status and Future Prospects [J]. Logistics engineering and management”, 2010, (03): pp. 40-42

[10] Jianxin You, Yiping CAI, “A Framework Model of Engineering Projects Logistics [J]”. Industrial engineering and management, 2006, (06): pp.49- 52

[11] Jin, Tongdan; Tian, Yu, “Optimizing reliability and service parts logistics for a time-varying installed base”, European Journal of Operational Research, v 218, n 1, April 1, 2012, pp. 152-162.

[12] Volovoi, Vitali ; Peterson, David K. “Coupling reliability and logistical considerations for complex system of systems using stochastic Petri nets”, Proceedings - Winter Simulation Conference, 2011, Proceedings of the 2011 Winter Simulation Conference, WSC 2011, pp. 1746-1757

[13] Panda, Chinmayananda ; Patro, Surya Narayan; Das, Pradipta Kumar; Gantayat, Pradosh Kumar,"Node reliability in WDM optical network" , International Journal of Computer Science Issues, v 9, n 2 2-3, 2012, pp. 315-320

[14] Singh, Ak. Ashakumar; Thingujam, Momtaz,"Fuzzy ID3 Decision Tree Approach for Network Reliability Estimation" ,International Journal of Computer Science Issues, v 9, n 1 1-1, 2012, pp. 446-450

Xiaoping BAI Associate Professor of Xi’an University of Architecture and Technology. His research interests include computer engineering, Operations research, system engineering, industrial engineering, and etc. He is Corresponding author of this paper. E-Mail:xxpp8899@126.com.

Xiaomin GU Master candidate of Xi’an University of Architecture and Technology, Her research interests include computer engineering, operations research, and system engineering. E-Mail: 375043640@qq.com

Empirical Studies on Community Structure for Networked Web Services

Yuanbin Han[1], Shizhan Chen[1,*] and Zhiyong Feng[1]

[1] **School of Computer Science and Technology, Tianjin University,
Tianjin, 300072, China**

Abstract

This paper presents studies on detecting community structure in web services formed network, which can significantly explore and understand the underlying functionality and behavior of interactions among web services, as well as facilitate the state of art service computing. The community structure in this paper focusing on two typical social characteristics for networked web services: competition and collaboration. Competition-oriented community structure is based on the functional semantics (i.e. the *inputs/outputs* of web services), in which we group web services sharing common interests. Collaboration-oriented community structure is computed by the topological analysis, so that we can cluster web services that interact densely. We present empirical analysis on our dataset and the generated communities for capturing the insight dynamics for web services formed network. Besides, we also present some potential utilities which can accelerate service-oriented computing.

Keywords: *Web service, Network analysis, Semantic, Community structure.*

1. Introduction

Service-oriented computing (SOC) has attracted much attention during recent years[1], which allows people reusing loosely-coupled software applications, by means of service discovering and composing. As the new related innovations continued to emerge, such as cloud computing [2, 3], software as a service, Internet of services [4], service computing plays an increasingly important role to date. Graph-based web service network opens new possibilities for handling the tremendous increase of web services. Graph-based web service networks are based on the interactions among web services. Since using networked web services can capture the pre-computing of some potential composing patterns, it can efficiently construct composite services, which yield lower time complexity, compared with most of other AI-based approaches [5].

Due to advances in network analysis techniques, recently, there has been a considerable amount of efforts focusing on the network analysis for web services formed network [6-12], which can facilitate the state of art SOC. Community structure is the common feature in complex networks [13-15], which describes nodes (actors) who interact heavily or share some common interests. Hence, detecting community structure in web services formed network can explore and understand the underlying functionality and behavior of interactions among web services, as well as simplify large-scaled service-based networks, which can be essentially helpful for service computing.

To accelerate the SOC, *a key issue is how to help users easily understand and navigate the behavior of web service ecosystem*. The objective of this paper is to detect significant communities, as well as explore and understand the underlying functionality and behavior web services, as well as facilitate the state of art service computing. In this paper, we proposed two approaches for mining community structure in web services formed network, which combine semantic techniques and network topological characteristics. As we will show in this paper, these findings can explore both the competition and collaboration features for networked web services.

The roadmap of this paper is structured as follows. Section 2 describes the related work. Section 3 shows the dataset we used, as well as the web service network model. Section 4 presents two approaches for mining community structure, as well as the experimental results, and demonstrates the dynamics of the proposed community structure. Section 5 presents discussions about the proposed methods, while Section 6 presents conclusion of this paper, as well as the limitation and future direction.

2. Related Work

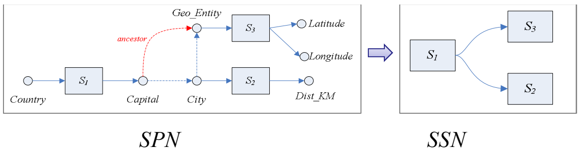
Previous studies have shown that using web services formed network can significantly benefit service composing process, particularly in terms of efficiency. [25, 26] proposed composing methods based on a graph model, and stated that the methods can achieve an acceptable performance. Shin et al. [5] extended the dependency graph by considering the services also as essential

[*] Corresponding author

stakeholder in the dependency relations, and leveraging a two-layered graph model for indexing functional semantics. Experimental results showed their work yielded lower complexity than most of other AI-based approaches.

As already stated in this paper, there has been a tremendous amount of efforts on network analysis for networked services, which are briefly following two research lines [1]: the bottom-up network analysis[6-9,11,12,18] focusing on the network formed by the services repository, and the top-down manner network analysis which the service-based network are triggered by the wisdom of crowds[10]. These efforts are generally motivated by the achievements in the area of complex network and social network, for the purpose of exploring meaningful phenomena and laws of service-based network. Additionally, recently we have witnessed some efforts on network analysis on a new kind of emerged web service called Open API and its composited application Mashup [27]. For instance, [28, 29] proposed to study the network properties for Open API and Mashup Ecosystem by leveraging 2-mode network models, in which the derivative 1-mode model can reveal the insight collaborative laws of Open APIs.

As to the community structure for web services, the Self-Serv project [30, 31] considered community as grouping of services with similarity measures. In [32], the authors presented a framework for gathering services with similar function into communities by combining argumentative agents. By the similar idea with [30, 31], Liu et al. [33, 34] comprised the "*service pool*" and task template in constructing service communities for bridging end-users and services computing. Additionally, the authors also implement a composition approach based on their proposed service community structure. We regard the above efforts as the similar idea of "*abstract services layer*" in our previous project [11], in which the abstract services are functional indices of concrete services by collecting functional similar concrete services. The work presented in this paper can be viewed as a second abstracting of the above mentioned communities by considering the hierarchy of ontology. Moreover, we also studied the topology-based community detection which captures the collaboration-oriented social characteristics.



Service	Inputs	Outputs	Description
S_1	*Country*	*Capital*	Provide the capital of a country
S_2	*City*	*Dist_KM*	Calculate the distance between 2 cities
S_3	*Geo_Entity*	*Latitude, Longitude*	Inform Latitude and longitude for a geographical entity

Fig. 1. An illustration of *SPN* & *SSN*.

3. Data sets and Models

3.1 Data sets.

The dataset used in our work is OWL-S Service Retrieval Test Collection (*OWLS-TC3*) [2]. We select the most popular 4 domain services in *OWLS-TC*, which related to communication, economy, education and travel domain, since they are most concerned with real-life applications. However, we also abandoned 77 parsed services since they either have no *inputs* or have no *outputs*. The sub-ontology extracted from the 23 different ontologies is used for specifying the semantics of web services, as summarized in Table 1. In doing network analysis, we comprised both *Pajek*[16] and *igraph*[17] tools for visualization and analysis of community.

Table 1: Dataset

# *parsed services* (along with *I/O*s)	867
Extracted sub-ontology size(# *classes*)	271
# *parsed services* used in this work	790
# *parameters* (both *inputs* and *outputs*)	250

3.2 Network Model

In general, functional semantics of a web service can be specified by its *inputs/outputs* parameters (in our dataset, services are stateless, which means that they have no *precondition* and *effect* information). In view of this, we formed the service dataset into two graph models: (1) Service-parameter network (*SPN*), a directed graph which describes the dataflow among web services, and (2) Service-service network (*SSN*) describes the direct relation between services, which can be transformed from *SPN*. The two network models are illustrated in Fig. 1.

[1] A detailed statements about literatures related with service-based network analysis were presented in our previous work [18].

[2] *semwebcentral.org*: http://projects.semwebcentral.org. The dataset was also used in our previous work [18], in which readers can refer to more details.

The construction process and network analysis for the two network models were presented in our previous work [18], readers can refer to our previous work for more detailed information.

4. Methodology

4.1 Community structure based on functional semantics

Community structure based on functional semantics was inspired by the visualization of *SPN*, as shown in Fig. 2, in which we use the "*dissimilarities*" [16] distance as the lines value. From Fig. 2, we can clearly witness the clustering phenomenon, where concepts (i.e. *I/O*s) tend to gather into different groups, as the circled areas point out. To further explore what exactly these clusters are, we magnify of the circled area 5 times, as the yellow circled area shown in Fig. 3(a), and we observed that the concepts in the cluster are all "*Time*" related in semantics. More specifically, they are bound up with their "*locations*" in the ontology. As the example shown in Fig. 3(b), the concepts in the same cluster are concept "*Time*" and the descendants of "*Time*".

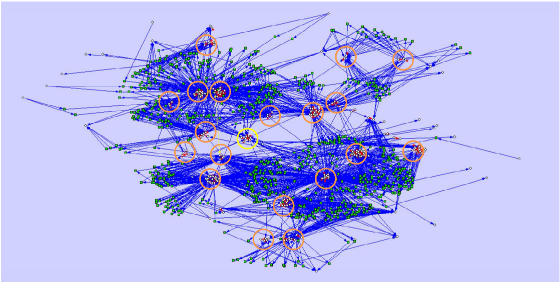
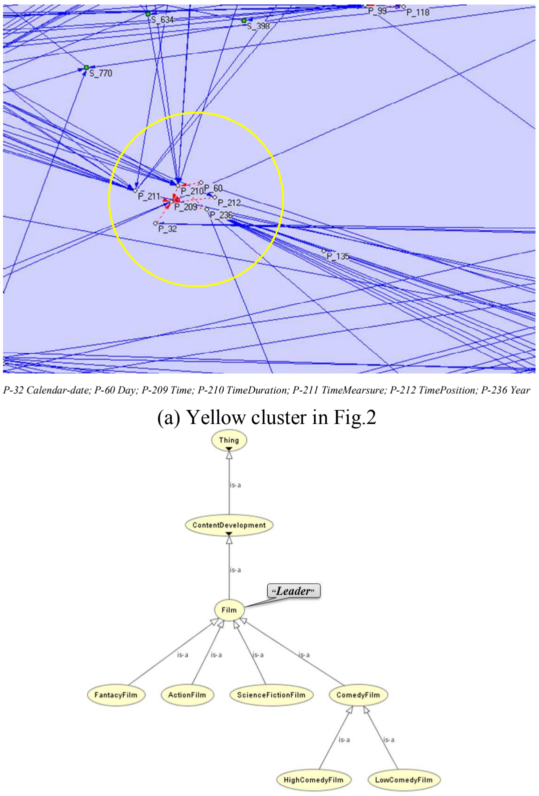


Fig. 2. Visualizations of Clusters in *SPN*, where green boxes denote services, white circles denote *I/O*s.

In view of this, we attempt to construct communities by mapping clusters in *SPN* into *SSN*, where each community implies that its members (i.e. web services) share common interests. In another angle, members in the same community also mean that they have similar function, or they can achieve similar goals while constructing web service discovering and composition for a certain user requirement. Therefore, we argue that web services in the same community hold competitive relations from the perspective of society, and we call community structure based on the functional semantics as "*competition-oriented*".

As mentioned before, concepts in the same cluster hold kinship and share a common ancestor. The idea of mining functional semantics-based service communities is simply followed as:

(1) We first divide all *output* parameters in our dataset into groups from the *I/O* ontology, by finding the *output* parameters that have no parents in the *output* set, and we denote the common ancestor shared by concepts in the same cluster as "*Leader*" for each community.
(2) Secondly, we compute all the descendants for each "*Leader*" parameter from the ontology, and form the concept community, as the example shown in Fig. 3(b) (visualized by protégé [19]), in which the "*Leader*" of the concept community is "*Film*".
(3) Finally, by constructing an inverted indexing from the service dataset, which is in terms of "*output-services*", we can compute the service community based on the concept community and mapping from the "*output-services*" inverted indexing.



P-32 Calendar-date; P-60 Day; P-209 Time; P-210 TimeDuration; P-211 TimeMearsure; P-212 TimePosition; P-236 Year

(a) Yellow cluster in Fig.2

(b) Concept community of (a)

Fig. 3 Insight of the cluster in *SPN*, with 5 times magnifying. As can be seen, all the members are time-related concepts.

By using this method, we construct 43 communities, which are marked by the "*Leader*" concepts. Fig. 4 shows the 43 resulted communities, as well as the number of concepts (including the "*Leader*") and services they maintain. Note that the total number of services is 1049, which is more than the number of 790 in the dataset, since some services may belong to more than one community.

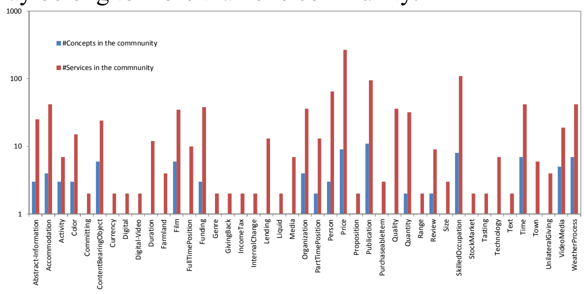


Fig. 4. Communities Based on Functional Semantics.

Competition-oriented community structure based on functional semantics can be helpful when computing composite services, since it can simplify the composing process by inter- and intra-community searching. Moreover, since we also decide the "*Leader*" of each community, it is often beneficial for refining users' requirements. For example, users often require "*Hotel*" at the beginning rather than "*3-star Hotel*".

4.2 Community structure based on topological information

Detecting community structure has been a flourishing research in the flied of social network, and there have been a large number of efforts focusing on efficiently mining communities based on topological information. In this section, we performed one of the efficient community structure detection method proposed by Pons et al. [20] to see how community structure looks like in our dataset.

The network structure is based on the *SSN* model mentioned in Section 3.1. We removed all the loops and isolated nodes (totally 97 isolated services) of *SSN* formed by the dataset in Table 1, and visualized it in Fig. 5.

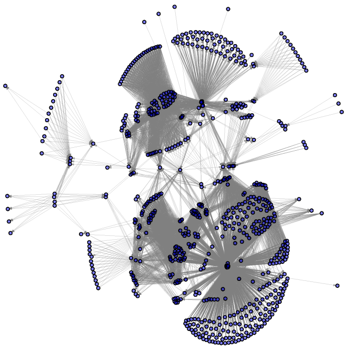
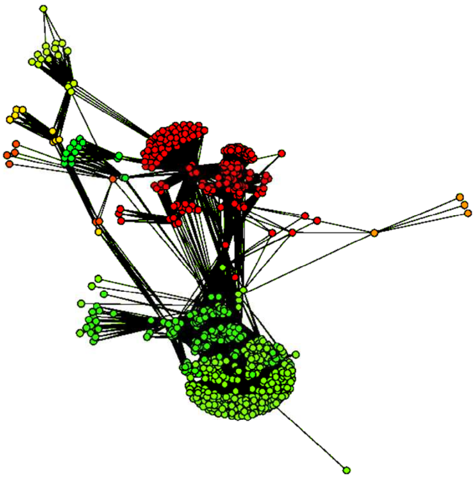


Fig. 5. Visualization of *SSN*

As can be seen from Fig. 5, the *SSN* presented in this paper is a connected direct graph, in which there are 693 web services, 13818 directed links (Note the basic statistics are a little different from the results in our previous work [18] since we removed loops and isolated nodes in *SSN*). The density and average node degree of *SSN* are 0.029 and 39.879 respectively.



Communities	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
Size	194	7	4	9	17	275	172	15

Fig. 6. Community Structure Based on *Walktrap* Method

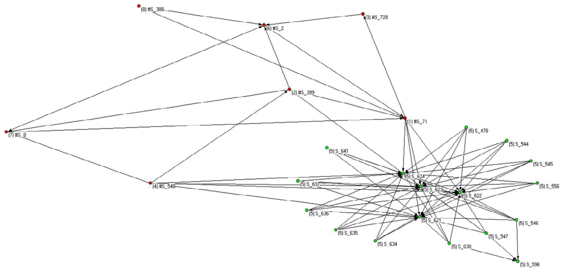


Fig.7. Microanalysis on community C_5. The red circles denotes other communities (i.e., C_1, C_2, C_3, C_4, C_6 C_7, C_8,) which are not detailed (other communities are abstracted into single vertices), green circles are the concrete members(web services) in C_5.

Walktrap [20] was proposed for mining community structure by using random walks in a graph, which is following the fundamental rule that vertices with short random walks tend to form communities. According Pons et al. [20], *Walktrap* method suffers a fairly good complexity of O (mn^2) and space O (n^2) in the worst case, and a time complexity of O (n^2log n) in the most real-world circumstances, where n and m are the numbers of vertices and links respectively in the graph. As to *SSN* we mentioned in the previous paragraph, communities can be computed in seconds. Fig. 6 is the visualization of communities (with different colors) determined by *Walktrap* method, in which 8 communities are computed.

The studying of performance about communities detecting approaches is not the primary task of this paper, since we are more interested in the insight of the function of the formed communities. To do so, we take a microanalysis on the communities by revealing the details.

Table 2. Concrete members of C_5.

WSs	*Inputs*	*Outputs*
S_{478}	Educational-Organization	Lecturer-In-Academia
S_{544}	Higher-Educational-Organization	Lecturer-In-Academia
S_{545}	Higher-Educational-Organization	Lecturer-In-Academia
S_{546}	Higher-Educational-Organization	Professor-In-Academia
S_{547}	Higher-Educational-Organization	Professor-In-Academia
S_{556}	Learning-Centred-Organization	Lecturer-In-Academia
S_{596}	Professor-In-Academia	Address
S_{621}	Researcher	Abstract-Information
S_{622}	Researcher	Address
S_{623}	Researcher	Address
S_{624}	Researcher	Postal-Address
S_{634}	University	Academic-Support-Staff
S_{635}	University	Lecturer-In-Academia
S_{636}	University	Lecturer-In-Academia
S_{637}	University	Lecturer-In-Academia
S_{638}	University	Lecturer-In-Academia
S_{641}	University	Senior-Lecturer-In-Academia

As Fig. 7 shows, services are densely connected inside the community C_5, while having less links with other communities. From Table 2, we can see that services are densely connected by sequential ties. For instance, there are four most popular services, i.e. "*S_621*", "*S_622*", "*S_623*" and "*S_624*", which have more degrees than other services in C_5, since they have the input parameter "*Researcher*" that can be called by other services with the output parameter "Researcher" (Note concepts "Lecturer-In-Academia", "*Professor-In-Academia*", "*Academic-Support-Staff*" and "*Senior-Lecturer-In-Academia*" are semantically related with "*Researcher*", thus four most popular services can be called by other services, as defined in our network model), which means that they are the "*succeed*" services of other services in the community. This demonstrates that community structure based on *Walktrap* method reveals the fact that services in the same community usually cooperate frequently, which is the feature of "*Collaboration-oriented community structure*" we mentioned.

Fig. 8 is the visualizing of another community C_6, which has the most member population among the 8 communities. It shows that C_6 also follows the fact that services are densely connected inside, while sparsely connected outside.

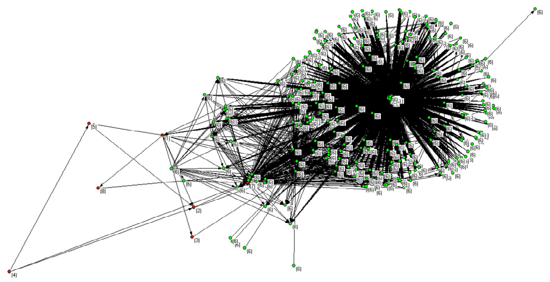


Fig.8. Visualization of Community C_6 (with 275 members).

5. Discussions

In this section, we discuss both pros and cons of the two proposed community detecting approaches, and present some potential utilities for facilitating SOC.

In the case of detecting community structure based on the functional semantics in Section 4.1, we emphasized that it is competition-oriented, since these communities group web services with similar functionality or corresponding to common requirements. Therefore, a possible usage for competition-oriented communities is that we can facilitate composing process by dividing composing into two phases: inter-community service composing and intra-community service discovery. One of our essential ongoing work is focusing on designing hierarchical service composing based on the model illustrated in Fig. 9. In the first step, we construct abstract composition by inter-community searching from users' requirements, which can quickly construct an abstract composite flow, and we can also refine users' requirements by leveraging user interactions with ontology. In the second step, service discovery is proposed for binding concrete services within the communities. Thus we can obtain the composite services by combining the two steps.

In the matter of limitations about community structure based on the functional semantics, services with same parameters may functionally differ from each other since semantics based on *input* and *output* parameters are short of sufficient context information. Therefore, services in the same community may be functionally different in some circumstances. This is serious existing in the community leading by "*Price*" summarized in Fig. 4. For instance, two services related with coffee and cars maintaining the same output "*Price*" are grouped into the same community, though they functionally differ from each other in reality context. We attempt to address this issue by extending the functional semantics with text descriptions and tags of services (Our ongoing work for annotating text descriptions and tags is by leveraging DBpedia [21], Spotlight [22] and Yago[23] ontology).

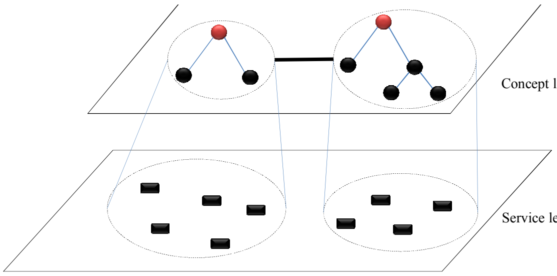


Fig.9. Two layered model for service selection and composition, where the red nodes in concept level can be used for refining users' fuzzy requirements.

For detecting services community structure based on *Walktrap* method stated in Section 3.2, we mentioned that services in the same community are densely connected inside, while sparsely connected outside. This can significantly benefit service composing, since the priority of service selection for a certain composing can be determined by communities. Therefore, a reasonable complexity can be achieved by confining services in the same community, which can characterize the priorities in composing. Furthermore, collaboration-oriented community structure can be also helpful to service recommending. We leave this as our future work.

Although mining community structure based on topological methods is gaining momentary attention in the field of complex network and social network, there have been still enormous challenges in leveraging it for real-world networks. For this paper, *Walktrap* method heavily depends on the topological information of services formed network, but seldom considers the practical semantics of networks, which will lead to the result that some small but semantically meaningful communities might be covered by large communities. It has been stated that considering semantic aspect of information in community structure detecting can achieve desirable results for practical context [24]. We leave this as our future direction for mining semantically meaningful "*Collaboration-oriented community structure*".

6. Conclusions

In this paper, we have suggested two meaningful community detecting methods, which followed two basic lines: (1) the competition-oriented community structure based on functional semantics which derived from the behavior and semantic property of web services, as well as (2) collaboration-oriented community structure computed by topological analysis.

We showed that these findings had a series of meaningful implications for service computing. Firstly, service community structure can track the challenge of constantly growth of web services, for which we can using the idea like the "*Autonomous System*" within the Internet (for web services, i.e. the community) to administer the vast amount of services, as well as to search composite services efficiently. Besides, the community structure based on functional semantics provided a gateway for user requirements refinement. Ultimately, it would be interesting to study service recommending based on the underlying attracting nature of the collaboration-oriented community structure. All these are our priority concerns in the immediate future.

As stated previously, one of the shortcomings is that we only consider the semantics of parameters, which are insufficient for representing the functional information of web services. Our ongoing work is planning to address this issue by adding the semantic information of text description, as well as tags from the wisdom of crowds. Our another interesting is concerning service communities in heterogeneous networks for OpenAPIs and Mashups, in which we consider heterogeneous types of entities(multi-mode networks) in Mashup Ecosystem, such as providers, users, APIs, Mashups, tags, data formats, protocols.

Acknowledgments

This work was funded by NSFC grant 61173155, National High-Tech Research and Development Program of China grant 2007AA01Z130 and the 985 Project of Tianjin University grant 2010XG-0009.

References

[1] M. P. Papazoglou, P. Traverso, S. Dustdar and F. Leymann, "Service-oriented computing: State of the art and research challenges", IEEE Computer, Vol.40, No.11, 2007, pp.38–45.

[2] Y. Wei, M.B. Blake, "Service-oriented Computing and Cloud Computing: Challenges and Opportunities", IEEE Internet Computing, Vol.14, 2010, pp.72-77.

[3] A. A. T. Innocent, "Cloud Infrastructure Service Management -A Review", International Journal of Computer Science Issues, Vol. 9, No 2, 2012, pp.287-292.

[4] J. Cardoso, K. Voigt, M. Winkler, "Service Engineering for the Internet of Services", in 10th International Conference on Enterprise Information Systems, 2008, pp.15-27.

[5] D. Shin, K. Lee, T. Suda, "Automated Generation of Composite Web Services Based on Functional Semantics", Journal of Web Semantics, Vol.7, No.4, 2009, pp.332-343.

[6] H. Cai, "Scale-free Web Services", in Proc. of ICWS, 2007, pp.288–295.

[7] H. Kil, S. Oh and D. Lee, "On the Topological Landscape of Web Services Matchmaking", in Proc. VLDB Workshop on Semantic Matchmaking and Resource Retrieval, 2006, pp.178: 19-34.

[8] H. Kil, S. Oh , E. Elmacioglu, W. Nam, D. Lee, "Graph Theoretic Topological Analysis of Web Service Networks", World Wide Web, Vol.12, No.3, 2009, pp.321-343.

[9] S. Oh, D. Lee, and S.R.T. Kumara, "Effective Web Service Composition in Diverse and Large-Scale Service Networks", IEEE Transactions on Services Computing, Vol.1, No.1, 2008, pp.15-32.

[10]W. Tan, J. Zhang, and I. Foster, "Network Analysis of Scientific Workflows: a Gateway to Reuse", IEEE Computer, Vol.43, No.9, 2010, pp.54-6.

[11]S. Chen, Z. Feng, H. Wang, "Service Relations and its Application in Services-Oriented Computing", Chinese Journal of Computers, Vol.33, No.11, 2010, pp.2068-2083.

[12]S. Chen, Y. Han, Z. Feng, "Social Network Analysis on the Interaction and Collaboration Behavior among Web Services", in Proc. AAAI Spring Symposium, 2012, pp. 9-15.

[13]M. Girvan, M.E.J. Newman, "Community structure in social and biological networks", in Proc. National Academy of Sciences, 2002, pp.99:7821-7826.

[14]M.E.J. Newman, "Detecting community structure in networks", European Physical Journal, Vol.38, No.2, 2004, pp.321-330.

[15]M.A. Porter, J.P. Onnela, P.J. Mucha, "Communities in Networks", Notices of the American Mathematical Society, Vol.56, No.9, 2009, pp.1082-1097.

[16]W. de Nooy, A. Mrvar, and V. Batagelj, “Exploratory Social Network Analysis with Pajek,” Cambridge University Press, 2005.

[17]G. Csardi and T. Nepusz, “The igraph software package for complex network research,” InterJournal, Vol. Complex Systems, 2006, pp. 1695.

[18]Y. Han, S. Chen, Z. Feng, “Optimizing Service Composition Network from Social Network Analysis and User Historical Composite Services”, in Proc. AAAI Spring Symposium, 2012, pp. 39-45.

[19] http://protege.stanford.edu

[20]P. Pons, M. Latapy, “Computing Communities in Large Networks Using Random Walks,” J. Graph Algorithms Appl. Vol.10, No.2, 2006, pp.191-218.

[21]C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, “DBpedia-A crystallization point for the Web of Data,” Journal of Web Semantics, Vol.7, No.3, 2009, pp.154-165.

[22]P. N. Mendes, M. Jakob, A. García-Silva and C. Bizer, “DBpedia Spotlight: Shedding Light on the Web of Documents,” in Proc. 7th International Conference on Semantic Systems (I-Semantics '11) , 2011, pp.1-8.

[23]F. M. Suchanek, G. Kasneci, G. Weikum, “Yago: a core of semantic knowledge,” in Proc. 16th international conference on World Wide Web, 2007, pp.697-706.

[24]W. Wu, Y. Xiao, Z. He, W. Wang, and T. Yu, “Mining Hidden Communities in Social Networks Based on Weight Information,” Journal of Computer Research and Development, Vol.46, No.z2, 2009, pp. 540-546.

[25]S.V. Hashemian, F. Mavaddat, “A Graph-Based Framework for Composition of Stateless Web Services,” in Proc. 4th IEEE European Conference on Web Services, 2006, pp.75-86.

[26]H. Elmaghraoui, I. Zaoui, D. Chiadmi, L. Benhlima, “Graph based E-Government web service composition”, in International Journal of Computer Science Issues, Vol.8, No 1,2011, pp.103-110.

[27]http://www.programmableweb.com/

[28]S. Yu, and C. J, Woodard, “Innovation in the Programmable Web: Characterizing the Mashup Ecosystem,” in Proc. 2nd International Workshop on Web APIs and Services Mashups, 2008, pp.136-147.

[29]G. Ji, J. Zhang, Z. Zhao, and Y. Han, “Service Collaboration Network: A Novel Mechanism for Web Service Management,” in Proc. 12th International Asia-Pacific Web Conference (APWEB '10), 2010, pp. 85-91.

[30]B. Benatallah, Quan Z. Sheng, M. Duma, “The self-serv environment for web service composition,” IEEE Internet Computing, 2003, 7: 40–48

[31]H. Y. Paik, B. Benatallah, F. Toumani, “Towards self-organizing service communities,” IEEE transactions on systems, man and cybernetics, Vol. 35, No.3, 2005, pp.408-419.

[32]J. Bentahar, Z. Maamar, D. Benslimane, P. Thiran, “An Argumentation Framework for Communities of Web Services,” IEEE Intelligent Systems, Vol. 22, No.6, 2007, 75-83.

[33]X. Liu, G. Huang, H. Mei, “A Community-Centric Approach to Automated Service Composition,” Science in China Series F: Information Sciences, Vol.53, No.1, 2010, pp. 50-63.

[34]X. Liu, G. Huang, W. Pei, H. Mei, “Discovering Homogeneous Web Service Community in the User-Centric Web Environment,” IEEE Transactions on Services Computing, Vol. 2, No.2, 2009, pp.167-181.

Yuanbin Han is currently a PhD student in Tianjin University, China. His current research interests include service computing and semantic web.

Shizhan Chen received his PhD from Tianjin University in 2010 and is a lecturer at Tianjin University, China. His current research interests include service computing and SOA.

Zhiyong Feng is a full tenured professor and head of the School of Computer Science and Technology, Tianjin University, Tianjin, China. His current research interests include knowledge engineering, service computing, and security software engineering.

The Combine Effect of Synchronous and Asynchronous E-Learning on Distance Education

Iqrar Ahmad[1], M.U. Bokhari[2]

[1]Lecturer
College of Science & Arts, King Khalid University
Mohayel Aseer—61913, Kingdom of Saudi Arabia

[2]Associate professor & Chairman
Department of Computer Science, Aligarh Muslim University
Aligarh, U.P—202002, India

Abstract

The aim of this paper is to highlight the combine effect of synchronous as well as asynchronous in E-learning environment. Both type of learning have importance in different scenario. Not one of them can fulfill whole requirement individually. Synchronous training is done in real-time with a live instructor facilitating the training. Everyone logs in at a set time and can communicate directly with the instructor and with each other. It lasts for a set amount of time - from a single session to several weeks, months or even years. Asynchronous is e-learning in the more traditional sense of the word. It involves self-paced learning, CD-ROM-based, Network-based, Intranet-based or Internet-based. It may include access to instructors through on-line bulletin boards, on-line discussion groups and e-mail.

Keywords: *Asynchronous e-Learning, real-time, self-paced learning, Synchronous e-Learning*

1. Introduction

Currently majority (over 80 %) of Distance Education Courses are paper based [1][2]. Internet based trainees are typically limited to the hypertext and graphics but not to high quality streaming video due to connection speeds obligation For example WebELS and Smart EDU e-learning platforms [3][4] supports a special type of contents i.e. slide with synchronized audio and cursor in addition to traditional multimedia contents and conceptually it is a fusion of synchronous and asynchronous e-Learning system. WebELS provides a web-based, multimedia enabled multi-platform tool, by which traditional instructors can archive their learning materials on the web and students can do their personal learning over the Internet. Uploaded contents can be used either in standalone or group learning in real-time with discussion. It supports wide range of multimedia contents including text, images, audio, video and slides with synchronized audio and cursor. Majority of the presentation video can be simulated with audio and cursor synchronized slides except introduction and conclusion parts where cameraman focuses on the face of the presenter. With audio and cursor synchronized slides, it is possible simulate presentation video which drastically reduces data volume and improves contents visualization. Such type system supports both synchronized online presentation as well as asynchronous off-line viewing SMART EDU integrates CBT and WBT functionalities with additional tools as follows:

Synchronous (Trainees and Trainer meet in the same time):
TV quality Live Video and Audio
Recorded Video and Audio
PC Screen Tests
Questions to Trainer
Chat

Asynchronous (Trainees and Trainer does P Discussion Forum not has to be present in the same time):
Knowledge Base Large File transmission

From the psychological point of view advantage of TV like quality video is tremendous for training purposes. It breaks several psychological barriers like isolation in self-learning environment and/or passivity. It is often being understood as a videoconferencing system. SMART EDU differs from videoconference in quality of provided video. Due to the Satellite communication It is able to provide trainees with TV quality video (2 mbps) that videoconference (ISDN or Internet based) is not capable to achieve. It results in lower picture resolution, and speed of move (images per second). On the other hand videoconferencing system assures feedback through video channel. Video feedback is planned to he incorporated in the Meeting mode while Conferencing and Training mode supports text based feedback. SMART EDU on top of videoconference system provides training tools of WBT, CBT not available by ordinary videoconference systems.

2. Background

2.1 Terminology of e-Learning

A novel world of learning is opening up in the knowledge economy of computer-based learning, online learning, e-learning, and distance learning. In a review of the literature (on terms such as e-learning), technology based learning and Web-based learning are defined and used differently by different organizations and user groups. However, these only form parts of the many modes of learning that will be increasingly essential as education and training becomes part of the lifelong experience of people working in knowledge based projects [5]. With regards to e-learning, Clarke classifies it as the delivery of content via all electronic media, including the Internet, intranets, extranets, satellite broadcast, audio/video tape, interactive TV, and CD-ROM. Moreover it used synonymously with technology based learning.

*2.*2 Synchronous Training

Synchronous training is done in real-time with a live instructor facilitating the training. Everyone logs in at a set time and can communicate directly with the instructor and with each other. It lasts for a set amount of time - from a single session to several weeks, months or even years. This type of training usually takes place via Internet Web sites, audio- or video-conferencing, Internet telephony, or even two-way live broadcasts to students in a classroom.[6]

2.3Asynchronous Training

This is e-learning in the more traditional sense of the word. It involves self-paced learning, CD-ROM-based, Network-based, Intranet-based or Internet-based. It may include access to instructors through on-line bulletin boards, on-line discussion groups and e-mail. Or, it may be totally self-contained with links to reference materials in place of a live instructor.[6]

3. Learning Objects

Learning objects are “windows” containing different information and each Learning Object present different media format. These objects can be mixed into various Screen Layouts. Learning objects can be handled in 2 different ways, in a Live Sessions or as Uploaded Educational materials, are:

3.1“Live session only” learning objects

Supported Learning objects in a Live Session are well explained at a picture below inside the Remote Studio application, where they are used.

Live Video (Streaming Video)

Source of Live Video is a Video Camera. Video is compressed in Video encoder (Lossy compression). Audio is normally part of a Video signal multiplexed and synchronized with video. Microphone from a Video Camera or external equipment must be connected to the sound card of Audio Encoder. Video Audio encoder performs precise time stamping to enable synchronization of video with audio at Trainee Site.

Live Audio

Another Live Audio channel *can* transmitted when needed e.g. in a case of simultaneous translations.

PC Screen

“PC Screen” object enables to transmit images of the Trainer’s PC Screen. In *this* way Trainer can show to trainees various objects a**s** Slides, Spreadsheets, Text files etc. with a slow motion requiring clean resolution of characters and numbers. PC screen uses lossless compression what is important to support clear readability of text and numbers at Trainee sites. PC Screen can be accompanied with Audio. Source of this audio can be Trainer’s speech but also PC application audio - for example when used in Power Point presentations.

Text

Text is a simple text box allowing putting a text message at the screen or describe other objects

Question

Question object is used to display currently answered (selected) Trainee’s question from interactive module. It helps Trainees to see the topic discussed.

Subtitling

Subtitling is a text which is designed to help better understand Trainer’s speech. It’s goal is to support intelligibility for example of unusual pronunciation (e.g. English speaking French man). Written text provides helpful additional information in case even simple words are not understood clearly. This helps overcome intercultural differences in language understanding.

3.2Learning objects which can be used in a live session as well as in educational materials:

Prerecorded video or audio

Pre-recorded video or audio files used in a live session or Uploaded via “Educational Material” tool are to be compressed in a supported format. There is a tool for reformatting other format types to SMART EDU supported. SMART EDU uses XVid (MPEG-4) video compression and MP3 audio compression.

Pictures

Pictures are frequently used tool for supporting explanation. All generally accepted picture formats are supported (e.g. *.jpg, *.gif, *bmp etc). It is up to the

Trainer what combination of Learning Objects she/he uses in her/his training. Screen layout can consist from a single Learning Objects such as Live Video with Audio or PC Screen with Audio only up to more objects. At a picture bellow is a sample of Screen Layout using background picture and recorded video.

“*Educational Material” Learning Objects*

Via this tool Trainer can transmit any educational materials to support training process. Educational material can present any file format recognized by Trainee PC. These can be books, slides, sample source codes, animations, simulators etc.

4. Uses and Application

The development of the global information society places new demands on the creation and delivery of learning materials and educational services. Education systems must learn to harness ICT to access a wider knowledge base and to help develop a new technology of learning. Yet in this still rapidly-changing sphere, the education world struggles to respond adequately to successive demands [7] The report “ Europe and the global information society” (May 1994)**,** produced by the high-level group chaired by Mr. Bangemann, stressed, ”that throughout the world, information and telecommunications technologies are bringing about a new industrial revolution which already looks to be as important and radical as those which preceded it, increased use of subcontracting, the development of work in teams, are some of the consequences of information technology. Information technology is contributing to the disappearance of routine and repetitive work which can be codified, programmed and automated. Work content will increasingly be made up of intelligent tasks requiring initiative and the ability to adapt. ”It is estimated that every year in the EU at least 20% of the economically active population is engaged in continuing vocational training/ education of various kinds for two weeks on average. According to a survey carried out in 1993 in 12 Member States, some 5% of male employees and 6% of female employees aged over 25 had undergone vocational training in the four weeks prior to the survey.” [8] There are detailed tables proving high grow of ICT spending for ICT in Education in OECD countries. It is based on statistics gathered from OECD countries. Most of the ICT spending is represented by Internet Access and PC purchase, but very low/ or no spending are reported for special educational ICT system offering broadband access to high quality video materials in real-time or pre- recorder. Average Internet connection speed in schools if any is very low not allowing to provide acceptable ways of tele-education. Good examples of SMART EDU applications are in life- long and vocational training. In fact SMART EDU can be used effectively in:

- Training teachers for improving ICT skills
- Creating teams of top specialists (e.g. medicine, auditors, accountants, lawyers.)
- Virtual Programming University (e.g. JAVA University)
- Sales agent trainings (e.g. in Insurance Companies)
- Retail chain trainings (e.g. electronic equipment suppliers: cameras, printers, faxes etc.)
- Medicine (theoretical + live sessions from hospitals, from surgery rooms, etc., for all specializations such as Hearth, Tropical diseases, Neurosurgery, Genetics, Cancer, etc.)
- Technical trainings **(** Telecommunication: ADSL, Satellite communication; Cryptology; OS administration: Windows, Unix, etc.; Databases)
- SW customer support trainings (e.g. for banks introducing new SW)
- Financial products trainings
- Technologies (Installation of civil and industrial plants; Car services: repair from car manufacturer; Electrical; Antitheft and alarms; Environmental protection in industries; Fire prevention.. .)
- E-Government (e.g. Ministries of Construction, Interior, Environment, Education, Tax offices, Defense & army, Custom at borderlines.)
- Tourism (e.g. Hotel chains staff trainings)
- Professional Associations (Auditors, Notaries, Accountants, etc.)
- Chamber of Commerce
- Non-profit organizations
- International networks
- Development Aid Organizations (UNDP, UNESCO etc.)
- Unemployment offices (e.g. Requalification courses)

5. Design Concepts and Goal

- E-Learning system designed to meet the requests of postgraduate education. PhD program is the main target of use
- Supporting seamless service of asynchronous and synchronous e-Learning: Asynchronous self-learning and synchronous Internet meeting using same contents
- Multi-OS system: Supports both Windows users and Mac users at same quality
- Powerful authoring features for end-users: One click uploading editor for Power Point, pdf, audio/video contents
- Multi-language interface to support international use: Automatic selection of English (standard),Japanese, Chinese, etc
- “Anywhere, anytime & anybody” system:

6. Major Objectives

- Provide a general purpose e-Learning environment for distributed and internationalized post-graduate education:

Distance learning, distance meeting, Multilanguage, synchronous/asynchronous

- Provide a variety of distance learning functions such as Internet interview, Internet conference, annotation system, on-line whiteboard.
- Provide a powerful authoring feature to assist the lack of other e-Learning platforms i.e. Integration to such as Moodle, Blackboard, WebCT etc.
- Independent offline viewing system

7. Results

The students were asked to choose an answer between A-F, according to how strongly they agree with the statement of the questions [9]. The answers are shown in Fig. 1. A total of 108 students participated in this survey. Response to survey questions 1 trough 8 5 reveal that students like to utilize the online course material as supplementary material because of the multimedia features, animations, graphics and simulations used. However, there seems to be resistance on part of the students to the idea of replacing the traditional face-to-face classes with learning fully online. A recent news report released by Cnet [10], one of the leading on-line publishing companies, suggested that an Internet-based learning system: “use a wide range of technology to make learning as easy and collaborative as possible. While the level of sophistication varies, standard to most Web courses are communication systems, such as email, real-time chat rooms, and threaded discussion groups that let students interact with instructors and each other online.” More specifically, this kind of Internet-based learning system should have the following learning and teaching features:[11]As results shown in Table 1, we had 73 student’s models in Fall 07, 113 student’s models for Winter 08, 357 student’s models for Spring 08 and 173 student’s models for Summer 08.[12]

Table 1: Student’s Usage Preferences

	Fall 07	***Winter 08***	***Spring 08***	***Summer 08***
# of Recording	388	47	756	378
# of Students Models	73	113	357	173
# of Viewings (streaming)	3542	1477	7579	2579
# of Downloads	356	356	433	434
# of Podcasts	81	81	136	136
# of Mp3	13	32	40	40
Avg. of reviewing/ student	10	13	10	15

Table 2: The features of online education

Electronic lecture notes	- Providing with student-customized Materials
Message system	Connecting the course participants so as to achieve communication and collaboration purposes
Discussion	Enabling real –time chat or threaded discussions
Interactive quizzes and self-assessment	Generating on line quizzes which are marked by the server
Course creation	Allowing the instructors to construct or modify their materials
Course management	Having a database management system which helps to organize the course materials
Student management	Having a database management system which helps to organize the students information and to track the individual user so that customized services can be provided.

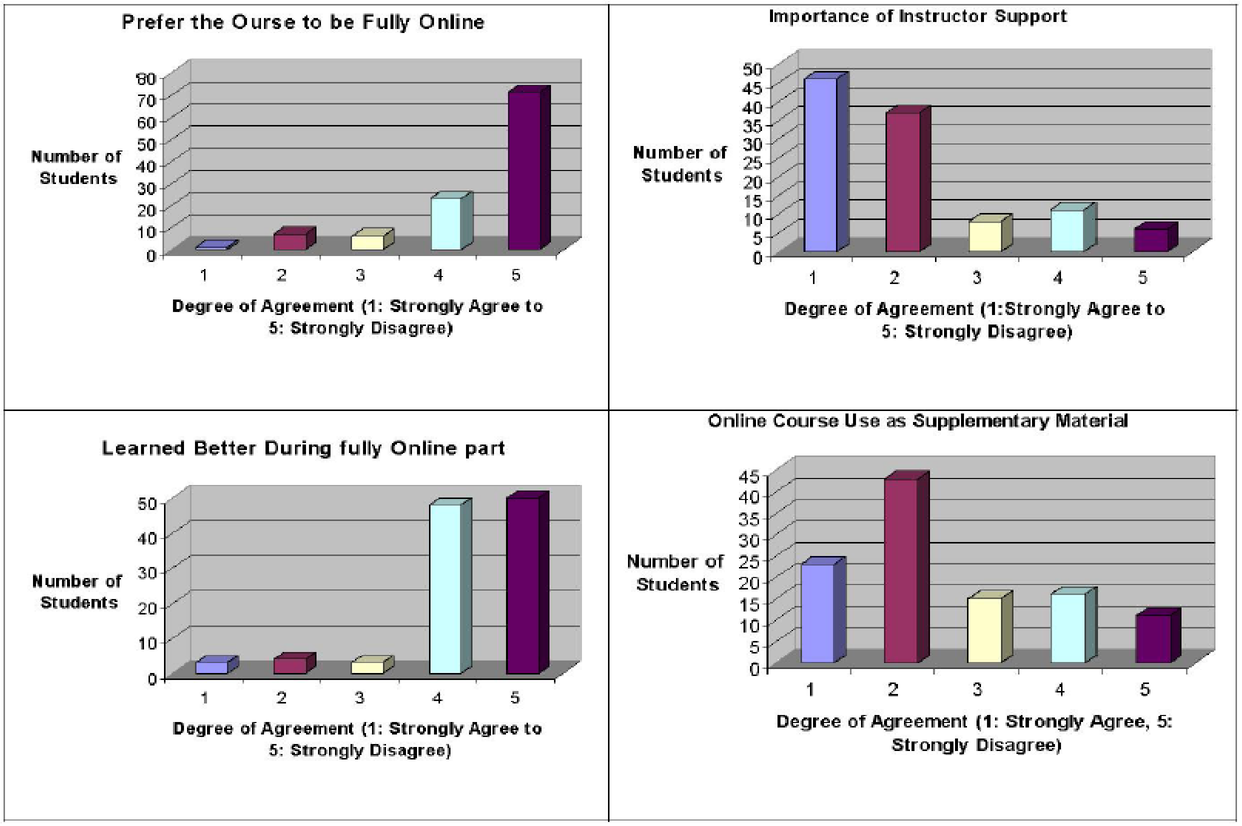


Fig. 1: Survey on 108 students

8. Conclusion

This paper discusses multimedia e-learning environment in education and conceptually it is a fusion of synchronous and asynchronous e-Learning system. It provides a web-based, multimedia enabled multi-platform tool, by which traditional instructors can archive their learning materials on the web and students can do their personal learning over the Internet. Uploaded contents can be used either in standalone or group learning in real-time with discussion. It supports wide range of multimedia contents including text, images, audio and video as well as audio and cursor synchronized slides which can be considered a special type of content and simulate video with drastic reduction in volume.

References

[1] Bokhari, M. U. and Kuraishy, S. “Enhancing the Effectiveness of ESOL Teaching with E-Learning”, Journal of Management and Technology EVOLUTION, March 2009. Published by RAMANUJAN College of Management, An ISO 9001:2000 Certified Institution, Approved by AICTE, Ministry of HRD, Govt. of India, Affiliated to Maharishi Dayanand University, Rohtak

[2] World Bank Statistics.

[3] mahfuzur, zheng, sato, vuthichai,”webels e-learning system: online and offline viewing of audio and cursor syncronised slides” in IEEE 2007.

[4] DuSan**, Statelov**, Martin “SMART EDU A new TV video enabled interactive e-learning platform” in IEEE 2003.

[5] I. Clements and Z. Xu, ELCAT: an e-learning content adaptation toolkit, *Campus-Wide Information Systems*, vol. 22(2), pp.108-122, 2005.

[6] Kalpana, Veni “Future Trends in E-Learning” in IEEE 2010 4th International Conference on Distance Learning and Education (ICDLE).

[7] Teaching and learning towards learning society White paper on education and training, Edith Cresson & Pidraig Flynn.

[8] Education Policy analysis OECD 1999.

[9] Mohandes, Dawoud, AlAmoudi, Abul ”Online Development of Digital Logic Design Course” in IEEE 206.

[10] The Net - Technical difficulties. http://www.news.com/Specialfeatures/0,5,8360,00.html.

[11] Cheng, Yen”Virtual Learning Environment (VLE): A Web-based Collaborative Learning System” in IEEE 1998.

[12] Leyla, Elizabeth, Robert” The Effectiveness of Personalization in Delivering E-learning Classes” in IEEE 2009 Second International Conferences on Advances in Computer-Human Interactions.

A New Method for Medical Image Clustering Using Genetic Algorithm

Akbar Shahrzad Khashandarag[1], Mirkamal Mirnia[2] and Aidin Sakhavati[3]

[1] **Department of Mechatronic Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran.**

[2] **Department of Mathematics Science, Tabriz University, Tabriz, Iran.**

[3] **Department of Electrical Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran.**

Abstract

Segmentation is applied in medical images when the brightness of the images becomes weaker so that making different in recognizing the tissues borders. Thus, the exact segmentation of medical images is an essential process in recognizing and curing an illness. Thus, it is obvious that the purpose of clustering in medical images is the recognition of damaged areas in tissues. Different techniques have been introduced for clustering in different fields such as engineering, medicine, data mining and so on. However, there is no standard technique of clustering to present ideal results for all of the imaging applications. In this paper, a new method combining genetic algorithm and K-Means algorithm is presented for clustering medical images. In this combined technique, variable string length genetic algorithm (VGA) is used for the determination of the optimal cluster centers. The proposed algorithm has been compared with the K-Means clustering algorithm. The advantage of the proposed method is the accuracy in selecting the optimal cluster centers compared with the above mentioned technique.

Keywords: Medical Image, Clustering, Genetic Algorithm, K-Means.

1. Introduction

Segmentation is vital in the analysis of medical images which can be useful in many applications such as distinguishing the arteries borders from each other in angiography, the size of tumor and its response to treatment, interpretation of operation, the study of brain growth and recognition of tumor and so on. Therefore, it is obvious that segmentation is often used as the first and most important phase in the recognition and treatment of a disease in analyzing medical images like MRI images, and by the intuitive nature of the image, segmentation can be totally different. Segmenting an image refers to the technique of segmenting the space of an image into meaningful homogeneity areas with no overlaps which are the same in some features like, intensity or tissues [1]. The segmenting techniques are classified into two categories:

1. Edge-based techniques
2. Area-based techniques

In the former, both the borders presented in the images and their surroundings are found. While in the latter, the purpose is to initiate from image histogram, and is based on the intensity of pixel intensity whether it is less or more than the given amount of the given value. Clustering as a technique of segmentation, is a segmenting process in which series of information, being usually multidimensional, are divided into groups, so that the members of groups are the same in some criteria, while the members of different groups are different.

Clustering involves finding a structure in a cluster with data having no label. Clearly labels are often observable in medical images like MRI which are analyzed by the physician. When the given labels are not clear, the computer should be applied in labeling. The process of labeling and segmenting can be either synchronously or separately.

There have been different clustering techniques for the image segmentation such as, K-Means [2], Fuzzy C-means [3] and Average link [4] algorithms. These algorithms play important roles in the analysis of imaging in medicine, engineering and etc. In the most techniques mentioned above, the numbers of classes are applied as initiate input, where the clustering is defined as the distribution of N sample, in the space of n dimension in the group K [5][6]. One of the problems in these methods to determine the number of optimal classes for each of the images. Genetic algorithm as an optimal searching technique in the length of searching process is used for the optimal search of cluster center in medical images regarding the high volumes of image pixels [7-9]. One of the well-known clustering methods is K-Means algorithm which is implementable easily. Clustering by K-Means algorithm has high speed convergence; however, its accuracy is not satisfied for abnormal brains (such as tumor, swelling and etc). Unfortunately, its original version has some objections like its dependence on the initial values of centers and convergence to the local optimal response [10]. Using genetic algorithm, the problems mentioned above have been removed. However, by combining two algorithms, incredible results have been obtained [11][12]. Different clustering algorithms present different results and evaluating these results is very important. Thus, cluster validity is an important challenge. Two main criteria of combination and separation are used for evaluating optimal clusters. Different accuracy criteria have already been proposed such as, Rand Index [13], Dunn’s separation measure [14], Davies Bouldin [15], C-index [16], Adjusted Rand Index [17] and etc.

2. The Proposed method

The process of clustering of the proposed model is shown step by step as follows:

Input: Medical image.
Step1: Application of a smoothing filter with a 3×3 neighborhood to reduce the radio frequency noise and small effects of image.
Step2: Computing the number of maximum chromosomes ($k_{\max}$) using noiseless image histogram.
Step3: Application of variable string length genetic algorithm (VGA) for obtaining optimal cluster centers.
Step4: Application of K-Means algorithm clustering on the images using the optimal cluster centers.
Output: Clustered medical image.

2.1 Image pre-processing

MRI image is often the results of noise in the imaging environment of MRI device. One of these noises is the noise of radio frequency (RF) which is reduced by a smoothing filter with the 3×3 neighborhood. In Fig 1, the smoothing filter with the 3×3 neighborhood is shown.

2.2 Evaluation of cluster maximum

In this section, the number of maximum chromosomes ($k_{\max}$) is computed by using noiseless image histogram as one of the features of image. So that $k_{\max} = (rand()\%k^*) + 2$ and k^* is the number of given peaks appearing in image histogram.

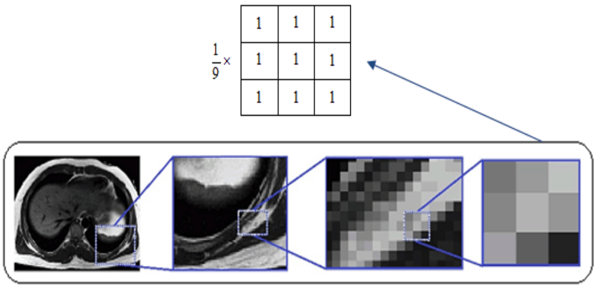


Fig. 1 Smoothing filter with the 3 × 3 neighborhood.

2.3 Variable string length genetic algorithm

In this paper, a genetic algorithm with a variable string length of chromosome has been presented. In the following, the implementation of the proposed genetic algorithm and the applications used in which will be explained.

2.3.1. Chromosome representation

In the proposed genetic algorithm, chromosomes are indicate the cluster centers. Supposing that the MRI images are two dimensions, l_i as the ith chromosome length which is $2 \times k_i$, and k_i as the number of clusters.

2.3.2. Generating initial population

For each i chromosome, the initial population is produced randomly from the surrounding borders of the image. After generating initial population, the qualification of each chromosome is obtained from the fitness function.

2.3.3. Constructing fitness function

The qualification of each member of community is determined by using fitness function after creating initial community. By using this fitness function, a number is assigned to each chromosome which is the value of that chromosome. This number is used as a merit determining the presence of this chromosome for the next generation.
To compute the fitness, suppose $V = \{v_1, v_2, v_3, \ldots, v_k\}$ be a chromosome that each of its members is the center of each cluster. Thus, fitness function is computed in equation (1).

$$F(v,x)=\frac{\left|v_i-v_j\right|}{\sum_{t=1}^{n_1}\left|v_i-x_l\right|+\sum_{p=1}^{n_2}\left|v_j-x_h\right|} \qquad (1)$$

$$\forall i,j \quad i\neq j,\ 1\leq t\leq n_1,\ 1\leq p\leq n_2$$

Where $\left|v_i-v_j\right|$ is the distance between the cluster centers of i and j, $\sum_{t=1}^{n_1}\left|v_i-x_l\right|$ and $\sum_{p=1}^{n_2}\left|v_j-x_h\right|$ are the sum of differences between i and j cluster members and the cluster centers, n_1 and n_2 are the number of i and j cluster members. To obtain the optimal clustering, the purpose is an increase of the fitness is required which is equivalent to clustering with minimum distribution into clustering and minimum separation between cluster centers.

2.3.4. Defining genetic operators

Selection: In this paper, for selecting parental chromosomes, some are selected randomly among the chromosomes with low fitness functions. Then, two chromosomes among them with the highest quality are selected as parental chromosomes.

Crossover: After selecting parental chromosomes, the process of composing must be done. Let P_1 and P_2 be the parental chromosomes and C_1 and C_2 be the chromosomes obtained from the process of composing. From a random number (α), the process of combining will be applied on the parental chromosomes as shown in the equation (2) and (3).

$$C_1=\alpha\times P_1+(1-\alpha)\times P_2 \qquad (2)$$

$$C_2=\alpha\times P_2+(1-\alpha)\times P_1 \qquad (3)$$

The process of composing is shown in Fig 2.

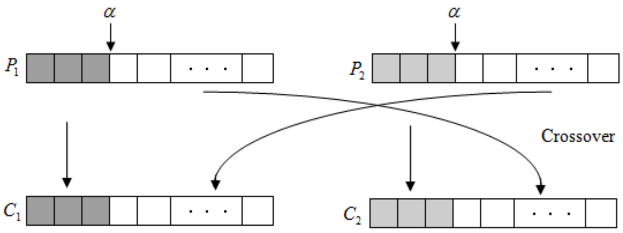


Fig. 2 Crossover operator.

2.3.5. Full genetic algorithm

The different processes of genetic algorithm with variable string length are shown as a follow:

Input: Medical image, maximum generation (G), size of population (P), crossover probability (P_c) and mutation probability (P_m).

Step1: Determining the maximum numbers of clusters ($k_{\max}$) through image histogram to create chromosomes using $k_{\max}$.

Step2: The initial values of chromosomes. Consider the number of generation (g) as 1.

Step3: Computing the fitness function for each chromosome in initial population.

Step4: Rearrange the population and assigning a degree for each chromosome in the population.

Step5: Selection among population chromosomes and then transferring them into mating pool. Apply the process of composing and mutation on the chromosomes which are in the mating pool to create child chromosomes.

Step6: Compute fitness function for each child chromosome.

Step7: Apply elitism by combining child and parent chromosomes, rearrange and finally, selecting the best chromosome to create the next generation. Add one unit to the number of generation ($g=g+1$).

Step8: If the maximum Iterative steps G is not reached ($g\leq G$), go to Step 4.

Output: Select the qualified chromosome in the final population.

Being found the qualified chromosome with the optimal cluster centers, the process of clustering must be applies on the medical images using K-Means clustering algorithm.

2.3.6. K-Means clustering algorithm

K-Means algorithm is considered as a technique based on gravity center, which is a basic method for many clustering methods such as fuzzy clustering, because of its simplicity. Different forms have been presented for the K-Means algorithm. But all of them have the same repetitive procedures which try to estimate only a fixed number of clusters as follows:

- Obtaining some points as the cluster centers. In fact, these points are the averages of points in each cluster.
- Assigning each sample data to a cluster in which the given data has the least distance from the center of the cluster.

Alternatively, corresponding to the number of needed clusters, some points are selected randomly. Then, data is assigned to these clusters regarding the amount of similarities. So some new clusters are obtained. By repeating the same procedures, it is possible to compute new centers in each iterative through rearranging the given data and then assigning data to new clusters.

It is possible to explain the stopping in K-Means algorithm as follows:

Let $v_i^{(n)}$ be the center of i in step n and $v_i^{(n-1)}$ be the center of i in step $(n-1)$, in step 4 the equation of $\left|v_i^{(n)} - v_i^{(n-1)}\right| < \delta$ is investigated. If the difference between two quantities of some hierarchical center is less than the value of a pre-determined initial δ, then the algorithm will stop. Otherwise; the process will be repeated as many times as it is needed and finally, the number of repeats reaches to a certain level.

Step1: The point K is selected as cluster centers.

Step2: Each sample data is assigned to the cluster center with least distance from the given data.

Step3: After assigning all data to all clusters, a new point is computed for each cluster as a center one by one (The average of points related to each cluster).

Step4: If there is no change in the cluster centers ($\left|v_i^{(n)} - v_i^{(n-1)}\right| < \delta$), go to step 2.

----

3. Cluster validity measure

In this paper, Dunn index has been used to measure validity of cluster. Dunn index, which is one of the internal criteria for the accuracy of clustering, is obtained from equation (4).

$$D = \min_{i=1\ldots n_c}\left\{ \min_{j=1\ldots n_c,\, j\neq i}\left(\frac{d(K_i, K_j)}{\max\limits_{h=1\ldots n_c}(\Delta(K_h))} \right) \right\} \qquad (4)$$

where $d(K_i, K_j)$ and $\Delta(K_i)$ are obtained from equations (5) and (6).

$$d(K_i, K_j) = \min_{x\in K_i,\, y\in K_j}\{d(x,y)\} \qquad (5)$$

$$\Delta(K_i) = \max_{x,y\in K_i}\{d(x,y)\} \qquad (6)$$

in which $d(K_i, K_j)$ is the distance from clusters i and j, $\Delta(K_h)$ is the maximum hth cluster diameter, $d(x,y)$ is the distance between the given x and given y and n_c is the number of clusters.

By combining criteria of clusters and separating criteria between clusters, the greater value for Dunn index will be an optimal value.

4. Experimental results

In the simulation of the proposed model, MRI images of brain and heart [18] have been used. For mutation and selection degrees, different values are considered. The maximum numbers of generations (g) is considered as 12 and the minimum numbers, is considered to be 4. The maximum numbers of chromosomes in each generation is considered to be 150 and the minimum numbers considered to be 30. The obtained results are shown in the Table1 and the Table2.

Table 1: The proposed model results for MRI brain image by different parameters.

#	P	G	P_m	P_s	$F(v,x)$		
					worst	***averag***	***best***
1	30	4	0.2	0.2	0.00657	0.00968	0.02486
2	30	8	0.15	0.5	0.31253	0.34973	0.72302
3	80	4	0.2	0.2	0.34522	0.45995	0.85545
4	80	8	0.15	0.5	0.00563	0.00789	0.15649
5	100	4	0.2	0.2	0.43050	0.55914	0.87859
6	100	8	0.15	0.5	0.32957	0.40699	0.91176
7	120	8	0.2	0.2	0.55791	0.65192	1.10000
8	120	4	0.15	0.5	0.29416	0.33681	0.93287
9	150	8	0.2	0.2	0.55396	0.64968	0.99788
10	150	4	0.15	0.5	0.30648	0.35442	0.90580

For example, the fourth test results are shown in Fig 3 and 4.

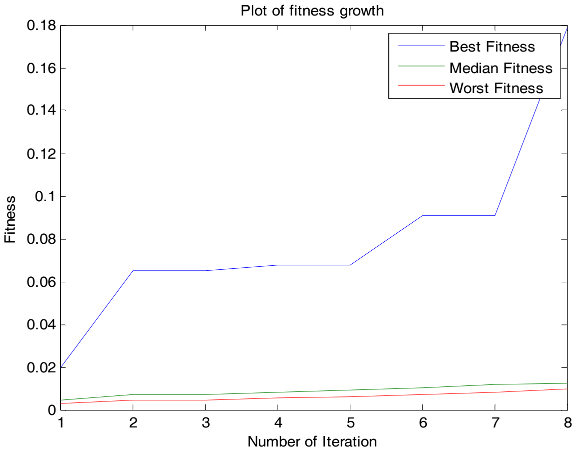
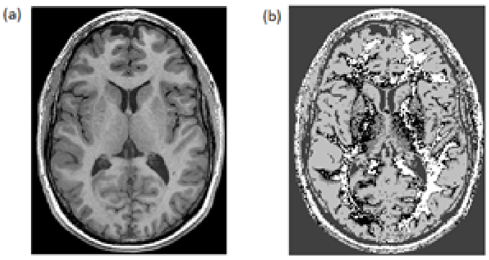


Fig. 3 Growth charts fitness function



Original image Clustered image

Fig. 4 MRI brain image

Table 2: The proposed model results for MRI heart image by different parameters.

#	P	G	P_m	P_s	$F(v,x)$		
					worst	**averag**	**best**
1	30	4	0.2	0.2	0.088837	0.130050	0.19349
2	30	8	0.15	0.5	0.031170	0.040669	0.23242
3	80	4	0.2	0.2	0.054473	0.071793	0.18693
4	80	8	0.15	0.5	0.048480	0.083582	2.40290
5	100	4	0.2	0.2	0.123110	0.167100	1.76150
6	100	8	0.15	0.5	0.039079	0.055477	0.33636
7	120	12	0.2	0.2	0.160360	0.257860	0.82045
8	120	4	0.15	0.5	0.023779	0.034712	0.28146
9	150	8	0.2	0.2	0.162350	0.252040	1.22620
10	150	12	0.15	0.5	0.083363	0.125290	4.47320

For example, the sixth test results are shown in Fig 5 and 6.

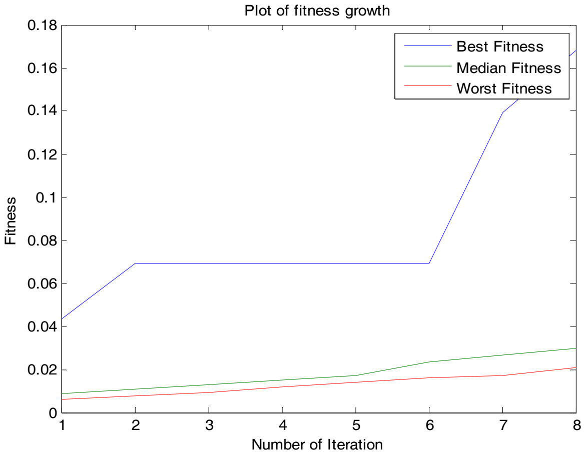
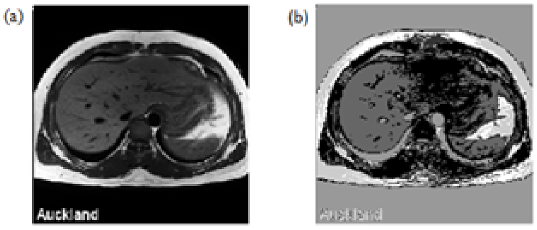


Fig. 5 Growth charts fitness function



Original image Clustered image

Fig. 6 MRI heart image

Different clustering algorithms show different results, and evaluating those results is very important. Thus, cluster validity has become a very important challenge. In this paper, the proposed model has been compared with the different clustering algorithms such as K-Means clustering algorithm. The comparison is based on the use of criteria and Dunn index, and its results are shown in Table3 and Fig 7. It is important to note that the greater value of Dunn index shows an optimal clustering.

Table 3: Clustering validation results for MRI images.

#	**K-Means**		**Proposed Method**	
	K	**Dunn index**	**K**	**Dunn index**
Image 1	7	0.2381	6	0.70000
Image 2	8	0.42857	6	0.66071
Image 3	7	0.73077	8	0.81818
Image 4	7	0.32203	7	0.40625

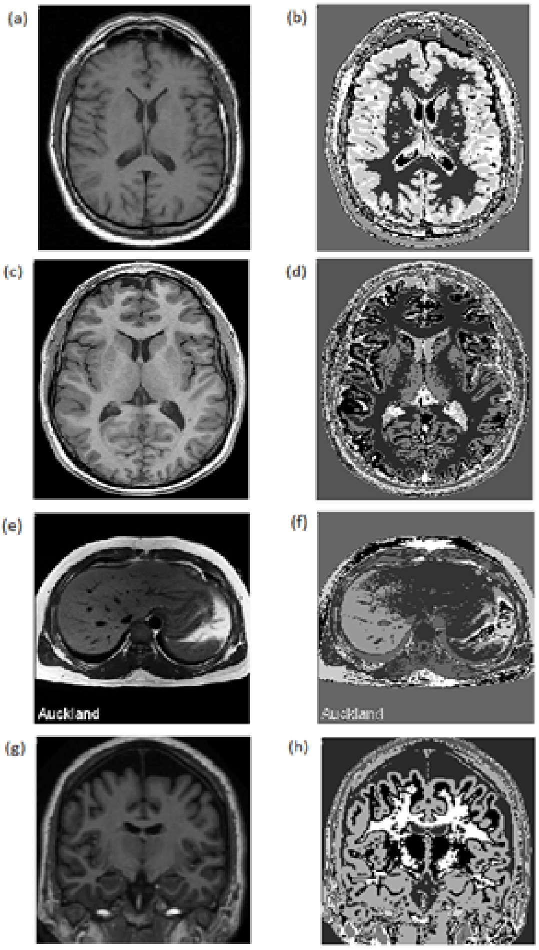


Fig. 7 (a,c,e,g) Original MRI images, (b,d,f,h) Clustered MRI images.

5. Conclusion

The results of investigation on clustering in medical images showed that K-Means algorithm and other similar algorithms are relatively more efficient in clustering normal brain MRI images with no noise and getting the best centers of cluster for clustering purposes must be done by random. Thus, using such methods are not suitable to measure the volume of tumor and its response to treatment, interpretation of operation, the study of brain growth, recognition of tumor and so on. Therefore, to overcome to this problem, a technique is required to be presented. In this paper, a new method has been introduced for clustering medical images, with respect to the importance and application of clustering in the segmentation of images. The advantage of the proposed method is obtaining the optimal cluster centers using variable string length genetic algorithm.

References

[1] R.C. Gonzalez, R.E. Woods, “Digital Image Processing”, Addision-Wesley, Massachusette, 1992.

[2] T. Kanungo, D.M. Mount, N.S. Netanyahu, C. Piatko, R. Silverman, A.Y. Wu, “An efficient k-means clustering algorithm: Analysis and implementation“, IEEE Trans. Patt. Anal. Mach. Intell. 24 (2002).

[3] R. L. Cannon, J. V. Dave, AND J. C. Bezdek, “Efficient Implementation of the Fuzzy c-Means Clustering Algorithms“, IEEE Trans. Patt. Anal. Mach. Intell. 8, 2 (1986).

[4] P. Filzmoser, R. Baumgartner, E. Moser, “A hierarchical clustering method for analyzing functional MR images“, Magnetic Resonance Imaging, Volume 17, Issue 6, July 1999, Pages 817-826.

[5] E. Forgey, “Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classification“, Biometrics, vol. 21, p. 768, 1965.

[6] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. “A local search approximation algorithm for k-means clustering“. Comput. Geom., 28(2-3):89–112, 2004.

[7] S. Cagnoni, A.B. Dobrzeniecki, R. Poli, J.C. Yanch, “Genetic algorithm-based interactive segmentation of 3D medical images“. Image and Vision Computing, 17(12), 881-895, 1999.

[8] E. Angelié, P. J. H de Koning, H. C van Assen, M. Danilouchkine, G. Koning, R.J van der Geest, J.H.C Reiber, “Automatic tuning of left ventricular segmentation of MR images using genetic algorithms“. International Congress Series, Volume 1256, June 2003, Pages 1102-1107.

[9] F. Xie, A. C. Bovik, “Automatic segmentation of dermoscopy images using self-generating neural networks seeded bygenetic algorithm“. Pattern Recognition, In Press, Corrected Proof, Available online 29 August 2012.

[10] J. B. MacQueen "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297 (1967).

[11] D. X. Chang, X. D. Zhang, C. W. Zheng, “A genetic algorithm with gene rearrangement for k-means clustering“, Pattern Recognition, Volume 42, Issue 7, July 2009, Pages 1210-1222.

[12] Y. Liu, X. Wu, Y. Shen, “Automatic clustering using genetic algorithms”, Applied Mathematics and Computation, Volume 218, Issue 4, 15 October 2011, Pages 1267-1279.

[13] W. M. Rand, “Objective Criteria for the Evaluation of Clustering Methods“. Journal of the American Statistical Association, 1971, 66(336): 846-850.

[14] J. C. Dunn, “A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters“. Journal of Cybernetics, 1973, 3(3): 32-57.

[15] D. L. Davies, D. W. Bouldin, “A Cluster Separation Measure“, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1 (2): 224–227.doi:10.1109/TPAMI.1979.4766909 (1979).

[16] L. J. Hubert and J. R. Levin. “A general statistical framework for assessing categorical clustering in free recall“. Psychological Bulletin, 1976, 83, 1072-1080.

[17] P. N. Tan, M. Steinbach and V. Kumar, “Introduction to Data Mining“. 2006.

[18] http://brainweb.bic.mni.mcgill.ca/brainweb/.

Akbar Shahrzad Khashandarag was born in Tabriz, Iran, in 1978. He received the B.Sc. degree in computer engineering from the Islamic Azad University Bonab Branch, Iran in 2009. He is currently M.Sc. student in Mechatronics engineering at the Islamic Azad University Tabriz Branch. His research interests include image processing, signal processing and wireless sensor network.

Dr. Mirkamal Mirnia received the B.Sc. degree in mathematical Physics from Ferdowsi University, Mashhad, Iran in 1967, and M.Sc. degree in pure mathematics from teacher training University of Tehran in 1969 also M.Sc. degree in numerical analysis and computing from Owen university of Manchester, UK in 1975 and PhD. in applied mathematics (optimization) from university of St.Andrews, UK in 1979. He is a member of the Mathematical society of Iran, Iranian operations researches, institute of applied mathematics, UK and member of SIAM.

Dr. Aidin Sakhavati was born in Orumieh, Iran, in 1978. He received his BS, MS and Ph.D. degrees in 2000, 2003 and 2010 from Islamic Azad university-Tabriz branch (IAUT), and Tabriz University, Tabriz, Iran, and Islamic Azad University, science and research branch, Tehran, Iran, respectively. He has been holding the Assistant Professor position at IAUT since 2010. He is the author of more than 30 journal and conference papers. His teaching and research interest include power system and transformers transients and power electronics applications in power systems. His researching interests include power systems dynamic and control application of Quantitative Feedback Theory, Particle Swarm Optimization and Genetic Algorithm in FACTS devices and Load Frequency control.

Splitting of traffic to maximize the data transmission in mobile ad hoc network under different constraints

Sushil Chandra Dimri[1], Kamlesh Chandra Purohit[2] and Durgesh Pant[3]

[1]**Graphic Era University, Dehradun, Uttarakhand-India 248001**

[2]**Graphic Era University, Dehradun, Uttarakhand-India 248001**

[3]**Uttarakhand Open University, Dehradun, Uttarakhand-India 248001**

Abstract

Mobile ad hoc network (MANET) is a set of wireless mobile computer forming a temporary network with out any wired infrastructure, due to dynamic nature of topology and other constraints transmission routing is a challenging task in MANET. The splittable routing establishes multi paths between source nodes and destination nodes; this scheme provides better performance of the network under different constraints. This paper presents a L.P.P. model for the routing problem in MANET and generates the optimal solution in term of route identification for transmission and the amount of traffic per route to maximize the data transmission under various constraints.

Keywords: *MANET, Single path routing, Maximization of splittable traffic flow*

1. Introduction

Mobile ad hoc network consists of the nodes which change their position over time, the movement of the nodes is random and unpredictable, and nodes may leave and join the network at any point of time. The node works as a source, which generates data (traffic), destination and the router. Each node has limited memory, processing power and battery energy. The transmission and receiving range of a node is limited so multiple hops are required to transmit the data to the other distant nodes, which makes routing a crucial issue in mobile ad hoc network. There are many factors associated with MANET which makes routing a difficult issue in this mobile environment. [1, 2 and 3].

The main characteristics of the MANET are:

a) Temporary topology which changes over time
b) Limitation on resources
c) Wireless transmission
d) Network partitions
e) Limited bandwidths
f) Multi functionality of a node as source, destination router and rely transmitter Mobile ad hoc network is totally different from the wired network, in MANET, mobility of the nodes is the biggest issue; the mobility is absolutely random in term of time and direction. The mobility of nodes is mainly responsible for various problems associated with mobile ad hoc network.

The main challenges of mobile environment are:

a) Packet loss due to transmission error and limited processing power of a node
b) Variable capacities of links
c) Limited communication bandwidth
d) Broad cast nature of communication
e) battery life limited
f) Node mobility
g) Wireless medium

1.1 Single Path Routing:

In MANET single path routing is not an effective routing technique specially when there are many constraints. Single path routing sends entire traffic via a single route from source to destination. In MANET, the link capacity (bandwidth), the memory and processing power of the nodes are limited so that they can not handle high amount of traffic, which generally leads to congestion ,packet loss

, formation of hot spots in the network ,as a consequences the end to end delay and unreliability of the network increases. In single path routing, if a link breaks or a node fails it leads to the network failure, i.e. no transmission occurs between source S and destination T.

1.2 Splitting of Traffic to Multiple Routes in Manet:

Splitting the traffic to multiple routes can provide better load balancing, fault tolerance and higher aggregate bandwidth. Splitting of the traffic can be helpful in reduction of congestion and bottle necks; this also improves network resource utilization and bandwidth optimization [4, 5, 6 and 7].

Suppose if node S (source) has 3 routes to the destination T, If S sends data packets along all the three paths, as long as all the routes are not failed, node T will receive the data and the transmission survives. In single path routing failure of a link/node of the route will stop the entire transmission; thus this approach of splitting of traffic increases the strength of network survival.

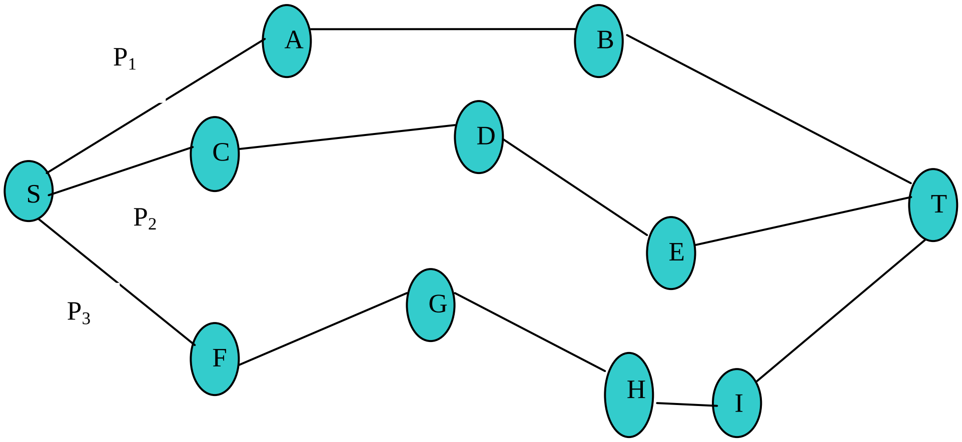


Fig. 1- Three disjointed routes between source S and destination T

There are 3 disjointed routes between source S and destination T, out of these paths P_1 is the shortest one containing merely two intermediate node and 3 links, let us suppose the probability of failure of link and nodes are 0.02 and 0.01 respectively, these probabilities are same for all nodes and links in the given network.(Figure-1). Then the probability of failure of path $P_1(S-A-B-T)$, $P\ (FP_1)$ is given by

$$P(FP_1)=3\times0.02+2\times0.01=0.08.$$

If we are applying splittable approach to send the data from S to T, and using routes P_1, P_2 and P_3 to send data then the probability that all the routes will fail simultaneously can be determined by the relation

$$P\ (FP_1\cap FP_2\cap FP_3\ )$$
$$=P(FP_1)\times P(FP_2)\times P(FP_3)$$

$$P(FP_2)=4\times0.02+3\times0.01=0.11$$
$$P(FP_3)=\ 5\times0.02\ +4\times0.01=0.14$$

Thus the probability of failure of all the routes is given by

$$P\ (FP_1\cap FP_2\cap FP_3\ )$$
$$=P(FP_1)\times P(FP_2)\times P(FP_3)$$
$$P\ (FP_1\cap FP_2\cap FP_3\ )$$
$$=0.08\times0.11\times0.14=0.001232$$

Clearly the $P(FP_1\cap FP_2\cap FP_3)\ <P(FP_1)$

Thus splittable routing enhances the network reliability and survivability.

The bandwidth in mobile ad hoc network is limited and single path routing can not provide enough bandwidth to maximize the data transmission , since nodes are battery supported and have limited power, high traffic will reduce the energy level of a node very quickly, so in case of single path routing the consumption of energy is high. Splittable routing will be the best answer to all these problems, in this scheme the traffic will be splitted among the multiple paths. These paths may or may not be disjointed. The splittable routing scheme provides better load balancing, fault tolerance, as it distributes the traffic to multiple routes and increase the aggregate bandwidth, further this approach also decreases the intensity of traffic load per path, per nodes.

1.3 Route Establishment and Route Maintenance:

The process of establishing routes consists of finding multiple routes between a source node and the destination node [8 and 9]. (Fig. 2).

- In MANET a route between any two nodes is established sending a route request by the source node to the neighboring nodes, from where this request it reaches to the destination node if the routes exist.
- This mechanism discovers multiple routes in the network between source and destination node.
- The route request first to arrive is accepted by the target. The target then responds on that route and intimates the initiator what the source routes are.
- In this way routes are established and then used to send the data traffic.

1.3.1 Maintenance of Routes:

This process in ad hoc network is used to repair the broken routes or finding alternative routes in case of routes failure.

- When the links lies on the alive routes between source and destination breaks then the existing routes does not work.
- When a node detects a broken link while attempting to forward a packet to the next hop, it generates an error message that is sent to all the sources using the broken link. If a source receive an error message and route to the destination is still required, it initiates a new route discovery process. Routes are also deleted from the routing table if they are unused for a certain amount of time.
- With help of route discovery mechanism an alternative route has established.

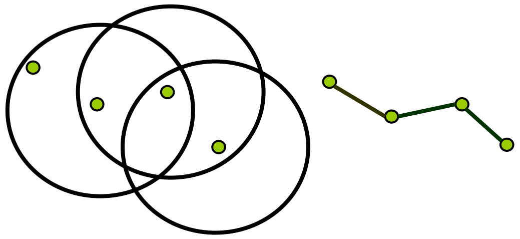


Fig. 2- The MANET and route establishment of routes in MANET

1.4 Distribution of Traffic:

Once source node establishes a set of paths to the destination, it can begin sending data to the destination along the paths. The distribution of traffic to different routes is based on the link capacity and cost of the routes. In case of equal capacity and equal cost routes ,the distribution of traffic will be same i.e. equally divided but when the cost of routes, capacity of the routes, energy consumption of the routes are different then the distribution of traffic among the different routes under various constraints is a difficult task. To solve this problem in this paper we present a L.P.P. model for distribution of traffic so as to maximize the data transmission between source and destination under various constraints. [7, 8, 9 and 10]

2. Network Model:

Assuming a path P_i (i =1, 2--p) consists of I_i intermediate relaying node. Suppose a traffic flow with average rate λ exist between source S and destination T, this traffic is then splitted in p routes, let the traffic along path P_i is $\lambda_{i\,,}(i=1,2,---,p)$, the distribution of traffic $\lambda_i,(i=1,2,---,p)$ is Poisson distributed.

$$\sum_{i\,\in\,(1,\ \ldots,\ p)}\lambda_i = \lambda.$$

Let μ be the average processing rate of the each node which is sufficiently high and ignore the background traffic at any node on any path. Maintaining several routes for every source-destination pair would balance the traffic more evenly across the network and would alleviate the effect caused by congested link. The load between any source destination pair is to be evenly divided among all available routes. Although we are studying the problem of ad hoc network in static environment, however it has been realized that node mobility pattern has significant impact on the connectivity of the network.

3. Formulation of Problem:

Let S is the source and T is the target in the given MANET and the path under consideration for flow between S and T

are p, these paths are not necessarily all disjointed [2, 3 and 6].
Let P_1, P_2, P_3… P_p are the routes between S and T and traffic along these paths are $\lambda_1,\ \lambda_2,\ \lambda_3,\ \ldots\ldots\ldots\ \lambda_p$ respectively.

Under a fixed given cost B>0(Budget) for per unit of time, different link capacity, energy amount available per unit of time and with given cost of link (C_e) per unit of traffic, we have to maximize the (S, T) traffic flow. It is possible to transform this problem in to a linear programming problem with the objective of maximization of total traffic flow between S and T with the restriction of the link capacities and node energy availability and under a given cost (budget), that is the traffic flow value in a link cannot exceed the capacity of the link and the total flow cost cannot be higher than the given cost [11,124, and 13].
Let energy consumption per unit of flow per node is D_n mW and the energy limitation per unit of time is E unit (mW- Mili Watt).Clearly the energy consumption of a node is directly proportional the amount of traffic passing through that node per unit of time.

The problem as L.P.P. is

$$Max\ \lambda\ =\ \lambda_1 +\ \lambda_2 +\ \lambda_3 +\ \ldots\ldots\ldots\ +\ \lambda_p$$

Subject to

$$\sum_{i\ \in\ \{1,\ \ldots,p\}\ ,\ e\ \in\ Pi} \lambda_i\ \le\ U_e$$

U_e is capacity of link e. for all link e in the network

$$\sum_{i\in\{1,\ \ldots,p\}} \lambda_{i.} Cp_i\ \le B,$$

Cost

$B>0$

$$\sum_{i\ \in\ \{1,\ \ldots,p\}} \lambda_{i.} Ep_i \le E,$$

Energy available $E>0$

$$\lambda_i \ge 0\ for\ all\ i \in \{1\ldots p\}$$

Where cost of path P_i, $C_{p_i} = \sum_{e\ \varepsilon\ P_i} C_e$ C_e is the cost of the link where $e \in Pi$

$E_{pi} = I_{pi} \times D_n$ Where I_{pi} is the number of intermediate nodes on path P_i between S and T.

The constraints set contains constraints such that the sum of the traffic flow values on the paths containing the link e must be bounded by U_e (capacity of edge e), the cost of total transmission per unit of time remain less than B, and energy consumption per unit of time bounded above by E [14, 15 and 16].

3.1 Example: Maximization of Splittable Traffic Flow Under Above Mentioned Constraints for P=4, For Given Network:

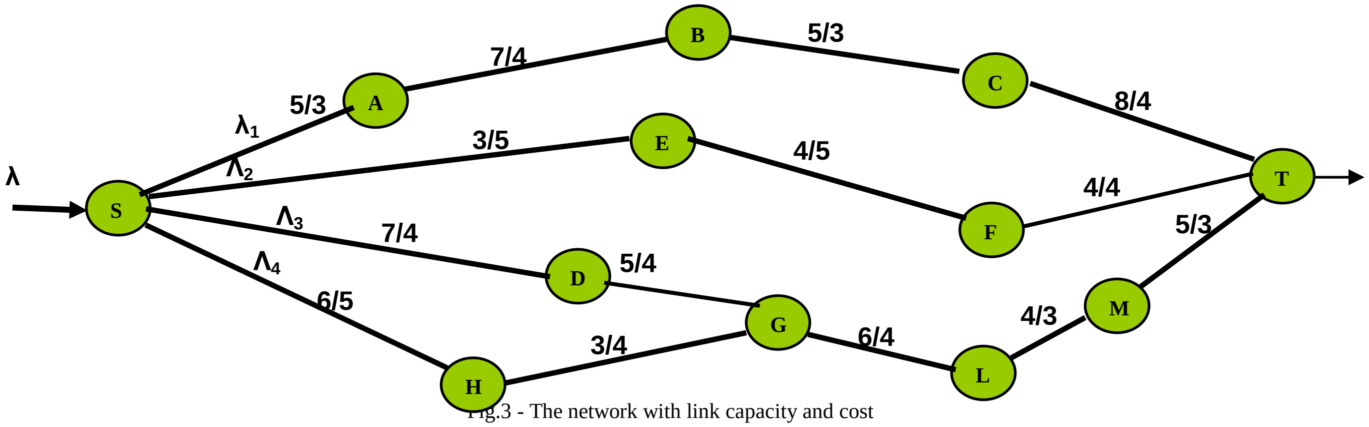


Fig.3 - The network with link capacity and cost

In Figure-3, a MANET is given with link capacity and cost to use that link per unit of traffic flow, S is the source node and T , is the destination, the objective is to maximize the traffic between S and T under cost, capacity and energy constraints.

Routes considered between S and T are:

$$P_1 \equiv S\ -A\ -B-\ C-T$$

$P_2 \equiv S - E - F - T$

$P_3 \equiv S - D - G - L - M - T$

$P_4 \equiv S - H - G - L - M - T$

λ_1 *traffic flow along* P_1

λ_2 *traffic flow along* P_2

λ_3 *traffic flow along* P_3

λ_4 *traffic flow along* P_4

The maximization of transmission with p = 4

Let the given cost (budget) B per unit of time is Rs200 and the available energy limit E is 50 unit (mw-mili-watt) the consumption of energy per unit of flow per node D_n =2 unit (mW-mili-Watt)

The different routes in the given network are

$P_1 \equiv S - A - B - C - T$

$P_2 \equiv S - E - F - T$

$P_3 \equiv S - D - G - L - M - T$

$P_4 \equiv S - H - G - L - M - T$

Also the cost of routes

$C_{P1} = 3+4+3+4=14$

$C_{P2} = 5+5+4=14$

$C_{P3} = 4+4+4+3+3=18$

$C_{P4} = 5+4+4+3+3=19$

The L.P.P for the problem.

$$Max\ Z = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4$$

The transmission constraints

$\lambda_1 \leq 5$

$\lambda_2 \leq 3$

$\lambda_3 + \lambda_4 \leq 4$

$\lambda_4 \leq 3$

$\lambda_3 \leq 5$

$14\lambda_1 + 14\lambda_2 + 18\lambda_3 + 19\lambda_4 \leq 200$ // Cost constraints//

$6\lambda_1 + 4\lambda_2 + 8\lambda_3 + 8\lambda_4 \leq 50$ // Energy constraints//

Using the slack variable $\lambda_{5,}\lambda_{6,}\lambda_{7,}\ \lambda_8,\ \lambda_9,\lambda_{10}$ and λ_{11} to convert inequalities in to equalities

$$Max\lambda = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + 0\lambda_5 + 0\lambda_6 + 0\lambda_7 + 0\lambda_8 + 0\lambda_9 + 0\lambda_{10} + 0\lambda_{11}$$

$$\lambda_1 + 0\lambda_2 + 0\lambda_3 + 0\lambda_4 + 1\lambda_5 + 0\lambda_6 + 0\lambda_7 + 0\lambda_8 + 0\lambda_9 + 0\lambda_{10} + 0\lambda_{11} = 5$$

$$0\lambda_1 + 1\lambda_2 + 0\lambda_3 + 0\lambda_4 + 0\lambda_5 + 1\lambda_6 + 0\lambda_7 + 0\lambda_8 + 0\lambda_9 + 0\lambda_{10} + 0\lambda_{11} = 3$$

$$0\lambda_1 + 0\lambda_2 + \lambda_3 + \lambda_4 + 0\lambda_5 + 0\lambda_6 + \lambda_7 + 0\lambda_8 + 0\lambda_9 + 0\lambda_{10} + 0\lambda_{11} = 4$$

$$0\lambda_1 + 1\lambda_2 + 0\lambda_3 + \lambda_4 + 0\lambda_5 + 1\lambda_6 + 0\lambda_7 + \lambda_8 + 0\lambda_9 + 0\lambda_{10} + 0\lambda_{11} = 3$$

$$0\lambda_1 + 0\lambda_2 + 1\lambda_3 + 0\lambda_4 + 0\lambda_5 + 0\lambda_6 + 0\lambda_7 + 0\lambda_8 + 1\lambda_9 + 0\lambda_{10} + 0\lambda_{11} = 5$$

$$14\lambda_1 + 14\lambda_2 + 18\lambda_3 + 19\lambda_4 + 0\lambda_5 + 0\lambda_6 + 0\lambda_7 + 0\lambda_8 + 0\lambda_9 + 1\lambda_{10} + 0\lambda_{11} = 200$$

$$6\lambda_1 + 4\lambda_2 + 8\lambda_3 + 8\lambda_4 + 0\lambda_5 + 0\lambda_6 + 0\lambda_7 + 0\lambda_8 + 0\lambda_9 + 0\lambda_{10} + 1\lambda_{11} = 50$$

$$\lambda_1,\ \lambda_{2,}\lambda_{3,}\lambda_{4,}\ \lambda_5,\lambda_6,\lambda_{7,}\lambda_{8,}\ \lambda_{9,}\ \lambda_{10,}\ \lambda_{11} \geq 0$$

FORMING THE SIMPLEX TABLE

Table 1: The initial Simplex Table

		C_J	1	1	1	1	0	0	0	0	0	0	0
C_B	X_B	B	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}
0	λ_5	5	1	0	0	0	1	0	0	0	0	0	0
0	λ_6	3	0	1	0	0	0	1	0	0	0	0	0
0	λ_7	4	0	0	1	1	0	0	1	0	0	0	0
0	λ_8	3	0	0	0	1	0	0	0	1	0	0	0
0	λ_9	5	0	0	1	0	0	0	0	0	1	0	0
0	λ_{10}	200	14	14	18	19	0	0	0	0	0	1	0
0	λ_{11}	50	6	4	8	8	0	0	0	0	0	0	1

		Δ_J	-1	-1	-1	-1	0	0	0	0	0	0	0

Table 2: λ_6 is outgoing and λ_2 is incoming vector

		C_J	1	1	1	1	0	0	0	0	0	0	0
C_B	X_B	B	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}
0	λ_5	5	1	0	0	0	1	0	0	0	0	0	0
1	λ_2	3	0	1	0	0	0	1	0	0	0	0	0
0	λ_7	4	0	0	1	1	0	0	1	0	0	0	0
0	λ_8	3	0	0	0	1	0	0	0	1	0	0	0
0	λ_9	5	0	0	1	0	0	0	0	0	1	0	0
0	λ_{10}	158	14	0	18	19	0	-14	0	0	0	1	0
0	λ_{11}	38	6	0	8	8	0	-4	0	0	0	0	1
		Δ_J	-1	0	-1	-1	0	1	0	0	0	0	0

Table 3:- λ_7 is outgoing and λ_3 is incoming vector

		C_J	1	1	1	1	0	0	0	0	0	0	0
C_B	X_B	B	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}
0	λ_5	5	1	0	0	0	1	0	0	0	0	0	0
1	λ_2	3	0	1	0	0	0	1	0	0	0	0	0
0	λ_7	1	0	0	1	0	0	0	1	-1	0	0	0
1	λ_4	3	0	0	0	1	0	0	0	1	0	0	0
0	λ_9	5	0	0	1	0	0	0	0	0	1	0	0
0	λ_{10}	101	14	0	18	0	0	-14	0	-19	0	1	0
0	λ_{11}	14	6	0	8	0	0	-4	0	-8	0	0	1
		Δ_J	-1	0	-1	0	0	1	0	1	0	0	0

Table 4:- λ_{11} is outgoing and λ_1 is incoming vector

		C_J	1	1	1	1	0	0	0	0	0	0	0
C_B	X_B	B	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}
0	λ_5	5	1	0	0	0	1	0	0	0	0	0	0
1	λ_2	3	0	1	0	0	0	1	0	0	0	0	0
0	λ_3	1	0	0	1	0	0	0	1	-1	0	0	0
1	λ_4	3	0	0	0	1	0	0	0	1	0	0	0
0	λ_9	5	0	0	0	0	0	0	-1	1	1	0	0
0	λ_{10}	83	14	0	0	0	0	-14	-18	-1	0	1	0
0	λ_{11}	6	6	0	0	0	0	-4	-8	0	0	0	1
		Δ_J	-1	0	0	0	0	1	1	0	0	0	0

Table 5:- λ_3 is outgoing and λ_7 is incoming vector

		C_J	1	1	1	1	0	0	0	0	0	0	0
C_B	X_B	B	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}
0	λ_5	4	0	0	0	0	1	2/3	4/3	0	0	0	$-1/6$
1	λ_2	3	0	1	0	0	0	1	0	0	0	0	0
1	λ_3	1	0	0	1	0	0	0	1	-1	0	0	0
1	λ_4	3	0	0	0	1	0	0	0	1	0	0	0
0	λ_9	4	0	0	0	0	0	0	-1	1	1	0	0

0	λ_{10}	69	0	0	0	0	0	$-14/3$	$2/3$	-1	0	1	$-14/6$
1	λ_1	1	1	0	0	0	0	$-2/3$	$-4/3$	0	0	0	$1/6$
		Δ_J	0	0	0	0	0	$1/3$	$-1/3$	0	0	0	$1/6$

Table 6:- λ_5 is outgoing and λ_8 is incoming vector

		C_J	1	1	1	1	0	0	0	0	0	0	0
C_B	X_B	B	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}
0	λ_5	8/3	0	0	$-4/3$	0	1	$2/3$	0	$4/3$	0	0	$-1/6$
1	λ_2	3	0	1	0	0	0	1	0	0	0	0	0
0	λ_7	1	0	0	1	0	0	0	1	-1	0	0	0
1	λ_4	3	0	0	0	1	0	0	0	1	0	0	0
0	λ_9	5	0	0	1	0	0	0	0	0	1	0	0
0	λ_{10}	205/2	0	0	$-2/3$	0	0	$-14/3$	0	$-1/3$	0	1	$-14/6$
1	λ_1	7/3	1	0	$4/3$	0	0	$-2/3$	0	$-4/3$	0	0	$1/6$
		Δ_J	0	0	$4/3$	0	0	$1/3$	0	$-1/3$	0	0	$1/6$

able 7:- all $\Delta_J \geq 0$ Optimal solution Table

		C_J	1	1	1	1	0	0	0	0	0	0	0
C_B	X_B	B	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}
0	λ_8	2	0	0	-1	0	$3/4$	$1/2$	0	1	0	0	$-1/8$
1	λ_2	3	0	1	0	0	0	1	0	0	0	0	0
0	λ_7	3	0	0	0	0	$3/4$	$1/2$	1	0	0	0	$-1/8$
1	λ_4	1	0	0	1	1	$-3/4$	$-1/2$	0	0	0	0	$1/8$
0	λ_9	5	0	0	1	0	0	0	0	0	1	0	0
0	λ_{10}	619/6	0	0	-1	0	$1/4$	$-9/2$	0	0	0	1	$-57/2$ 4
1	λ_1	5	1	0	0	0	1	0	0	0	0	0	0
		Δ_J	0	0	0	0	$1/4$	$1/2$	0	0	0	0	$1/8$

Solving these equations with help of Simplex method we get the optimal traffic flow as

$\lambda_1 = 5,\ \lambda_2 = 3,\ \lambda_3 = 0$ and $\lambda_4 = 1.$

The maximum amount of traffic that can be routed per unit of time between source and destination under the constraints is given by $\lambda_{max} = 9.$ unit, also the selected paths to route the data are P_1, P_2 and P_4 only.

4. CONCLUSION:

It is found that the traffic between source S and destination T with given link capacity, cost constraints and energy constraints need not use all available paths but selects few specific paths so as to maximize the transmission under the constraints. In MANET, the nodes are having limited powers so it is very important to route the traffic along the different paths so that the energy consumption remains minimum and network survives for long time. There are several set of source and destination pair in MANET and the splittable routing definitely increase the performance of the network by improving the load balancing, end to end delay and optimization of bandwidth and by minimization the of energy consumption. The L.P.P. formulation of the routing problem in MANET certainly solves the problem of selection of routes and distribution of traffic under constraints. This study of MANET is in static environment but in actual mobility play a great role in MANET.

5. Acknowledgment

We would wish to thank our President Honorable Professor Kamal Ghanshala, for providing such a research oriented platform for us. Thanks are also due to professor

Dr. R. C. Joshi, Honorable Chancellor Graphic Era University, for and Professor Dr. Sanjay Jasola, Honorable Vice - Chancellor Graphic Era Hill University, for their valuable support and guidance.

References

[1]C. E .Perkins, Ad Hoc Networking, Addison Wesley, New York, 2001.

[2]W. Liang, "Minimizing energy and maximizing network life time multicasting in wireless ad hoc network ", in IEEE International Conference on Communication, pp. 3375-3380, 2005.

[3]G.Chakarbarti, S. Kulkarni, "Load balancing and resource reservation in mobile ad hoc networks ", Ad hoc Networks, pp. 186-203, 2006

[4]L.Wang, L.F.Zhang, Y.T. Shu, M. Dong, O.W.W. Yang, "Adaptive multipath source routing in wireless Ad hoc networks" IEEE ICC 98, Helsinki, Finland, June, 2001.

[5]M. K.Marina, S. R. Das, "Adhoc on-demand multipath distance path routing", ACM SIGMOBILE mobile computing and communication review, Vol6, No.3, July 2002.

[6] P.P. Pham, S. Perreau, "Performance analysis of reactive shortest path and multipath routing mechanism with load balance", in INFOCOM2003, March2003.

[7]S.C. Dimri, K.C. Purohit, D. Pant, "Improvement of performance of mobile ad hoc network using k-path splittable traffic flow scheme" International Journal of Computer Technology and Application, Vol. 2 (6), 1911-1917 IJCTA | NOV-DEC 2011.

[8]S.J Lee, M.Gerla, "Split multipath routing with maximally disjointed paths in ad hoc network", in ICC 2001, pp. 867-871, June2001.

[9]Y. Ganjali, A. Keshavarzian, "Load balancing in Adhoc networks, single path routing vs. Multipath routing." INFOCOM2004, March2004.

[10]G. I. Iavscu, S. Pierre, A.Quintero ,"QoS routing with traffic distribution in mobile ad hoc networks" In proceeding of ACM journal Computer communication ,Vol.32, No.2, Feburary, 2009.

[11]J.H. Chang, L.Tassiulas, "Maximum lifetime routing in wireless sensor network", Transaction on Networking IEEE/ACM, Vol, 12. No.4. pp. 609-619, 2004.

[12]E.Horowitz, S. Sahni, S.Rajasekaran, Computer Algorithms, Galgotia Publications Pvt. Ltd. 2003

[13]J.K. Sharma, Operation Research Theory and Applications, MacMillan India limited 2004

[14]D. Bertseakas, R.Gallager, Data Networks, Prentice Hall, 1992.

[15]D.B. Johnaon, D.A. Maltz, "Dynamic Source routing in Adhoc wireless Networks" Mobile Computing, pp.153-181, 1996

[16]Qin, F. and Liu, Y. "Multipath routing in mobile adhoc Network", in proceeding of the international symposium on information processing, Huangshan, China, pp. 237-240, 2009

Sushil Chandra Dimri

He is Professor of the Department of Computer Applications in the Graphic Era University. He received his Ph.D. in Computer Science from the Kumaun University, India, a Master's degree in Computer Science from the ISM, Dhanbad, India, His main research areas are network services in mobile networks, cyberspace and cyber security, Analysis and Designing of Algorithms , and so on.

Kamlesh Chandra Purohit

He is an Assistant Professor of the Department of Computer Applications in the Graphic Era University. He received his Master's in Computer Science from the Graphic Era University, India. His main research areas are network Security in mobile networks, cyber security and so on.

Durgesh Pant

He is Professor of the Department of Computer Applications in the Uttarakhand Open University. He has more than 20 Years Experience in Academics. His main research areas are network services in mobile networks, Software Engineering and Network resource optimization.

Energy Efficient Investigation of Oceanic Environment using Large-scale UWSN and UANETs

Swarnalatha Srinivas*, Ranjitha P†, R Ramya‡ and Narendra Kumar G§

*Dept. of Electronics & Communication, UVCE, Bangalore University
Bangalore, Karnataka 560001, INDIA
swarnalatha.ss@gmail.com

†Dept. of Electronics & Communication, UVCE, Bangalore University
Bangalore, Karnataka 560001, INDIA
ranjitha040391@gmail.com

‡Dept. of Electronics & Communication, UVCE, Bangalore University
Bangalore, Karnataka 560001, INDIA
ramya161091@gmail.com

§Dept. of Electronics & Communication, UVCE, Bangalore University
Bangalore, Karnataka 560001, INDIA
gnarenk@yahoo.com

Abstract—**Investigating coastal oceanic environment is of great interest in pollution monitoring, tactical surveillance applications, exploration of natural undersea resources and predicting wave tides. Deployment of underwater sensor networks for real time investigation is the major challenge. Acoustic communication intends to be an open solution for continuous wireless sensor network in underwater scenarios. In this paper large-scale underwater Sensor Networks (UWSN) and Underwater Ad-hoc Networks (UANETs) using Solar-Powered Autonomous Underwater Vehicles (SAUV) to explore the oceanic environment is proposed. A kong wobbler carrying base station with acoustic communication devices is considered, which locates the pre-deployed underwater sensor modules through acoustic communication. The sensor modules are installed with various sensors and video capturing devices to study the underwater resources as well as for surveillance needs for predicting the environmental conditions. The simulation results are encouraging as this approach is extremely helpful in surveillance as the intruders are tracked and real-time video streaming is done.**

Keywords: *Underwater Ad-hoc Networks (UANET's), Underwater Sensor Networks (UWSN), Solar-Powered Autonomous Underwater Vehicles (SAUV), Acoustic Communication, Underwater Acoustic Sensor Networks (UW-ASN), Geographic Adaptive Fidelity (GAF) Protocol, Kong Wobbler.*

I. Introduction

Underwater environment investigation is vital in predicting wave tides, pollution monitoring, oceanic data collection, tactical surveillance applications, disaster prevention and exploring natural resources. The largely unexplored vastness of the ocean, covering about 79% surface of the earth, has fascinated human race for a very long time. The traditional approach for ocean-bottom or ocean column monitoring is to deploy underwater sensors that record data during the monitoring mission, and then recover the instruments finds disadvantages:

- Real time monitoring is critical especially in surveillance or in environmental monitoring applications such as seismic monitoring wherein the recorded data cannot be accessed until the instruments are recovered, which may happen several months after the beginning of the monitoring program.
- No interaction is possible between onshore control systems and the monitoring instruments which impedes adaptive tuning of the instruments nor it is possible to reconfigure the system.
- In case of failures, it is not possible to detect them before the instruments are recovered which leads to the complete failure of a monitoring mission.
- The amount of information that can be recorded during the monitoring mission by every sensor is limited by the capacity of the onboard storage devices in the instrument.[1]

To assess the aqueous environment, its role and function call for the need of large-scale, long term and distributed information collection networks for periodic oceanic monitoring. The large scale aquatic applications demand us to build large-scale underwater Sensor Networks (UWSN) and Underwater

Ad-hoc Networks (UANETs) to explore the inhibited oceans. A large number of sensor nodes are used for long-term exploration of oceanic environment and gathering scientific data in collaborative monitoring missions. Energy efficient routing protocols are the most important criteria for the design of underwater sensor networks since the sensor nodes will be powered by batteries with limited power capacity. Power failure of a sensor node not only affects the node itself but also its ability to forward data packets to the other nodes. It is seen that in simple acoustic propagation models that multi-hop routing saves energy in underwater networks with respect to single hop communications, especially with distances of the order of some kilometers. In the proposed work Underwater Acoustic Sensor Networks (UW-ASN) consisting of variable number of sensors that are deployed to perform collaborative monitoring tasks over a given area which is implemented adopting GAF (Graphical Adaptive Fidelity) protocol which proves to be an energy efficient routing protocol as the transmitting power is altered according to the distance of the neighboring nodes.

II. Kong Wobbler

A wireless access point with kong wobbler structure is made to float on the surface of sea around the area of investigation to which the structure of the base station is fixed on the top. The weight of the base station is less than the wobbler base structure, maintaining a ratio of 1:4 to prevent the whole structure from toppling. Kong wobbler is made up of non toxic FDA approved polypropylene chosen for its overall strength, impact absorption, sound deadening and non toxicity on the surface of sea made up of high strength polymer and is PBA and phthalate free. Vacuum is maintained inside the structure that allows it to float and uses play grade sand for its weight. The sand remains in a compartment that has been permanently sealed using ultrasonic technology. The kong wobbler is made to float on the surface of sea and anchored to the sea bed through cables to keep in position such that due to its unique structure it floats vertically. They are hand-launched over the side of a ship or air dropped in the area of investigation. The column of fluid has greater pressure at the bottom of the column of ocean than at the top. This difference in pressure results in a net force that tends to accelerate the kong wobbler structure upwards. The magnitude of that force is equal to the difference in the pressure between the top and the bottom of the column, and is also equivalent to the weight of the fluid that would otherwise occupy the column. For this reason, if the density of the structure is greater than that of the fluid in which it is submerged, tends to sink. The buoyancy of the kong wobbler exceeds its weight and tends to rise. Density is maintained lesser than the liquid and shaped appropriately so that force can keep the whole structure afloat. The floating kong wobbler tends to restore itself to an equilibrium position after a small displacement. It has vertical stability, in case it is pushed down slightly, which will create a greater buoyancy force and unbalanced by the weight force, will push the object back up. Rotational stability is of great importance as given a small angular displacement, the structure returns to its original position, (stable).

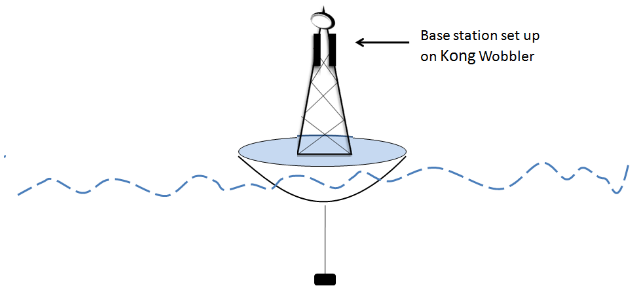


Fig. 1. Kong Wobbler structure

III. Solar-Powered Autonomous Underwater Vehicles (SAUV)

The SAUV is a solar powered AUV designed for long endurance missions that require monitoring, surveillance, or station keeping, with real time bi-directional communications to shore.

- The SAUV is a solar-powered autonomous vehicle which is equipped with rechargeable lithium ion batteries to allow maximum mission endurance even under conditions where minimal solar radiation is available.
- Operate autonomously at sea for extended periods of time from weeks to months. Typical missions require operation at night and solar energy charging of batteries during daytime.
- In case of failure of Kong Wobbler Sauv's Communicate with a remote operator on a daily basis via Satellite phone, RF radio, or acoustic telemetry.
- Operate at speed up to about 3 knots when needed and cruise at speed of about 1 knot.
- Battery system is to provide a total capacity of about 1500 whrs.
- Capability to acquire GPS updates when on the ocean surface and compute SAUV position at all times using GPS when on surface.
- Capability to maintain fixed depth and fixed altitude and to smoothly vary depth or altitude profile.
- Capability to log and upload all sensor data correlated in time and SAUV geodetic position.
- Provide sufficient volume, power, interfaces, and software hooks for future payload sensors.
- Allow user to program missions easily using a Laptop PC and provide for graphical display of mission.[13]

IV. Acoustic Communication

Wireless underwater communication is challenging task with the growing need for underwater surveillance and

Fig. 2. SAUV

develop persistent long-term ocean observation has led to many underwater wireless technologies. Present underwater communication involves transmission of data in the form of optical waves, electromagnetic or sound waves. Optical waves involved in underwater communication are generally limited to very short ranges because of the strong backscatter from the suspended particles in the ocean, severe absorption by water at optical frequencies and high level of ambient light in the upper part of the water column. Even the clearest water has 1000 times the attenuation of clear air and turbid water has more than 100 times the attenuation of densest fog. Acoustic communication is the most versatile and widely used technique in underwater wireless communication which has low attenuation of sound in water used as the primary carrier for underwater wireless communication systems that holds well in thermally stable and deep water settings.[2]

A. AD-HOC Networks

A wireless ad hoc network is a system of self-directed nodes which form a decentralized communications network. Wireless communication allows for a dynamic network topology where new nodes can be rapidly deployed and likewise rapidly removed. Nodes act as both host and router, performing tasks and forwarding information to each other. Mobile nodes can form dynamic networks where they are linked with their nearest neighboring node and when they move too far from their neighboring nodes might lose connection but come into contact with other nodes to begin interacting and changing the network topology. Efficient routing protocols is needed to communicate new data over multi-hop paths consisting of possibly several links to cope with noise and interference as well as sharing limited bandwidth. A class of Ad hoc networks, Underwater Ad-hoc Networks (UANET) are used in underwater explorations.

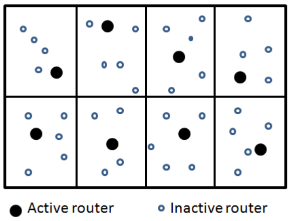
1) UWSN and UANETs: Large scale Underwater Ad-hoc Networks (UANET) and Underwater Sensor Networks (UWSN) are essential to explore large uninhibited oceans. In the characteristics of these new networks, the propagation delay, floating node mobility, and limited acoustic link capacity are hugely different from ground based mobile ad-hoc networks (MANET) and wireless sensor networks (WSN). UANET and UWSN rely on low-frequency acoustic communications because RF radio does not propagate well due to underwater energy absorption. Unlike wireless links amongst land-based ad hoc nodes, each underwater acoustic link features large-latency and low bandwidth. Most ground sensor nodes in a WSN are typically stationary and large portion of UWSN sensor nodes, except some fixed nodes mounted on the sea floor are with low or medium mobility (3-5 knots) due to environmental water current. The large-scale aquatic applications demand to build large-scale Underwater Ad-hoc Networks (UANET) and Underwater Sensor Networks (UWSN) to explore the large uninhabited oceans. The difference between UANET and UWSN is due to controlled mobility and associated implementation cost. In a UANET, mobile nodes can be implemented by Solar-Powered Autonomous Underwater Vehicles (SAUV) and Autonomous Underwater Vehicles (AUV) or Remotely Operated Vehicles (ROV), which are high cost robots that can move under the water by following pre-programmed or autonomous motion patterns. On the other hand, UWSN only incurs a fraction of implementation cost of UANET at the same network scale. All sensor nodes in a UWSN are of low-cost.[3]

The advantages of the new UANET and UWSN paradigm are:

- Localized and coordinated sensing and attacking is far more precise than the existing remote telemetry technology, eg, those relying on directional frequency and ranging (DIFAR) sonobuoy or magnetic anomaly detection (MAD) equipment.
- Scalability of UWSN ensures that a large area can be covered for time-critical applications.
- Casualty ratio is expected to be zero if unmanned UANET and UWSN platforms are used.
- Implementing reusable underwater nodes reduces the deployment and maintenance cost. Each underwater sensor unit can be bundled with an electronically controlled air bladder device. Once the network mission is accomplished, the command center issues commands to trigger all air-bladder devices and all sensor units float to surface to be recollected for next mission.[1]

V. GAF Protocol

In underwater applications, it is vital to let every underwater node know its current position and the synchronized time with respect to other coordinating nodes. GAF protocol uses Global Positioning System (GPS) to get the node location.As Global Positioning System (GPS) is unavailable under the water surface as the high-frequency radio waves used by Global Positioning System (GPS) is quickly absorbed by water, hence cannot propagate deeply under the water surface. Therefore underwater networks rely on Doppler Instrumentation or distributed GPS-free localization and time synchronization schemes to let the sensor nodes know their positions and the network clock value. In other words, before the network can use geo-routing schemes, it needs a multi-hop packet delivery service, which must be GPS-free. Geographic adaptive fidelity protocol is an energy effective position based routing protocol. Position based protocols are also referred to as geographic routing protocols as the sensor nodes are addressed by means of their locations instead of the information that they carry. Location information is needed in order to calculate the distance between two particular nodes so that energy consumption can be estimated. In GAF protocol, each node uses location information to associate itself with a virtual grid so that the entire area is divided into several square grids, and the node with the highest residual energy within each grid becomes the master of the grid. Only a single node from a cell of a given virtual grid is chosen to be active at any given time. Nodes will select one sensor node to stay awake for a certain period of time which is responsible for monitoring and reporting data to the sink on behalf of the other nodes in the zone is known as the master node. Other nodes in the same grid can be regarded as redundant with respect to forwarding packets, and thus they can be safely put to sleep without sacrificing the routing fidelity.



GAF virtual grid. Only one active node per cell.

Fig. 3.

VI. System Architecture

The general architecture of underwater sensor network is reviewed before describing the specific applications. The rough capabilities of a sensor node are estimated on its interaction with the environment, other underwater nodes, and applications. At the lowest layer is the large number of sensor nodes to be deployed on the sea floor which has computing power, and storage capacity. They collect information through their sensors and communicate with other nodes through short-range acoustic communication. In large networks, there exists a type of nodes, called supernodes, having access to higher speed networks and can relay data to the base station very effectively with rich network connectivity creating multiple data collection points. Battery power and the ability to carefully monitor energy consumption are essential for the sensor node. All components of the system operate at as low a duty cycle as possible which is enabled by examining each layer of system software to minimize energy consumption and in addition nodes entirely shut off for very long periods of time, up to hours or days when not in use. In a harsh, underwater environment, some nodes will be lost over long deployments due to fishing trawlers and underwater life affecting cables or node which needs redundancy in communication and sensing as loss of a node will not have wider effects. In addition, multiple failures can be recovered, either with mobile nodes, or with human deployment of replacements.[4]

A. Sensors

All nodes are integrated with temperature, vibration, pressure, viscosity, turbidity, seismic, proximity, chemical & gas, fluid flow, speed, water level, altitude and visibility sensors to measure the different parameters in the marine environment which are small, robust, inexpensive, low power consuming yet efficient.

- Temperature sensors: Temperature measurements in marine applications make high demands on sensor reliability. They may be subject to changes during their lifetime. A periodic calibration of temperature sensors is required to make sure that they display a correct value of temperature.
- Turbidity sensor: Turbidity is measured in terms of the amount of scattering and absorption of light rays by small particulate matter suspended. It is an ecologically important factor because a high value of turbidity means a high amount of suspended particles which can affect the aquatic life. Thus this sensor works by the illuminating the medium by infrared light emitted by two LEDs to a common centre and received by a symmetrically placed photodiode of the required wavelength. The amount of backscattered light is the measure of the turbidity of the medium measured in NTU (Nephelometric Turbidity Unit).
- Seismic sensors: Seismic sensors detect the vibrations or sounds. A standard piezo sensor is used to detect vibrations/sounds due to pressure changes. They are very sensitive and can detect vibrations caused by any movement. So it can be used to monitor area of investigation.
- Pressure sensors: Pressure sensors are used for permanent immersion in freshwater and sea-water to measure the level of rivers, seas and tidal waters.

Real-time readings are taken from all the sensors, when the readings cross the threshold values necessary actions are taken.

VII. Implementation

Initially the UWSNs and UANETs are deployed in the ocean. Sensor nodes with limited battery power are deployed to record the environmental changes underwater. The recorded data is transferred to the surface of the earth through the nearest access point, the base station fixed to the specially designed kong wobbler floating on the surface of the water by the multi-hop network of sensor nodes which results in energy savings and increased network capacity. The base stations

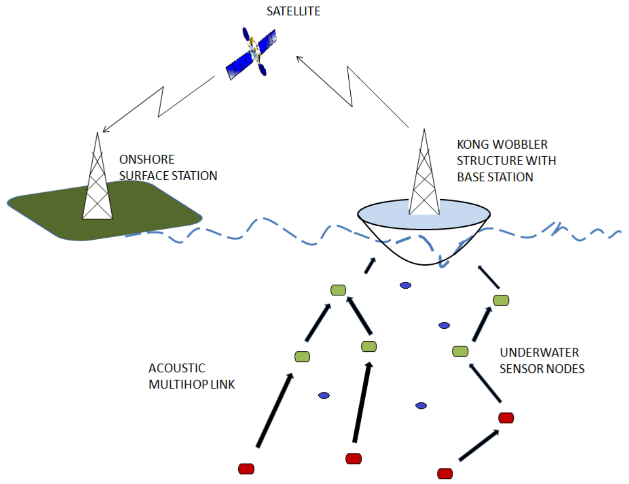


Fig. 4. Logical diagram of System Architecture.

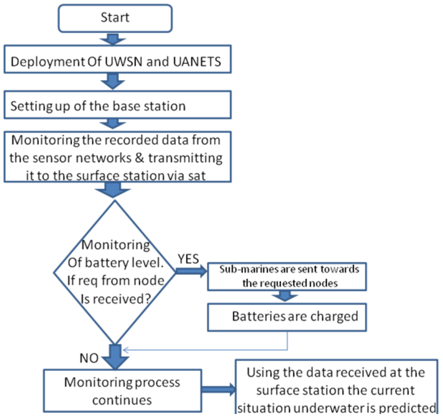

Start
Deployment Of UWSN and UANETS
Setting up of the base station
Monitoring the recorded data from the sensor networks & transmitting it to the surface station via sat
Monitoring Of battery level. If req from node Is received?
YES
Sub-marines are sent towards the requested nodes
Batteries are charged
NO
Monitoring process continues
Using the data received at the surface station the current situation underwater is predicted

Fig. 5. Flow Chart of the system

monitor the entire sensor nodes within the network and receive as well as stores the recorded sensor data within its area of range. These data are further relayed to an onshore surface station via satellite transceiver. Acoustic communication proves to be efficient in data transfer underwater from the sensor nodes to the base station. During the process, the base station monitors the battery power level of all the nodes. Upon receiving the request from the nodes to recharge its batteries, the base station then guides the submarines towards the requested nodes which is continued with all the sensor modules deployed underwater. The sub-marines would in turn charge themselves at the base stations, which are installed with the solar panels. The power required by the base station for the reception of the data from the sensor modules and transmission is supplied by the solar panels. Once the surface station has finished collecting data from each sensor module the processing and analysis of data is performed to get the real time study of the underwater scenario. As the nodes have limited battery power, it is essential to implement an energy efficient routing protocol that conserves power during transmission and reception of data. Power failure of a node not only affects the node itself but also its ability to forward packets on behalf of other nodes and thus the overall network lifetime. GAF is an energy efficient routing protocol as the transmitting power is altered according to the distance of the neighboring nodes. The flow chart explains the process undertaken in the investigation of underwater environment.

The investigation of the underwater resources is thus collected at real time. Compilation of all the recorded sensor data is done in the surface station which acts as the command center and thus predicting the current scenario of the environment underwater.

VIII. Simulation and Results

Continuous network development and higher functionality requirements have created the need for tools that could monitor network transmissions and analyze them. Network Simulator (NS) for communication networks works under UNIX and Windows system platforms and is mainly used for network research. The simulation is performed using NS2 (version 2.34) running on LINUX platform (ubuntu 11.04). The graphical representation of this simulation is shown with Network Animator(nam-1.14). 20 nodes are considered for each kong wobbler base station on the surface of ocean bed. The kong wobbler carrying base station act as the center that monitors the entire sensor node within the network and receives as well as stores the recorded sensor data. UANETs along with kong wobbler structure is hand dropped by the side of the ship. Kong wobbler structure starts to float on the surface of the ocean and the UANETs start moving randomly. The black fixed nodes are anchored to the ocean bed at specific co-ordinates, Fig:6. Some of the nodes in UANET are mobile which are implemented by Solar-Powered Autonomous Underwater Vehicles (SAUV) and Autonomous Underwater Vehicles (AUV) or Remotely Operated Vehicles (ROV), which are high cost robots that move under the water by following pre-programmed or autonomous motion patterns. And remaining nodes in UWSN are stationary, mounted on the sea floor are with low or medium mobility (3-5 knots) due to environmental water current. The trace file and nam file results provided by the ns2 gives enormous amount of information. It specifies position of the node, number of nodes within the network of access point and also visualizes in detail about the packet transmission among the nodes and the kong wobbler carrying base station is simulated. The red nodes represent nodes in UANET on the ocean bed which are moving around collecting data. The black nodes are fixed nodes anchored to oceanic bed. The blue node indicate the Kong wobbler structure with base station which covers the area under investigation.

UANETs continue their random motion gathering data. Both stationary UWSNs and mobile UANETs collect information and relay it to base station which is carried by kong wobbler structure, Figs:7,8,9,10,11,12.

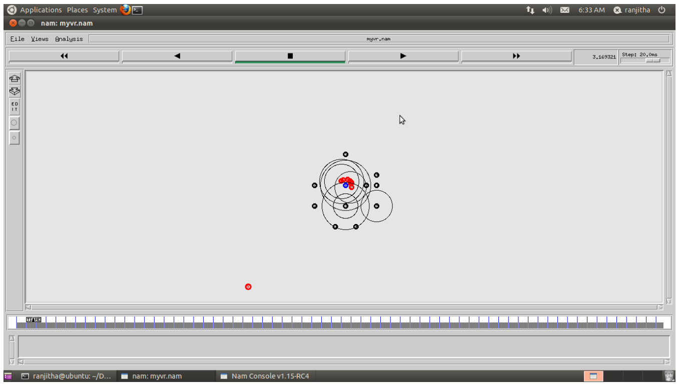


Fig. 6. Simulation in ns2

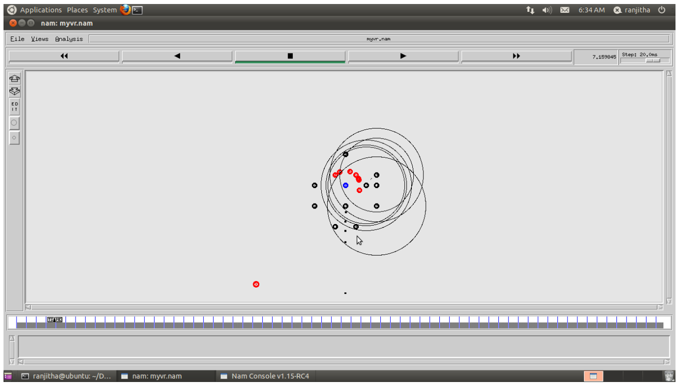


Fig. 7. Simulation in ns2

IX. Conclusion

This paper has summarized our ongoing research in underwater sensor networks, including applications and research challenges. It is explained that traditional approach to deploy underwater sensors that record data during the monitoring mission, then recovering the instruments is not a feasible and the need of large-scale long-term and distributed information collection networks for periodic oceanic monitoring is essential. GAF (Graphical Adaptive Fidelity) protocol was adopted as it proves to be an energy efficient routing protocol. it is also explained that acoustic communication is the most versatile technique in underwater wireless communication. The applications of large scale Underwater Ad-hoc Networks (UANET) using Solar-Powered Autonomous Underwater Vehicles (SAUV) and Underwater Sensor Networks (UWSN) and their reliability in implementing a localized, precise, and large-scale networking efficiently than any existing small-scale Underwater Acoustic Network (UAN) is described. The main objective of the paper is to develop advanced communication techniques for efficient real time investigation of large uninhibited oceans. Development of underwater communication and networking for enhanced oceanic monitoring is also essential for pollution monitoring, tactical surveillance, exploration of natural undersea resources, predicting wave tides and various applications.

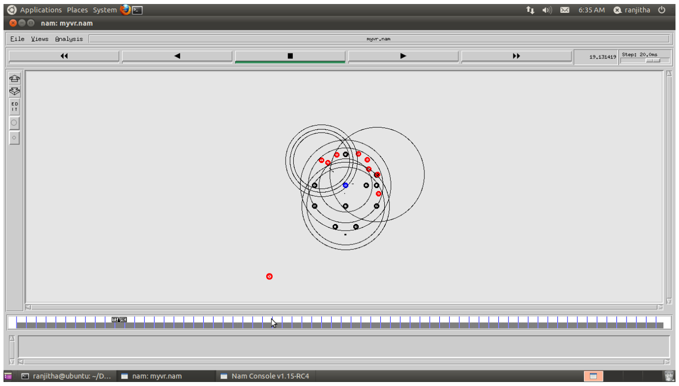


Fig. 8. Simulation in ns2

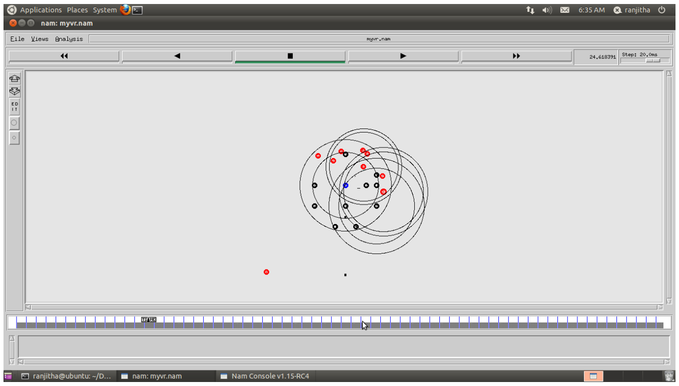


Fig. 9. Simulation in ns2

References

[1] Ian F. Akyildiz, Dario Pompili, Tommaso Melodia. *Challenges for Efficient Communication in Underwater Acoustic Sensor Networks*.ACM Sigbed Review.

[2] Jun-Hong Cui, Jiejun Kong, Mario Gerla, Shengli Zhou. *Challenges: Building Scalable and Distributed Underwater Wireless Sensor Networks (UWSNs) for Aquatic Applications*. Special Issue of IEEE Network on Wireless Sensor Networking, May 2006.

[3] Jiejun Kong, Jun-hong Cui, Dapeng Wu, Mario Gerla.*BUILDING UNDERWATER AD-HOC NETWORKS AND SENSOR NETWORKS FOR LARGE SCALE REAL-TIME AQUATIC APPLICATIONS*.IEEE MILCOM, 2005.

[4] John Heidemann, Yuan Li, Affan Syed, Jack Wills, Wei Ye.*Underwater Sensor Networking: Research Challenges and Potential Applications*.The IEEE Wireless Communications and Networking Conference, Las Vegas, Nevad, USA, April 2006.

[5] Sudhakar Pillai M, Pranav P D, and Narendra Kumar G.*MANET Based Dynamic Power Conscious Emergency Communication Module*.LISS, 2010, China.

[6] Giuseppe Anastasi, Marco Conti, Mario Di Francesco, Andrea Passarella.*Energy Conservation in Wireless Sensor Networks: a Survey*.

[7] Robert Been, David T. Hughes, Arjan Vermeij.*Heterogeneous underwater networks for ASW: technology and techniques*. OCEANS 2010 IEEE Sydney.

[8] Lanbo Liu, Shengli Zhou, and Jun-Hong Cui. *Prospects and Problems of Wireless Communication for Underwater Sensor Networks*. Journal: Wireless Communications & Mobile Computing - Underwater Sensor Networks: Architectures and Protocols.

[9] Ian F. Akyildiz , Dario Pompili, Tommaso Melodia.*Underwater acoustic sensor networks: research challenges. Telecommunications (ICT)*.2010 IEEE 17th International Conference.

[10] Raja Jurdak, Cristina Videira Lopes, Pierre Baldi. *BATTERY LIFETIME ESTIMATION AND OPTIMIZATION FOR UNDERWATER SENSOR NETWORKS*.

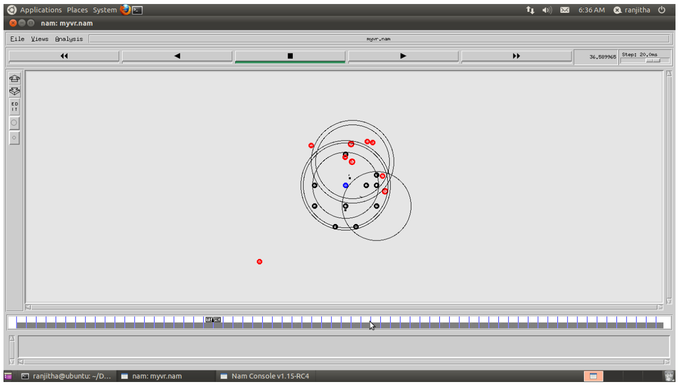


Fig. 10. Simulation in ns2

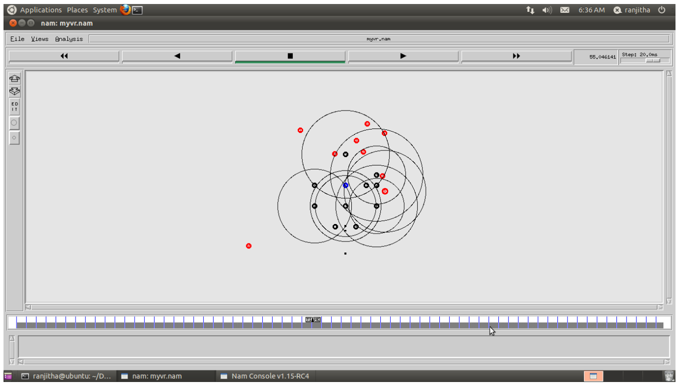


Fig. 11. Simulation in ns2

[11] Sudhakar Pillai M, Pranav P Deshpande, Chetan B M, Smitha Shekar B and Narendra Kumar G.*Efficient Performance Of MANETs In Coal Mines*, International Multi-Conference on Informatics and Cybernetics, IMNCIC 2011, Florida, USA march 2011.

[12] Swarnalatha Srinivas, Ranjitha P, R Ramya and Narendra Kumar G. *Investigation of Oceanic Environment using Large-scale UWSN and UANETs*, The 8th International conference on Wireless Communications, Networking and mobile computing, Shanghai, China, sep 21-23, 2012.

[13] James Jalbert, John Baker, John Duchesney, Paul Pietryka, William Dalton Acoustikos Div. Falmouth Scientific, Inc. D.R. Blidberg, Steve Chappell Autonomous Undersea Systems Institute Robert Nitzel, Technology Systems, Inc. Dr. Ken Holappa, Private Consultant. *SOLAR-POWERED AUTONOMOUS UNDERWATER VEHICLE DEVELOPMENT*, JalbertEtal2003.

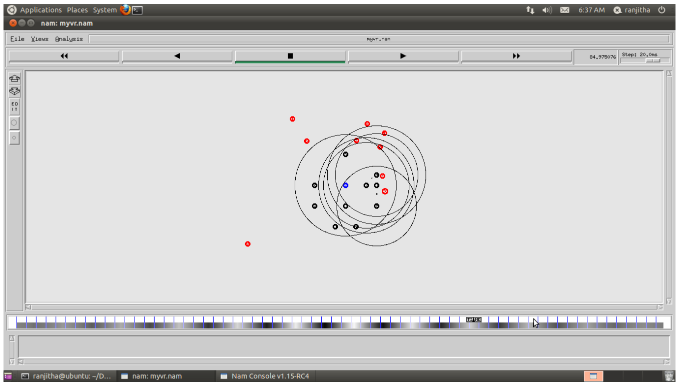


Fig. 12. Simulation in ns2

Biography

Swarnalatha Srinivas, born in Bangalore on 22nd October, 1964. Obtained Bachelors Degree in Electrical Engineering from University Visvesvaraya College of Engg., Bangalore, Karnataka, India in 1988. Obtained Masters degree in Power Systems, University Visvesvaraya College of Engg., Bangalore, Karnataka, India in 1992.
Currently Associate Professor in the Department of Electrical Engineering, Bangalore Institute of Technology, VTU, Bangalore, held the positions of Associate Professor, Lecturer, currently pursuing PhD under the guidance of Dr Narendra Kumar G.

Dr. Narendra Kumar G, born in Bangalore on 5th February, 1959. Obtained Masters Degree in Electrical Communication Engineering, (Computer Science & Communication) from Indian Institute of Science, Bangalore, Karnataka, India in 1987. Was awarded PhD in Electrical Engineering(Computer Network) from Bangalore University, Bangalore, Karnataka, India in 2006.
Currently Professor in the Department of Electronics & Communication Engineering, University Visvesvaraya College of Engg., Bangalore University, Bangalore, held the positions of Associate Professor, Lecturer and Director of Students Welfare. Research interests include Mobile Communication, Wireless Communication, E-Commerce, Robotics and Computer Networks.

Ranjitha P and R Ramya are students doing research under the guidance of Prof. Narendra Kumar G.

A New Algorithm for Calculating the Daytime Visibility Based on the Color Digital Camera

Xiaoting Chen [1], Changhua Lu [1,2], Wenqing Liu[2], Yujun Zhang[2]

[1] School of Computer and Information, Hefei University of Technology, Hefei, Anhui, China, 230009

[2] Anhui Institute of Optics and Fine Mechanics, Chinese Academy Sciences, Hefei, Anhui, China, 230031

Abstract

There are some deficiencies in the traditional daytime visibility calculation method using CCD digital camera: visibility observation value is accurate while using the artificial objects， but expensive; nonzero internal reflection coefficient of the target lead to inaccurate visibility observation results when using natural objects as the target. A new daytime visibility algorithm based on color CCD digital camera is proposed in this paper: we use the color images obtained by the CCD digital camera, then estimate transmission of color digital image by using the knowledge of dark channel prior, and obtain the atmospheric attenuation coefficient form transmission to calculating visibility value. The experimental data show that the max-error of the proposed algorithm is less than 20% which conform to the error provisions of WMO on the meteorological visibility instrument, and simple operation, without artificial target, low cost.

Keywords: *Color digital camera, Daytime visibility, A new visibility Algorithm, Estimated transmission.*

1. Introduction

In recent years, with the development of computer technology and CCD digital camera technology, design and research of CCD digital camera visibility meter rapid development. The United States began research in calculating the daytime visibility based on Video image in 1998[1]. The traditional daytime visibility calculation method based on CCD digital camera mainly includes the frequency domain method and time domain calculation method[2,3], and most of them use gray image—by calculating contrast between the target and the sky background, or calculating frequency change of the target to obtain the visibility value. There are some deficiencies here: visibility observation value is accurate while using the artificial objects，but expensive; nonzero internal reflection coefficient of the target lead to inaccurate visibility observation results when using natural objects as the target. For correcting this error, we need a long-term experiment to fitting reflection coefficient, and recalibrate observations in the fixed environment[1], [4], [5], [6], [7].

Based on several factors of the atmospheric environment impacting atmospheric visibility, according to the atmospheric transport model, a new daytime visibility algorithm based on color CCD digital camera is proposed in this paper: we use the color images obtained by the CCD digital camera, then estimate transmission of color digital image by using the knowledge of dark channel prior, and obtain the atmospheric attenuation coefficient from transmittance to calculating visibility value. The proposed algorithm completes daytime visibility calculation based on color CCD digital camera from the perspective of the image processing. The experimental results show that the new algorithm is feasible, and simple operation, without artificial target, low cost.

2. Algorithm design

2.1 Visibility calculation model

The atmospheric visibility dropped in the atmosphere mainly due to three reasons: the light of object (including self-luminous and reflective) absorbed by the atmosphere before it reaching at the detector; the light of ob scattered by all kinds of atmospheric particulate matter so that it can not reaching at the detector; other background light is scattered into the detector because of the atmospheric particulate matter on the path from the object to the detector. Atmospheric absorption and scattering result in light attenuation jointly. We ignore atmospheric absorption to achieve the proposed algorithm because most of the atmospheric optical attenuation mainly roots in atmospheric scattering.

According to the atmospheric scattering model, the process of atmospheric optical transmission that the light of target reaching at the detector is represented as[8,9]:

$$I(x) = J(x)t(x) + A(x)(1-t(x)) \qquad (1)$$

where x is the plane coordinates of target, I is the scene radiance, J is target light intensity, A is the atmospheric light, t is the atmospheric transmission describing.

The transmission can be calculated with the atmospheric extinction coefficient[8,9]:

$$t(x) = e^{-\sigma(x)r(x)} \qquad (2)$$

Where σ is the scattering coefficient of the atmosphere, r is the scene depth. This formula assumed that the atmospheric extinction coefficient is homogeneous on the path from the object to the detector.

According to the Koschmieder's Law [10], the atmospheric visibility can be expressed as:

$$V = -In\varepsilon / \sigma \qquad (3)$$

Where ε denotes threshold of visual contrast. According to the rules of WMO, we set the contrast threshold for 0.02[14], which is equivalent to the distance that the target disappears. Formula (3) is expressed as:

$$V = -In0.02 / \sigma = 3.912 / \sigma \qquad (4)$$

So, we can get visibility value with transmission t.

2.2 Estimated transmission[11]

We use dark channel prior to estimate transmission. For image J, the dark channel prior[11]:

$$J_{dark}(x) = \min_{c \in \{r,g,b\}}(\min_{y \in \Omega(x)}(J_c(y))) \qquad (5)$$

Where J^c is a color channel of J and $\Omega(x)$ is a local patch centered at x. Taking the min operation[11]:

$$\min_{y \in \Omega(x)}(I^C(y)) = t(x)\min_{y \in \Omega(x)}(J^C(y)) + \min_{y \in \Omega(x)}(1-t(x))A^C \qquad (6)$$

The min operation is performed on three color channels independently, this equation (6) is equivalent to[11]:

$$\min_{y \in \Omega(x)}(\frac{I^c(y)}{A^c}) = t(x)\min_{y \in \Omega(x)}(\frac{J^c(y)}{A^c}) + (1-t(x)) \qquad (7)$$

We take the min operation among three color channels on the above equation and obtain[11]:

$$\min_c(\min_{y \in \Omega(x)}(\frac{I^C(y)}{A^C})) = t(x)\min_c(\min_{y \in \Omega(x)}(\frac{J^C(y)}{A^C})) + (1-t(x)) \qquad (8)$$

As A^c is always positive, this leads to $A^c > 0$, then we have[11]:

$$t(x) = 1 - \min_c(\min_{y \in \Omega(x)}(\frac{I^C(y)}{A^C})) \qquad (9)$$

So, we can gain region object transmission from image J，then calculate atmospheric visibility. We pick the top 0.1% brightest pixels in the dark channel [8]. Among these pixels, the pixels with highest intensity in the input mage J is selected as the atmospheric light.

It is noteworthy that if the image contains large brightness higher sky area, due to we can not find pixel intensity close to zero in the dark channel, the dark channel prior is not established in this area. J.G. Jiang et al estimated transmission of the bright area by using the tolerance mechanism solely in order to adjust the transmission of the bright areas[12]. But for the proposed algorithm, as the selection of the target area is the dark objects or surfaces, rather than the bright objects or surfaces, the transmission of this dark objects or surfaces is changeless no matter the tolerance mechanism is used or not.

2.3 The selection of target area

The selection of target area in the image needs to according to different situations in this paper. Generally, for obtaining accurate results of choosing target area, two steps judgment should be done: (1) selecting dark objects or surfaces as the target area; (2) judging if there is a bright single channel in target areas. The dark objects or surfaces mean that a channel of the target area will have very low intensity value, which accord with the application conditions of using the dark channel prior. Exiting a bright single channel in target areas means that position of the target area is far, and the baseline is long, so that the range of visibility observable is far. Fig. 1 shows the selection process of target area：

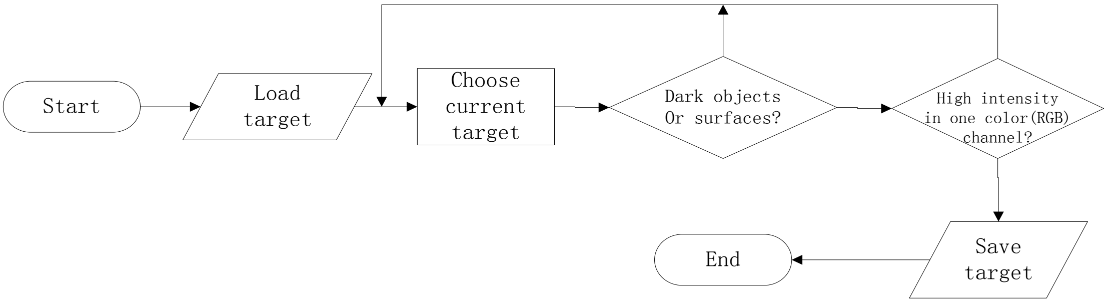

Start
Load target
Choose current target
Dark objects Or surfaces?
High intensity in one color(RGB) channel?
Save target
End

Fig. 1: Flow chart for the selection of target area

Those who need a specification is, condition 2 is not established when visibility is poor. Either channel intensity may be not very high in the situation of poor visibility. And this means that the current visibility range is low. In this case, if we can not find the target area which can satisfy the two conditions at the same time, we abandon condition 2, and only judge condition 1 to discover the optimal target object.

3. Data Processing and A Short Time Experiment

In the experiment, we capture images by using CCD digital camera on the roof of Hefei Institute of Physical Science. The experimental results are compared with the forward scatter visibility meter (FD12). In order to ensure the accuracy of the experiment, CCD digital camera and forward scatter visibility meter keep consistent direction and angle of view in the experiment. For example, the image was taken at 9:00 am on November 6, 2012. Fig. 2 show an original image. The yellow area is the target area, and the area in the red box is enlarged the target details. Fig. 3 show the estimated transmission map from an input image. The experimental procedure as follows: capturing the image by the CCD digital camera, choosing a target area, estimating transmission of the target area, and then calculating the visibility value.

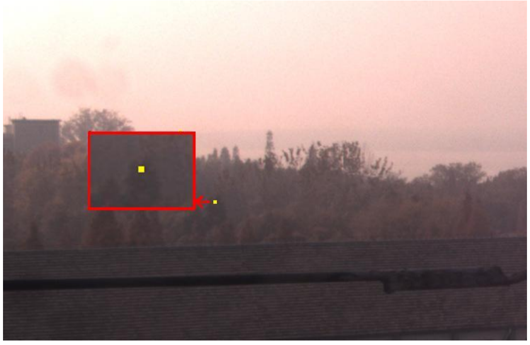


Fig. 2: Original image

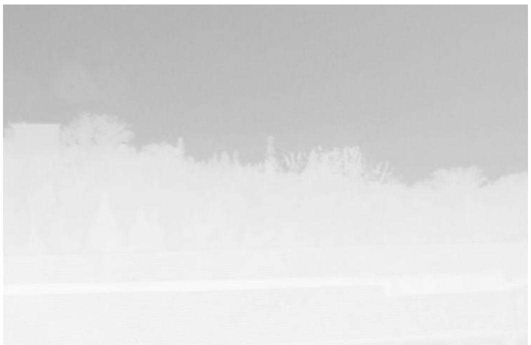


Fig. 3: Estimated transmission map

For Fig. 2, we gain visibility value is 1538.2 meters. Visibility value of the forward scatter visibility meter is 1604.8 meters, and the error is 4.15%.

We have collected data from 8:30 am to 11:30 am on November 2, 2012. The weather was clear turning to cloudy with the temperature was 17-23^{o}C, and the wind was east 3-4 level. The lens of digital camera toward the southwest, and resolution was set to 740*480. Every five minutes carried out an information collection and each minute got ten images. We treated average results of the ten images processing value as the visibility value. Fig. 4 shows the comparison of visibility between the proposed algorithm and the forward scatter visibility meter. X-axis represents time while the Y-axis represents the visibility values in meters. The dotted line represents the results of the forward scatter visibility meter, and the solid line shows the results of the proposedthe proposed algorithm.

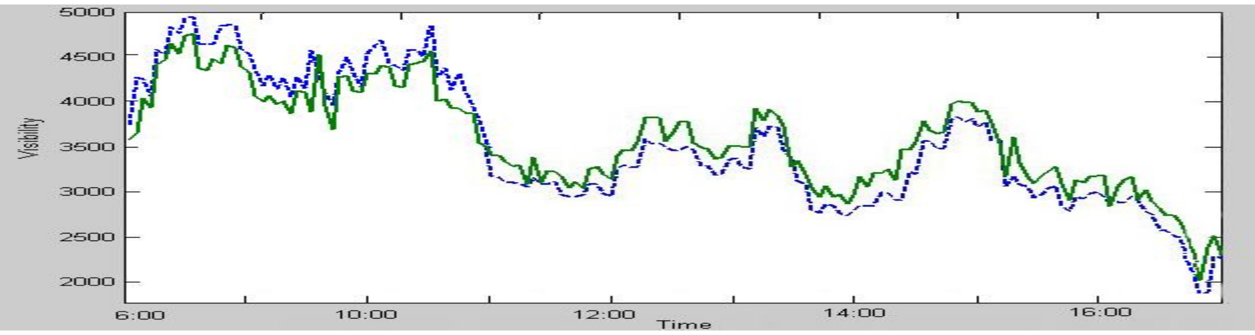


Fig. 4: The comparison of visibility on Nov.2

In Fig. 4, the negative error value indicates that the visibility value of the proposed algorithm is larger than the results of the forward scatter visibility meter. From the experimental results, we can know the correlation coefficient between the proposed algorithm and the results of the forward scatter visibility meter. Their average standard relative error is 6.13%, and the biggest relative error is 8.56%. The correlation coefficient of two results is 0.8124, so that Our results and forward scatter visibility meter's results have a good correlation. Relative RMSE of two results is 6.69%, and that means the difference of distribution between two result is small.

4. Long Time Experiment In the Natural

In order to verify the effect of the proposed algorithm further, a long time experiment was done from November 24 to November 26 in Beijing. The site we chosen is in the south of Beijing with very few human building, and the comparison of visibility between the proposed algorithm and the forward scatter visibility meter is shown in Fig. 5, Fig. 6, Fig. 7.

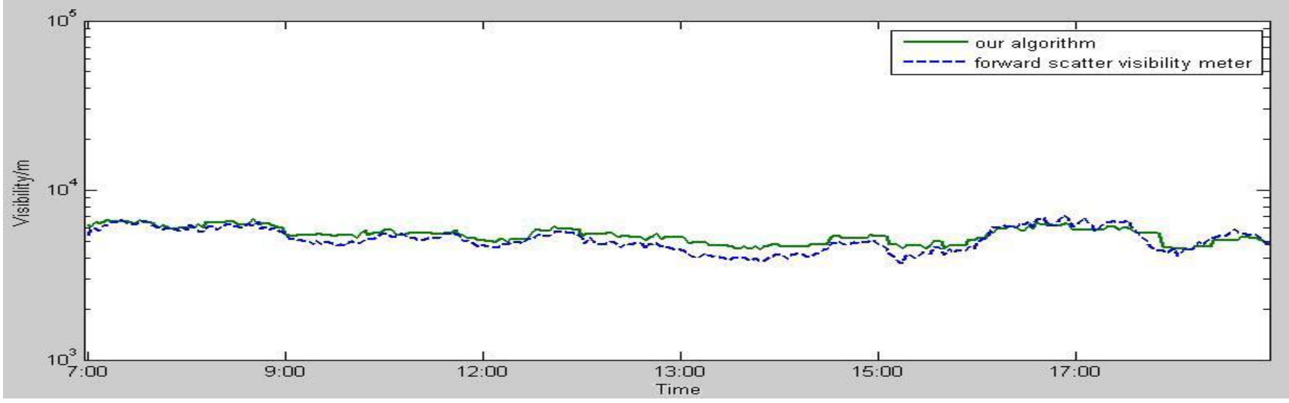


Fig. 4: The comparison of visibility on Nov.24

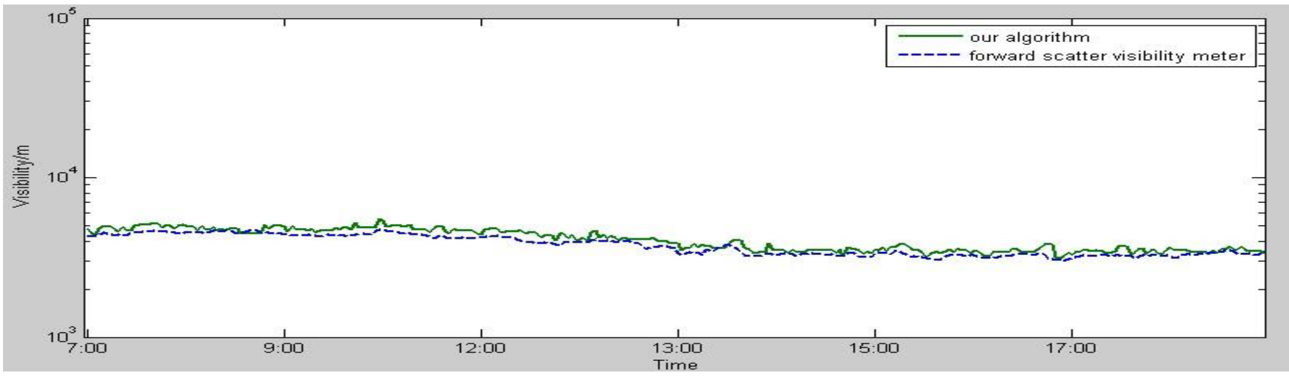


Fig. 4: The comparison of visibility on Nov.25

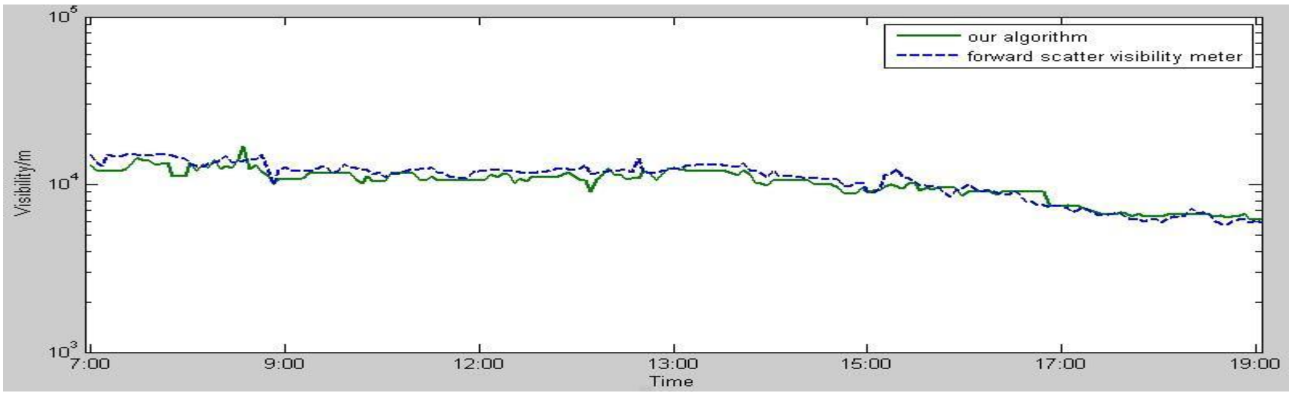


Fig. 4: The comparison of visibility on Nov.26

The correlation coefficient and relative RMSE between the proposed algorithm and the forward scatter visibility meter is shown in Table 1:

Table 1 The correlation coefficient and relative RMSE

Date	Correlation Coefficient	Relative RMSE
2012.12.24	0.7942	7.62%
2012.12.25	0.8120	7.93%
2012.12.26	0.8432	8.18%

According to the experimental results we can know, we known the correlation coefficient the proposed algorithm and the forward scatter visibility meter is less than 0.85, and the difference of distribution between two results is less than 8.2%. Therefore the distribution of the proposed algorithm has good consistency with forward projection instrument results. Cause the error range parameter of FD12 is 10%, so the maximum error of the proposed algorithm is less than 20% which satisfied the error provisions of WMO on the meteorological visibility instrument[13,14], and It proves that the proposed algorithm is feasible. The data processing of this experiment is the same as last experiment.

4. Conclusion and Future Work

A new daytime visibility algorithm based on color CCD digital camera is proposed in this paper. The proposed algorithm can deal natural images directly to gain visibility value, without artificial object and low cost. The proposed algorithm eliminates the error which caused by the no-blackbody characteristics of the target, and obtains good experimental results in visibility observation range 1500-3000m. It's worth to discuss some questions for this algorithm:

1.The atmospheric light we chosen can not accurately reflect the really atmospheric light. How to gain the accurate atmospheric light from the color image is a problem which is worth studying.

2. In acquiring images with a CCD camera, light levels and sensor temperature are major factors affecting the amount of noise in the resulting image[15]. We further research work is how to de-noise for a more accurate visibility value.

Acknowledgments

The help of P. D. Dai, the doctor of Anhui Institute of Optics and Fine Mechanics, Chinese Academy Sciences is gratefully acknowledged in the experiment of this paper[15].

References

1 T.M. Kwon, "An automatic visibility measurement system based on video cameras". Publication MN/RC-1998-25. Minnesota Department of Transportation.

2 W. T. Lu, S. C. Tao, Y. F. Liu, Y. B. Tan, B. G. Wang, "Measuring Meteorological Visibility Based on Digital Photography-Dual Differential Luminance Method and Experimental Study", Chinese Journal of Atmospheric Sciences, Vol.28, No.4, 2004, pp.559-570.

3 D. Baumer, S. Versick, B. Vogel, "Determination of the visibility using a digital panorama camera". Atmospheric Environment, Vol.42(11), 2008, pp. 2593-2602.

4 D. Allard , I. Tombach, "The effects of non-standard conditions on visibility measurement", Atmos. Environ , Vol.15, 1981 , pp.1847-1857.

5 T. J. Mao, J. G. Li, "Visibility and telephoto meter", Chinese Journal of Atmospheric Sciences, Vol.8 , 1984, pp.170-177.

6 W. T. Lu, S. C. Tao, Y. F. Liu, et al, "Application of Pr

Blackbody Technique to Digital Photography Visiometer System", Journal of Applied Meteorological Science, Vol.14 , 2003, pp.691-699.

7 R. G. Hallowell, M. P. Matthews, P. A. Pisano, "Automated Extraction of Weather Variables from Camera Imagery," Mid Continent Transportation Symposium, Iowa State University, Ames, Iowa, 2005.

8 R. Tan, "Visibility in bad weather from a single image", in IEEE Conference on Computer Vision and Pattern Recognition, June 2008, pp.1–8.

9 S. G. Narasimhan, S. K. Nayar, "Vision and the atmosphere" International Journal of Computer Vision, Vol.48, 2002, pp. 233–254 .

10D. H. Lenschow, V. Viezee, R. Lewis, et al, Atmospheric boundary layer detection, China Meteorological Publising, 1990.

11K. M. He, J. Sun, X. O. Tang, "Single image haze removal using dark channel prior", IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp.1956-1963.

12J. G. Jiang, T. F. Hou, M. B. Qi, "Improved algorithm on image haze removal using dark channel prior", Journal of Circuits and Systems, Vol6, No2, 2011, pp.7-12.

13J. L. Wang, L. L. Cheng, X. F. Xu, "Digital camera measurement visibility instrument system than the experimental", Meteorological Science and Technology, Vol30, No6, 2002, pp.353-357.

14D. J. Griggs, D. W. Jones, M. Ouldridge, et al, "The first WMO Intercomparison of Visibility Measurements, United Kingdom 1988/1989", in Instruments and Observing Methods Report, Final Report NO.41, WMO/ TD- NO.401, 1990.

15Rubeena Vohra, Akash Taya, "Image Restoration Using Thresholding Techniques on Wavelet Coefficients", International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, 2011,pp.400-404.

Xiaoting Chen received the B.E. Degree from Guangxi Normal University, Guilin, China, in 2005. Now she is a student in the Intelligent Detection Laboratory at Hefei University of Technology to study for a Master Degree in Engineering. Her main research interests include signal and information processing, intelligent detection algorithm.

Changhua Lu received a Master Degree in Engineering from Harbin Institute of Technology in 1988, and a Ph. D. degree from Chinese Academy Sciences in 2001. He is now a professor and doctoral tutor at Hefei University of Technology, and also a doctoral tutor in Hefei Institutes of Physical Science. His research interests cover signal detection and processing, computer application, DSP technology, photoelectric information processing and automatic test system.

A Bio-Inspired Algorithm based on Membrane Computing for Engineering Design problem

Jian-hua Xiao[1*], Yu-fang Huang[2], Zhen Cheng[3*]

**[1] Logistics Research Center, Nankai University
Tianjin 300071, China**

**[2] College of Mathematics, Southwest Jiaotong University
Chengdu 610031, China**

**[3] College of Computer Science and Technology, Zhejiang University of Technology
Hangzhou 310023, China**

Abstract

Membrane computing is an emergent branch of natural computing, which has been extensively used to solve various NP-complete and intractable problems. In this paper, a bio-inspired algorithm based on membrane computing (BIAMC) is proposed to solve the engineering design problem. BIAMC is designed with the framework and rules of a cell-like P systems, and particle swarm optimization with the neighborhood search. Simulation and experimental results demonstrate that the improved algorithm is valid and outperforms other evolutionary algorithms for engineering design problems.

Keywords: *Engineering Design Problem; Membrane Computing; Particle Swarm Optimization; Neighborhood Search*

1. Introduction

Membrane computing (P systems) was initiated by Paun [1] in 1998, which is a class of new computing model abstracted from the structure and functioning of living cells, as well as from the interactions of living cells in tissues or higher order biological structures. In recent years, many variant of membrane computing models have developed rapidly, and also have turned out that membrane computing has significant potential to be applied to various computationally hard problems in feasible time, such as PSPACE-complete problem [2], 0-1 knapsack problem [3], maximum clique problem [4], Hamilton path problem [5], tripartite matching problem [6].

Inspired from framework and function of living cells, membrane algorithm was firstly proposed by Nishida [7, 8], and used solving the traveling salesman problem. In those membrane algorithms, the nest membrane structure was used together with ideas from genetic algorithms. After Nishida, Huang et al. [9] proposed a membrane algorithm combining the nested membrane structure and conventional genetic algorithm to solve multi-objective numerical optimization problems. Leporati et al. [10] developed a polynomial time membrane algorithm that computed approximate solutions to the instances of min storage. The quantum-inspired evolutionary algorithm based on P systems was also developed to solve the knapsack problem [11], the satisfiability problem [12] and the radar emitter signals problem [13]. Zhao et al. [14] used a bio-inspired algorithm based on membrane computing to optimize gasoline blending scheduling. In 2012, Yang et al. [15] developed a P systems based hybrid optimization algorithm to estimate the parameters of FCCU reactor regenerator model. Xiao et al. [16] used the membrane evolutionary algorithm to select the DNA sequences in DNA Computation.

In the paper, a bio-inspired based on membrane computing with neighborhood search strategy is proposed to solve the engineering design problems (EDP). The rest of this paper is organized as follows. In section 2, a hybrid membrane evolutionary algorithm will be proposed. The simulation results and analyses will be given in section 3. Section 4 is the conclusion and further remark.

2. The Bio-inspired Algorithm based on Membrane Computing

2.1 Cell-like P systems

P systems can be classified into the cell-like P systems, the tissue-like P systems and neural-like P systems [17]. In a cell-like P systems, the membrane structure is a hierarchical arrangement of membranes

[*] Corresponding author (jhxiao@nankai.edu.cn)

embedded in the skin membrane. A membrane without any other membranes inside is said to be elementary membrane. Each membrane has a region containing a multiset of objects and a set of evolutionary rules. The multisets of objects evolve in each region and move from a region to a neighboring one by applying the rules in a nondeterministic and maximally parallel way. The membrane structure of a cell-like P systems is shown in Fig. 1.

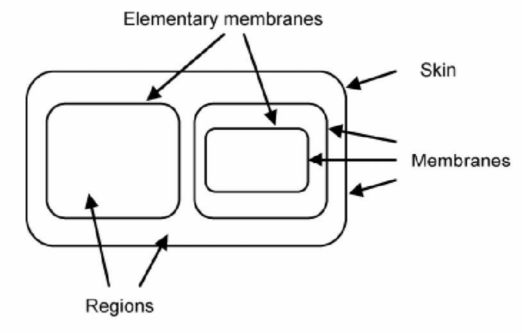


Fig. 1 The membrane structure

The membrane structure of a cell-like P system can be formally defined as follows [16].

$$\Pi = (O, T, u, s_1, \cdots, s_n, R_1, \cdots, R_n, i_0)$$

where:

(i) O is the alphabet of objects;

(ii) T is the output alphabet, $T \subseteq O$;

(iii) μ is a membrane structure consisting of n membranes, and the membranes labeled with $1, 2, \cdots, n$; n is called the degree of the system $\prod$;

(iv) s_i $(1 \le i \le n)$ are strings which represent multisets over O associated with the region $1, 2, \cdots, n$ of μ.

(v) R_i $(1 \le i \le n)$ are the evolution rules over O^*, R_i is associated with region i of μ, and it is of the following forms.

(a) $[_i s_1 \rightarrow s_2]_i$, where $i \in \{1, 2, \cdots, n\}$, and $s_1, s_2 \in O^*$.

(Evolution rules: a rule of this type works on a string objects by the local search algorithm or various evolutionary operator, and the new strings object are created in region i.)

(b) $s_1[_i\ ]_i \rightarrow [_i s_2]_i$, where $i \in \{1, 2, \cdots, n\}$, and $s_1, s_2 \in O^*$.

(Send-in communication rules; a string object is send in the region i.)

(c) $[_i s_1]_i \rightarrow [_i\ ]_i s_2$, where $i \in \{1, 2, \cdots, n\}$, and $s_1, s_2 \in O^*$.

(Send-out communication rules; a string object is sent out of the region i.)

(vi) i_0 is the output membrane.

P systems, regarded as a model of computation, also is called as membrane algorithm, which is composed of a series of computing steps between configurations. Each computation starts from the initial configuration, and halts when there are no more rules applicable in any region. In the computing process, the system will go from one configuration to a new one by applying the rules associated to regions in a non-deterministic and maximally parallel manner. The result of the computation is obtained in region i_0. The basic pseudocode of the membrane evolutionary algorithm is shown in Fig. 2. For more details about the cell-like P systems, please refer to [17].

```
Begin
  Initialize the membrane structure and parameters;
  gen=0;
    While (Not termination condition) do
        Execute the evolution rules in all elementary membranes;
        Calculate the fitness by the fitness function;
        Execute the communication rules;
        Record the current best solution;
        gen=gen+1;
    end while
End begin
```

Fig. 2 The pseudocode of the membrane algorithm

2.2 The basic idea of PSO

Particle swarm optimization (PSO) is an effective optimization method that belongs to the category of swarm intelligence methods, originally developed by Kennedy and Eberhart [18]. Since the simple concept, easy implementation and effectiveness, PSO has become popular in evolutionary optimization community. In recently years, various heuristic algorithms have been developed to solve hard benchmark and real-world optimization problems, such as the traveling salesman problem (TSP) [19], the production-planning problem [20] and the economic load dispatch [21].

In PSO, each particle in the swarm is attracted by its previous best particle ($pbest$) and the global best particle ($gbest$), and is moved toward the optimal point by adding a velocity with its position. For a search problem in N-dimensional space, the velocity $v_{ij}(t)$ and position $x_{ij}(t)$ of the j-th dimension of the i-th particle are updated as follows in t-th generation.

$$\begin{aligned} v_{ij}(t+1) = v_{ij}(t) + c_1 \times rand(\,) \times (pbest_{ij}(t) - x_{ij}(t)) \\ + c_2 \times Rand(\,) \times (gbest_j(t) - x_{ij}(t)) \end{aligned} \quad (1)$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \quad (2)$$

where $pbest_{ij}(t)$ represents the best location in the search space ever visited by particle i and $gbest_j(t)$ is the best location discovered so far; c_1 and c_2 are the acceleration constants; $rand(\,)$ and $Rand(\,)$ are the uniform random value in the range [0, 1]. The flow chart of the basic particle swarm optimization is shown in Fig. 3.

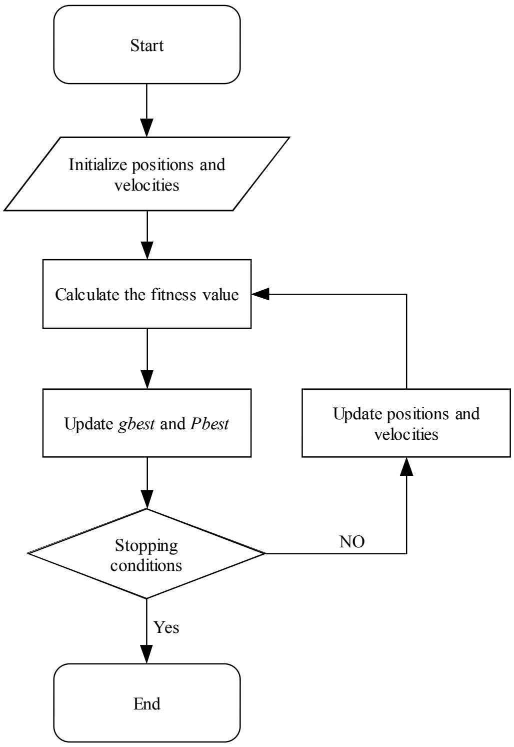

Start
Initialize positions and velocities
Calculate the fitness value
Update *gbest* and *Pbest*
Update positions and velocities
Stopping conditions
NO
Yes
End

Fig.3 The flow chart of the particle swarm optimization

In PSO, proper selection of c_1 and c_2 is crucial to improve the search ability during the optimization process. However, it is difficult to obtain the optimal values because the different optimization problems have the different values. In [22], Krohling and Coelho implemented the Gaussian probability distribution to generate the accelerating coefficients of PSO, and got good performance. In the paper, the velocity equation of PSO is modified as follows.

$$\begin{aligned} v_{ij}(t+1) &= |\,randn(\,)\,| \times (pbest_{ij}(t) - x_{ij}(t)) \\ &+ |\,Randn(\,)\,| \times (gbest_j(t) - x_{ij}(t)) \end{aligned} \quad (3)$$

where $|\,randn(\,)\,|$ and $|\,Randn(\,)\,|$ are positive random numbers generated using $abs(\,N(0,1)\,)$.

3.3 The bio-inspired algorithm based on computing for EDP

In this subsection, the bio-inspired algorithm based on membrane computing will be proposed by using the concepts and mechanism of both P systems and PSO with the neighborhood search. In this algorithm, a P systems-like framework is introduced to arrange objects and evolution rules, and two neighborhood searches are employed to enhance the ability of exploration and exploitation. The procedure of the hybrid membrane evolutionary algorithm based on PSO is described as follows.

Step 1: Initialize membrane structure and $X(t)$, $V(t)$. Specify one level membrane structure $[_0\,[_1\,]_1, [_2\,]_2, \cdots, [_n\,]_n\,]_0$ which composed of a skin membrane denoted by 0 and n regions inside the skin membrane; randomly generate m individuals in each elementary membrane. Multisets are initialized as follows:

$$\begin{aligned} s_0 &= \lambda \\ s_1 &= b_{1,1} b_{1,2} b_{1,3} \cdots b_{1,m} \\ s_2 &= b_{2,1} b_{2,2} b_{2,3} \cdots b_{2,m} \\ s_3 &= b_{3,1} b_{3,2} b_{3,3} \cdots b_{3,m} \\ &\cdots \cdots \\ s_n &= b_{n,1} b_{n,2} b_{n,3} \cdots b_{n,m} \end{aligned}$$

where m is the population size of each elementary membrane, and $b_{i,j}$, $1 \le i \le n$, $1 \le j \le m$ is an individual.

Step 2: Evolution rules in each of the region 1 to n are implemented. The particle swarm optimization (PSO) based on Gaussian distribution will be performed in each elementary membrane simultaneously.

Step 3: Implement the send-out communication rules, the strings are sent to skin membrane from each elementary membrane;

Step 4: To improve the disadvantage of the premature convergence problem, the local and global neighborhood searches are implemented in the skin membrane to improve the ability of exploration and exploitation. The equations of local neighborhood search are defined as follows [23].

$$LX_i = r_1 \cdot X_i + r_2 \cdot pbest_i + r_3(X_c - X_d) \qquad (4)$$

$$LV_i = V_i \qquad (5)$$

where X_i is the position vector of the i-th particle, $pbest_i$ is the previous best particle of P_i; X_c and X_d are the position vectors of two random particles in the k-neighborhood radius of P_i, $c, d \in [i-k, i+k] \wedge c \neq d \neq i$; r_1, r_2 and r_3 are three uniform random numbers within (0,1), and $r_1 + r_2 + r_3 = 1$.

The equations of global search are shown as follows [23].

$$GX_i = r_4 \cdot X_i + r_5 \cdot gbest + r_6 \cdot (X_e - X_f) \qquad (6)$$

$$GV_i = V_i \qquad (7)$$

where *gbest* is the global best particle, X_e and X_f are the position vectors of two random particles chosen from the entire swarm, $e, f \in [1, N], e \neq f \neq i$; r_4, r_5 and r_6 are three uniform random numbers in [0, 1], and $r_4 + r_5 + r_6 = 1$.

Step 5: Calculate the fitness of each string object by fitness function, and save the current best strings;

Step 6: Implement the send-in communication rules between the skin membrane and each elementary membrane simultaneously. The detail description is as follows.

(i) First, the best strings and $m-1$ strings with the worst fitness are sent to the elementary membrane 1;

(ii) Subsequently, in the remaining strings, the current best strings and $m-1$ strings with the worst fitness are sent to the elementary membrane 2;

(iii) The above process is executed constantly until the strings from the skin membrane back to each region;

Step 7: If the stopping condition is met, then output the results; otherwise, return to step 2.

4. Experimental Results

In this section, we will carry out numerical simulation based on several well-known engineering design problems to test the effectiveness and efficiency of the proposed algorithms.

4.1 Parameters Setting

In our experiment, the bio-inspired algorithm based on membrane computing for the engineering design problems is executed with Matlab 7.0. The parameters of the algorithm used are shown in Table 1. For each engineering design problem, 30 independent runs are carried out.

Table 1: Parameters used in the proposed algorithm

Parameters	Value	Meaning
$iter_{\max}$	500	The maximum number of iterations
N_s	8	The swarm size of each elementary membrane
N_m	20	The number of the elementary membrane
N_R	30	The run times independently for each test function
ε	0.0001	Tolerated equality constraint violation
k	5	The neighborhood radius

4.2 Welded beam design problem (Example 1)

The welded beam design problem is taken from [24], which is designed for minimum cost subject to constraints on shear stress (τ), bending stress in the beam (σ), buckling load on the bar (P_c), end deflection of the beam (δ), and side constraints. There are four design variables ($h(x_1)$, $l(x_2)$, $t(x_3)$ and $b(x_4)$) as shown in Fig. 4.

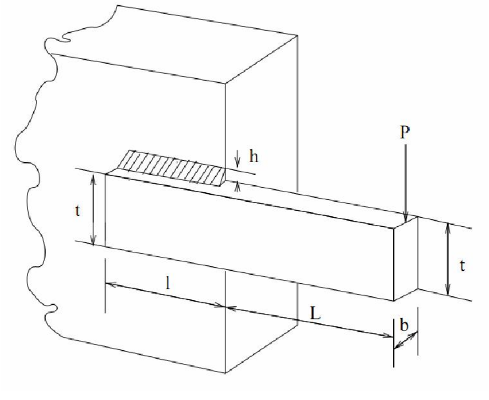


Fig.4 The welded beam design problem (Example 1)

The problem can be mathematically formulated as follows [25].

Minimize:

$$f(X)=1.10471x_1^2x_2+0.04811x_3x_4(14+x_2) \quad (8)$$

Subject to:

$$g_1(X)=\tau(X)-13600\le 0 \quad (9)$$

$$g_2(X)=\sigma(X)-30000\le 0 \quad (10)$$

$$g_3(X)=x_1-x_4\le 0 \quad (11)$$

$$g_4(X)=0.10471x^2{}_1+0.04811x_3x_4(14+x_2)-5\le 0 \quad (12)$$

$$g_5(X)=0.125-x_1\le 0 \quad (13)$$

$$g_6(X)=\delta(X)-0.25\le 0 \quad (14)$$

$$g_7(X)=6000-P_c(X)\le 0 \quad (15)$$

where:

$$\tau(X)=\sqrt{(\tau')^2+2\tau'\tau''\frac{x_2}{2R}+(\tau'')^2} \quad (16)$$

$$\tau'=\frac{6000}{\sqrt{2}x_1x_2} \quad (17)$$

$$\tau''=\frac{MR}{J} \quad (18)$$

$$M=6000(14+\frac{x_2}{2}) \quad (19)$$

$$R=\sqrt{\frac{x_2^2}{4}+(\frac{x_1+x_3}{2})^2} \quad (20)$$

$$J=2\left\{\sqrt{2}x_1x_2\left[\frac{x_2^2}{12}+(\frac{x_1+x_3}{2})^2\right]\right\} \quad (21)$$

$$\sigma(X)=\frac{504000}{x_4x_3^2} \quad (22)$$

$$\delta(X)=\frac{2.1952}{x_3^3x_4} \quad (23)$$

$$P_c(X)=64746.022(1-0.0282346x_3)x_3x_4^3 \quad (24)$$

Various methods were proposed to solve this problem, such as co-evolutionary particle swarm optimization (CPSO) [25], GA-based co-evolution model [26], and the culture algorithm (CA) [27]. In the paper, the variable regions are defined as follows: $0.1\le x_1, x_4\le 2$, $0.1\le x_2, x_3\le 10$. The comparison of the experiment results for the welded beam problem shown in Table 2, where the highlighted boldface is the better results. Table 3 also presents the statistical results.

Table 2: Comparison of the best solution for the welded beam design problem

Design Variables	Our Algorithm	Deb [28] (1991)	Coello [29] (2000)	Coello [26] (2002)	CA [27] (2012)
x_1	0.205675	0.248900	0.208800	0.205986	0.202369
x_3	3.470993	6.173000	3.420500	3.471328	3.544214
x_3	9.040587	8.178900	8.997500	9.020224	9.048210
x_4	0.205728	0.253300	0.210000	0.206480	0.205723
$g_1(X)$	-2.640579	-5758.603777	-0.337812	-0.074092	-12.839796
$g_2(X)$	-26.062354	-255.576901	-353.902604	-0.266227	-1.247467
$g_3(X)$	-0.000053	-0.004400	-0.001200	-0.000495	-0.001498
$g_4(X)$	-3.432268	-2.982866	-3.411865	-3.430043	-3.429347
$g_5(X)$	-0.080675	-0.123900	-0.083800	-0.080986	-0.079381
$g_6(X)$	-0.235559	-0.234160	-0.235649	-0.235514	-0.235536
$g_7(X)$	-1.588096	-4465.270928	-363.232384	-58.666440	-11.681355
$f(X)$	**1.725507**	2.433116	1.748309	1.728226	1.728024

Table 3: Statistical results of different methods for the welded beam design problem

Optimization Method	*Best*	*Mean*	*Worst*	*Std.*
Our Algorithm	**1.725507**	**1.726594**	**1.728121**	**7.246e-04**
CA (2012)	1.728024	N/A	N/A	N/A
CPSO (2007)	1.728024	1.748831	1.782143	0.012926
Coello (2002)	1.728226	1.792654	1.993408	0.074713
Coello (2000)	1.748309	1.771973	1.78535	0.011220
Deb (1991)	2.433116	N/A	N/A	N/A

From Table 2, it can be seen that the best feasible solution found by our algorithm is better than other evolutionary algorithms. From Table 3, it is also clear that the proposed algorithm performs better than other methods according to all statistical values for the welded beam design problem. Furthermore, the proposed algorithm provides smaller standard deviation in 30 independent runs.

4.3 Pressure vessel design problem (Example 2)

The objective of the pressure vessel design problem is to minimize the total cost, including the cost of the material, forming and welding. Four variables (thickness of the shell $T_s(x_1)$, thickness of the head $T_h(x_2)$, inner radius $R(x_3)$ and length of cylindrical section of the vessel $L(x_4)$, not including the head) are shown in Fig. 5.

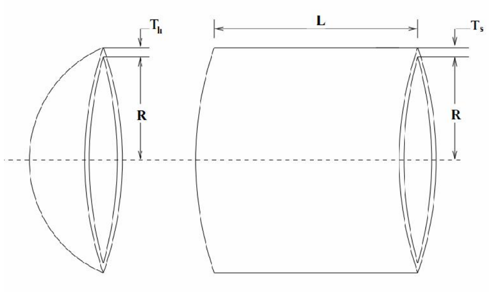


Fig.5 The pressure vessel design problem (Example 2)

The problem can be mathematically formulated as follows [25]:

Minimize:

$$f(X) = 0.6224x_1x_3x_4 + 1.7781x_2x_3^2 + 3.1661x_1^2x_4 + 19.84x_1^2x_3 \quad (25)$$

Subject to:

$$g_1(X) = -x_1 + 0.0193x_3 \leq 0 \quad (26)$$

$$g_2(X) = -x_2 + 0.00954x_3 \leq 0 \quad (27)$$

$$g_3(X) = -\pi x_3^2 x_4 - \frac{4}{3}\pi x_3^3 + 1296000 \leq 0 \quad (28)$$

$$g_4(X) = x_4 - 240 \leq 0 \quad (29)$$

In recent years, various approaches are used to optimize the pressure vessel design problem, such as feasibility-based co-evolutionary PSO [29], genetic adaptive search [30], and quantum-behaved PSO [31]. In the paper, the variable regions are defined as follows: $1 \leq x_1, x_4 \leq 99$, $10 \leq x_2, x_3 \leq 200$. Table 4 and Table 5 present the comparison and statistical results of the experiment, respectively.

Table 4: Comparison of the best solution for the welded beam design problem

Design Variables	Our Algorithm	Deb [30] (1997)	Coello [29] (2000)	Coello [26] (2002)	CPSO [25] (2007)	QPSO [31] (2010)
x_1	0.827599	0.937500	0.812500	0.812500	0.8125	0.8125
x_3	0.413794	0.500000	0.437500	0.437500	0.437500	0.4375
x_3	42.703137	48.329000	40.323900	42.097398	42.091266	42.0984
x_4	169.965254	112.679000	200.00000	176.650405	176.746500	176.0984
$g_1(X)$	-0.003429	-0.004750	-0.034324	-0.000020	-0.000139	-8.7999e-07
$g_2(X)$	-0.006406	-0.038941	-0.052847	-0.035891	-0.035949	-3.5881e-02
$g_3(X)$	-3897.842439	-3652.87638	-27.105845	-27886075	-116.382700	-0.2179
$g_4(X)$	-70.034746	-127.321000	-40.00000	-63.345953	-63.253500	-63.3628
$f(X)$	**6029.181059**	6410.3811	6288.7445	6059.9463	6061.0777	6059.7208

Table 5: Statistical results of different methods for the welded beam design problem

Optimization Method	*Best*	*Mean*	*Worst*	*Std.*
Our Algorithm	**6029.1811**	**6136.7807**	**6288.2630**	56.76628
QPSO (2010)	6059.7209	6839.9326	8017.2816	479.2671
CPSO (2007)	6061.0777	6147.1332	6363.8041	86.4545
Coello (2002)	6059.9463	6177.2533	6469.3220	130.9297
Coello (2000)	6288.7445	6293.8432	6308.1495	**7.4133**
Deb (1997)	6410.3811	N/A	N/A	N/A

From Table 4 and Table 5, it can be observed that the proposed algorithm is robust and find solutions which are better than other evolutionary algorithm. Furthermore, BIAMC finds a better "*Best*", "*Mean*" and "*Worst*" results, except for "*Std*" result.

4.4 Tension/compression string problem (Example 3)

In this case, we will consider the design of a tension/compression spring to be designed for minimum weight subject to constraints on minimum deflection, shear stress, surge frequency, limits on the outside diameter, and on design variables [31]. Three variables (mean coil diameter $D(x_1)$, the wire diameter $d(x_2)$, the number of active coils $N(x_3)$) are shown in Fig. 6.

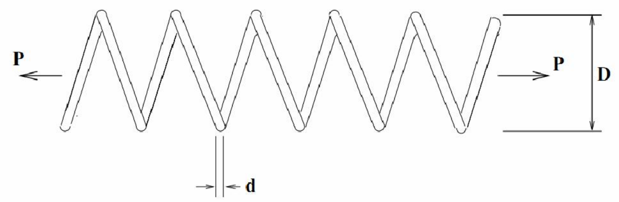


Fig.6 The tension/compression string problem (Example 3)

The problem can be mathematically formulated as follows.

Minimize:

$$f(X) = (x_3 + 2)x_1x_2^2 \tag{30}$$

Subject to:

$$g_1(X) = 1 - \frac{x_1^3 x_3}{71785 x_2^4} \le 0 \tag{31}$$

$$g_2(X) = \frac{4x_1^2 - x_1x_2}{12566(x_1x_2^3 - x_2^4)} + \frac{1}{5108x_2^2} - 1 \le 0 \tag{32}$$

$$g_3(X) = 1 - \frac{140.45x_2}{x_1^2 x_3} \le 0 \tag{33}$$

$$g_4(X) = \frac{x_1 + x_2}{1.5} - 1 \le 0 \tag{34}$$

where $0.05 \le x_1 \le 2$, $0.25 \le x_2 \le 1.3$, and $2 \le x_3 \le 15$. Table 6 and Table 7 present the comparison and statistical results of the experiment, respectively.

Table 6: Comparison of the best solution for the tension/compression string problem

Design Variables	Our Algorithm	Coello [29] (2000)	Coello [26] (2002)	QPSO [31] (2010)	CA [27] (2012)
x_1	0.051988	0.051480	0.051989	0.051515	0.015728
x_3	0.363936	0.351661	0.363965	0.352529	0.357644
x_3	10.892225	11.632201	10.890522	11.538862	11.244543
$g_1(X)$	-0.000021	-0.002080	-0.000013	-4.8341e-5	-0.00845
$g_2(X)$	-0.000024	-0.000110	-0.000021	-3.5774e-5	-1.2600e-5
$g_3(X)$	-4.061229	-4.026318	-4.061338	-4.0455	-4.051300
$g_4(X)$	-0.722717	-4.02318	-0.722698	-0.73064	-0.727090
$f(X)$	0.012681	0.0127048	0.0126810	**0.012665**	0.0126747

Table 7: Statistical results of different methods for the tension/compression string problem

Optimization Method	*Best*	*Mean*	*Worst*	*Std.*
Our Algorithm	0.012681	**0.012687**	**0.012694**	**2.9393e-06**
CA (2012)	0.0126747	N/A	N/A	N/A
QPSO (2010)	**0.012669**	0.013854	0.018127	0.001341
CPSO (2007)	0.0126747	0.012730	0.012924	5.1985e-05
Coello (2002)	0.0126810	0.0127420	0.012973	5.900e-05
Coello (2000)	0.0127048	0.012769	0.012822	3.939e-05

From Table 6 and Table 7, it can be seen that the best feasible solution found by QPSO is better than the best solutions found by our algorithm. Nevertheless, our proposed algorithm performs better than other evolutionary algorithms for “Mean” and “Worst” results. Moreover, the standard deviation of our algorithm is very small in 30 independent runs.

5. Conclusion

In this paper, a bio-inspired algorithm based on membrane computing was presented to solve the constrained engineering design optimization problems. The method proposed combined the neighborhood search strategy in to particle swarm optimization to improve the exploration and exploitation the ability of the membrane algorithm. By comparing with other evolutionary algorithms, our algorithm can get the good solutions for some well-known engineering design problems.

However, we have some further work to do. The dynamic membrane evolutionary algorithm based on the divide rule of P systems will be considered. We will apply the improved algorithm to solve other optimization hard problems and the real engineering design problems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 60903105, 70673046, 61100055 and 61202204), and the Fundamental Research Funds for the Central Universities (Grant No. NKZXB1110).

References

[1] G. H. Paun. “Computing with Membranes”. Technical Report. Finland: Turku Center for Computer Science, 1998

[2] A. Alhazov, C. Martin-Vide, L. Q. Pan. “Solving a PSPACE-complete problem by recognizing P systems with restricted active membranes”. Fundamenta Informaticae, 2003, 58: 67-77

[3] L. Q. Pan, C. Martin-Vide. “Solving multidimensional 0-1 knapsack problem by P systems with input and active

membranes". Journal of Parallel and Distributed Computing, 2005, 65: 1578-1584

[4] M. Garcia-Arnau, D. Manrique, A. Rodriguez-Paton, et al. "A P system and a constructive membrane-inspired DNA algorithm for solving the maximum clique problem". BioSyetems, 2007, 2(5): 1-11

[5] L. Q. Pan, A. Alhazov. "Solving HPP and SAT by P systems with active membrane and separation rules". Acta Inform, 2006, 43: 131-145

[6] Y. Y. Niu, L. Q. Pan, M. J. Perez-Jimenez, et al. "A Tissue P Systems Based Uniform Solution to Tripartite Matching Problem". Fundamenta Informaticae, 2011, 109: 1-10

[7] T. Y. Nishida. "An Application of P-System: A New Algorithm for NP-Complete Optimization Problems". The 8th World Multi-Conference on Systems, Cybernetics and Informatics, 2004, pp. 109-112

[8] T. Y. Nishida. "An approximate algorithm for NP-complete optimization problems exploiting P systems". The Brainstorming Workshop on Uncertainty in Membrane Computing, 2004, pp185-192

[9] L. Huang, X. X. He, N. Wang, et al. "P systems based multi-objective optimization algorithm". Progress in Nature Science, 2007, 17, 458-465

[10] A. Leporati, D. Pagani. "A membrane algorithm for the min storing problem". Proceedings of Membrane Computing, International Workshop, WMC7, Leiden, The Netherlands, 2006, pp397-416

[11] G. X. Zhang, M. Gheorghe, C. Z. Wu. "A quantum-inspired evolutionary algorithm based on P systems for knapsack problem". Fund Inform, 2008, 87: 93-116

[12] G. X. Zhang, C. X. Liu, M. Gheorghe, et al. "Solving satisfiability problems with membrane algorithm". Fourth International Conference on Bio-Inspired Computing, 2009, 1-8

[13] G. X. Zhang, C. X. Liu, H. N. Rong. "Analyzing radar emitter signals with membrane algorithms". Math Comput Model, 2010, 52: 1997-2010

[14] J. H. Zhao, N. Wang. "A bio-inspired algorithm based on membrane computing and its application to gasoline blending scheduling". Computers and Chemical Engineering, Computers and Chemical Engineering, 2011, 35: 272-283

[15] S. P. Yang, N. Wang. "A P systems based hybrid optimization algorithm for parameter estimation of FCCU reactor regenerator model". Journal of Chemical Engineering, 2012, 508-518

[16] G. Paun. "Tracing some open problems in membrane computing". Romanian Journal of Information Science and Technology, 2007, 10: 303-314

[17] J. H. Xiao, X. Y. Zhang, J. Xu. "A membrane evolutionary algorithm for DNA sequence design in DNA computing". Chinese Science Bulletin, 2012, 57(6): 698-706

[18] J. Kennedy, R. C. Eberhart. "Particle Swarm Optimization". The IEEE International Conference on Neural Networks, Perth, Australia, 1995, 1942-1948

[19] K P Wang, L Huang, C G Zhou, et al. "Particle swarm optimization for traveling salesman problem". The 2nd ICMLC, 2003, pp. 1583-1585

[20] Y. Y. Chen, J. T. Lin. "A modified particle swarm optimization for production planning problems in the TFT Array process". Expert Systems with Applications, 36 (2009), 12264-12271

[21] L. D. S. Coelho, V C Mariani. A novel chaotic particle swarm optimization approach using Henon map and implicit filtering local search for economic load dispatch". Chaos, Solitons & Fractals, 2009, 39: 510-518

[22] R. A. Krohling, L. S. Coelho. "Coevolutionary particle swarm optimization using Gaussian distribution for solving constrained optimization problems". IEEE Transactions on Systems Man and Cybernetics, Part B: Cybernetics, 2006, 36(6): 1407-1416

[23] H. Wang, H. Sun, C. H. Li, et al. "Diversity enhanced particle swarm optimization with neighborhood search". Information Sciences, 2013, 223: 119-135

[24] S. S. Rao. "Engineering Optimization". Wiley, New York, 1996.

[25] Q. He, L. Wang. "An effective co-evolutionary particle swarm optimization for constrained engineering design problems". Engineering Applications of Artificial Intelligence, 2007, 20: 89-99

[26] C. A. C. Coello. "Use of a self-adaptive penalty approach for engineering optimization problem". Computer in Industry, 2000, 4(2): 113-127

[27] X. S. Yan, W. Li, W. Chen, et al. "Cultural algorithm for engineering design problem". International Journal of Computer Science Issues, 2012, 9(6): 53-61

[28] K. Deb. "Optimal design of a welded beam via genetic algorithms". AIAA Journal, 1991, 29(11): 2013-2015

[29] C. A. C. Coello. "Theoretical and numerical constrained-handling techniques used with evolutionary algorithms: a survey of the state of the art". Computer Methods in Applied Mechanics and Engineering, 2002, 191(11-12): 1245-1287

[30] K. Deb. "GeneAS: a robust optimal design technique for mechanical component design". Evolutionary Algorithm in Engineering Applications. Berlin: Springer-Verlag, 1997

[31] L. S. Coelho. "Gaussian quantum-behaved particle swarm optimization approaches for constrained engineering design problems". Expert Systems with Application, 2010, 27, pp. 1676-1683

Jian-hua Xiao is received the Ph. D. degree in System Engineering from Huazhong University of Science and Technology, Wuhan, China, in 2008. He is currently a lecture at Nankai University, Tianjin, China. His research interests include combinatorial optimization, Bio-inspired computation and logistics optimization etc.

Yu-fang Huang received the Ph.D. degree from Huazhong University of Science and Technology in 2010. She is now on the Post Doctor research of the Department of Control Science and Engineering at Huazhong University of Science and Technology from 2011. Currently, she is a teacher of the College of Mathematics at Southwest Jiaotong University. Her research interests include combinatorial optimization and molecular computation.

Zhen Cheng received the PH.D degree from Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2010.She is currently a lecturer at Zhejiang University of Technology, Hangzhou, China. Her research interests are combinatorial optimization, Bio-inspired computation.

Statistical Approach for Predicting Factors of Mood Method for Object Oriented

Firas Jassim[1], Fawzi Altaani[2]

**[1] Management Information Systems Department,
Irbid National University, Irbid, Jordan**

**[2] Management Information Systems Department,
Irbid National University, Irbid, Jordan**

Abstract

Object oriented design is becoming more popular in software development and object oriented design metrics which is an essential part of software environment. The main goal in this paper is to predict factors of MOOD method for OO using a statistical approach. Therefore, linear regression model is used to find the relationship between factors of MOOD method and their influences on OO software measurements. Fortunately, through this process a prediction could be made for the line of code (*LOC)*, number of classes (NOC), number of methods (NOM), and number of attributes (NOA). These measurements permit designers to access the software early in process, making changes that will reduce complexity and improve the continuing capability of the design.

Keywords: *Software engineering, Software metric, Object Oriented, MOOD.*

1. Introduction

Software metrics are most often proposed as the measurement tools of choice in empirical studies in software engineering, and the field of software metrics is the most often discussed from the perspective referred to as measurement theory. Software Metrics can be defined by measuring quality or characteristic of a software objects in any complex software project. Object oriented approach is capable of classifying the problem in terms of objects and provide many benefits like reliability, reusability, decomposition of problem into easily understood object and aiding of future modifications [2]. Nowadays, a quality engineer can choose from a large number of object–oriented metrics. The question posed is not the lack of metrics but the selection of those metrics which meet the specific needs of each software project. A quality engineer has to face the problem of selecting the appropriate set of metrics for his software measurements. A number of object–oriented metrics exploit the knowledge gained from metrics used in structured programming and adjust such measurements so as to satisfy the needs of object–oriented programming. On the other hand, other object–oriented metrics have been developed specifically for object–oriented programming and it would be pointless to apply them to structured programming [6]. Recently, many companies have started to introduce object-oriented (OO) technologies into their software development process. Many researchers have proposed several metrics suitable for measuring the size and the complexity of OO software. Some of them are in terms of Function Point (FP), others are in the terms of Lines of Code (LOC). Traditional metrics such as (FP) are unsatisfactory for predicting software size. On the other hand, LOC are quit satisfactory because it can be used to measure the software size [1, 7].

2. MOOD Method

The MOOD (Metrics for Object-Oriented Design) method is a collection of metrics which is used to evaluate the main abstraction of OO [4], such as *inheritance*, *encapsulation*, *coupling*, and *information hiding* or *polymorphism* and finally how to reuse that, together, for the increase in software quality. MOOD includes the following metrics [3, 5, 6, 13]:

- Method Hiding Factor (MHF)
- Attribute Hiding Factor (AHF)
- Method Inheritance Factor (MIF)
- Attribute Inheritance Factor (AIF)
- Coupling Factor (CF)
- Polymorphism Factor (PF)

These metrics are intended to presents the presence or the absence of a certain property or attribute. Mathematically speaking, it can be viewed as probabilities ranging from 0 (total absence) to 1 (total presence).

Objects are an encapsulation of information that is relative to some entity. The ***class*** can be viewed as an abstract data type (ADT), which includes two types of features: methods and attributes, where the number of defined methods in a class C_i is given as:

$$M_d(C_i) = M_v(C_i) + M_h(C_i) \qquad (1)$$

M_d (represents defined methods), M_v (represents visible methods), and M_h (represents hidden methods).

Then we define the Method Hiding Factor (*MHF*), as follows:

$$MHF = \frac{\sum_{i=1}^{TC} M_h(C_i)}{\sum_{i=1}^{TC} M_d(C_i)} \quad , \qquad (2)$$

$$Tc = Total\ Classes$$

Conversely, the number of attributes defined in class C_i (using the same manner above) is given by:

$$A_d(C_i) = A_v(C_i) + A_h(C_i) \qquad (3)$$

$$AHF = \frac{\sum_{i=1}^{TC} A_h(C_i)}{\sum_{i=1}^{TC} A_d(C_i)} \qquad (4)$$

And all other factors are calculating using similar mathematical formulas. So, *MIF* and *AIF* can be defined through equations (5) and (6), as:

$$MIF = \frac{\sum_{i=1}^{TC} M_i(C_i)}{\sum_{i=1}^{TC} M_d(C_i)} \qquad (5)$$

AIF is defined as the ratio of the sum of inherited attributes in all classes of the system under consideration to the total number of available attributes (locally defined plus inherited) for all classes

$$AIF = \frac{\sum_{i=1}^{TC} A_i(C_i)}{\sum_{i=1}^{TC} A_d(C_i)} \qquad (6)$$

PF is defined as the ratio of the actual number of possible different polymorphic situation for class C_i to the maximum number of possible distinct polymorphic situations for class C_i, and can be defined as:

$$PF = \frac{\sum_{i=1}^{TC} M_o(C_i)}{\sum_{i=1}^{TC} [M_n \times DC(C_i)]} \qquad (7)$$

where M_o represents overridden methods, M_n for new methods, and DC for descendants methods.

Polymorphism arises from inheritance and [10] suggest that in some cases overriding methods could contribute to reduce complexity and therefore to make the system more understandable and easier to maintain. While, [14] have shown that this metric is a valid measure within the context of the theoretical framework.

Finally, CF is defined as the ratio of the maximum possible number of couplings in the system to the actual number of couplings not imputable to inheritance.

$$CF = \frac{\sum_{i=1}^{TC} \left[\sum_{j=1}^{TC} is\_client(C_i, C_j) \right]}{TC^2 - TC} \qquad (8)$$

where:

TC^2-TC = maximum number of coupling in a system with TC classes.

$$is\_client(C_i, C_j) = \begin{cases} 1 & iff\ C_i \Rightarrow C_j \wedge C_i \neq C_j \\ o & otherwise \end{cases} \qquad (9)$$

Coupling Factor (CF) has a very high positive correlation with all quality measures [11]. Therefore, as coupling among classes increases, the defect density and normalized rework is also expected to increase. This result shows that coupling in software systems has a strong negative impact on software quality and then should be avoided during design. In fact, many authors have noted that it is desirable that classes communicate with as few others as possible because coupling relations increase complexity, reduce encapsulation and reuse.

3. Estimation of Factors

MOOD method used widely to measure many target OO programs and many studies have compare it with other methods. Mainly, our focus will be on line of code (LOC), number of classes (NOC), number of methods (NOM), and number of attributes (NOA), so to reach this; we have collect our data from 33 systems [9, 10, 12, 14] to be suitable for normal distribution curve[1]. Results obtained using SPSS package.

Table 1: Product metrics from 33 commercial samples

	NOL	NOC	NOM	NOA
1	15837	65	1446	537
2	23570	57	1535	876
3	47106	91	2141	1178
4	23154	51	1420	538
5	20747	154	2814	1113
6	44930	92	2224	1132
7	28582	71	1978	839
8	19254	69	1815	675
9	20085	74	1876	700
10	57086	140	322	81
11	92231	201	481	124
12	167541	355	735	204
13	261260	562	1193	297
14	838128	1966	3227	611
15	2062982	5107	6735	2297
16	2129555	5035	7292	2294
17	1948354	4566	5975	2095
18	64492	222	210	81
19	70514	243	229	88
20	113919	349	325	132
21	177356	565	516	185
22	6593	324	1310	60
23	1023	25	103	220
24	1729	20	134	185
25	50000	46	2025	510
26	300000	1000	11000	10960
27	500000	1617	37191	17141
28	9189	339	1993	4022
29	7102	45	711	482
30	830	10	175	89
31	1602	26	180	247
32	3451	18	170	145
33	549	15	33	172
Total N	33	33	33	33

According to table (1), we can plot the relation between LOC (in the x-axis), and NOC, NOM, and NOA (in the y-axis), as shown in fig.1.

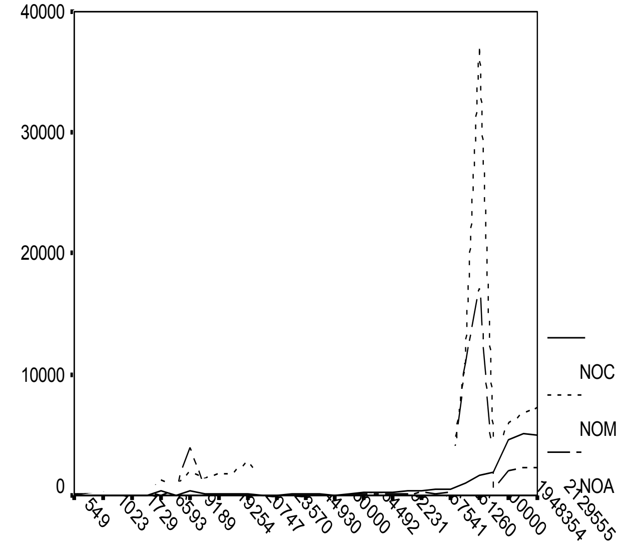


Fig. 1 The relationship between LOC and (NOC, NOM, and NOA)

Now, by implementing log transform to avoid large number scale we can plot the data again as fig. 2.

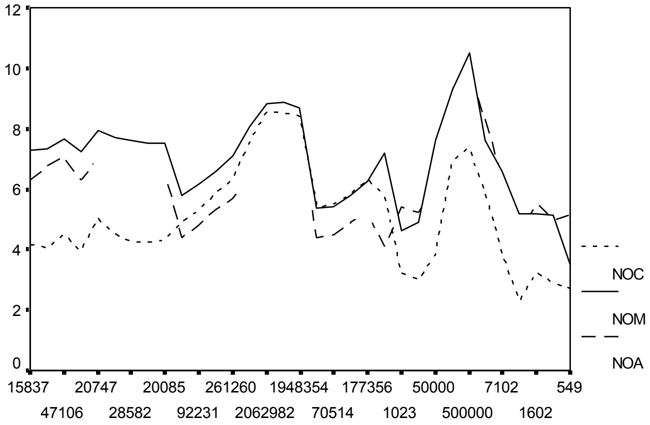


Fig. 2 The logarithmic relationship between LOC and (NOC, NOM, and NOA)

The main contribution in this article is to use statistics, especially regression; to predict number of classes needed for the software, also number of attributes and methods needed. Hence, linear regression model is used to find the relationship between factors and their influences on OO software measurements. Fortunately, through this process we can predict the suitable number of *LOC*, classes (objects), methods, and attributes we need to satisfy the software metrics using MOOD.

[1] Normal distribution needs more than thirty observation, while t distribution needs less than thirty observations, see [11].

4. Regression Analysis

Actually, we can use linear regression model to predict the *LOC*, *NOC*, *NOM*, and *NOA* needed. Statistically speaking, In order to investigate the correlations and relationships between the object-oriented metrics and software quality we conducted a correlation and a multiple linear regression analysis. The mathematical formula for the model is as follows:

$$LOC = \beta_0 + \beta_1 NOC + \beta_2 NOM + \beta_3 NOA \qquad (10)$$

$$NOC = \beta_0 + \beta_1 LOC + \beta_2 NOM + \beta_3 NOA \qquad (11)$$

$$NOM = \beta_0 + \beta_1 LOC + \beta_2 NOC + \beta_3 NOA \qquad (12)$$

$$NOA = \beta_0 + \beta_1 LOC + \beta_2 NOC + \beta_3 NOM \qquad (13)$$

Each time we have used one variable as an independent variable while the others as the dependent variables. To reach the fact that, each one of these variables responsible for the efficiency of the MOOD method. The regression analysis shows the values of the coefficients of the model (β_0,β_1,β_2, *and* β_3).

The independent variable in an experiment is the variable that is systematically manipulated by the investigator. In most experiments, the investigator is interested in determining the effect that one variable; has one or more effect on the other variables. On the other hand, the dependent variable in an experiment is the variable that the investigator measures to determine the effect of the independent variable.

First, we consider LOC as the dependent variable and the other factors as the independent variables, equation 10, table (2) shows the value of (β_0,β_1,β_2, *and* β_3), and the significances (p-value).

Table 2: Results of β_0,β_1,β_2, *and* β_3 when LOC is the dependent variable

	Regression coefficients	p-value
β_0	-9458.918	0.220
β_1	421.994	0.000
β_2	3.025	0.327
β_3	-16.009	0.008

So, if we want to use the values of the coefficients above, we may re-write the regression line as:

LOC = -9458.918 + 421.994 NOC + 3.025 NOM - 16.009 NOA

Therefore, if we want to predict the value of *LOC* we can substitute the given values of *NOC, NOM*, and *NOA* in the above formula and get an estimated (predicted) value for *LOC*. Also, from the values of p-value we can see that the values of (β_1 *and* β_3) only are less than 0.05, so we can conclude that *LOC* are mainly affected by *NOC* and *NOA*. On the other hand, *NOM* does not affect *LOC* too much.

There is some statistical measures used to measure the goodness of fit and it is an indicator of how well the model fits the data. The higher the value of R square, the more accurate the model is. These values can be seen in table (3).

Table 3: The value of R square and adjusted R square for the regression model

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.998[a]	.996	.996	37024.69

Since the value of significant (p-value) is less than 0.05. This means that *LOC* mainly affect the other factor according to table (4), which shows the **ANOVA** (**AN**alysis **O**f **VA**riance).

Table 4: ANOVA results for LOC as the dependent variable

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.12E+13	3	3.749E+12	2734.947	.000[a]
	Residual	3.98E+10	29	1370827508		
	Total	1.13E+13	32			

a. Predictors: (Constant), NOA, NOC, NOM

b. Dependent Variable: NOL

Second, we consider *NOC* as the dependent variable and the other factors as the independent variables, table (5) shows the value of (β_0,β_1,β_2, *and* β_3), and the significances (p-value).

Table 5: Results of β_0,β_1,β_2, *and* β_3 when NOC is the dependent variable

	Regression coefficients	p-value
β_0	24.439	0.179
β_1	0.002	0.000
β_2	-0.006	0.397
β_3	0.037	0.011

Also, if we want to use the values of the coefficients above, we may re-write the regression line as:

NOC = 24.439 + 0.002 LOC - 0.006 NOM + 0.037 NOA

Therefore, if we want to predict the value of *NOC* we can substitute the given values of *LOC*, *NOM*, and *NOA* in the above formula and get an estimated (predicted) value for *NOC*. Also, from the values of p-value we can see that the values of (β_1 and β_3) only are less than 0.05, so we can conclude that *NOC* are mainly affected by *LOC* and *NOA*. On the other hand, *NOM* does not affect *LOC* too much. As previously mentioned the values of R square and the ANOVA table are shown in tables 6 &7.

Table 6: The value of R square and adjusted R square for the regression model when NOC is the dependent variable

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.998[a]	.997	.996	87.55

Table 7: ANOVA results for NOC as the dependent variable

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	64118680	3	21372893.45	2788.435	.000[a]
	Residual	222280.2	29	7664.835		
	Total	64340961	32			

a. Predictors: (Constant), NOA, NOL, NOM

b. Dependent Variable: NOC

Similarly, we can do the same thing for *NOM* and *NOA*, put we mainly focused on the *LOC* and *NOC* because of their main role in MOOD method [8].

5. Conclusions

A simple and easy technique has been constructed to use statistics for predicting the values of MOOD factors, in the same manner one can use this technique to estimate other factors rather than *LOC*, *NOC*, *NOM*, and *NOA*, which can be used to evaluate software quality. Additionally, using linear regression model can be extended to non-linear model and multivariate analysis to add more complicated model to give more accurate estimation for these factors and also use another statistical estimation approaches such as maximum Likelihood Estimator (MLE) to give better estimation than regression model, and to be standards for MOOD method and to give more accurate measurements for object-oriented metrics.

Acknowledgments

Author Firas A. Jassim pays his regards to Mrs. Hind Emad Qassim for giving her moral support and help to carry out this research work.

References

[1] ‘Cost Estimating & Assessment Guide: Best Practices for Developing and Managing Capital Program Costs’, US Government Accountability Office, March 2009, GAO-09-3SP, obtainable from www.gao.gov/new.items/d093sp.pdf.

[2] A. Deepak, K. Pooja, T. Alpika, S. Shipra and S. Sanchika, “Software Quality Estimation through Object Oriented Design Metrics”, IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.4, April 2011.

[3] A. Fernando B, E. Rita and G. Miguel, “The Design of Eiffel Program: Quantitative Evaluation Using the MOOD metrics”, Proceeding of TOOLS’96 USA, Santa Barbara, California, July 1996.

[4] A. Fernando B: “Design metrics for OO software system”, ECOOP’95, Quantitative Methods Workshop, 1995.

[5] A. Shaik, C. R. K. Reddy, Bala Manda, C. Prakashini, K. Deepthi, “Metrics for Object Oriented Design Software Systems: A Survey”, Journal of Emerging Trends in Engineering and Applied Sciences (JETEAS), vol. 1, no.2, pp: 190-198, 2010.

[6] C. Neelamegam and M. Punithavalli, “A Survey - Object Oriented Quality Metrics”, Global Journal of Computer Science and Technology, Vol 9, No 4, 2009.

[7] F. Brito, E Abreu and W. Melo, “Evaluating the impact of object-oriented design on software quality”. In Proc. METRICS' 96, Berlin, Germany, March 1996. IEEE

[8] http://www.ercim.org/publication/Ercim_News/enw23/abreu.html

[9] http://www.jot.fm/issues/issue_2005_11/article1

[10] http://www.sourceforge.net/projects/metrics

[11] M. Xenos, D. Stavrinoudis, K. Zikouli and D. Christodoulakis, “Object-Oriented Metrics – A Survey”, Proceedings of the FESMA, Federation of European Software Measurement Associations, Madrid, Spain, 2000.

[12] Muktamyee Sarker, “An overview of Object Oriented Design Metrics”, Master Thesis, Department of Computer Science, Umeå University, Sweden, June 23, 2005.

[13] P. Ponmuthuramalingam and M. Yamunadevi, “An Effective Analysis of Object Oriented Metrics in Software Quality”, International Journal of Computing Technology and Information Security, Vol.1, No.2, pp.43-47, December, 2011.

[14] R. Harrison, J. Steve, “An Evaluation of the MOOD Set of Object-Oriented Software Metrics”, IEEE Transactions on Software Engineering, VOL. 24, NO. 6, JUNE 1998

[15] R. V. Hoggs and A. Elliot, “Probability and Statistical Inference”, 2nd edition, Macmillan publishing Co., 1983.

Firas Jassim received the BS degree in mathematics and computer applications from Al-Nahrain University, Baghdad, Iraq in 1997, and the MS degree in mathematics and computer applications from Al-Nahrain University, Baghdad, Iraq in 1999 and the PhD degree in computer information systems from the University of Banking and Financial Sciences, Amman, Jordan in 2012. His research interests are Image processing, image compression, image enhancement, image interpolation and simulation.

Fawzi Altaani received the BS degree in public administration from Al-Yermouk University, Irbid, Jordan 1990, and Higher Diploma in health service administration from university of Jordan, 1991 and the MS degree in health administration from Red Sea University, Soudan in 2004 and the PhD degree in managment information systems from the University of Banking and Financial Sciences, Amman, Jordan in 2010. His research interests are management information system and public administration, and Image processing.

Using a KMERP Framework to Enhance Enterprise Resource Planning (ERP) Implementation

Hamdan M. Al-Sabri[1], and Saleh M. Al-Saleem[2]

**Department of Information Systems, College of Computer and Information Sciences, King Saud University
Riyadh, Saudi Arabia**

Abstract

Enterprise Resource planning (ERP) systems mainly aims to develop information sharing between different sections within the organization, and consider as a way of continuous improvement. Implementation and use of ERP systems require a tremendous amount of knowledge and experience. There are many failure reasons in the implementation of ERP systems and facing many challenges to apply of knowledge within organizations. This paper proposed KMERP framework to manage various knowledge within the ERP software. The proposed framework is composed of five dimensions (KM life cycle, ERP life cycle, System Development life cycle, information systems project management, and organization's Knowledge). The KMERP framework allows the organization to identify relevant knowledge to ERP systems as well as management of diverse sources of knowledge. It also helps to link explicit knowledge that stored in ERP repository with tacit knowledge that has been converted to an explicit knowledge and storage Knowledge Management repository. Moreover, the Implementation of KMERP Framework on ERP is discussed.

Keywords: *Enterprise resource planning (ERP), Knowledge Management (KM), System Development Life Cycle (SDLS), Information Systems Project Management (ISPM), KMERP Framework.*

1. Introduction

Enterprise resource planning (ERP) is an integrated structure to help the organization in business process development and ERP systems used to automate processes to support and control on repository, sales and purchases, client relationship and accounts, financial and human resources, etc. [1]. ERP systems in the organization are used to process the business process for organization, reduce the threats related to inconsistencies and redundancy increases the chances of integration, and ideal to achieve the objectives of the organization. The main approach to support and continue improvement in ERP is knowledge management that contributes significantly to its success. Organization should use knowledge management to create the knowledge during the implementation of ERP systems as well as the process of capture and distribution [2]. Knowledge Management (KM) is combination of management and technology that provide integrated management strategies for the organization. Also the activity of KM is to create, store, transfer, and apply knowledge, as well as provide the employees of the organization with the necessary knowledge to accomplish the tasks and achieve the organization's goals [4]. Implementation of knowledge management in any information system project is one of the most important challenges of the organization and it requires a deep look to makes the organization as an integrated and holistic system [3]. There is a strong relationship between knowledge management (KM) and resource planning systems (ERP), where it is considered as a reciprocal relationship (KM for ERP, ERP for KM). In this paper, we present the framework to enhance the ERP Implantation, which is called (KMERP Framework). KMERP framework consists of five dimensions (ERP life cycle, KM life cycle, organization Knowledge, system development life cycle, Information system project management life cycle). This framework is used to transfer the organization's knowledge from the individual to the group and the organization. Subsequently it helps to share knowledge, and ultimately leads to the success of the ERP implementation.

The rest of this paper is organized as follows; Section 2 will present background and related work and overview on KM, ERP. In section 3, The Methodology for ERP Implementation with KM will be explained; The Methodology for KM Implementation and Motivation for developing KMERP Framework will be presented in sections 4 and 5. The proposed framework for Knowledge Management Enterprise Resource Planning (KMERP) and their dimensions will be explained in section 6, while conclusion and future research work will be discussed in section 7.

2. Background and Related Work

An Enterprise System (ES) is set of integrated applications that interact together to perform the functions of the organization [6, 7]. Enterprise resource Planning (ERP) as integrated, customized, packaged application software solutions that is employed by organizations to interact with a range of processes and functions into a holistic view of the business from a single IT architecture [8, 9]. Knowledge Management (KM) is an attempt to put the existing expertise, knowledge of the Organization staff in specific place to be easily re-used and applied [10]. The Knowledge capture, codification, transfer, sharing and use of ERP implementation knowledge by large consulting firms conform to this particular definition.

Knowledge Management System should be consists of four elements, Knowledge creation and capture, Knowledge sharing and enrichment, Information storage and reuse, and Knowledge dissemination [11].

Information System analysis and design is a method used for the organizations of various sizes to establish and maintain the systems. As the stages of system development life cycle (SDLC) is a systems planning and selecting, systems analysis, systems design, and implementation and operations [5]. Information Systems Project Management is an important process for developing information systems and it requires good analysis skills. Also the focus on project management is used to ensure that the development of information systems harmony with requirements of the organization and developed the system within specific budget and time [5]. As the stages of Information Systems Project Management is an initiation, planning, executing, and closed IS project.

In paper [1] the authors focused on two areas according to knowledge management, a tacit knowledge management, and issues related to process-based nature of organizational knowledge. Paper [2] develop a continuous improvement model for the life cycle of ERP with the integration each phase in ERP with of knowledge management life cycle. Another approach presents ideas how knowledge management help to reduce the failure rate of implementation of ERP systems in organizations [12]. [13] In another paper, they proposed framework of the different types of knowledge required to manage software ERP systems. In [14] the authors a report evaluate the effectiveness of the implementation of the Enterprise Resource Planning (ERP) from Knowledge Management perspective. Knowledge Management Enterprise Resource Planning (KMERP) was proposed. The framework supports the ERP life cycle, KM life cycle, organization's Knowledge, SDLC, and Information Systems Project Management to develop the level of knowledge within the organization from the individual to the group and convert the tacit knowledge into an explicit knowledge.

3. The Methodology for ERP Implementation with KM

Enterprise System Planning (ERP) is an integrated structure of the organization to assist in the development of business process, ERP systems used to automate processes that support all functions of the organization. ERP allows the organization's information to integrate through a centralized database and to use the integrated business processes [16, 17, 18]. Knowledge about the functions and objectives of the organization are required by the users of ERP system. Knowledge management techniques used within the life cycle of ERP to facilitate the sharing of knowledge. There are four stages in ERP life cycle i.e. analysis, design, construction, and deployment, and each stage has deliverables and outcomes [2]. In fact, selection, implementation, use, continuing change in ERP systems desperately needs to experience and different kinds of knowledge during ERP life cycle [13].

3.1 Benefits of ERP systems

There are many benefits of ERP systems and these benefits can be divided into two parts: the tangible benefits and intangible benefits [27]. Tangible benefits are Rapid response, meet customer demands, increase quality and performance, improve the use of resources, and improving the accuracy of the information for decision-maker. Intangible benefits are cooperation and integration form, meet customer satisfaction.

3.2 Failure reasons of the ERP Systems

In fact, there are many failure reasons for ERP systems in organizations; some of them are as follows [12, 25]:

1. Changes issues during ERP implementation,
2. Issues of communication and coordination between team members in ERP project,
3. Budget issues during the implementation of ERP systems,
4. Customization issue, increased customization leads to reduce the ERP features, and
5. Lack of experience, Due to increased tacit knowledge and non-sharing of knowledge etc.

To address the failure reasons that mentioned previously, we need to merge knowledge management (KM) with Enterprise Resource Planning (ERP), and building a knowledge repository for storing most knowledge that addressing the problems mentioned above. For example: ERP management projects knowledge, accurately requirements for organization. This will ultimately leads to the process of knowledge sharing, coordination between the members of the project team, ERP project management correctly, and increase experience sharing.

4. The Methodology for KM Implementation

Knowledge management has become important factor for the success of organizations. The importance of knowledge management is because its focus is on people, product and services and it is used to support the integration and development of productivity. Knowledge Management (KM) is the process of creating, capturing, and using the knowledge to improve the performance of the organization [24]. According to the literature review knowledge management can be classified into two types' i.e. explicit knowledge and tacit knowledge. Explicit knowledge can be expressed in words and numbers and can be shared, transferred and stored but tacit knowledge is personal knowledge and difficult to forming, share with others [22]. There are six basic processes of knowledge required to manage the knowledge of the organization, and the integration of knowledge sees as an important process to build the capacity of the organization [15]. The individual knowledge must convert to group knowledge by sharing knowledge within the organization; this requires knowledge management. Knowledge Management mainly aims to collect knowledge into knowledge repositories, which means storage practices and experiences and shared among teams in the Organization [19, 20, 21]. Knowledge in the organization retains in three main levels: individual, group, and organization. There are also four tracks between the two types of knowledge within the organization: socialization, externalization, combination, and internalization. Socialization is the information available to the team within the organization but it is a kind of tacit knowledge, Externalization is to convert the tacit knowledge to explicit knowledge through contact between groups. Combination means coordinating teamwork and Internalization is learning-by work and get ideas from several experiments and it is a kind of explicit knowledge [2, 23].

4.1 The relationship between KM and ERP

There are two types of knowledge that are transferred during the ERP implementation: transfer knowledge related to work procedures of the organization (organization requirements), transfer knowledge to users about how to use ERP systems [26]. The relationship between KM and ERP can see as mutuality relation, we can express this relationship (ERP for KM) and (KM for ERP) [13]. ERP for KM means implementation of ERP systems in organizations and it is a major source of explicit knowledge to the members of the organization. And KM for ERP means management and implementation of ERP requires expertise and extensive knowledge, so we use KM to identify all kinds of special knowledge that help us to manage ERP correctly. The other relationship between KM and ERP can see as complementary relation, Enterprise Resource Planning (ERP) systems used to integrate information within the organization between the different sections and put the explicit knowledge in central databases, Knowledge Management (KM) will work to manage the tacit knowledge and this will create balance and integration between KM, ERP.

4.2 The knowledge challenges in ERP implementation

In fact, there are many challenges when applying knowledge management in ERP: that are as follows:

1. Knowledge that captured and addressed during the life cycle of ERP systems may fade after the implementation of ERP systems, and this means that the knowledge will not be available to all members of the organization [2],
2. In most organization, there is no clear methodology to make sure that knowledge is captured, shared and stored for the future,
3. Increasing the size of knowledge in the organization could leads to knowledge loss if it's not captured and stored quickly in organizations,
4. Converting tacit knowledge to explicit knowledge within the organization. It means that to store experience, skills, and understand the individuals within the repositories of knowledge [12], and
5. Disregarding the importance of knowledge management in most organizations, where organizations are struggling to get knowledge [13].

To address the above challenges, the organizations needs to capture knowledge during all stages of the ERP implementation and will transport it to knowledge repository in order to share and distribute knowledge at various levels.

5. Motivation to develop KMERP Framework

Without the implementation of KMERP framework for knowledge management through all stages of the ERP life cycle, it could leads to the failure in the appropriate implementation of ERP system within the organization. We need cohesion between knowledge management elements and ERP life cycle to avoid a lack of knowledge within the organization. In this paper we proposed the

KMERP framework to address the knowledge challenges in ERP implementation and reasons for the failure of ERP. The KMERP framework consists of five dimensions as show in figure 1:

1. System Development Life Cycle (SDLC) dimension which consists of four different stages (systems selection and planning, systems analysis, Systems Design, and Implementation and operation). During these stages all the necessary information and knowledge about the system and organization's objectives will be extracted.
2. Information Systems Project Management dimension which consists of four different stages (Project Initiation, Planning the project, executing the Project, and Closing down the Project). During these stages the project schedule and proper feasibility study will be managed.
3. ERP Life Cycle dimension which consists of four different stages (Selecting, Using, Implementing, and Changing).
4. Knowledge Management Life Cycle dimension which consists of five stages (Identifying, Creating, Transferring, Storing, and Reusing).
5. Knowledge of the Organization (Business K, Technical K, ERP Knowledge, Organization K, and Project Management K).

6. Proposed Framework

The main objective of KMERP framework is to store specific knowledge in five dimensions that are mentioned previously where it will be re-used in the life cycle of ERP systems. As KMERP framework also allows to manage the ERP from the point of view of knowledge management. This framework focuses on the definition of the different kinds of knowledge as well as knowledge management through the ERP life cycle.

The KMERP Framework consists of five dimensions (SDLC, KM Life Cycle, ERP Life Cycle, IS Project Management and Knowledge for the Organization) as show in figure 2.Each element in ERP life cycle will be surrounded by elements of software development and project management. Using a Joint Application Design (JAD) method in software development life cycle will be a right way of communication and coordination between the ERP team: and the elements of project management will facilitate the process to customize ERP correctly, completely and to understand the organization requirements to extract it properly.

Henceforward, knowledge repository will capture important knowledge and will re-use in the rest of the stages associated with the other dimensions. All explicit knowledge of the organization will be captured, distributed, and will stored in ERP repository.

In terms of tacit knowledge, it will determine the required knowledge of the organization and is convert it into explicit knowledge that will be further distributed and stored in KM repository. There are reciprocal, complementarities relationships between KM repository and ERP repository, which ultimately leads to the distribution of knowledge in various levels of the organization and to make knowledge transfer from individual to the group and organization levels. The framework proposed in this paper is a starting point for analyzing and structuring the knowledge available and required within organizations.

This framework will focus on the teamwork's needs in ERP implementation, where each stage in ERP implementation will produce an outcome or deliverable. The skills and experience will be documented, converted to explicit knowledge and stored in a knowledge management repository, which will be available later for all members in the organization.

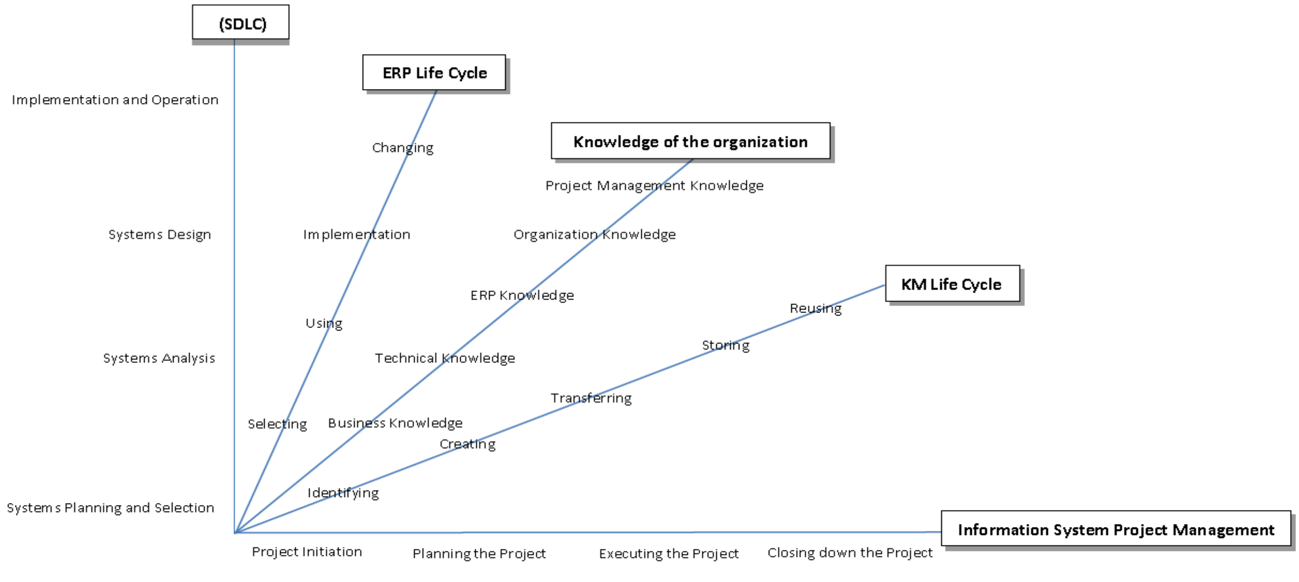


Fig. ١ Dimensions for KMERP framework

6.1 Implementation of KMERP Framework on ERP

KMERP framework that is discussed previously consists of five dimensions that will be used to surround every application in ERP as shown in figure 3. ERP applications (Inventory Management, Payroll, CRM, Purchasing, Accounting, Sales, Vendor Integration, and E-Commerce) will be analyzed and managed properly and will store all kinds of knowledge in a KM repository. This method will lead to integrate ERP, KM to succeed the ERP implementation that means the success of the organization.

6.2 Type of knowledge required to manage ERP

In fact, the lack of ERP knowledge for organization's staff is lead to the reduction of development the business process and procedures within the organization. For that, any employee must be answering these questions: What is the required knowledge to manage ERP? What is knowledge that should be collected and stored?, In KMERP framework we mentioned the knowledge of the organization dimension which identified areas of knowledge required to manage ERP, as the five areas of knowledge as described in Table 1.

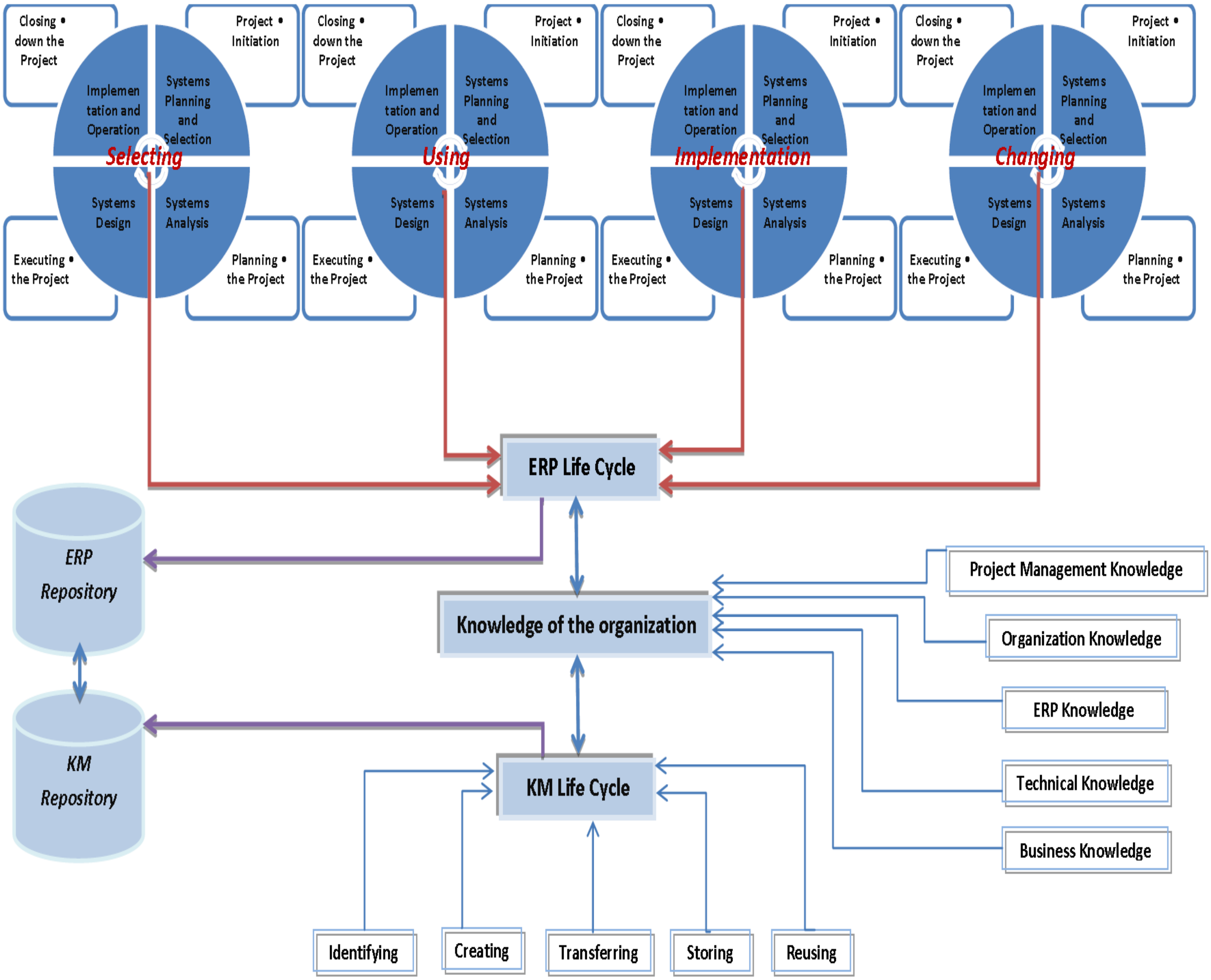


Fig. ٢. Proposed framework (KMERP)

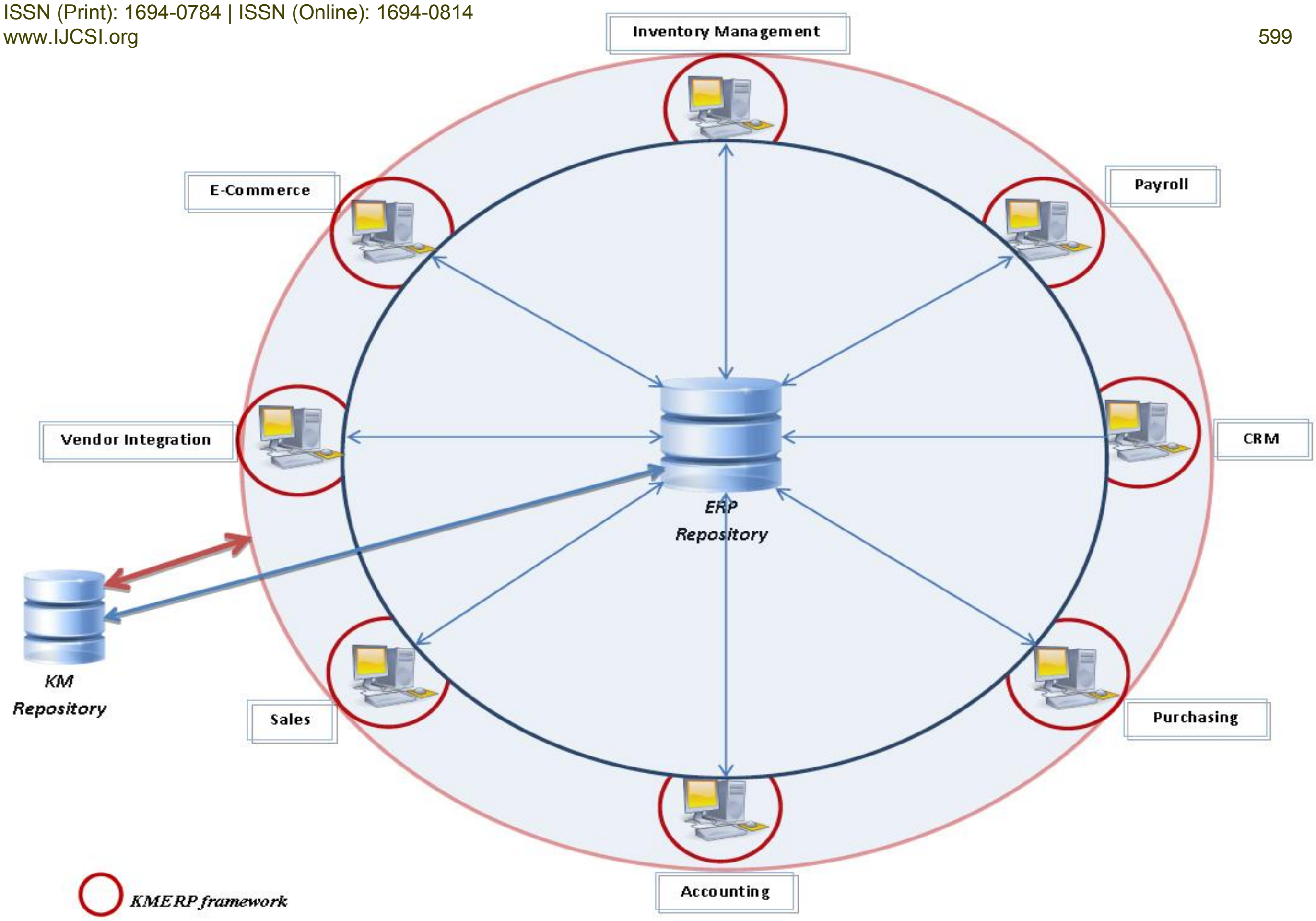


Fig. ٣ Implementation KMERP framework on ERP Systems

Table ١.Type of knowledge required to manage ERP

		Knowledge of the organization				
Life Cycle	***Stages***	***Business K***	***Technical K***	***ERP K***	***Organization K***	***Project Management K***
SDLC	***Selecting and planning***		Criteria for evaluating projects	Select right package	Organization's Objectives	Determining the appropriate project
	Analysis	Determine work procedures	Business Process Redesign, Joint Application Design	Process Modeling, Data Modeling	Organization's requirements	
	Design		Interface, Database	Interface, Database		
	Implementation and Operation		User training, documentation, maintenance	User training, documentation, maintenance	Using the system	Management system and distribution privileges
IS project Management	***Project intention***	CRM	Solutions the technical issues	Team management	Change, Conflict Management	Leadership and managerial
	Planning the project	Project Charter	Using Gantt chart, network diagram, Economic feasibility, Expected Time Durations Using PERT	Identify risks	Organization's Resource, Objectives	Time, Scheduling Management
	Executing the project		Monitor and measure the productivity of the project		Execute Basic planproject	Changes, risks Management
	Close down the project				Meeting, workshops skills	Reports, documentation skills
ERP	***Selecting***	Business Process		Select right package	Flow Quality Management	
	Using	Business Process in Organization				
	Implementation		Technical support	Implement ERP	Subdivisions Organization	System management
	Changing	Development work procedures	Business Process Redesign		Development organization work	
KM life Cycle		***Identifying***	***Creating***	***Transferring***	***Storing***	***Reusing***
		⇩	⇩	⇩	⇩	
			Stored all types of Knowledge in KM Repository			

7. Conclusions and further work

In this paper, a new framework that handles all kinds of knowledge within ERP implementation has been proposed. The KMERP framework dimensions and organization's knowledge have been identified and show how to implement in ERP systems. The integrated work flow between KM and ERP has been presented. The all kinds of knowledge required to manage ERP is identified and show the relationship between ERP repository and KM repository. As for a future work, we plan to use the quality of services (QoS) to test the performance of all dimensions with the KMERP framework. Also, we plan to apply how to separately control the unexpected knowledge and the dimensions rules from the separated view.

Acknowledgments

This work is part of research Chair of Enterprise Resource Planning& Business Process Management which has been fully funded by a grant through King Saud University (KSU) in the Kingdom of Saudi Arabia.

References

[1] Usman Musa Zakari Usman, Mohammad Nazir Ahmad, "KNOWLEDGE MANAGEMENT IN SUCCESS OF ERP SYSTEMS," International Journal of Advances in Engineering & Technology, March 2012.

[2] Thomas, Zhenyu huang, "Incorporation of Knowledge Management into ERP continuous improvement: A Research Framework", Issues in Information Systems, Nov 2 2004.

[3] Alsadhan, A., Zairi, M. and Kamala, "KM System Implementation: A Best Practice Perspective and Proposed Model", The European Centre for Total Quality Management (ECTQM), Report No. R-06-10, October 2006.

[4] Alsadhan, A.O., "The implementation of knowledge management systems: an empirical study of critical success factors and a proposed model" Unpublished Ph.D dissertation, School of Informatics Department of Computing.University of Bradford, 2007.

[5] Joseph Valacich, Joey George, Jeff Hoffer "Essentials of System Analysis and Design", 3th Edition, 2006.

[6] Hernandez .J ,"The SAP R/3 Handbook", New York, Mcgrqw-Hill, 2000.

[7] IDC, "Enterprise Resource Management Application Market Forecast and Analysis," IDC Software Research, June, 2000-2004.

[8] Watson E., Schneider H., "Using ERP in Education", Communications of the AIS, Vol, No 9, February, 1999.

[9] Klaus H., Rosemann M., Gable G.,"What is ERP?", Information Systems Frontiers, Vol2, PP. 141-162, 2000.

[10] IM I., Hars A.,"Knowledge Reuse- Insights from Software Reuse", Proceedings of the Nineteenth International Conference on Information Systems, 13-16 Dec, Helsinki, Finland, 1998.

[11] Filemon A., Uriarte JR., "Introduction to Knowledge Management", Japan. ASEAN Solidarity Fund, 2008.

[12] Anubhav Kumar, P C Gupta, " Implementation Of Knowledge Management To Minimize ERP Based System's Failure Of An Organization: A Survey", IJRFM Volume 1, (ISSN 2231-5985), Issue 3 July, 2011.

[13] Michael R., Roy Chan,"Structuring and Modeling Knowledge in the Context of Enterprise Resource Planning", Brisbane, Australia.

[14] Eric W.L. Chan, Derek H.T. Walker and Anthony Mills, "Using a KM framework to evaluate an ERP system implementation", Journal of Knowledge Management, VOL. 13 NO. 2, pp. 93- 109, 2009.

[15] A.F. Buono, &F. Poulfelt, "Challenges and Issues in Knowledge Management", Information Age Publishing, Greenwich, CT, USA, 2005.

[16] C. Argyris, &D.A. Schön, "Organizational learning II: theory, method and practice", Organization Development Series, Addison Wesley, Reading, MA, USA, 1996.

[17] P. Brossler, "Knowledge management at a software engineering company – an experience report", Workshop on Learning Software Organizations, Kaiserslautern, Germany, pp. 77–86., 1999

[18] R. Baskerville, S. Pawlowski,& E. McLean, "Enterprise resource planning and organizational Knowledge", ICIS conference, 2000.

[19] M. Beer, &N. Nohria, "Cracking the code of change", Harvard Business Review, p.p133–141, 2000.

[20] M. Alavi, &D.E. Leidner, Review. "Knowledge management and knowledge management systems", Conceptual foundations and research issues, MISQuarterly, pp.107–136, 2001.

[21] L. Argote, B. McEvily, &R. Reagans, "Managing knowledge in organizations: an integrative framework and review of emerging themes", Management Science, pp 571–582, 2003.

[22] Nonaka W. and Konno N. ,"The concept of building a foundation for knowledge creation", California management review. Pp.40-54, 1998.

[23] Nonaka I., "The Knowledge-Creating company", Harvard Business Review, PP. 96-104, 1991.

[24] Wigg, K. ,"Knowledge Management Foundation: thinking about thinking how people and organization create, represent and use knowledge", Arlington, TX: Schema Press, 1993.

[25] "Why ERP fails: most common reason", an article at website: http://articleseo.org/most-common-reason-why-erp-system-is-not-successfully implemented/

[26] Z., Lee & J., Lee ,"An ERPs implementation case study from a knowledge transfer perspective", Journal of Information Technology, Vol. 15, no. 4, pp. 281-288, 2000.

[27] Ellen & Wagner , "Concept in Enterprise Resource Planning", Course Technology , 2008.

Hamdan M. Al-Sabri received his B.Sc. degree in computer science in 2009 and his Master degree from King Saud University in 2011, and he is currently doing his PhD at the department of information Systems, college of computer and information sciences, King Saud University, Saudi Arabia. Al-Sabri has published papers in SOA, Knowledge Management, ERP, and Computer Security.

Saleh M. Al-Saleem Dr. Saleh Al-Saleem , Associate Professor in College of Computer and Information Science, King Saud University. He received his PhD from Wayne State University, Michigan, USA, 2001, in the field of computer science (Evolutionary Computation). He received his Master degree in computer science from Ball State University, IN, USA 1996, and His BS degree in computer science from College of education, King Saud University, Saudi Arabia 1991. He served as the dean of admission & registration in Shaqra University, and also served as the head of IT and e-Learning in Shaqra University. Previously he worked as head of Information Technology department at the Arab Open University, and before that he worked as the head of Computer Technology department and faculty member in Riyadh College of Technology. Dr. Al-Saleem current research interests includes: evolutionary computation, Text Classification, ERP, BPM, e-Learning, and Open Source.

The Evidential Reasoning Approach for Multiple Decision Analysis Using Normal Cloud Model

Li-Min Zhang[1], Shi-Jie Bao[1] and Chao Li[1]

**[1] Math and Computer Department, HengShui University, HengShui, 053000
Hebei, China**

Abstract

In this paper, normal cloud model and evidential reasoning (E-R) approach is used in multiple attribute decision analysis (MADA) problems. Different attributes Belief function are represented by cloud model interval. Using cloud model generating algorithm, belief degree interval is obtained without numerical computation. In addition, it is reasonable and it accords with human's mind. Evidential reasoning algorithm is also used to incorporate different attributes interval in different ranks. Maximum and minimum in belief degree interval is computed by software. Then aggregative index number of attribute value is computed. Thing's rank is decided by the index number. In the example, truck's integrated performances are analysed. Simulation results further illustrate the effectiveness of the design method.

Keywords: *Multiple attribute decision analysis (MADA), Evidential reasoning (E-R) approach, Cloud model, Cloud model belief degree.*

1. Introduction

Multiple attribute problem of quantitative and qualitative attribute are both exist in practice. Recently, it is a hot topic. There are different attributes in an object, which can be divided into two categories: data attribute and quality attribute. Data attribute is quantitative and quality attribute value is qualitative. Various factors should be taken into account in object analysis and evaluation. We use evidence reasoning method in multiple attribute decision analysis (MADA) problem[1]. Two attributes are in the same framework of MADA problem. We deal with the two attributes by unifying level estimating reliability structure and uncertainty of fuzzy linguistic variables [2].

MADA problem is mainly based on D-S theory, so it is lack of flexibility. In this paper, we use D-S 、E-R theory combined with normal cloud model to analyze multiple attribute problem. The advantage of this method is belief degree fuzzification. Fuzzy belief degree interval much more accords with human's mind than unfuzzed one. In belief degree calculating, X cloud model generating algorithm is used, which could adapt to MADA problem[3-7].

This work is organized as follow, the normal cloud model theory and evidential reasoning theory are reviewed in Section 2. Section 3 introduces effects of evidential reasoning using normal cloud model. Section 4 shows the simulation and results. Finally, Section 5 concludes the study of future work.

2. Review of Related Works

2.1 Normal Cloud Model

Definition 1:

Let U be the set, $U = \{x\}$,as the universe of discourse and T a linguistic term T , $C_T(x)$ is a random variable with a probability distribution $C_T(x)$ takes values in [0, 1]. A membership cloud is a mapping from the universe of discourse U to the unit interval [0,1], that is

$$C_T(x) \ : \ U \rightarrow [0,1]$$

$$\forall \ x \ \forall \ x \in U \ , x \rightarrow C_T(x)$$

In the society and science, the expected curve of membership cloud approach normal distribution, so we usually study the quality of normal membership cloud[1].Normal cloud curve can be describe using three important parameters $(Ex,\ En,\ He\ )$, Ex represents fixed quality conception or expected value, which is the center of normal cloud; En is entropy, which is the expected value and center value of He at the same time, is the scale in measuring the fuzzy degree and the only standard in measuring bandwidth[1]. He is super entropy (entropy's entropy), which represents the uncertain degree of En and shows the sparse degree of cloud. The three characteristics is the frame of cloud theory. Using the three characteristics, fixed conception could be represented by cloud model[3].

Normal half-ascended cloud model generated algorithm

(1) Give the expected value Enx, deviation Hex
(2) Generate a n dimensional normal random Enx, whose expected value is $E^{'}nx$, deviation is Hex
(3) Generate a n dimensional normal random $x = x_i$, whose expected value is Ex, deviation is $E^{'}nx$.
(4) Compute:

$$C_T = \exp[-\frac{1}{2}\frac{(x_i - Ex_i)^2}{E^{'}nx_i^{2}}]$$

(5) Repeat (1) ~ (4) till the number of cloud model drops is enough, if $x = x_i$ is given, the algorithm is X cloud model algorithm.

2.2 Generation and Structure of Cloud Model Belief Degree

If M Objects are estimated, there are L attributes in every object, and there are N ranks in every attribute, which are independent respectively. The rank of object α in attribute e_i is H_n . Belief degree is $\beta_{n,i}(\alpha_l)$. The estimation is $S(e_i(\alpha_l))$ ={ H_n, $\beta_{n,i}(\alpha_l)$ $(n = 1 \cdots N)$ }, $\beta_{n,i}(\alpha_l) \geq 0$

If there are N cloud model rank H_n $(n = 1 \cdots N)$ in M objects , then they are independent. Cloud model belief of H_n in attribute e_i of object α_l $(l = 1 \cdots M)$ is: [inf $\beta^-_{n,i}(\alpha_l)$, sup $\beta^-_{n,i}(\alpha_l)$] ∪ [inf $\beta^+_{n,i}(\alpha_l)$,where sup $\beta^+_{n,i}(\alpha_l)$], sup $\beta^+_{n,i}(\alpha_l) \geq$ sup $\beta^-_{n,i}(\alpha_l)$, inf $\beta^+_{n,i}(\alpha_l) \geq$ inf $\beta^-_{n,i}(\alpha_l)$.

Cloud model belief ($\beta^-_{n,i}(\alpha_l)$ and $\beta^+_{n,i}(\alpha_l)$ are interval values, which are generated by X cloud model algorithm. It is based on attribute interval [x_i^-, x_i^+], as $S(e_i(\alpha_l))$ ={ H_n ,[inf $\beta^-_{n,i}(\alpha_l)$, sup $\beta^-_{n,i}(\alpha_l)$] ∪ [inf $\beta^+_{n,i}(\alpha_l)$, sup $\beta^+_{n,i}(\alpha_l)$], $n = 1 \cdots N$ }, where $\beta^-_{n,i}(\alpha_l) \geq 0$. If inf $\beta^-_{n,i}(\alpha_l) \equiv$ inf $\beta^+_{n,i}(\alpha_l)$, sup $\beta^+_{n,i}(\alpha_l) \equiv$ sup $\beta^-_{n,i}(\alpha_l)$, then attribute value is precise. As follow, we give the definition:

Definition 2 :
$S(e_i(\alpha_l))$ ={ H_n,[inf $\beta^-_{n,i}(\alpha_l)$, sup $\beta^-_{n,i}(\alpha_l)$] ∪ [inf $\beta^+_{n,i}(\alpha_l)$, sup $\beta^+_{n,i}(\alpha_l)$], is cloud model estimated vector of attribute value, if [inf $\beta^-_{n,i}(\alpha_l)$, sup $\beta^-_{n,i}(\alpha_l)$] ∪ [inf $\beta^+_{n,i}(\alpha_l)$, sup $\beta^+_{n,i}(\alpha_l)$] satisfies : $\exists$ $\beta_{n,i}(\alpha_l) \in$ [inf $\beta^-_{n,i}(\alpha_l)$, sup $\beta^-_{n,i}(\alpha_l)$], $\sum_{n=1}^{N} \beta_{n,i}(\alpha_l) \leq 1$, then $S(e_i(\alpha_l))$ is valid, otherwise it is invalid.

Definition 3:
$S(e_i(\alpha_l))$ ={ H_n,[inf $\beta^-_{n,i}(\alpha_l)$, sup $\beta^-_{n,i}(\alpha_l)$] ∪ [inf $\beta^+_{n,i}(\alpha_l)$, sup $\beta^+_{n,i}(\alpha_l)$], $n = 1 \cdots N$ } is attribute
cloud distribution estimation vector, if belief interval [inf $\beta^-_{n,i}(\alpha_l)$, sup $\beta^-_{n,i}(\alpha_l)$], $\exists$ $\beta^-_0(\alpha_l) \in$ [inf $\beta^+_{n,i}(\alpha_l)$, sup $\beta^+_{n,i}(\alpha_l)$], $\beta^+_0(\alpha_l) \in$ [inf $\beta^+_{n,i}(\alpha_l)$, sup $\beta^+_{n,i}(\alpha_l)$], $\exists \, \forall$ $\beta_{n,i}(\alpha_l) \in$ [$\beta^-_0(\alpha_l)$, $\beta^+_0(\alpha_l)$], where $\sum_{n=1}^{N} \beta_{n,i}(\alpha_l) = 1$, $S(e_i(\alpha_l))$ is called complete estimated vector, or incomplete.
In complete cloud distribution estimation, there is only one rank estimation in α_l. The other belief distributes are in the whole set H, if cloud distribution is not incomplete.

Definition 4:
$S(e_i(\alpha_l))$ = { H_n,[inf $\beta^-_{n,i}(\alpha_l)$, sup $\beta^-_{n,i}(\alpha_l)$] ∪ [inf $\beta^+_{n,i}(\alpha_l)$, sup $\beta^+_{n,i}(\alpha_l)$], $n = 1 \cdots N$ }
is not incomplete cloud model estimated vector. Belief degree
$\beta_{H,i}(\alpha_l)$ is assigned to H [4-5].

$$\inf \beta^-_{H,i}(\alpha_l) = \max(0, 1 - \max \sum_{n=1}^{N} \beta^-_{n,i}(\alpha_l)),$$

$$\sup \beta^-_{H,i}(\alpha_l) = \max(0, 1 - \inf \sum_{n=1}^{N} \beta^-_{n,i}(\alpha_l)) \quad (1)$$

$$\inf \beta^+_{H,i}(\alpha_l) = \max(0, 1 - \max \sum_{n=1}^{N} \beta^+_{n,i}(\alpha_l)),$$

$$\sup \beta^+_{H,i}(\alpha_l) = \max(0, 1 - \inf \sum_{n=1}^{N} \beta^+_{n,i}(\alpha_l)) \quad (2)$$

after L attributes in M objects are estimated[8-11], cloud model belief degree Decision-making matrix is given: $D=(S(e_i(\alpha_l)))_{L\times M}$

3. Attribute Dater Representation in Cloud Model Belief Degree

3.1 Attribute Data Representation in Cloud Model Belief Degree

Dater attribute usually could be divided into two parts

(1)Accurate data attribute representation
Attribute value is usually represented by accurate data. To deal with MADA problem in E-R method, we make all the cloud model rank Figure and get the belief interval of data. In order to describe the Evaluation rank of data attribute, we should know effect of every rank. H_1 is impossible rank and H_N the highest[12-13].

(2)Interval data attribute representation
Because interval data crosses over many ranks, representation of cloud model is much more complex. If $[x_i^-, x_i^+]$ crosses two ranks H_n, H_{n+1}, the other is the same. Belief degree $\beta_{n,i}$ is generated by X cloud model algorithm. Belief degree interval on H_n, H_{n+1} is

$\{H_n, [\inf\beta_{n,i}^-(\alpha_l), \sup\beta_{n,i}^-(\alpha_l)]\cup[\inf\beta_{n,i}^+(\alpha_l), \sup\beta_{n,i}^+(\alpha_l)], n=1\cdots N\}$; $\{H_{n+1}, [\inf\beta_{n+1,i}^-(\alpha_l), \sup\beta_{n+1,i}^-(\alpha_l)]\cup[\inf\beta_{n+1,i}^+(\alpha_l), \sup\beta_{n+1,i}^+(\alpha_l)], n=2\cdots N\}$

3.2 Data Integration of Attribute Cloud Model Distribution Belief Degree

E-R analysis algorithm can fully use and synthesize the evidence. Cloud model theory can strengthen capability of processing uncertain evidence data. Cloud model theory belief is transformed into mass function using formula(3)~(6)

$$m_{n,i}=m_i(H_n)=\omega_i\beta_{n,i}(\alpha_l), n=1\cdots N ; i=1\cdots L \quad (3)$$

$$m_{H,i}=m_i(H)=1-\sum_{n=1}^{N}m_{n,i}=1-\omega_i\sum_{n=1}^{N}\beta_{n,i}(\alpha_l), \quad i=1\cdots L \quad (4)$$

$$\bar{m}_{H,i}=\bar{m}_i(H)=1-\omega_i, i=1\cdots L \quad (5)$$

$$\tilde{m}_{H,i}=\tilde{m}_i(H)=\omega_i(1-\sum_{n=1}^{N}\beta_{n,i}(\alpha_l)), \ i=1\cdots L \quad (6)$$

$$m_{H,i}=\bar{m}_{H,i}+\tilde{m}_i(H), \text{ and } \sum_{i=1}^{L}\omega_i=1$$

The possibility of set H is m_H and it is divided into two parts $\tilde{m}_H, \bar{m}_H$. Multiple attribute mass function are integrated in (7)~(12)

$$\{H_n\}: m(H_n)=k\{\prod_{i=1}^{L}[m_i(H_n)+m_i(H)]-\prod_{i=1}^{L}m_i(H)\}, n=1\cdots N \quad (7)$$

$$\{H\}: \tilde{m}_H=k\{\prod_{i=1}^{L}m_i(H)-\prod_{i=1}^{L}\bar{m}_i(H)\} \quad (8)$$

$$\{H\}: \bar{m}_H=k[\prod_{i=1}^{L}\bar{m}_i(H)] \quad (9)$$

$$k=\{\sum_{n=1}^{L}\prod_{i=1}^{L}[m_i(H_n)+m_i(H)]-(N-1)\prod_{i=1}^{L}\bar{m}_i(H)\}^{-1} \quad (10)$$

$$\{H_n\}: \beta_n=\frac{m(H_n)}{1-\bar{m}_H}, n=1\cdots N \quad (11)$$

$$\{H\}: \beta_H=\frac{\tilde{m}(H)}{1-\bar{m}_H} \quad (12)$$

Multiple attribute cloud model belief degree is based on (13~15)

$$m_{n,i}=m_i(H_n)\in[\inf m_{n,i}^-, \sup m_{n,i}^+]=[\omega_i\inf\beta_{n,i}^-, \omega_i\sup\beta_{n,i}^+], n=1\cdots N \quad i=1\cdots L \quad (13)$$

$$\bar{m}_{H,i}=\bar{m}_i(H)=1-\omega_i, \ i=1\cdots L \quad (14)$$

$$\tilde{m}_{H,i}=\tilde{m}_i(H)\in[\inf\tilde{m}_{H,i}^-, \sup\tilde{m}_{H,i}^+]=[\omega_i\inf\beta_{H,i}^-, \omega_i\sup\beta_{H,i}^+] \quad (15)$$

$$\text{and} \sum_{n=1}^{N}m_{n,i}+\bar{m}_{H,i}+\tilde{m}_{H,i}=1, \sum_{i=1}^{L}\omega_i=1$$

3.3 Cloud Model Distribution Expectation Effect

$$u(S(e_i(\alpha_l)))=\sum_{n=1}^{N}u(H_n)\beta_n(\alpha_l), l=1\cdots M$$

$u(S(e_i(\alpha_l)))$ is cloud model distribution expectation effect and $u(H_n)$ is the effect of H_n. $\beta_{n,i}(\alpha_l)$ is belief degree of α_l on H_n. If distribution is complete,

$\beta_H(\alpha_l)$ =0 and If distribution is complete, incomplete $u(S(e_i(\alpha_l)))$ has maximum and minimum: (16-17)

$$\begin{aligned} &u_{\max}(\alpha_l)=\sum_{n=1}^{N\text{-}1}u(H_n)\beta_n(\alpha_l)\\ &(\beta_N(\alpha_l)+\beta_H(\alpha_l))u(H_N),l=1\cdots M \qquad (16)\end{aligned}$$

$$\begin{aligned} &u_{\min}(\alpha_l)=\sum_{n=2}^{N}u(H_n)\beta_n(\alpha_l)+\\ &(\beta_1(\alpha_l)+\beta_H(\alpha_l))u(H_1),l=1\cdots M \qquad (17)\end{aligned}$$

$$u_{ave}(\alpha_l)=\frac{u_{\max}(\alpha_l)+u_{\min}(\alpha_l)}{2}$$

4. Experiment Result

Description of truck's attributes

There are many factors in truck's comprehensive estimation: acceleration time (s), braking (m), power (kw), gear-box property, weight ω_i =0.25 $(i=1\cdots 4)$.

rank of truck: top(T), excellent(E) , good(G), average(A) , poor(P) , worst(W).

H_j={ $H_j \mid j=1\cdots 7$ }={'top', 'excellent', 'good', 'average', ' poor', ' worst'}

Table 1: Attribute values

attribute	car1	Car 2
acceleration time	4.4	4.0
braking	[19.2,19.26]	[19.11,19.2]
power	288	223
gear-box property	5.4	[6,7]

Table 2: Belief degree of attributes in different ranks

attribute	truck 1	truck 2
acceleration time	*P* [0.03, 0.09] *A* [0.8, 0.86]	*G* [0.25, 0.42] *E* [0.25, 0.42]
braking	*G* [0, 0] *E* [1, 1] *E* [0.78,0.90] *G* [0,0.15]	*T* [0.02,0.16] *E* [0.81,0.88] *T* [0, 0] *E* [1, 1]
power	*E* [0.1, 0.17] *T* [0.54, 0.68]	*P* [0.31, 0.45] *A* [0.21, 0.34]
gear-box property	*A* [0.1, 0.3] *P* [0.5, 0.6]	*G* [1, 1] *A* [0, 0]

Data in Table 1 is transferred into belief degree interval in Table 2 using cloud model theory . Every attribute value corresponds to a belief interval.

Every different attribute has its rank (Fig.1~Fig.4)

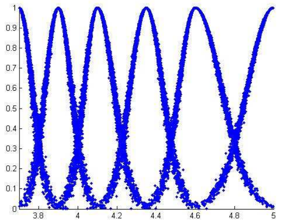


Fig.1 Cloud rank of acceleration time

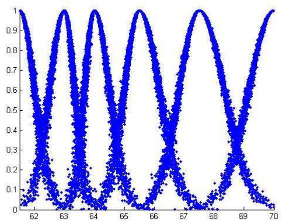


Fig.2 Cloud rank of braking

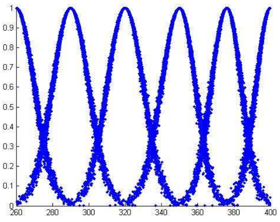


Fig.3 Cloud rank of power

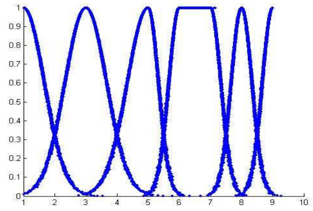


Fig.4 Cloud rank of gear-box property

Use (4)~(11) formula of dater integration

Truck 1, acceleration time:

$m_{1,1}$ =0, $m_{2,1}$ =0, $m_{3,1}$ =0, $m_{4,1} \in$ 0.25*[0.8,0.86], $m_{5,1} \in$ 0.25*[0.03,0.09], $m_{6,1}$ =0, $m_{1,H} \in$ 1−0.25 *[0.83,0.95]=[0.76,0.79], $\bar{m}_1(H) \in$ [0.75, 0.75], $\tilde{m}_1(H) \in$ [0.01, 0.75]

braking:

$m_{1,2}$ =0, $m_{2,2} \in$ 0.25*[0.78,1], $m_{3,2} \in$ 0.25*[0, 0.15], $m_{4,2}$ =0, $m_{5,2}$ =0, $m_{6,2}$ =0, $m_{2,H} \in$ 1−0.25*[0.78,1]=[0.75,0.8], $\bar{m}_2(H) \in$ [0.75,0.75], $\tilde{m}_2(H) \in$ [0, 0.06]

power:

$m_{1,3} \in$ 0.25*[0.54, 0.68], $m_{2,3} \in$ 0.25*[0.1, 0.17], $m_{3,3}$ =0, $m_{4,3}$ =0, $m_{5,3}$ =0, $m_{6,3}$ =0, $m_{3,H} \in$ 1−0.25*[0.64, 0.85]=[0.79,0.84], $\bar{m}_3(H) \in$ [0.75, 0.75], $\tilde{m}_3(H) \in$ [0.04, 0.09]

gear-box property:
$m_{1,4} \in 0.25*[0.54, 0.68], m_{2,4} \in 0.25*[0.1,0.17],$
$m_{3,4}=0, m_{4,4} \in 0.25*[0.1,0.3], m_{5,4} \in 0.25*[0.5,0.6],$
$m_{6,4}=0\ m_{4,H} \in 1-0.25*[0.6,0.9]=$
$[0.77, 0.85], \overline{m}_4(H) \in [0.75,0.75], \tilde{m}_4(H) \in [0.02,0.1]$
Truck2:
$\overline{m}_H \in [0.32,0.5], \tilde{m}_H \in [0.01,0.34], \beta_1 \in [0,0.76],$
$\beta_2 \in [0.18,1], \beta_3 \in [0.2,1], \beta_4 \in [0.03,0.84], \beta_5 \in$
$[0.04, 0.88], \beta_6=0, \beta_H \in [0.02,0.68]$

5. Conclusion

This paper has proposed a new method of evidence reasoning based on normal cloud model and introduced an method that belief degree is represented in interval value. To overcome the drawbacks of evidence reasoning, we adopted fuzzy method. Example in truck shows that the method could achieve better estimating effect than generic evidence reasoning, and own a good performance in truck quanlity estimation. We will further consider the selection of cloud model parameter in future work.

References

[1] J.B Yang, M.G. Singh, "An evidential reasoning approach for multiple attribute decision making with uncertainty", IEEE Tansaction on Systems, Man, and Cybernetics, Vol. 24, No. 1, 1994, pp. 1-18.

[2] J.B. Yang, Y.M. Wang, "The evidential reasoning approach for MADA under both probabilistic and fuzzy uncertainties", European journal of operational research, Vol. 171, 2006, pp. 310-312.

[3] D.Y Li, "membership clouds and membership cloud generators", Computer research and development, Vol. 6, No. 32, 1996 , pp. 15-20.

[4] Dempster, A.P., "Upper and lower probabilities induced by a multi-valued mapping", Annals of Mathematical Statistics, 1967.

[5] Shafer, G., A Mathematical Theory of Evidence, NJ : Princeton University Press, 1976.

[6] Bauer, M. "Approximation algorithms and decision making in the Dempster–Shafer theory of evidence—an empirical study", International Journal of Approximate Reasoning, Vol. 17, 1997, pp. 217–237.

[7] M. Beynon, D. Cosker, D. Marshall, "An expert system for multi-criteria decision making using Dempster Shafer theory", Expert Systems with Applications, Vol.20, 2001, pp. 357–367.

[8] M. Beynon, B. Curry, P. Morgan, "The Dempster–Shafer theory of evidence: An alternative approach to multicriteria decision modeling", Omega, Vol. 28, 2000, pp. 37–50.

[9] M. Beynon, "DS/AHP method: A mathematical analysis, including an understanding of uncertainty", European Journal of Operational Research, Vol. 140, No. 1, 2002, pp. 148–164.

[10] E. Binaghi, I. Gallo, P. Madella, "A neural model for fuzzy Dempster–Shafer classifiers", International Journal of Approximate Reasoning, Vol. 25, No. 2, 2000, pp. 89–121.

[11] E. Binaghi, P. Madella, "Fuzzy Dempster–Shafer reasoning for rule-based classifiers", International Journal of Intelligent Systems, 1999, pp. 559–583.

[12] G. Biswas, M. Oli, A. Sen, "An expert decision support system for production control", Decision Support Systems, 1988, pp. 235–248.

[13] L.H. Chen, "An extended rule-based inference for general decision-making problems", Information Sciences, 1997, pp. 111–131.

First Author received a Master Degree in Engineering from Hebei University of Science and Technology, Shijiazhuang, China, in 2008. He is now a instructor in Hengshui University, Hengshui, China. His research interests cover the knowledge representation&data mining.

Second Author received a Master Degree in math from Hebei University, Baoding, China, in 2012. He is a instructor in Hengshui University, Hengshui, China.

Third Author works in Hengshui University.

Genetic Fuzzy Logic Control Technique for a Mobile Robot Tracking a Moving Target

Karim Benbouabdallah and Zhu Qi-dan

College of Automation, Harbin Engineering University
Harbin, 150001, China

Abstract

Target tracking is a crucial function for an autonomous mobile robot navigating in unknown environments. This paper presents a mobile robot target tracking approach based on artificial intelligence techniques. The proposed controller calculates both the mobile robot linear and angular velocities from the distance and angle that separate it to the moving target. The controller was designed using fuzzy logics theory and then, a genetic algorithm was applied to optimize the scaling factors of the fuzzy logic controller for better accuracy and smoothness of the robot trajectory. Simulation results illustrate that the proposed controller leads to good performances in terms of computational time and tracking errors convergence.

Keywords: *Mobile robot, Target tracking, Fuzzy logic controller, Genetic algorithm.*

1. Introduction

An autonomous mobile robot is a programmable and multi-tasks mechanical device, capable to navigate freely or execute different functions such as obstacles avoidance, target tracking …etc.

The last few decades have witnessed ambitious research efforts in the areas of mobile robotics, due to their wide range of use in different fields like military and industrial applications. These research works aim to improve their operational capabilities of navigation and interaction with its surrounding work environments through the different kinds of sensors.

Target tracking is one of the basic and an interesting function for a mobile robot, and researchers have worked to propose different control approaches to improve target tracking performances. Several control approaches like PID controllers [1], non linear controllers based Lyapunov stability analysis [2] and sliding mode control [3] have been developed to make the mobile robot track easily a moving target. However, these approaches have showed some problems, due to the complexity of the mobile robot surrounding environment to be modeled or to the simplification assumptions taken for the elaboration of the mathematical model of the robot and control law.

To overcome these drawbacks and the need to exhibit robust performances while operating in highly uncertain and dynamic environments, artificial intelligence approaches have been attracting considerable research interest in recent years. Different techniques such as reinforcement learning [4], neural networks [5], fuzzy logics [6] [7] and genetic algorithms [8] have been applied to synthesize control laws enabling the mobile robot to follow a moving target freely. Artificial potential field approaches [9] [10] have been also developed to solve the problem of mobile robots target tracking.

The intelligent control does not require knowing exactly the mathematical model of the mobile robot or its surrounding work space, since it uses human reasoning and decision making in spite of uncertainty and imprecise information provided by the different perception sensors such as cameras, ultrasonic and infrared sensors.

A fuzzy logic controller is a control strategy whose decisions are made by using a fuzzy inference system, which is a rule-based or knowledge-based system containing a collection of fuzzy if-then rules based on human experts. It utilizes heuristic knowledge to develop perception-action for mobile robots to achieve different tasks (target tracking, obstacle avoidance, path tracking). Furthermore, fuzzy logic controllers' methodology is very helpful because it deals with uncertainties in real word.

For more robustness and effectiveness of FLCs, many researchers have explored the use of genetic algorithms to tune FLC in order to optimize its different parameters (termsets, rules, scaling factors) [11] [12]. The basic concepts of GA were developed by Holland [13], and subsequently in several researches work [14]. Genetic algorithms are a robust optimization technique because they ensure a gradual increasing of a good solution. Thus, GA does not need gradient or any prior information of the search space which is usually not available for the designer. Therefore, GA has lower chance to converge into local minima because they operate over a population of points.

Hence, the use of genetic algorithms is more suitable for this kind of situations.

In this paper, a control approach for target tracking by a mobile robot is presented, based on fuzzy logics and optimized with a genetic algorithm for more efficiency and effectiveness. The paper is organized as follows. A general description of the system model and problem formulation are given in section 2. The adopted fuzzy logic controller is presented in section 3. Section 4 describes the genetic algorithm used to tune the FLC. To validate the proposed approach, simulation results are discussed in section 5. Conclusion and future works are given in section 6.

2. Model System and Problem Formulation

The objective of this paper is to make an autonomous mobile robot track a moving target. To realize that, we shall assume that:

- The target and the robot move on the same plane in the absence of obstacles.
- The target moves along an unknown trajectory and its position is denoted as $o_T(x_T, y_T)$.
- The robot has the kinematics of a unicycle. It is described by the following equations:

$$\begin{cases} \dot{x} = V\cos\theta \\ \dot{y} = V\sin\theta \\ \dot{\theta} = w \end{cases} \tag{1}$$

where (x, y) are the robot coordinates in the world frame XOY, θ is its orientation with respect to the same frame, and V and w are the robot linear and angular velocities.

- The robot is subject to non-holonomic constraint:

$$\dot{y}\cos\theta - \dot{x}\sin\theta = 0 \tag{2}$$

The linear and angular positions of the mobile robot to the target to be tracked are denoted as D and α, respectively and can be written by the following equations:

$$\begin{cases} \alpha = \theta - \beta \\ \beta = \tan^{-1}\left(\dfrac{y_T - y}{x_T - x}\right) \\ D = \sqrt{(x_T - x)^2 + (y_T - y)^2} \end{cases} \tag{3}$$

Roughly speaking, to achieve the target tracking objectives, the main idea is to design an intelligent controller enabling the robot to move freely such that it maintains a desired distance to the target and steer in order to point the target's direction.

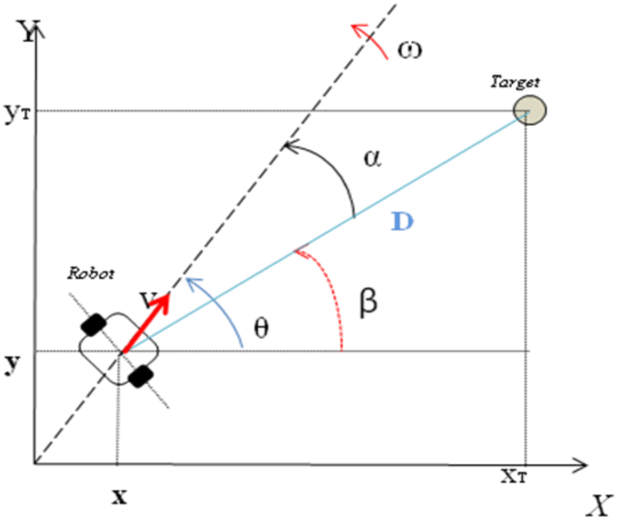


Fig. 1 Schematic model of a mobile robot tracking a target.

3. Design of the Fuzzy Logic Controller

The fundamental objective is to synthesize a robust control law able to force the mobile robot to follow a moving target as closely as possible.

We define the distance and angle errors and the change in angle error variables to mathematically formulate the control objective as follows:

$$\begin{cases} e_D = D - D_d \\ e_\alpha = \alpha - \alpha_d \\ \Delta e_\alpha = e_\alpha(k) - e_\alpha(k-1) \end{cases} \tag{4}$$

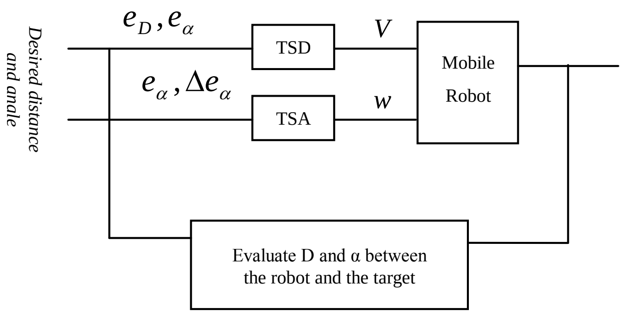

Desired distance and angle
e_D, e_α
$e_\alpha, \Delta e_\alpha$
TSD
TSA
V
w
Mobile Robot
Evaluate D and α between the robot and the target

Fig. 2 FLC block diagram

It's clear that to achieve the control objective, it suffices to make the tracking errors tend to zero. Thus, two fuzzy logic Takagi-Sugeno controllers of order zeros have been designed: distance controller TSD and angle controller TSA.

3.1 Distance controller TSD

This controller aims to keep a desired distance D_d between the mobile robot and the target when it navigates in its work space. The control input variables are chosen as the distance and angle errors e_D and e_α . The linear velocity of the mobile robot V is defined as the output of TSD, so we can write:

$$V = K_V.TSD(K.e_D, K'.e_\alpha) \qquad (5)$$

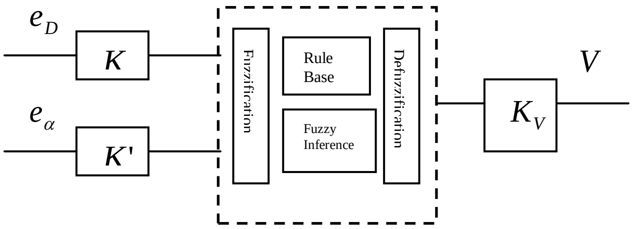


Fig. 3 Distance controller architecture

The inputs/output linguistic variables are defined as:

$$\begin{cases} e_D : \{N, ZE, PS, PM, PB\} \\ e_\alpha : \{NB, NM, NS, ZE, PS, PM, PB\} \\ V : \{N, ZE, PS, PM, PB\} \end{cases}$$

Triangular distributions in [-1, 1] interval were chosen as membership functions for the scaled inputs e_D and e_α and singletons for V .

Table 1 shows the rules' base of TSD. For example, if the errors are big, it means that the robot is so far from the target, so it should be much faster to reduce the distance to D_d .

Table 1: TSD Rules' Base

Linear Velocity		***Distance Error***				
		N	***ZE***	***PS***	***PM***	***PB***
Angle Error	***NB***	N	PS	PM	PB	PB
	NM	N	ZE	PS	PM	PB
	NS	N	ZE	PS	PM	PB
	ZE	N	ZE	PS	PM	PB
	PS	N	ZE	PS	PM	PB
	PM	N	ZE	PS	PM	PB
	PB	N	PS	PM	PB	PB

3.2 Angle controller TSA

This controller aims to point the mobile robot in the moving target's direction, which means reduce the angle error to zero. It has two input variables; the angle error e_α and the change of error Δe_α . It returns the angular velocity of the mobile robot w , so we can write:

$$w = K_w.TSD(K'.e_\alpha, K''.\Delta e_\alpha) \qquad (6)$$

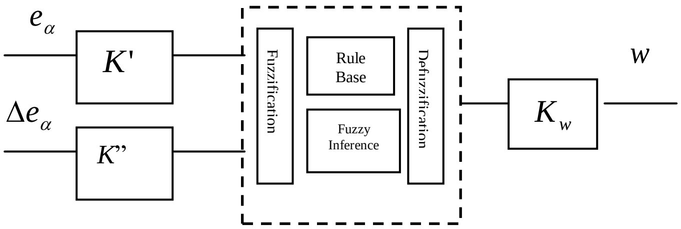


Fig. 4 Angle controller architecture

The inputs/output linguistic variables are defined as:

$$\begin{cases} e_\alpha : \{NB, NM, NS, ZE, PS, PM, PB\} \\ \Delta e_\alpha : \{NB, NM, NS, ZE, PS, PM, PB\} \\ w : \{NB, NM, NS, ZE, PS, PM, PB\} \end{cases}$$

Table 2: TSA Rules' Base

Angular Velocity		***Angle Error***						
		NB	***NM***	***NS***	***ZE***	***PS***	***PM***	***PB***
Change In Error	***PB***	ZE	NS	NB	NB	NB	NB	NB
	PM	PS	ZE	NS	NM	NM	NB	NB
	PS	PB	PS	ZE	NS	NS	NM	NM
	ZE	PB	PM	PS	ZE	NS	NM	NB
	NS	PB	PM	PS	PS	ZE	NS	NB
	NM	PB	PB	PM	PM	PS	ZE	NS
	NB	PB	PB	PB	PB	PB	PS	ZE

Triangular distributions in [-1, 1] interval were chosen as membership functions for the scaled inputs e_α and Δe_α and singletons for w . The rules' base of the controller TSA is depicted in Table 2.

Note that the inputs' scaling factors of the two fuzzy logic controllers $(K, K', K")$ will be optimized with the genetic algorithm. The outputs' scaling factors (K_V, K_w) are not optimized because they do not affect the controllers' performances. This appears to be due to the scaling down of the linear velocity to not exceed the maximal velocity of the mobile robot.

4. The adopted Genetic Algorithm

In this part, a genetic algorithm is applied to tune the FLCs inputs scaling factors in order to guarantee a fast and smooth robot trajectory by reducing computational time and control errors.

A fitness function is used as a convergence criterion to satisfy to measure the best values of the FLCs inputs' scaling factors. The fitness function J is chosen as:

$$J = \frac{1}{2}\int ((\frac{e_D}{e_D(1)})^2 + (\frac{e_\alpha}{e_\alpha(1)})^2)dt \qquad (7)$$

where $e_D(1)$ and $e_\alpha(1)$ are the tracking errors at the initial positions of the robot and the target.

Once the genetic representation of the population and the fitness function are defined, the genetic algorithm proceeds to initialize a population of solutions, which is usually generated randomly and then try to improve it through repetitive application of operators of selection, mutation and crossover.

Fig. 5 shows the fitness function evolution according to the number of generation of the genetic algorithm.

The steps involved in the proposed genetic algorithm are summarized with the following steps:

Step 1: Generate an initial random population of individuals for a fixed size according to the variation range of each scaling factor.

Step 2: Evaluate the fitness of each individual in the population.

Step 3: Select fittest individuals of the population for reproduction.

Step 4: Breed new individuals through reproduction, crossover and mutation operators.

Step 5: Repeat step 2 until the convergence criterion is achieved.

Once the genetic algorithm is completed, it will return three parameters corresponding to the inputs' scaling factors of the two designed fuzzy logic controllers TSD and TSA; which are the best population found during the execution of GA.

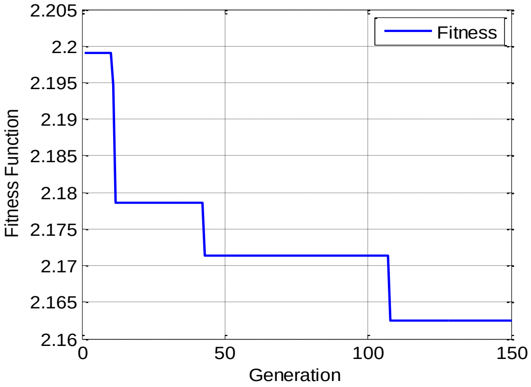


Fig. 5 Fitness Function's Evolution

The population's size and generations' number are the most significant parameters of GA since they have direct influence on the convergence of the GA to the optimal solution. Table 3 and 4 show the genetic algorithm results for different size of population and generations' number.

Table 3: GA's Results with Population Size S=30

Generations	K	K'	K''	*J*
50	3.8844	0.6251	0.3247	2.1605
100	3.7341	0.6287	0.3389	2.1615
150	3.9797	0.6208	0.3219	2.1625

Table 4: GA's Results with Generations G=50

Population size	K	K'	K''	*J*
10	3.7493	0.6212	0.4950	2.1851
30	3.8844	0.6251	0.3247	2.1605
50	3.4694	0.6333	0.4043	2.1702

5. The Simulation Results and Discussions

To perceive the effectiveness of the control scheme proposed in this paper, we have simulated both the fuzzy logic controller and its optimization with the genetic algorithm using Matlab 7.0 ; when the mobile robot pursue a target moving along a circular path during the interval time [0, 55s], then a straight line path for the interval [55s, 90s]. The target's trajectory can be described by the following equations:

$$\begin{cases}(x_T-2.5)^2+(y_T-2.5)^2=9\\ y_T=-0.5x_T+1\end{cases} \tag{8}$$

The initial positions of the mobile robot and target are given by (-1, 1) and (5.5, 2.5) respectively. The linear velocity of the target is set to be 0.2 m/s. The desired values of the distance and angle from the robot to the target are set to be $D_d = 0.2m$ and $\alpha_d = 0$.

The simulation results are depicted in Fig. 6 to Fig. 10. The trajectories of the target and the robot applying the two designed control laws are plotted in Fig. 6. Fig. 7 and Fig. 8 show the variation of the tracking errors during the motion of the target and the robot. Fig. 9 and Fig. 10 show the evolution of the mobile robot's linear and angular velocities.

Table 4 presents the time of convergence of the tracking errors for both FLC and FLC-GA controllers.

Table 5: Time of Convergence of the Two Controllers

Convergence time (s)	***Distance Error***	***Angle Error***
FLC	15.6	32.5
FLC-GA	10	15

From Fig. 6, it can be observed that the robot track easily the moving target and catches up with it as well the tracking errors tend to zero.

In term of robot's velocities, the robot moves forward with high speeds and evolve to stabilize then around the target's velocities as the tracking errors are converging to maintain the control's objectives (Fig. 7 to Fig. 10).

In the other hand, in term of comparison between the two controllers' performances, it's obvious that the FLC-GA controller is much better than the FLC controller. In fact, the trajectory of the mobile robot under FLC-GA is much smoother and shorter than that under FLC, which presents deviations in its trajectory before it matches with that of the target. The FLC's outputs exhibit fluctuations at various time instants with high magnitudes than those of the FLC-GA. As shown in Table 5, the FLC-GA is much effective than FLC in term of the speed of convergence of the tracking errors.

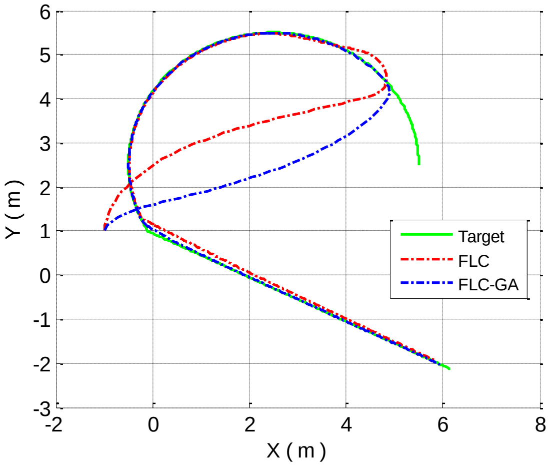


Fig. 6 Robot and Target Trajectories

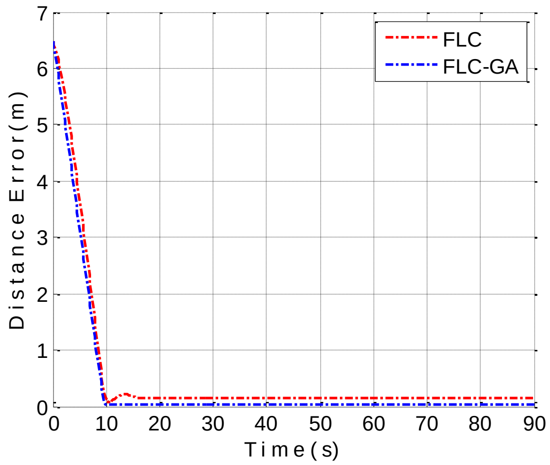


Fig. 7 Distance Error Evolution

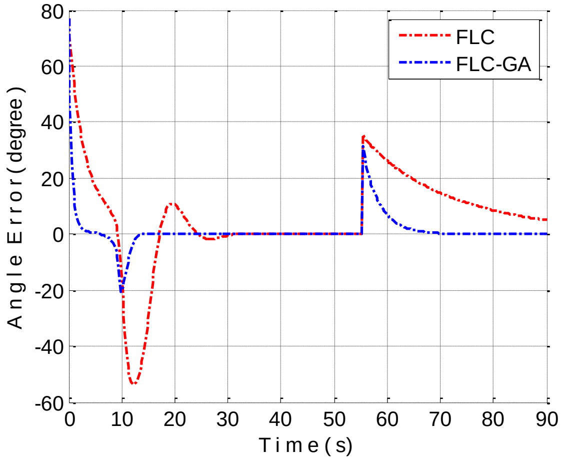


Fig. 8 Angle Error Evolution

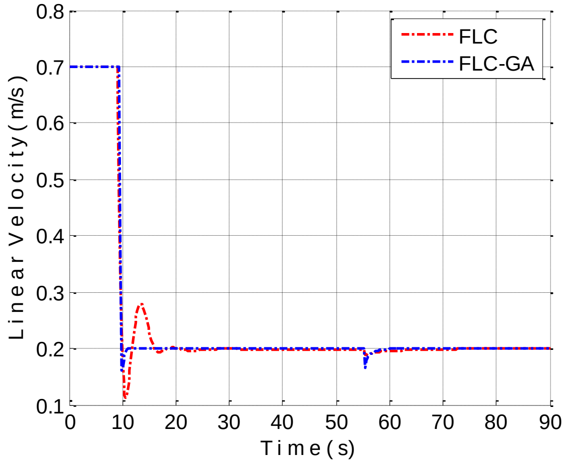


Fig. 9 Robot's Linear Velocity

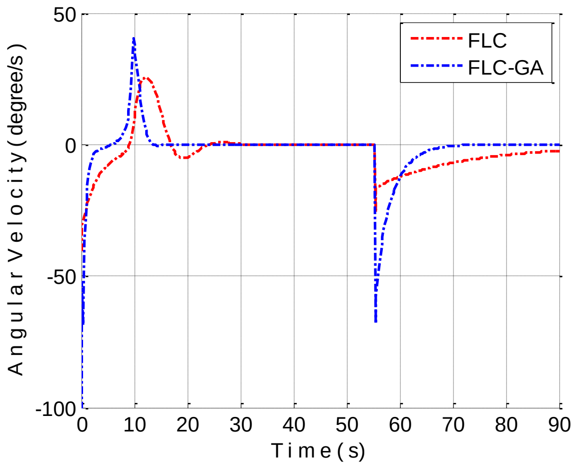


Fig. 10 Robot's Angular Velocity

We have also tested the target tracking under different outputs' scaling factors (K_V, K_w). The target and robot trajectories are depicted in Fig.11.

Table 5: Outputs' scaling factors

(K_V, K_w)	Case 1	Case 2	Case 3	Case 4
	(1.3, 3)	(2, 3.5)	(2.5, 4)	(3, 4.5)

The trajectory of the robot under the different cases are almost the same and catches up the target at the same point within the convergence time; which can be explained by the saturation block added in the output of the TSD controller to scale down the linear velocity when it exceeds 0.7 m/s. Thus, the optimization of these factors is not necessary since they don't affect the controllers' performances.

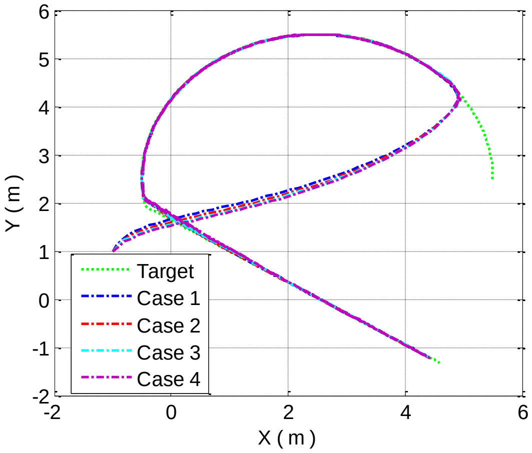


Fig. 11 Robot and Target Trajectories under different outputs' scaling factors.

6. Conclusions

This paper addresses the problem of the tracking of moving target along an unknown trajectory by a mobile robot. The proposed approach aims to design an intelligent controller based on fuzzy logic techniques and improved by a genetic algorithm.

The principal role of artificial intelligence techniques is their ability to design robust controllers with good performances in spite of the lack of information about the mobile robot or its surrounding environment models. Two fuzzy logic controllers based on Takagi-Sugeno model and have been adopted to determine the mobile robot's velocities to fulfill the control objectives. A genetic algorithm has been also implemented to improve the FLC by optimizing its inputs scaling factors for better efficiency and effectiveness.

The provided simulation results show that the proposed approach acts successfully and enable the robot to track the moving target easily with good performances in term of convergence time and accuracy. The use of a genetic algorithm to tune the inputs' scaling parameters of the FLC makes the robot's trajectory and the controllers' outputs much smoother and reduces considerably the convergence time and the tracking errors.

Further works may be directed to extend the result in an environment containing obstacles and considering the target velocities in the design of the control law.

References

[1] P. K. Padhy, T. Sasaki, S. Nakamura and H. Hashimoto, "Modeling and Position Control of Mobile Robot", in International Workshop on Advanced Motion Control, 2010, pp. 100-105.

[2] R. Carelli, C. M. Soria, and B. Morales, "Vision-Based Tracking Control for Mobile Robots", in IEEE International Conference on Advanced Robotics, 2005, pp. 148-152.

[3] J. Qiuling, X. Xiaojun, and L. Guangwen, "Formation Path Tracking Controller of Multiple Robot System by High Order Sliding Mode", in IEEE International Conference on Automation and Logistics, 2007, pp. 923-927.

[4] R. Calvo and M. Figueiredo, "Reinforcement Learning for Hierarchical and Modular Network in Autonomous Robot Navigation", in International Joint Conference on Neural Networks, 2003, pp. 1340-1345.

[5] X. Ma, W. Liu, Y. Li and R. Song, " LVQ Neural Network Based Target Differentiation Method for Mobile Robot ", in IEEE International Conference on Advanced Robotics, 2005, pp. 680-685.

[6] T-H. S. Li, S. J. Chang, and W. Tong, “Fuzzy Target Tracking Control of Autonomous Mobile Robots by using Infrared Sensors ", IEEE Transactions on Fuzzy Systems, Vol. 12, No. 4, 2004, pp. 491-501.

[7] L. Ming, G. Zailin and y. Shuzi, "Mobile Robot Fuzzy Control Optimization using Genetic Algorithm", Artificial Intelligence in Engineering, Vol. 10, No. 4, 1996, pp. 293-298.

[8] L. Moreno, J. M. Armingol, S. Garrido, A. De La Escalera and M. A. Salishs, "Genetic Algorithm for Mobile Robot Localization using Ultrasonic Sensors ", Journal of Intelligent and Robotic System, Vol. 34(2), 2002, pp. 135-154.

[9] M. Taher, H. E. Ibrahim, S. Mahmoud and E. Mostafa, "Tracking of Moving Target by Improved Potential Field Controller in Cluttered Environments ", International Journal of Computer Science Issues, Vol. 9(2), No. 3,2012, pp. 472-480.

[10] L. Huang, "Velocity Planning for a Mobile Robot to Track a Moving Target - a Potential Field Approach", Robotics and Autonomous Systems, Vol. 57(1), No. 3,2009, pp. 55-63.

[11] Q. Liu, Y-G. Lu and C-X. Xie, "Fuzzy Obstacle-avoiding Controller of Autonomous Mobile Robot Optimized by Genetic Algorithm under Multi-obstacles Environment", in World Congress on Intelligent Control and Automation, 2006, pp. 3255-3259.

[12] C. Rekik, M. Jallouli, and N. Derbel, "Optimal Trajectory of a Mobile Robot by a Genetic Design Fuzzy Logic Controller", in International Conference on Advances in Computational Tools for Engineering Applications, 2009, pp. 107-111.

[13] J. H. Holland, Adaptation in Natural and Artificial Systems, Ann Arbor: University of Michigan Press, 1975.

[14] D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, New York: Addison-Wesley Professional, 1989.

Karim Benbouabdallah received his first university cycle diploma from National Preparatory School for Engineering studies, Rouiba, Algeria in 2005 and state engineering diploma in Systems Control from Military Polytechnic School, Bordj El Bahri, Algeria in 2008. He then joined Harbin Engineering University, China as research staff working toward a Ph.D degree. His research interests include Vision in robotics, mobile robots navigation and intelligent control techniques.

Zhu Qi-dan is a professor at College of Automation, Harbin Engineering University, China. He received his M.Sc degree from Harbin Shipbuilding Engineering Institute in 1987 and his Ph.D degree in 2001 from Harbin Engineering University. He has supervised several master and Ph.D students. His research interests are mainly focused in the area of robotics, machine vision, omnidirectional vision and engineering control systems.

Further Research on Registration System with Vandermonde Matrix

Ning Huang[1], Xian-tong Huang[2], Jing-li Ren[3], Xian-wen He[2], Yang Liu[2]
[1]Center of Modern Educational Technology, Gannan Normal University
Ganzhou, 341000, China
hngzjx@qq.com
[2]College of Mathematics and Computer Science, Gannan Normal University
Ganzhou, 341000, China
[3]College of Communication and Media, Gannan Normal University
Ganzhou, 341000, China

Abstract

We propose an improved software registration system from our previous research. Our improvements are mainly as follows. (1) Changing basic field to make the scheme suitable for all characters. (2) Changing encryption and decryption formulae to make the scheme more complex. (3) Using the technique of letter decomposition and composition to make the scheme more deceptive to a possible adversary. (4) Using mobile phone in the system to enhance the security. Experimental results and analysis show that the improvements are successful and the scheme is viable and secure.

Keywords: *software, registration, Vandermonde matrix*

1. Introduction

Copy protection for computer software started a long cat-and-mouse struggle between publishers and crackers. These were (and are) programmers who would defeat copy protection on software as a hobby, add their alias to the title screen, and then distribute the ”cracked” product to the network BBSes or Internet sites that specialized in distributing unauthorized copies of software [1]. Research on the topic never ends. In [2], we proposed a scheme of registration with Vandermonde matrix in a Galois field $GF(p)$, which is an application of Hill cipher [8]. In this paper, we further discuss the improvements of the method. Our improvements are mainly as follows. (1) Using the Galois field $GF(2^m)$ instead of $GF(p)$ to make the scheme suitable for all characters. (2) Using matrix equation $Y = V^{-1}X+C$ instead of $Y = V^{-1}X$ to make the scheme more complex. (3) Using the technique of letter decomposition and composition to make the scheme more deceptive to a possible adversary. (4) Using mobile phone in the system to enhance the security. The rest of the paper is organized as follows. In Section 2, we briefly introduce our original research in $GF(p)$. In Section 3, we propose some novel ideas to improve our original scheme. In Section 4, we design the registration system. In Section 5, we give experimental results and analysis. We conclude the paper in Section 6.

2. Original Scheme

Agree on permission control character string such as $PROFESSIONALVERSION$ on both sides of the vendor and user. Then we take n different characters $\mu_1, \mu_2, , \cdots, \mu_n$, from hard id [7] to create a Vandermonde matrix in a Galois field $GF(p)$, where p is a prime. That is $V = V(\mu_1, \mu_2, \cdots, \mu_n)$

$$= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \mu_1 & \mu_2 & \cdots & \mu_n \\ \vdots & \vdots & \ddots & \vdots \\ \mu_1^{n-1} & \mu_2^{n-1} & \cdots & \mu_n^{n-1} \end{pmatrix} mod\ p \qquad (1)$$

Then we obtain a determinant formula

$$det(V) = \prod_{1\le i<j\le n} (p + \mu_i - \mu_j)\ mod\ p \qquad (2)$$

It follows from $\mu_i s$ are different from each other that V is invertible. The fast computation of the inverse of V is

$$A = V^{-1}(\mu_1, \mu_2, \cdots, \mu_n) = HL \ mod \ p \quad (3)$$

where H is an upper triangular matrix and L a lower triangular one. The elements of each can be obtained from recursive formulae. Let

$$X = \begin{pmatrix} x_1, & x_2, & \ldots & , x_n \end{pmatrix}^T$$

be the plain text, define a linear mapping:

$$X \mapsto Y = \begin{pmatrix} y_1, & y_2, & \ldots & y_n \end{pmatrix}^T = VX \ mod \ p \quad (4)$$

as an encryption algorithm, which is equivalent to the following linear functions:

$$\begin{cases} y_1 = x_1 + x_2 + \cdots + x_n \\ y_2 = \mu_1 x_1 + \mu_2 x_2 + \cdots + \mu_n x_n \quad mod \ p \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_n = \mu_1^{n-1} x_1 + \mu_2^{n-1} x_2 + \cdots + \mu_n^{n-1} x_n \end{cases} \quad (5)$$

The decryption algorithm is

$$X = V^{-1} Y = AY \ mod \ p \quad (6)$$

which is equivalent to the following linear functions:

$$\begin{cases} x_1 = a_{11} y_1 + a_{12} y_2 + \cdots + a_{1n} y_n \\ x_2 = a_{21} y_1 + a_{22} y_2 + \cdots + a_{2n} y_n \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_n = a_{n1} y_1 + a_{n2} y_2 + \cdots + a_{nn} y_n \end{cases} mod \ p \quad (7)$$

In order to minimize the set of necessary formulae on the user's side, we use $A = V^{-1}$ on the vendor's side as the encryption key while V on the user's side as the decryption key. On the vendor's side, we take the permission control string as the plain text X. We compute $Y = V^{-1} X \ mod \ p$ as the cipher text. On the user's side, we use $X = VY \ mod \ p$ to verify the registration. Our original scheme takes $p = 37$ as the modulus, choose 10 numbers and 26 capital letters and the symbol '\$'. Lower case letters are converted to upper ones. Other symbols are ignored. The key space is $37 \cdot 36 \cdots (37 - n + 1)$ for a given n.

3. Improved Scheme

3.1 Basic field of the scheme

To make the scheme universal, we define our computation in the field $GF(2^m)$. In fact, $GF(2^m)$ is congruent to $Z_2[x]/f(x)$, where $Z_2[x]$ is the polynomial ring over Z_2 , $f(x)$ is a primitive polynomial of $Z_2[x]$. In $Z_2[x]/f(x)$, the addition of polynomials $\alpha(x)$ and $\beta(x)$ is defined by $\sigma(x) = \alpha(x) + \beta(x)$.With the character of Z_2 being 2, for arbitrary $\alpha(x) \in Z_2[x]$, we have $\alpha(x) + \alpha(x) = 0$.The multiplication of polynomials and is defined by $\pi(x) = \alpha(x) + \beta(x) \quad mod \ f(x)$.See more details in [4, 5, 6]. If we take $m = 8$,$GF(2^8)$ is just the letter set of extended $ASCII$. This means we can use every symbol in a plain text. However, that will cause the problem of unprintable letters in the registration string. We use the decomposition of letters to solve this problem by splitting a letter into two, each is in the range of $\{0, 1, 2, \cdots, 9, A, B, C, D, E, F\}$. Furthermore, we plus each value with $'A'$ in the generation of a registration string. For example, if a letter with the value 0 is used for a registration string, we really see it as $'A'$. Similarly, we see $'B'$ for value 1, $'C'$ for value 2 through $'P'$ for value F. This can easily be realized in C language. Suppose c is in the cipher text, the decomposition and conversion is expressed by

$$c_1 = c/0x10 +' A';$$
$$c_2 = c\%0x10 +' A';$$

This method also increases the complexity of the encryption. Bravo! It actually butters both sides of our bread.

3.2 Fast computation of $V^{-1} = HL$

We develop the method in [3] to use it in $GF(2^m)$. Let L be the matrix whose rows are associated with the coefficients of the polynomials

$$\begin{cases} \psi_1(s) = 1 \\ \\ \psi_i = (s + \mu_j)\psi_{i-1}(s) \end{cases} \quad (8)$$

L can be denoted by
$L(1, s, \cdots, s^{n-1})^T = (\psi_1(s), \psi_2(s), \cdots, \psi_n(s))^T$

$$L = \begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1n} \\ l_{21} & l_{22} & \cdots & l_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{pmatrix} \quad (9)$$

$$\text{Let} \quad l_{ii} = 1, l_{ij} = 0 (i \neq j). \quad (10)$$

$$\begin{cases} l_{i+1,1} = (\mu_i) \cdot l_{i,j-1} \\ l_{i+1,j} = l_{i,j-1} \\ j = 2, 3, \cdots, i, i = 1, 2, \cdots, n \end{cases} \quad (11)$$

Let the initial vector $h_n = (c_1, c_2, \cdots, c_n)^T$ be determined from the partial fraction expansion

$$\frac{1}{(s+\mu_1)\cdots(s+\mu_n)} = \frac{c_1}{s+\mu_1} + \cdots + \frac{c_n}{s+\mu_n}.$$

Denote H by

$$H = (h_1, h_2, \cdots, h_n)^T \tag{12}$$

$$h_{ij} = \begin{cases} \frac{1}{\psi'_{j+1}(\mu_i)} & if\ i \le j \\ 0 \quad\quad\quad , & if\ i > j \end{cases} \tag{13}$$

Denote *d(s)* by

$$d(s) = (\mu_1 + s, \mu_2 + s, \cdots, \mu_n + s)^T \tag{14}$$

Let

$$h_{i-1} = h_i \otimes d(\mu_i), i = n, n-1, \cdots, 2 \tag{15}$$

ending at $h_1 = (1, 0, \cdots, 0)^T$. The right side of (14) is the inner product of h_i and $d(\mu_i)$, i.e., if $u = (u_1, u_2, \cdots, u_n)^T$,$v = (v_1, v_2, \cdots, v_n)^T$ then

$$h_i \otimes d(\mu_i) = (u_1v_1, u_2v_2, \cdots, u_nv_n)^T \tag{16}$$

It follows that the formulae to compute the elements in the upper triangular matrix H are given by

$$h_{ij} = \begin{cases} \frac{1}{\psi'_{j+1}(\mu_i)}, & if\ i \le j \\ 0 \quad\quad\quad , & if\ i > j \end{cases} \tag{17}$$

where

$$\psi'_{j+1}(\mu_i) = \prod_{k=1, k\neq i}^{j-1} (\mu_i + \mu_k)\ . \tag{18}$$

3.3 Improvements of the linear mappings

In the field $GF(2^8)$, Let

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \tag{19}$$

be the plain text, and

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nm} \end{pmatrix} \tag{20}$$

be the same size as X.Then

$$Y = V^{-1}X + C$$

is the matrix containing cipher text.

This is equivalent to m groups of linear mappings as follows.

$$\begin{cases} y_{11} = a_{11}x_{11} + \cdots + a_{1n}x_{n1} + c_{11} \\ y_{21} = a_{21}x_{11} + \cdots + a_{2n}x_{n1} + c_{21} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_{n1} = a_{n1}x_{11} + \cdots + a_{2n}x_{n1} + c_{n1} \end{cases} \tag{21}$$

$$\begin{cases} y_{12} = a_{11}x_{12} + \cdots + a_{1n}x_{n2} + c_{12} \\ y_{22} = a_{21}x_{12} + \cdots + a_{2n}x_{n2} + c_{22} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_{n2} = a_{n1}x_{12} + \cdots + a_{2n}x_{n2} + c_{n2} \end{cases} \tag{22}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\begin{cases} y_{1m} = a_{11}x_{1m} + \cdots + a_{1n}x_{nm} + c_{1m} \\ y_{2m} = a_{21}x_{1m} + \cdots + a_{2n}x_{nm} + c_{2m} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_{nm} = a_{n1}x_{1m} + \cdots + a_{2n}x_{nm} + c_{nm} \end{cases} \tag{23}$$

In the field of character 2, the decryption is expressed as follows.

$$X = V(Y - C) = V(Y + C).$$

This is equivalent to m groups of linear mappings as follows.

$$\begin{cases} x_{11} = y^*_{11} + \cdots + y^*_{n1} \\ x_{21} = \mu_1 y^*_{11} + \cdots + \mu_n y^*_{n1} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_{n1} = \mu_1^{n-1} y^*_{11} + \cdots + \mu_n^{n-1} y^*_{n1} \end{cases} \tag{24}$$

$$\begin{cases} x_{12} = y^*_{12} + \cdots + y^*_{n2} \\ x_{22} = \mu_1 y^*_{12} + \cdots + \mu_n y^*_{n2} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_{n2} = \mu_1^{n-1} y^*_{12} + \cdots + \mu_n^{n-1} y^*_{n2} \end{cases} \tag{25}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\begin{cases} x_{1m} = y^*_{1m} + \cdots + y^*_{nm} \\ x_{2m} = \mu_1 y^*_{1m} + \cdots + \mu_n y^*_{nm} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_{nm} = \mu_1^{n-1} y^*_{1m} + \cdots + \mu_n^{n-1} y^*_{nm} \end{cases} \tag{26}$$

where $y^*_{ij} = y_{ij} - c_{ij} = y_{ij} + c_{ij}$.

This cipher is obviously more complex compared with the original one. The security is enhanced.

3.4 Use of mobile phone as a receiver

In our original scheme, we assumed that the user got the registration string vie web service or email. Now we propose the registration string can also be received as a message via a mobile phone.

Step1 The user pays for the software and submits personal information, e.g. hard id, name, mobile phone number to the server via web;

Step2 The server checks the submission;

Step3 The sever creates a registration string;

Step4 The server sends the registration string to the user's mobile phone via wireless tunnel;

Step5 The user reads the message and register the software. The flow of the process is shown by Fig. Registration steps.

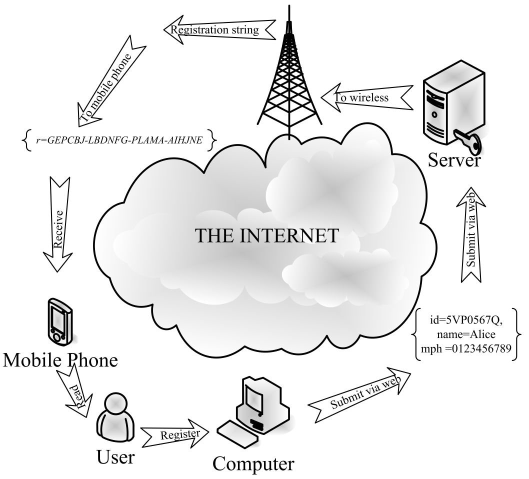


Fig. Registration steps

4. Registration scheme

Set preliminary conditions on both sides of the vendor and user:
A character string as a permission control string p_1 ; dimension n.

4.1 The Creation of registration string

The user submits the message as follows. Hard id(id for short), name, mobile phone number(mph for short).
On the vendor's side, the server computes as follows:
Input: $id, name, mph$;
Output: Registration string like
$r = xxxxxx - xxxxxx - \cdots - xxxxxx$

Algorithm:

Step1 Selects different elements from id, forms a new id_1;

Step2 Creates matrices of lower triangular L and upper triangular H from id_1 as shown in subsection 3.2;

Step3 Computes the inverse of Vandermonde matrix

$$V^{-1} = HL;$$

Step4 Appends necessary dots to the permission control string p_1 to fit the dimension, puts the result in p_2 ;

Step5 Puts p_2 into matrix X and creates a matrix C from $name$ which matches dimensions in step4, using the elements of $name$ circularly.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nm} \end{pmatrix}$$

Step6 Computes

$$Y = V^{-1}X + C;$$

Step7 Decomposes Y and adds $'A'$ to each element respectively to update Y;

Step8 Takes elements from Y to create a registration string like

$$r = xxxxxx - xxxxxx - \cdots - xxxxxx.$$

Then the server sends the registration string above to the user's mobile phone via wireless tunnel.

4.2 The use of registration string

Reading the registration string from mobile phone, the user keys in to register the software. **Input:** Registration string like

$$r = xxxxxx - xxxxxx - \cdots - xxxxxx.$$

Output:

$$TRUE/FALSE$$

Algorithm:

Step1 Takes elements from r to create a matrix Y

Step2 Minuses $'A'$ from each element of Y to update Y;

Step3 In the same way as step6 in last subsection, creates matrix C ;

Step4 In the field of character 2, computes
$Y - C = Y + C$;

Step5 Computes Vandermonde matrix V from local hardware id , id_1;

Step6 Computes
$X = V(Y - C) = V(Y + C)$
to get a possible plain text;

Step7 Obtains p_2 from X;

Step8 Removes redundant dots if there are any at the end to get p_1 ;

Step9 Compares p_1 with the preliminary one;

Step10 The result is $TRUE$ or $FALSE$. If the verification is successful, the registration will be approved, otherwise it will be defied. A registered software picks up the registration string automatically and verifies it according to the above routine to decide which functions the software can use.

5. Experimental results and analysis

Embeds in the software on both sides of the vendor and user:
$p_1 = VIPversion$
as a permission control string; Agree on dimension $n = 6$.

5.1 The Creation of registration string

The user submits the messages as follows.$id = 5VP0567Q, name = Alice, mph = 0123456789$.
On the vendor's side, the server computes as follows.

Input:
$id = 5VP0567Q, name = Alice$;

Output:
$r = GEPCBJ - LBDNFG - PIMAMA - AIHJNE$;

Algorithm:

Step1 Selects different elements from id , forms a new id_1 :
$id_1 = 5VP067Q$;

Step2 Creates lower triangular and upper triangular matrices:

$$L = \begin{pmatrix} 01 & 00 & 00 & 00 & 00 & 00 \\ 35 & 01 & 00 & 00 & 00 & 00 \\ 27 & 63 & 01 & 00 & 00 & 00 \\ BE & 22 & 33 & 01 & 00 & 00 \\ 19 & 3A & 2B & 03 & 01 & 00 \\ B0 & CC & DF & 27 & 35 & 01 \end{pmatrix}$$

and

$$H = \begin{pmatrix} 01 & 18 & 73 & 99 & 92 & 37 \\ 00 & 01 & 19 & BE & A2 & 8B \\ 00 & 00 & 18 & 95 & B3 & 39 \\ 00 & 00 & 00 & 73 & 88 & 17 \\ 00 & 00 & 00 & 00 & 99 & A5 \\ 00 & 00 & 00 & 00 & 00 & 92 \end{pmatrix}$$

Step3 Computes$V^{-1} = HL$

$$= \begin{pmatrix} 07 & A3 & C0 & A2 & 83 & 85 \\ A5 & 10 & BA & F5 & D3 & 06 \\ B1 & AA & F0 & D7 & FE & 73 \\ CD & F5 & 86 & 71 & 2D & 64 \\ 92 & 53 & 50 & 8A & F8 & 39 \\ 4D & BF & 5C & 7B & 7B & AD \end{pmatrix}$$

Step4 Appends 2 dots to the permission control string to fit the dimension,
$p_2 = VIPversion..$;

Step5 Puts $p_2, name$ into matrices:

$$X = \begin{pmatrix} V & s \\ I & i \\ P & o \\ v & n \\ e & . \\ r & . \end{pmatrix} \quad or \quad \begin{pmatrix} 56 & 73 \\ 49 & 69 \\ 50 & 6F \\ 76 & 6E \\ 65 & 2E \\ 72 & 2E \end{pmatrix}$$

then

$$C = \begin{pmatrix} A & l \\ l & i \\ i & c \\ c & e \\ e & A \\ A & l \end{pmatrix} \quad or \quad \begin{pmatrix} 41 & 6C \\ 6C & 69 \\ 69 & 63 \\ 63 & 65 \\ 65 & 41 \\ 41 & 6C \end{pmatrix}$$

which matche dimension in step4;

Step6 Computes

$$Y = \begin{pmatrix} 64 & F8 \\ F2 & C0 \\ 19 & C0 \\ B1 & 08 \\ 3D & 79 \\ 56 & D4 \end{pmatrix};$$

Step7 Decomposes Y and adds $'A'$ to each element respectively to get

$$Y=\begin{pmatrix} G & L & P & A \\ E & B & I & I \\ P & D & M & H \\ C & N & A & J \\ B & F & M & N \\ J & G & A & E \end{pmatrix};$$

Step8 Takes elements from Y to create a registration string: $r = GEPCBJ - LBDNFG - PIMAMA - AIHJNE$; Then the server sends the registration string to the user's mobile phone via wireless tunnel.

5.2 The use of registration string

Reading the registration string from mobile phone, the user keys in to register the software.

Input:
$r = GEPCBJ - LBDNFG - PIMAMA - AIHJNE$

Output:
$TRUE/FALSE$

Algorithm:

Step1 Takes elements from
$r = GEPCBJ - LBDNFG - PIMAMA - AIHJNE$
to get a matrix

$$Y=\begin{pmatrix} G & L & P & A \\ E & B & I & I \\ P & D & M & H \\ C & N & A & J \\ B & F & M & N \\ J & G & A & E \end{pmatrix};$$

Step2 Minuses $'A'$ from each element of Y to get a new matrix

$$Y=\begin{pmatrix} 64 & F8 \\ F2 & C0 \\ 19 & C0 \\ B1 & 08 \\ 3D & 79 \\ 56 & D4 \end{pmatrix};$$

Step3 In the same way as step5 in last subsection, creates

$$C=\begin{pmatrix} A & l \\ l & i \\ i & c \\ c & e \\ e & A \\ A & l \end{pmatrix} \quad or \quad \begin{pmatrix} 41 & 6C \\ 6C & 69 \\ 69 & 63 \\ 63 & 65 \\ 65 & 41 \\ 41 & 6C \end{pmatrix};$$

Step4 In the field of character 2, computes

$$Y-C=Y+C=\begin{pmatrix} 25 & 94 \\ 9E & A9 \\ 70 & A3 \\ D2 & 6D \\ 58 & 38 \\ 17 & B8 \end{pmatrix}$$

Step5 Computes Vandermonde matrix

$$V=\begin{pmatrix} 01 & 01 & 01 & 01 & 01 & 01 \\ 35 & 56 & 50 & 30 & 36 & 37 \\ 96 & D9 & CD & 87 & 93 & 92 \\ D0 & 5F & 33 & 05 & AF & 0B \\ DA & 69 & 52 & F0 & CB & CA \\ 33 & B4 & 6D & CD & 5D & A1 \end{pmatrix}$$

from local hardware
$id = 5VP0567Q, id1 = 5VP067Q;$

Step6 Computes a possible plain text
$X = V(Y - C) = V(Y + C)$

$$=\begin{pmatrix} 56 & 73 \\ 49 & 69 \\ 50 & 6F \\ 76 & 6E \\ 65 & 2E \\ 72 & 2E \end{pmatrix} \quad or \quad \begin{pmatrix} V & s \\ I & i \\ P & o \\ v & n \\ e & . \\ r & . \end{pmatrix};$$

Step7 Obtains
$p_2 = VIPversion..;$
from X;

Step8 Removes 2 redundant dots at the end to get $p_1 = VIPversion$

Step9 Compares p_1 with the preliminary one;

Step10 The result is $TRUE$.

5.3 Analysis of the results

In our novel scheme, the computations are performed in $GF(2^8)$ to make the scheme suitable for the whole extended $ASCII$ table. Meanwhile, the key space expands from $37 \cdot 36 \cdots (37 - n + 1)$ for a given $256 \cdot 255 \cdots (256 - n + 1)$. The uses of mobile phone, user's name, the decomposition and composition of letters also enhance the cipher. Experimental results show the success of the scheme.

6. Conclusions

It follows from Experimental results and analysis that the novel scheme is enhanced. We get better results in our further research.

Acknowledgements

1)The authors would like to express their thanks to the hard work of the editors and reviewers of this paper.
2)This work was supported by the fund from Natural Science of Jiangxi Province of China under Grant No.20114BAB201033. The authors would like to express their thanks to the Committee of the fund.

References

[1] Copy protection for computer software , http://en.wikipedia.org/wiki/Copy _ protection;2012.

[2] N. Huang, ”Permission control of software based on registration system with Vandermonde matrix in a Galois field ”,In Instrumentation & Measurement, Sensor Network and Automation (IMSNA), 2012 International Symposium on, 2012, Vol. 2, pp.487-490.

[3] S.H. Hou, and E. Hou, ”Triangular Factors of the Inverse of Vandermonde Matrices,” In Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 Vol 2, IMECS 2008, pp.19-21.

[4] Darrel Hankerson, ”Alfred Menezes and Scott Vanstone.Guide to Elliptic Curve Cryptography”, Berlin:Springer,2003.

[5] Roberto M. Avanzi, Henri Cohen, Christophe Doche,Gerhard Frey, Tanja Lange, Kim Nguyen, and Frederik Vercauteren.,”Handbook of elliptic and hyperelliptic curve cryptography”, London:Taylor & Francis Group,2006.

[6] William J.Gilbert and W.Keith Nicholson, ”Modern Alegbra with Applications”,Second Edition.New Jersy:John Wiley &Sons,Inc,2003.

[7] S.D.S. Monteiro and R.F. Erbacher, ”Exemplifying Attack Identification and Analysis in a Novel Forensically Viable Syslog Model”,In Washington:IEEE Computer Society,Proceedings of the Third International Workshop on Systematic Approaches to Digital Forensic Engineering,2008,pp.57-68.

[8] L. S. Hill, ”Cryptography in an Algebraic Alphabet,” The American Mathematical Monthly, Vol. 36, No. 6. (Jun. - Jul., 1929), pp. 306-312.

Ning Huang, born in 1958, received Master’s degree in applied mathematics and computer science from Jiangxi University, China in 1991, awarded senior engineer of the Industrial and Commercial Bank of China in 2001. He is now with Center of Modern Educational Technology, Gannan Normal University, Ganzhou, China,as an associate professor. His research interests include information security and digital campus.

Xian-tong Huang, born in 1966, received Doctor’s degree in computational mathematics from Hunan University, China in 2006. He is now with College of Mathematics and Computer Science, Gannan Normal University, Ganzhou, China as a professor. His research interests include computational mathematics and information security.

Jing-li Ren, born in 1980, received Master’s degree in computer science from Wuhan University of Technology, China in 2006. She is now with College of Communication and Media, Gannan Normal University, Ganzhou, China, as a lecturer. Her research interests include Intelligent computation and simulation.

Xian-wen He, born in 1974, received Master’s degree in computer science from Nanchang University, China in 2007. He is now with College of Mathematics and Computer Science, Gannan Normal University, Ganzhou,China, as an associate professor. His research interests include network security and image recognition.

Yang Liu, born in 1972, received Master’s degree in computer science from Nanchang University, China in 2007. He is now with College of Mathematics and Computer Science, Gannan Normal University, Ganzhou,China, as a lecturer. His research interests include algorithm optimization and image recognition.

Head and Neck Cancer Treatment with Particle Beam Therapy

Mehrzad Zargarzadeh

Faculty of Engineering, Islamic Azad University Science and Research Branch
Tehran, Iran

Abstract

In this century, cancer incidence has become one of the most significant problems concerning human. Conventional radiotherapy damage healthy tissue and in some cases may cause new primary cancers. This problem can be partially solved by hadron therapy which would be more effective and less harmful compared to other forms of radiotherapies used to treat some cancers. Although carbon ion and proton therapy both are effective treatments, they have serious differences which are mentioned in this paper and compared between the two methods. Furthermore, various treatments have been performed on head and neck cancer with hadrons so far will be discussed.

Keywords: *cancer; proton therapy; carbon ion therapy; Boron neutron capture therapy (BNCT); head; neck*

1. Introduction

Cancer is the major cause of mortality in economically developed countries claiming about 350,000 lives annually, and is a leading factor of death in developing countries. Head and neck cancer is ranked sixth and considered as the most prevalent cancers throughout the world. Global statistics show the fact that there are about 640,000 cases of head and neck cancer per year, resulting in nearly 350,000 deaths per year [1]. Treating with conventional radiation therapy often harms healthy tissue and organs near the tumor site. Proton therapy uses a highly precise beam in order to target radiation directly at the tumor site, minimizing damage to nearby healthy tissue and organs and greatly reducing the risk of both acute and long-term side effects. Cancers of the oral cavity and pharynx are the most prevalent type of head and neck cancer with approximately 480,000 cases per year [1]. Therapy by radiation, with surgery or chemotherapy can produce lasting locoregional disease control in a high percentage of patients with head and neck cancer. Radiotherapy is usually delivered using high-energy X-rays produced by linear accelerators. Theoretically, it would be possible to treat and cure most patients suffering head and neck cancer by using a high dose of radiation. However, in many treatment situations, the dose is limited and by the presence of adjacent radiosensitive normal tissues. The limitations of X-ray radiotherapy can be substantially overcome by using hadrons (i.e., protons and light ions such as helium, carbon, oxygen and neon, in particular carbon ions). Proton and ion therapy both started in 1954 with patient treatments at the 184-inch cyclotron at LBL Berkeley, first with protons and in later stages with helium beams. Until the closing of the accelerators in the year of 1992 more than 2000 patients went through the treatment and therapy process at Berkeley center. Proton therapy started at Uppsala/Sweden in 1957 while the same treatments activities started at Harvard cyclotron in the year of 1961 (more or less 9000 patients were treated and mostly cured until the year of 2002). Studies indicate that hadrons therapy can be replaced with conventional photon therapy specifically for tumors with low radio sensitivity and critical location. In this research proton, carbon ion therapy and BNCT will be described. In addition the number of treatments and clinical experiments conducted in head and neck cancer will be discussed and compared.

2. STATUS OF PROTON THERAPY

In Proton therapy a highly precise beam is used to target radiation directly at the tumor locating point, this act minimizes the possible damages to close healthy tissues and the rest organs and substantially reduces the risk of both severe and long-term side and subsequent effects. Therapy with traditional radiation treatment method often damages the healthy tissues and organs near the tumor site. Proton and ion therapy started in 1954 with patient treatments at the 184-inch cyclotron at LBL Berkeley, first with protons and in later stages with helium beams. Until the closing of the accelerators in 1992 more than 2000 patients were treated at Berkeley. This precision makes proton therapy useful specially for treating brain tumors, head and neck cancers, and tumors located near the spinal cord, heart or lungs. Since the energy emission of the proton beam is confined to the narrow Bragg peak, collateral damage to the surrounding tissues should be reduced, while an increased dose of radiation can be delivered to the tumor. Currently, there exist 10 hospital-based proton therapy centers working around the world and 15 others are being constructed or are in final completion phases. The PSI or Paul Scherer Institute is the only center in the world has the experience of treatment with intensity modulated proton therapy. Paul Scherer Institute, is the first proton therapy center, has the world's only gantry so far using so-called spot-scanning-technology. For most disease therapy sites and treatment centers, proton therapy treatments typically take about quarter to half an hour each day and are delivered five days a week for nearly four to seven weeks. The course of treatment and time duration per treatment each day differs based on each patient's individual case.

3. Carbon Ion Therapy

Since 1990s, the researchers have treated about 5,000 patients using carbon ions in Japan and about 440 patients in Germany. Heavy Ion Medical Accelerator in Chiba (HIMAC) was the central therapy site of the world's first carbon-ion treatment in 1994 and the facility has now treated approximately 3795 patients [2]. In the body, the 12C carbon isotope is able to exchange a nucleon in an interaction to convert it to 11C, after that the 11C decay starts via ß decay, giving off a positron which annihilates, emitting a pair of photons. Carbon ions can also be employed before surgery to shrink a tumor or immediately after surgery because, unlike x-rays or protons, they don't damage the skin. Furthermore, this method might offer a useful tool for assessing unpredictable deviations between planned and actual treatment. They can produce extreme damage to tumor cells by depositing their maximum energy in the Bragg peak. Another advantage in the use of carbon ions is that they can be formed as narrow focused and scanning pencil beams. Therefore, any parts of the tumor will be irradiated. Targeting critical areas such as back bone or spinal cord or optical nerves can be monitored with on-line positron emission tomography (PET) [3-5].

4. Biological Factors Related To Radiation

Linear Energy Transfer (LET) is a method used to express and describe beam quality. It is the rate of energy deposited or lost per distance travelled. Hadrons may have the property of low or high Linear Energy transfer. High LET radiation creates various and multi-dimensional damages and harms in DNA and other cellular structures, yielding tumor killing with fewer side effects on normal tissue [4,6]. Since the biological effect is not predicted by absorbed dose, a coefficient of relative biological effectiveness (RBE) is introduced to take in to account the dissimilarity in the effect of radiations of various types for the same physical dose. The effectiveness is defined as the ratio between the absorbed dose of a reference radiation and that of the test radiation required to produce the same biological effect. RBE depends on radiation quality of linear energy transfer, radiation dose, number of fractions, dose rate and the like [5]. Many chemicals can change the response of cells to radiation. The one chemical which has a very big effect and has possible importance in Radiation Therapy is Oxygen. It acts at the level of free radicals which are formed when radiation interacts with water molecules in the cell. Repairing the damage caused by the radiation can take place in the absence of Oxygen. The oxygen enhancement ratio or 'OER' is the ratio of the doses of radiation necessary to present or yield similar and identical biological impacts and effects in the nonexistence of oxygen and in its presence. The oxygen effect is vast and important for LOW LET radiation (x and gamma) [7].

5. Proton and Photon in Contrast With Carbon

Although Protons and carbon ions deeply penetrate tissues and emit most of their energy near the end of their range where the tumor is existed, they have some differences given as follows:

- Carbon ions disperse or scatter much less than protons and concentrate their radiation in a smaller area [8], [9].
- However protons and carbon ions both have sharp Bragg peaks as shown in fig.1, protons are characterized by low LET whereas carbon ions are characterized by high LET [6].
- A significant benefit of carbon ions is that, unlike x-rays and protons, they do not need oxygen to work and can, therefore, reach and kill or terminate hypoxic areas of tumors, which are notoriously hard to treat [9].
- Bone and soft tissues tumors can be treated and cured by carbon, but not even by protons and certainly not with x-rays [10].
- The OER is more than 3/2 times better than that of protons. Another distinction is that carbon ions bring about more irreparable harms and damages to the cancer cells. Protons and x-rays have about the same relative biological impact, which is an assessment of the damage from ionizing radiation while the RBE of carbon ions is three times higher; this means that they damage DNA in a way that is double-stranded and irreparable.
- Today, 24 proton facilities are working worldwide, and almost 20 more such sites are planned to be constructed, while there are only 3 carbon facilities currently treating patients of mentioned disease. One is situated in Germany and two others are working in Japan [11]

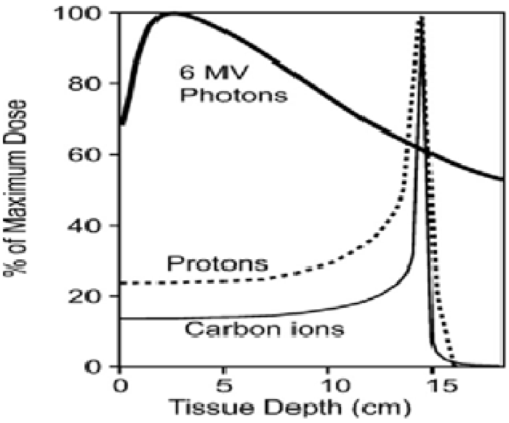


Figure.1. Comparison between Carbon and Proton Bragg peak

6. Boron Neutron Capture Therapy (BNCT)

BNCT is a combined treatment method including thermal and epithermal neutron beams. BNCT derived from ^{10}B nucleus tendency to capture thermal neutrons. As a result unstable 11B nucleus generates a lithium ion and an α particle. The yields of this reaction have high LET characteristics. Thus, it is able to selectively irradiate tumor cells which have received appropriate amount of ^{10}B. The initial idea to use neutron capture reactions in cancer treatment was broadcasted by Locher in 1936. In May 1999 the first patient was treated with BNCT at the Finnish Research Reactor. By this way cancer cells will be killed selectively as well as treating tumors by a cell-by-cell basis. BNCT would be effective in treating malignant melanoma (skin cancer), malignant brain tumor, head and neck cancer, lung cancer, liver cancer. In order to estimate the number of monitor MU which should be delivered to the patient in BNCT treatment, the on-line beam monitor system requires to be calibrated. Each patient was planned to receive two BNCT treatments in 3-5 weeks apart. Furthermore, CT (for constructing 3-dimensional model), MRI and PET (for specifying target volume) images can be assumed with each other.

7. Intensity-Modulated Radiation Therapy (IMRT)

IMRT is a high accuracy radiotherapy technique with the ability of releasing precise doses to virulent tumor or characteristic areas within the tumor. IMRT can be assumed for cancers in the nasopharynx, sinonasal region, parotid gland, tonsil, buckle mucosa, gingiva, and thyroid [12]. Due to its potency to spare healthy tissues, this method may be beneficial for re-treatment of formerly irradiated with head and neck cancers. In traditional radiotherapy, the doses given to the healthy tissues are the real limiting and restricting factor. IMRT increases the doses delivered to the healthy tissues and then dispenses it over a large mass in order to provide an enhanced dose for the tumor cells. Owing to some reasons IMRT would be appropriate for children. Child's body is highly sensitive to radiation. Moreover, scattered radiation created by radiotherapy is very serious in children [13]. In CT scanning and IMRT both rotating beam is used. Though, in IMRT beams delivers radiation. Conventional three dimensional conformal radiotherapy, modificated three-dimensional conformal radiotherapy and IMRT are compared in figure.2 [14]. Furthermore, side effects of conventional radiation therapy and IMRT are approximately the same. However, the proton therapy would cost about 2.4 times more than IMRT [15].

8. Clinical Experience in Head and Neck Cancer

Chondrosarcomas and Chordomas are uncommon tumors which respectively arise from notochordal remnants and primitive mesenchymal cells. Skull base Chordomas and chondrosarcomas are close to dose-limiting structures such as optic pathways, brainstem, and spinal cord. .A series of 621 cases of chordoma and chondrosarcoma of the base of skull treated at the Massachusetts General Hospital in Boston to a total dose ranging from 66 to 83 GyE indicated local control at 10 years of 54% and 94%, respectively [16]. In all cases, surgery was performed before radiotherapy to eliminate the tumor. Salivary gland tumors can also treated by particle beam. Because of the low radiosensitivity of these tumors, conventional radiotherapy is not effective for them. Three- dimensional conformal radiation therapy (3D-CRT) and IMRT can be used for these cancers [17]. Another approach is to use neutron beam radiation. In this way, high-energy neutrons assumed in place of using x-ray beam. Furthermore, in 18 patients with salivary gland carcinoma, survival rate of 59% were observed. Although therapies with carbon ions and neutrons may give the same results, treating with carbon ion has lower toxicity. Treating skin carcinomas with conventional radiotherapy is limited. 45 patients suffered from skin carcinomas were treated and cured with carbon ions RT from 2006 to 2009 which caused 1- and 3-year overall survival rates for 45 patients between 88.9% and 86%, respectively [18]. Five randomized studies of particle beam therapy in malignant glioma were compared. None of these trials detected a significant survival benefit for particle therapy. This study is divided into two types of therapies as the table.I represent, with neutrons (first four studies) and with photons or Pions (fifth study) [19]. Since 2001, 26 patients with salivary gland carcinomas, sarcomas, squamous cell carcinomas were treated with BNCT. All patients survived 1 up to 72 months after the treatments. The mean survival times were 13.6 months. Entirely, BNCT has the potential to be used for the reappeared tumors [20]. Another example of BNCT treatment effect is an old man (36 year old) with glioblastoma multeforme (GBM) which is shown in Figure 3[21]. Several paranormal head and neck cancers, such as laryngeal sarcomas, bone and soft tissue sarcomas and glomus tumors were treated with combined proton and photon radiotherapy. Moreover, hadron therapy has been applied for the treatment of numerable cases of carcinoma of pituitary, thyroid gland and ear.

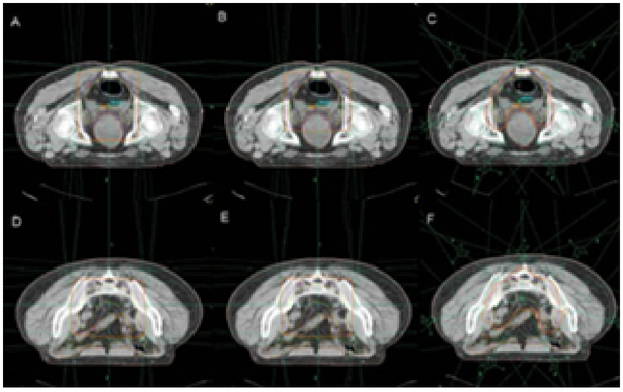


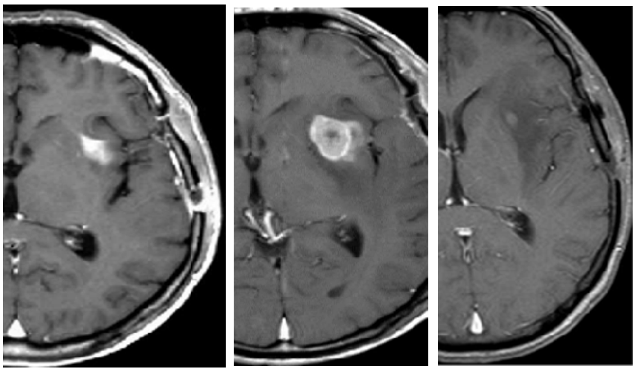
Figure 2. Three method planning in patient with a uT3N+ medial rectal cancer, conventional tridimensional conformal radiotherapy (A and D), modificated tridimensional conformal radiotherapy (B and E) and IMRT (C and F).

9. Facilities

Every year, several thousand cancer patients with both early and advanced tumors are treated in the proton therapy centers as Loma Linda Institute and PSI with encouraging results. Although in both centers many successful treatments carried out for meningiomas, chordomas, chondrosarcomas, sarcomas, eye tumors ,children cancers and etc. , ORL tumors (nose and throat region tumors) were only treated at PSI. Moreover, in 98% of aforementioned cancer types complete cessation of tumor growth has been reported. Loma Linda has the smallest variable-energy proton synchrotron which delivers adequate energy to the deep-seated tumors. Furthermore, it includes four treatment rooms with approximately 90 tones gantries. In contrast PSI applies the Spot-Scanning technique in order to uniformly spread out radiation dose overall tumor region. In addition, a new gantry called gantry2 is setting up at PSI and will be provided at the end of 2012 for patient treatment [22]. On the other hand, HIT (Heidelberg Ion Centre) has the capability of treating patients with protons as well as various heavy ions such as carbon, oxygen, and helium ions. Another considerable facility of this center is intensity-controlled rasterscan method leading to maximum accuracy in the three-dimensional radiation of tumors. Two treatment rooms of HIT devoted to fixed horizontal beamline and one other room hosts heavy ion gantry which would be able to rotate 360° around the patient. MIT BNCT includes fission converter epithermal neutron beam (FCB), thermal neutron beam and Prompt gamma neutron activation analysis (PGNAA) facilities .In order to treat a patient in less than one hour with BNCT, the proton-Lithium reaction will require a proton beam current between 10 and 100 mA at 2.5 MeV as well as proton-Beryllium reaction needs 5-10 mA at 20 MeV [23]. Moreover Carborane (composed of boron and carbon atoms nucleotides) can be beneficial to be a boron-10 delivery agent for BNCT.

10. Conclusion

This paper deals with hadron therapy as an operative treatment for critical organs. In particular, carbon ion and proton therapy are described and compared in terms of their features and preferences .Although hadron therapy can be effective method for decreasing damage to adjacent healthy tissues and treating with fewer fraction , due to high cost of equipment that requires such as magnet, huge gantry and synchrotron ring it has almost slow development. Accordingly, major investigation is required to reduce costs by finding alternative methods for gantry rotation. On the other hand, because of the extremely complex nature of cancer, extensive calculation must be performed to estimate absorbed dose rate of various organs. Also calculating the probability of new primary cancers and other side effects should not be ignored.



Before BNCT 1 month after BNCT 3months after BNCT

Figure 3. MRI image of patient with glioblastoma multeforme treated following BNCT

FIFTH STUDIES OF PATICLE THERAPY IN MALIGNANT

Study	Treatment	No. of patients randomized (analyzed)	Median survival (months)
Griiffin et al., 1983	50 Gy photon WBRT + 15Gy photon boost	83 (78)	8.6
	50 Gy photon WBRT + 15Gy photon boost	83 (80)	9.8
Duncan et al., 1986	47.5 Gy photon	30 (Not reported)	8
	5.1 Gy neutron + 28.5 Gy photon	31(Not reported)	4
Laramore et al., 1988	45 Gy photon WBRT + 3.6 Gy neutron boost	17 (17)	13.9
	45 Gy photon WBRT + 4.2 Gy neutron boost	13 (12)	Not reported
	45 Gy photon WBRT + 4.8 Gy neutron boost	29 (28)	Not reported
	45 Gy photon WBRT + 5.2 Gy neutron boost	53 (44)	8.6
	45 Gy photon WBRT + 5.6 Gy neutron boost	61(59)	Not reported
	45 Gy photon WBRT + 6.0 Gy neutron boost	30 (30)	Not reported
Duncan et al., 1986	47.5 Gy photon	16 (16)	11
	13.8 Gy neutron	18 (17)	7
Pickles et al., 1997	60 Gy photon	Not reported (41)	10
	33–34.5 Gy pion	Not reported (40)	10

References

[1] M. Parkin, F. Bray, J. Ferlay, P. Pisani, “Global cancer statistics, 2002,” CA Cancer J Clin, 2005, pp. 74-108.

[2] C. M. Anwar, “Production and potential implications of secondary neutrons within patients undergoing therapy with hadrons,” In Proc. AIP Conference, 2001. vol. 600, p.49.

[3] O. Jäkel, M. Krämer, C. P. Karger, “Treatment planning for heavy ion radiotherapy: clinical implementation and application,” Physics in Medicine and Biology, 2001, vol. 46, pp. 1101-1116.

[4] C.Allen, B.Borak, H.Tsujii, “Heavy charged particle radiobiology: using enhanced biological effectiveness and improved beam focusing to advance cancer therapy,” Department of Environmental and Radiological Health Sciences, 2011, vol. 711, pp. 150-157.

[5] S. E. Combs, J. Bauer, D. Unholtz, et. al, “Monitoring of patients treated with particle therapy using positron-emission-tomography (PET): the MIRANDA,” BMC Cancer study, Apr 2012.

[6] N. Yamamoto, C. Ikeda, T. Yakushiji, et. al, “Genetic effects of X-ray and carbon ion irradiation in head and neck carcinoma cell lines,” Bull Tokyo Dent Coll , Nov 2007.

[7] M. Majrabi, “Hadron Therapy for Cancer Using Heavy Ions,” University of Surrey, Sep 2009.

[8] H. Suit, T. DeLaney, S. Goldberg, “Proton vs carbon ion beams in the definitive radiation treatment of cancer patients,” Elsevier Ireland Ltd, 2010, vol.95, pp.3-22.

[9] V. Brower, “Carbon Ion Therapy to Debut in Europe,” JNCI Journal of the National Cancer Institute, 2009, vol.101, no. 2, pp.74-76.

[10] T. Kamada, H. Tsujii, H. Tsuji, et. al, “Efficacy and safety of carbon ion radiotherapy in bone and soft tissue sarcomas,” Journal of Clinical Oncology, 2002, vol. 20, pp.4466-4471.

[11] A. Eleanor, Y. Chang, Y .Polly, “Late effects from hadron therapy,” Lawrence Berkeley National Laboratory, In proc. Heavy charged particle in biology and medicine , 2004,vol.73, pp. 134-140.

[12] M. Chong, A. Hunt, “IMRT for head and neck cancer” Medical Physics Publishing, 2002.

[13] J. Hall “Intensity-modulated radiation therapy, Protons and the Risk of Second Cancers,” Int J Radiat Oncol Biol Phys, 2006, vol.65, No. 1, pp. 1-7.

[14] L. Arbea, L. Isaac Ramos, R. Martínez-Monge, “Intensity-modulated radiation therapy (IMRT) vs.3D conformal radiotherapy (3DCRT) in locally advanced rectal cancer (LARC): dosimetric comparison and clinical implications,” BioMed Central, 2010 Vol.5.

[15] M. Pijls-Johannesma, P. Pommier, Y. Lievens, “Cost-effectiveness of particle therapy: Current evidence and future needs,” Journal of Radio therapy and Oncology, 2008, vol.89, pp.127-134.

[16] B .Jereczek-Fossa, M. Krengli, R. Orecchia, “Particle beam radiotherapy for head and neck tumors: radiobiological basis and clinical experience,” Head Neck, 2007, vol. 28. No. 8, pp.750-760.

[17] Salivary Gland Cancer - American Cancer Society [online]. available: http://www.cancer.org/cancer/salivaryglandcancer/index

[18] H. Zhang, S. Li, X. H. Wang, Q. Li, et. al. “Results of carbon ion radiotherapy for skin carcinomas in 45 patients”, Br J Dermatol, 2008,vol. 166, No. 5. pp. 156-162.

[19] N. Laperriere, L. Zuraw, G. Cairncross, “Radiotherapy for newly diagnosed malignant glioma in adults: a systematic review,” Journal of Radio therapy and Oncology, 2002, vol.64, pp.259-273.

[20] “Boron Neutron Capture Therapy for Cancer Treatments,” Australian and New Zealand Head & Neck Society, 2007.

[21] H .Joensuu, L.Kankaanranta, T.Seppälä et. al, "Boron neutron capture therapy of brain tumors: clinical trials at the Finnish facility using boron phenylalanine," Journal of Neurooncol, 2003, pp.123-134.

[22] [online] available : http://www.klinikum.uniheidelberg.de/index.php?id=113005&L=1

[23] J R.Alonso, "Hadron particle therapy," proc. Particle Accelerator Conference, 1996, vol.1, pp.58-62.

Providing a Triangular Model for Gap Analysis Case Study: Iran Khodro Company

Maryam Nazaridoust[1], Behrouz Minaie Bidgoli [2] and Jalal Rezaeenoor[3]

[1] **Department of Computer Engineering and Information Technology, University of Qom, Qom, Iran**
P.O. Box 3719676333, No.52, 24th avenue, 30 metri Keyvanfar, Qom, Iran
***Corresponding author**

[2] **Department of Computer Engineering, University of Science and Technology, Tehran, Iran**

[3]**Department of Industrial Engineering, University of Qom, Qom, Iran**

Abstract

The investigation of current situation of an organization is one of the most important steps for implementation of knowledge management (KM). Also, gap analysis is among techniques proposed for evaluation of present state of the organization. This paper is an attempt to provide a new framework for gap analysis within implementation phase of KM project. In this way, we first introduce effects of three important factors of Organizational Culture (OC), Information Technology (IT), and KM mechanisms (human based tools) on gap analysis. Therefore, the status of each of these gaps will be investigated in Iran Khodro Company (IKCO) by using statistical methods and data mining techniques. Then, the triangular model of gap analysis is provided considering new researches in this area and also the findings of investigation of gap analysis in aforementioned company. Finally, a framework is proposed for planning and establishment of KM project in organizations.

Keywords: *Gap analysis,* knowledge management, *data mining, knowledge management implementation*

1. Introduction

Nowadays, more attentions have been paid to individuals' interaction, knowledge of creative human resources, and knowledge oriented workforces than tangible capitals. Therefore, smart mangers try to make better use of technologies and mechanisms for management of intellectual capitals and knowledge assets. This is for encountering with innovation in products and services, increasing cooperation, increasing customers and etc. Nowadays, achieving stable competitive advantages is possible only in case the companies take step towards development, transfer and sharing of knowledge [6]. KM consists of processes including identification, acquisition, production, organization and sharing of tacit and explicit knowledge of an organization. Its infrastructures include culture, physical environment, IT infrastructures and structure of the organization. So far, companies have accelerated achieving competitive advantage by utilizing IT and have been able to use it as a tool for KM. We believe that IT tools are not enough for moving towards a knowledge-based enterprise and implementation of an efficient knowledge management system (KMS). Because, these tools cannot carry out special KM processes about tacit knowledge transfer of individuals. Note that, facing toward systems like Enterprise Resource Planning, Customer Relationship Management and other information- based systems cannot be enough, by themselves, for achieving competitive advantage. Therefore, in addition to IT-based tools, special attention to culture development and utilizing human-based solutions, are necessary for implementation and acceleration of KM processes. Therefore, a set of three factors of culture, IT and KM mechanisms can cause development of organizational competitive advantage. On the other hand, endeavor for implementation of KMSs for more and more utilization of this competitive source is ever-increasing. While, in many cases, the implemented system and sometimes the resultant output are very different from what was meant initially by implementing the system and the achieved goals are very different from the defined goals in strategy of KM [1]. Therefore, it is necessary to identify probable gaps before any implementation and consider solutions for removing each of gaps during implementation in order to implement the KMS with the least distance from desired state and achieve the predefined goals.

In this research, we first touch KM and the major factors thereto. In second part, we deal with a review of the six gap model for implementing KMS provided by Lin and Tseng and identify theses gaps and enumerate the reasons for their creation. Then, we provide the new triangular gap analysis model relying on the state-of-the art researches carried out in this area. Considering the three major roles of KM, we investigate the gap between current and desired state in the organization and finally, we provide a framework for projecting and planning KM. For a more precision investigation in this model, using data mining techniques are proposed.

2. Culture, Human and Technology

For implementation of KM in an organization, two aspects should be emphasized: basis of KM and KM solutions. As may be observed in figure 1, KM solutions include processes and systems of KM . For having successful KM processes in the organization, appropriate technologies and mechanisms should exist.

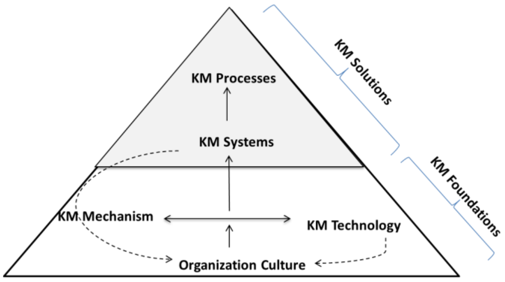


Fig.1 A scheme of bases and solutions to KM. [6]

2.1 knowledge management and Information Technology

The companies have accelerated achieving competitive advantage by application of KM and they have utilized it as a tool for KM . Some already believed that KM is part of information management. In addition to management of storage and access to documents, IT provides the possibility of knowledge creation, knowledge integration and in general KM by providing appropriate organizational architecture, Of course, in recent years, many of researchers have come to the belief that knowledge is not exclusively an outcome of IT and KM processes cannot be implemented just through its solutions, because we will come across problems for exchange of tacit knowledge.

2.2 knowledge management and Organizational Culture

As it was mentioned, KM gets its importance from the importance it attaches to the most valuable capital of human, namely intellectual capital. But what is vital for effective implementation of KM in the organization is an OC which is receptive to it. Success in KM area is closely dependent on OC . So that KM has turned to culture management to some extent. Some barriers that develop due to lack of appropriate OC are fear of innovation and knowledge sharing. Therefore, a solution should be found that organization improve its culture and its reward system so that the staff are encouraged to share their experience and knowledge as far as the organization collects knowledge as an asset. In general, by OC is meant a system of common sense that the members to an organization have towards it and the same feature makes it distinct from other organizations. Chris R Jeris regards OC a live system and defines it in the frame of behavior individuals' show in practice, the way they think and fell and the way they behave with each other actually.

2.3 knowledge management and its Mechanisms

KM mechanisms are human based solutions which are used with IT for implementation of KM . Mechanisms are regarded as a combination of organizational, social, structural and even sometimes IT arrangements which are applied for facilitation and promotion of major processes of KM . Different mechanisms fall in two groups of short term and long term. Some of these mechanisms like brainstorming sessions, occupational circulation, on-job training and exit interview are generally used in enterprises [6].

3. Gap Analysis

Gap analysis deals with analysis of the difference between current statues (as is) and the desired state (to be) in the future [10]. It is possible to found the real level of knowledge in the organization by analysis and comparison of current gaps at organization and comparing it with provided standards in this area [12].

4. A history of Gap analysis

In most researches, knowledge gap refers to the difference between the present state of the organization and the desired state of it for implementation of KM. It may be said that the first knowledge gap was discussed by Lavrich and Pears in 1984, which dealt with two types of knowledge gap for determining distance of social classes. Then, Zack (2002) discussed the gap between what an organization should do for competitiveness and what actually is implemented in real world and called it strategic gap. Relying on traditional approach of strategic management, Zack investigates the strengths, weaknesses, opportunities and threats in gaps. Strong and Weak points determine the current capabilities of the organization or the present state. While, opportunities and threats is representative of the way ahead and the things it should achieve. The strategy also shows how an organization should make a balance between these two. In addition, another potential gap called knowledge gap is developed beside strategic gap which relates to what the organization should know for implementing its strategies and what it really knows. The role of strategy in this knowledge gap is assisting the organization for bridging the gap relying on KM innovations and solutions. What is obvious is the alignment of knowledge gap and strategy gap, since knowledge gap stems directly from strategy gap. Also, Hall and Adriane investigated the knowledge gap in innovative enterprises in the same year. Then, knowledge gap was investigated more seriously by researchers. Relying on concepts of PZB, Tseng and Lin provided a model called KM gap in 2005. Initially, it consisted of 5 gaps due to managerial gaps in implementing KMS. These gaps were developed due to weaknesses in current managerial activities and incapability of the staff in planning, implementation and support of activities of KMS. Figure 2 indicates KM gaps in the model provided by Tseng and Lin [9].

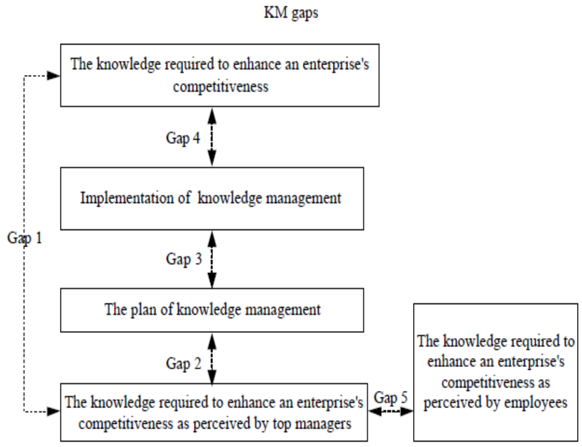

KM gaps
The knowledge required to enhance an enterprise's competitiveness
Gap 4
Implementation of knowledge management
Gap 1
Gap 3
The plan of knowledge management
Gap 2
The knowledge required to enhance an enterprise's competitiveness as perceived by top managers
Gap 5
The knowledge required to enhance an enterprise's competitiveness as perceived by employees

Fig.2 the 5 gap framework provided by Tseng and Lin in 2005[18].

Tseng also investigated important indicators in evaluation of organization performance and the effects of KM activities on them. They then investigated the effects of KM gap or lack of it on the activities of KM and finally on organization efficiency [1].Considering Holzapple's conceptual model of knowledge chain value and Nonaka knowledge cycle, Tseng investigated his framework from another aspect in the same year and added another gap to it [19]. As is shown in figure 3, these 6 gaps were investigated in terms of 4 different aspects of strategy, implementation, planning and perception [9, 16].

As per strategic aspects, the organizations should investigate their internal and external environments continuously for increasing their competiveness. Failure to do so may result in first gap. As per perception aspect, the manger may have not the capability to determine the knowledge that the organization really needs which may result in development of second gap.

In addition, it is possible that there may be some differences in their perception of the knowledge needed by the enterprise due to differences in the role, place and professional knowledge of mangers and staff. Also, there may be a gap between the knowledge needed for increase of competitiveness of enterprise and the knowledge needed on the basis of staff perception at time of implementation and administration of KMSs which is suggestive of gap 5. In terms of planning, if senior managers cannot consider the knowledge acquired from the environment at implementation of KMSs, gap 2 develops. If the staff is not able to understand the KM plans at the face of it, gap 3 develops. In terms of implementation, if implementation of KMSs is not coordinated with the programs planned for it, gap 3 develops. In addition, at the time of implementation, the staff should have a correct perception of the knowledge needed for increase of enterprise competitiveness. Otherwise, gap 4 develops [9, 16].The description of these gaps may be observed in table 1.

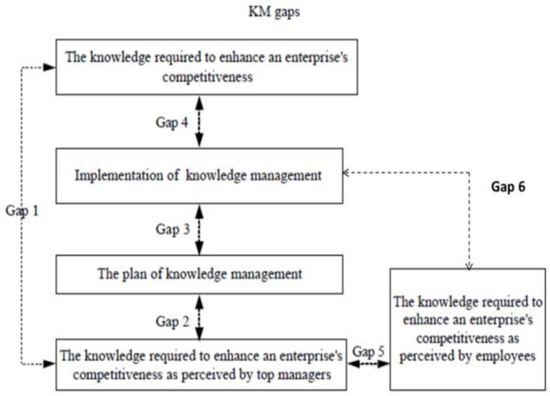

KM gaps
The knowledge required to enhance an enterprise's competitiveness
Gap 4
Implementation of knowledge management
Gap 6
Gap 1
Gap 3
The plan of knowledge management
Gap 2
The knowledge required to enhance an enterprise's competitiveness as perceived by top managers
Gap 5
The knowledge required to enhance an enterprise's competitiveness as perceived by employees

Fig.3 The 6 gap framework of KM provided by Tseng and Lin in 2005[9].

Tseng et Al (2008) investigated the effects of IT in improvement of status of gaps of implementation of KM on the basis of Tseng and Lin 5 gap framework. The reasons for appearance of KM gaps in some practical examples are also dealt with in this research. Finally, the effects of IT and tools based on it are analyzed for improvement of the status of these gaps in this research [20].One year later, in continuation of his research; in addition to IT, he also investigated the place of OC and the factors affected by it in improvement and bridging the KM gaps [21]. The findings of this research reveal that though IT is an essential factor for KM implementation, it cannot fully include factors affecting the implementation of an efficient and successful KMS. The organizations should also deal with other important factors, like, issues related to human resource and OC which play effective roles in the success of this system [18, 21]. Figure 4 shows the conceptual framework provided by Tseng.

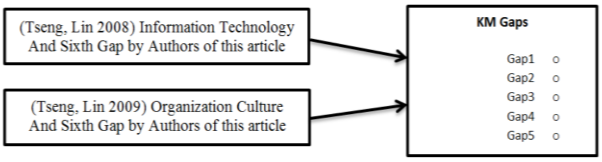

(Tseng, Lin 2008) Information Technology And Sixth Gap by Authors of this article
(Tseng, Lin 2009) Organization Culture And Sixth Gap by Authors of this article
KM Gaps
Gap1 ○
Gap2 ○
Gap3 ○
Gap4 ○
Gap5 ○

Fig.4 the conceptual model provided on the basis of IT and OC

5. Research Method

Questionnaire is a widely used tool and a direct method for acquiring research data in survey research. Therefore, considering the current literature regarding KM gaps, this method was selected. At present research, the statistical society consists of the managers, supervisors, experts and sales and marketing staff of Iran Khodro Company. Convenience sampling has been used in this research, since selection of the sample members have been performed on the basis of availability [2].

Table1: Description of every gap [18,19]

Gap1	The gap between the knowledge needed for promotion of competitive status of the organization in view of senior managers and the real knowledge needed for increasing competitive situation.
Gap2	The gap between the knowledge needed for promotion of enterprise competitive situation in view of senior management e and planning KMS.
Gap3	The gap between the plan provided by the senior managers for implementation of KM and the progress of KM implementation plan.
Gap4	The gap between knowledge received after implementation of the KMS and the knowledge needed for promotion of competitive situation of the enterprise.
Gap5	The gap between the knowledge needed for promotion of competitive situation of the enterprise in view of senior managers and in view of other staff
Gap6	The gap between the knowledge needed for promotion of the competitive situation of the enterprise in view of the staff and the real knowledge received after implementation of KMS.

For investigating effects of OC, IT and KM mechanisms, we distributed independent questionnaires in 3 steps and in two parts among the target group. 6 questions on the personal information of individuals were asked in the first part of the questionnaires, and the second part of questionnaires, asked independent questions. In fact, the first questionnaire consisted of 22 questions on investigating and analyzing OC in the enterprise and the second questionnaire consisted of 30 questions on analyzing gaps of KM mechanisms. Likert Scale is used for rating questionnaire data, which includes 5 status of too little, little; average, very and very much. The choice selected by the respondents on the company reveals the perception of the respondents of the current situation of the company considering the mentioned factors. For testing the reliability of the questionnaires, Cornbrash's alpha was calculated by SPSS. Cronbach's alpha for the questionnaires is valued at .92%, .77% and .78%, respectively. The frequency percentage of the respondents in the mentioned questionnaires is as table 2.

6. Research Conceptual framework

As it was mentioned before, Tseng and Lin were among the first to raise and investigate the issue of KM gaps and the factors affecting it. On the other hand, as per what is shown in Figure 5, technology (IT), culture (as the main element of the infrastructures) and KM mechanisms are regarded as foundation of KM implementation. Considering these factors, this research is an attempt to consider model provided by Tseng and Lin as the basic model and provide a new triangular model of gap analysis [8].

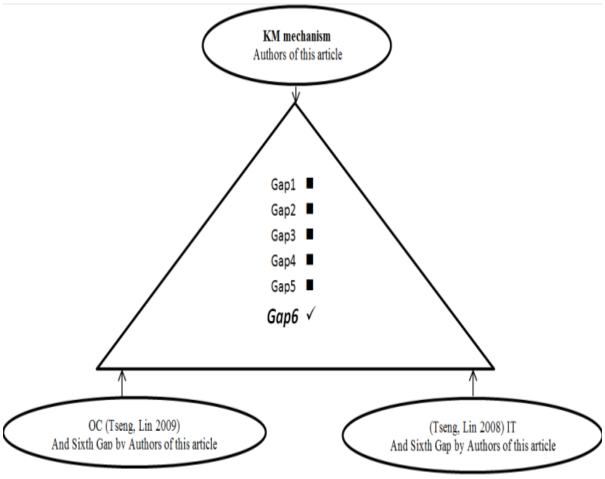


Fig.5 the theoretical framework of this research IT

In addition to taking into account the role of IT and OC in the sixth gap, the triangular model of gap analysis investigates the role of KM mechanisms in every of 6 gaps and then, it proposes an applied framework for projecting and planning KM .

Table 2: Frequency of respondents in each of three Questionnaire

		First Questionnaire	Second Questionnaire	Third Questionnaire
Gender	Male	69%	68%	60.6%
	Female	31%	32%	39.9%
Occupation	Deputyship	1%	0%	0%
	Manager	3.9%	1.3%	5.6%
	Chair	6.8%	14.7%	7%
	Director	12.6%	4%	11.3%
	Expert	61.2%	75%	66.2%
	Staff	14.6%	5%	9.9%
Education	Associate	20.4%	5%	9.9%
	Bachelor	54.4%	70.3%	63.4%
	Master	23.3%	22%	23.9%
	PhD	1.9%	2.7%	2.8%
Total Number		103	75	71

The first gap, which is the distance between the knowledge needed for promotion of competitive situation of the company in view of senior managers and the real knowledge needed for promotion of completive situation of the organization is created due to the following reasons: Senior managers play important roles in implementing KM activities by KM implementation, review of the internal and external environments of the organizations for identifying strengths, weaknesses, opportunities and threats. Also, the organization would be able to identify its strengths and weaknesses in terms of KM and adopt appropriate strategy by analysis of current situation and features [5].In addition, every organization has its own knowledge area and as such it faces problems which it can solve through its own special solutions. In such a situation, the key role of senior management is identifying and key knowledge for acquisition of competitive advantage and surviving in competitive market [11]. Since a competitive market is not stable and it changes permanently, the only factor which may assist the organization with tracking these changes and reacting appropriately against them is knowledge creation and storage. Considering the fact that the environment and features of KM are highly variable, it is possible that the expectations of senior managers of competitive advantage be very optimistic or pessimistic, considering KM for setting appropriate goals for KMSs [21].

The second gap, which is the distance between knowledge needed for promotion of competitive situation of the organization in view of senior management and design of KMS, is created due to the following reasons: If the senior managers understand the place of the organization in internal and external environment, they can plan more appropriately from KM implementation. Though senior managers have understood the necessity for the operation of knowledge acquisition, they cannot acquire their necessary knowledge due to their failure in correct and efficient need description [5]. In other words, the managers cannot identify the knowledge required by the organization for continuous implementation of KM implementation plan, which leads to creation of the second gap; the major reason of it is the non-conformity between perception of senior managers and the plan approved for KM implementation [21].

The third gap, which refers to the distance between the plan provided by senior managers for KM implementation and progress of implementation of KMS, is created due to the following reasons: Since there are different definitions for basic knowledge, the value and procedures for defining KMSs faces different barriers. As such, every organization should provide a logical master plan for the whole of the organization. Nevertheless, there may be some misunderstandings due to lack of full understanding of KMS and its nature by the staff and also the misconception that using this system and sharing their own knowledge may bear negative effects on their personal place and values. Reluctance of the staff for sharing knowledge or their failure in developing a correct understanding of KMSs, leads to creation of gap between internal and external processes of the organization at the time of implementation [21].

The fourth gap, which is the distance between the received knowledge after KM implementation and the knowledge needed for promotion of competitive position of the organization is created due to the following reasons: Effective implementation of KM strategies include definition and explanation of the knowledge required to be acquired and what motivating methods should be used for this purpose. In addition, there is the need for development of a comprehensive evaluation system for determining whether the organization can develop its competitive advantage after implementation of KM processes or not. KM includes evaluation of knowledge resources and processors. This process includes identification and understanding of resources and processors which create value added, evaluation and comparison of trends of KM implementation and evaluation of the effects of its implementation on organization performance. It is through this way that it is possible to correctly understand the present state of the organization. The organizations often fail in evaluation of the results of KM for determining whether their expectations are met or not. Therefore, the method of knowledge evaluation is always a disputable issue for organizations. Despite of different measurement ways, measurement of knowledge assets by using current financial systems is not easily possible due to the tacit and dynamic nature of knowledge [16].

The fifth gap, which refers to the distance between the knowledge needed for promotion of competitive advantage of the organization in view of the senior managers and the view of the staff, is created due to the following reasons: Creating new knowledge is a common responsibility of every section or expert group in knowledge based companies. Executive managers and directors should participate in this process. However, a gap may be created between the perception of senior managers and staff due to differences in their occupational position, role and professional knowledge in the organization. Different managerial levels consist of: executive managers involved in routine and operational problems that are at the lowest levels of managerial hierarchy. Middle managers who act as an intermediary between executive managers and senior managers and senior managers who are responsible for drawing up policies and general policies of the organization. Therefore, perception of the staff of the required knowledge is different and depends on their role and occupational place. As a result, the coordination among perception of all staff in differ occupational places for goals and plans approved by all of them for KMSs is one of key issues in KM implementation [21].

Sixth gap which refers to the distance between knowledge needed for promotion of competitive position of the organization in view of the staff and the real knowledge received after KM implementation is created due to the following reasons: The staff spends lots of time on promotion of their knowledge level for improvement of their performance in the organization. Therefore, the organization should pave the ground and encourage the staff from knowledge sharing and creation. If the staffs are not encouraged to do so, they shall have no participation in the KM implementation. As a result, the process of acquiring required knowledge of the organization faces problems. Executive managers are involved in the details of daily activities and they work high amounts of data. Therefore, they face with the problem of making information into useful knowledge and in this way lots of concepts are lost. Of course, even if we assume that meaningful concepts are developed, their sharing with other colleagues are not easily possible. Ordinarily, the staff defines knowledge on the basis of their perception and occupational positions. Therefore, the knowledge concepts permanently changes during publication process. In addition, knowledge workers are reluctant in sharing their intellectual assets, because this is a cause for competition among them. Since knowledge power stems from knowledge for knowledge workers, strong motivating systems are needed for encouraging them to share knowledge Otherwise, competition alone remains and knowledge sharing for achieving to organizational competitive advantage would be ignored. As such, the sixth gap creates in the organization [20].

7. Usage of the triangular model in IKCO

The provided model deals with the investigation of KM gaps in the company in terms of OC, IT and KM mechanisms. Therefore, considering what was mentioned before, the status of each gap should be investigated for improving KM implementation in IKCO in order to specify which gap exists and which one has more priority for bridging the gap. Therefore, independent questionnaires were distributed in 3 stages to 450 of the company staff for analysis of OC gap, IT gap and KM mechanisms gap. At each stage, more than 70 questionnaires were responded and returned. After investigating the reliability of each test, the data was investigated using Friedman statistical test. The results of rating and mean of every gap may be observed in table 3. As may be observed in this table, the higher the mean, the higher would be the rating, and the lower the gap, the higher it is in priority than other gaps. As it may be observed, in investigation of KM mechanisms and IT, the fifth and sixth gaps are in priority than the other 4 gaps. Also, in investigation of OC gap, the sixth and fourth gaps are of higher priority, respectively. The other issue to be pointed out is the fact that the status of KM mechanisms has been evaluated almost average in all of six gaps, and in the investigation of IT the first four gaps are in average status and fifth and sixth gaps are in rather poor status. Finally, the means in investigation of OC reveal that all six gaps are in rather poor status, among them the status of fourth and sixth are worse than other gaps. Therefore, it may be concluded that for optimization of its KM, the company should pay special attention to problems regarding OC and specially problems creating fourth and sixth gaps, and then bridge the gaps and synchronize them with goals and processes of KM on the basis of the lowest rate and mean.

In the continuation, for identifying the present state of each company for KM, we draw the distance between mean of present state and the desired state in a diagram. Using this diagram, it is possible to find which gap is to be optimized and invested on, first. The diagram of IKCO status for all under investigation gaps may be observed in figure 8.

Table 3: The results of rating and mean of every gap in the questionnaires

OC gaps			KM mechanism's Gaps			IT's Gaps		
mean	**Ranking**	**Gap's order**	**mean**	**Ranking**	**Gap's order**	**mean**	**Ranking**	**Gap's order**
2.90	4.23	**Gap1**	3.8	4.9	**Gap3**	3.8	5.5	**Gap1**
2.87	3.96	**Gap5**	3.6	4.1	**Gap1**	3.4	4.2	**Gap4**
2.85	3.90	**Gap2**	3.6	4.1	**Gap2**	3.2	3.5	**Gap2**
2.62	3.16	**Gap3**	3.4	3.5	**Gap4**	3.2	3.2	**Gap3**
2.60	2.98	**Gap4**	3.0	2.5	**Gap6**	2.9	2.6	**Gap6**
2.55	2.77	**Gap6**	3.0	2.0	**Gap5**	2.7	2.1	**Gap5**

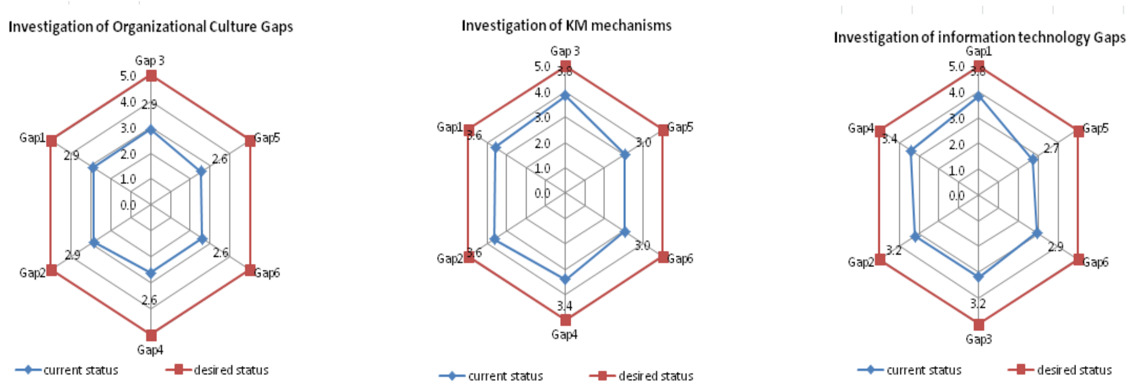


Fig.6 Investigation of the present state of the IT, KM processes and OC gaps

7.1 The Role of Data mining in improvement of Gap Analysis

In this research, we believe that we have been able to provide comprehensive feature for improvement of better KM implementation, using triangular model of gap analysis. This mode brings a multi-faceted investigation due to taking into account the roles of three important factors of OC, IT and KM mechanisms. The findings of this model will assist the company to know where it is located for every of these factors. But to achieve more precise results, we propose using data mining techniques[8]. The goal of data mining is heuristic gap analysis of the data, detecting models, rules and algorithms, predictive modeling and search of deviations [14]. For instance, as it was observed in gap analysis of KM mechanisms in Iran Khodro Company, the mean result for gap analysis of gap fifth is valued at 3 or the average. A more precise investigation by using data mining techniques, it is found that managers and some of the chairs, generally, use some mechanisms more than the staff and employees. In the continuation, we would mention the method of using data mining for in investigation of gap analysis findings as a practical instance and due to high amount of data in investigation of the fifth gap; we would mention the method of using data mining techniques in analysis of KM mechanisms.

7.1.1 Clustering

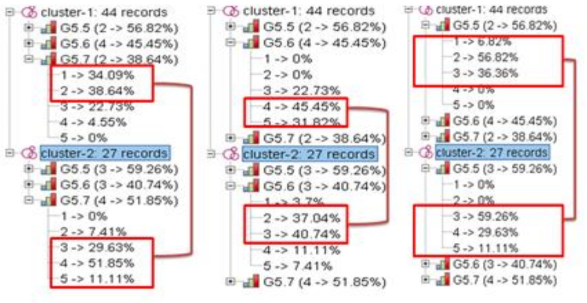
Clustering has been referred to as an algorithmic and conceptual rich frame for data analysis and interpretation which is to find the organization and detect the structure of the set of collected data. Often, clustering is regarded as synonymous to unguided learning. In clustering, without any prior classes, the heterogynous set of data is divided into some homogenous clusters. In this method, the data are merely grouped on the basis of investigation of the existing similarities or differences, like distance among the data points, and the final clusters should enjoy two features: (1) high homogeneity in every cluster and (2) heterogeneity among different clusters [3, 7, 13 and 15].

So far, diverse clustering algorithms have been introduced which bring different results. K-means is among segmentation based methods which uses variance minimum criterion for data organization. This algorithm constitutes one of the simplest unsupervised learning algorithms, which needs a pre-specified number of (K) for data grouping. Therefore, the main idea of this algorithm is the definition of a central K for every cluster. In this research, K is valued at 2 (on the basis of agreements or disagreements) [7, 15, and 17]. In investigation of the questionnaire data regarding the questions of fifth gap, we came to the conclusion that most people agree to somewhat on questions of 1-4, but the results were different regarding questions 5, 6 and 7. (The questions are shown in Appendix 1)The results of administering k-means algorithm may be observed in figure 7. The features of the clusters which are considered as a class consist of:

First class including 44 that: More than 63% of these clusters believe that group activities in the company are poorly supported and 77% out of this 44 have considered KM processes to be necessary just for the managers and only 72% of this cluster believe that there is no possibility for collective and self-motivated learning in the company.

Second Class which consists of 27 believes that: group activities in the company are supported rather desirably and 77% of them consider the KM processes to be necessary for all (not just the managers) and about 92% believe that there is the possibility of collective and self-motivated learning in the company.

Considering the above classes which are the results of administering k-means algorithm on responses of questions on fifth gap, as is shown in attachment 1, the difference between these two groups is tangible and noteworthy.



```
cluster-1: 44 records
   G5.5 (2 -> 56.82%)
   G5.6 (4 -> 45.45%)
   G5.7 (2 -> 38.64%)
      1 -> 34.09%
      2 -> 38.64%
      3 -> 22.73%
      4 -> 4.55%
      5 -> 0%
cluster-2: 27 records
   G5.5 (3 -> 59.26%)
   G5.6 (3 -> 40.74%)
   G5.7 (4 -> 51.85%)
      1 -> 0%
      2 -> 7.41%
      3 -> 29.63%
      4 -> 51.85%
      5 -> 11.11%
```

```
cluster-1: 44 records
   G5.5 (2 -> 56.82%)
   G5.6 (4 -> 45.45%)
      1 -> 0%
      2 -> 0%
      3 -> 22.73%
      4 -> 45.45%
      5 -> 31.82%
   G5.7 (2 -> 38.64%)
cluster-2: 27 records
   G5.5 (3 -> 59.26%)
   G5.6 (3 -> 40.74%)
      1 -> 3.7%
      2 -> 37.04%
      3 -> 40.74%
      4 -> 11.11%
      5 -> 7.41%
   G5.7 (4 -> 51.85%)
```

```
cluster-1: 44 records
   G5.5 (2 -> 56.82%)
      1 -> 6.82%
      2 -> 56.82%
      3 -> 36.36%
      4 -> 0%
      5 -> 0%
   G5.6 (4 -> 45.45%)
   G5.7 (2 -> 38.64%)
cluster-2: 27 records
   G5.5 (3 -> 59.26%)
      1 -> 0%
      2 -> 0%
      3 -> 59.26%
      4 -> 29.63%
      5 -> 11.11%
   G5.6 (3 -> 40.74%)
   G5.7 (4 -> 51.85%)
```

Fig.7 the results of k-means algorithm on fifth gap

7.1.2 Association Rules

Extracting Association Rules is one of unguided and important methods in data mining. It is possible to find interesting relations and dependencies in data set using this technique. Often, exploration of interesting and useful rules provides an information resource which assists with taking better decisions and practicing better clusters [3, 7, 4, and 13]. By administering Association Rules algorithm on the data, in way that particulars of the individuals are considered as algorithm input and type of class as target of the algorithm, 96 rules may be extracted. Though the number of input data is very low for extracting appropriate rules, it is possible to find good rules among them. For instance, the managers and directors together are located in class 1. Interestingly, individuals with 15+ work records or 50+ in age are also located in this class. So, the experts or younger people mostly fall in second class. However, it is not possible to consider the rules as criteria with this low numbers and low dispersion, according to the results it may be claimed that in case this procedure is administered in the whole company, the data would be more reliable and significant. Any way, it may be claimed that by application of data mining techniques after gap analysis, it is possible to have a deeper understanding of the structure and rules underlying the data to be able to behave more precisely for promotion and application of different mechanism (presented in the sample). This method brings about performance improvement and also causes cost decrease in KM implementation in the company.

8. Projecting and planning of knowledge management

So far different models and methods have been provided for Projecting and planning knowledge management. This paper is an attempt to provide a model by combining our triangular gap analysis and data mining techniques for organizations to be able to investigate their present state more Precision and comprehensive. It is obvious that the mere understanding of the gap between the present state and desired state, even with the precise added to it by data mining, is not enough for implementation of KM in the organization. Therefore, as is shown in figure 8, it is necessary to adapt the results of combining triangular model of gap analysis and data mining with strategic components or directive components. Consist of mission, vision and Guidelines of the company. In this way, it is possible to specify the required strategic goals for bridging the current gaps (KM mechanisms, OC and IT). The very existence of these goals makes the management of new knowledge more operational and also provides the possibility of measuring mission performance of knowledge in the organization. The set of explained strategic components makes more transparent the space for need to new knowledge and provides the possibility of explaining knowledge strategy. Knowledge strategy describes the general approach of the organization in the new required knowledge for meeting the strategic needs arising out of the future strategic movement of the organization. As the name indicates, knowledge strategy is concentrated on the contents of new knowledge and the general approach of directing it to the organization. After explaining the organization strategy, this framework needs specifying more details for measuring KM performance in the organization which includes items like qualitative purposes, measurement indicators and quantitative goals. Finally, it is possible to move for implementation of these strategies in the organization by having appropriate strategy and understating the gaps and the related details and also it is possible to design appropriate measures and plans. We believe that applying data mining and triangular model of gap analysis can be an appropriate way for effective and efficient implementation of knowledge management.

9. Conclusion

In this research, after reviewing factors affecting on establishment of KM in the organization, factors of OC, IT and KM mechanisms were considered as effective factors for successful implementation of knowledge management. Investigation of present state of the organization is also regarded as one of important stages for establishment of knowledge management. Therefore, after investigation of different resources on gaps of KM implementation, we selected the model provided by Tseng and Lin, consisting of 6 gaps, as the base of my model and in the continuation; we provided a triangular model of gap analysis. We believe that the triangular model of gap analysis enjoys comprehensiveness for gap analysis or the study of distance between the present states of the organization with the desired state. By investigation of this model in Iran Khodro Company, we arrived at the conclusion that in addition to preciseness, .the results of this gap need more comprehensiveness to be able to act with more precision for bridging her present gaps. Therefore, we applied data mining techniques to both identify the structure of data of the questionnaires and also analyses the rules and dependencies between individual characteristics of the respondents and their responses by using unguided algorithms of rules and dependencies. We found that applying triangular model of gap analysis and combining it with data mining techniques can provide more comprehensiveness in addition to the important feature of precision and presented the results as a framework for Projecting KM in the company to put together the strategic components or directive components with the results of the above combination and to make necessary plans and actions for implementation of KM in the organization.

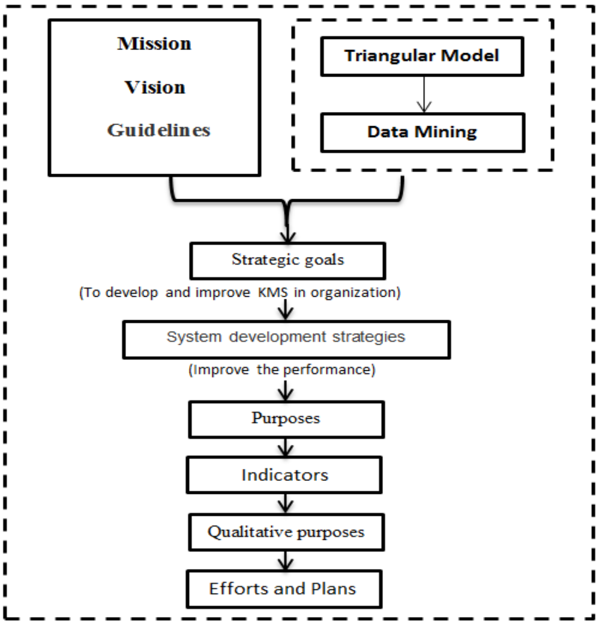


Fig.8 our framework for Planning and projecting KM

400 questionnaires of gap analysis of KM mechanisms, gap analysis of IT and gap analysis of KM mechanisms were distributed among the staff of marketing and sales department of Iran Khodro in 3 stages. About 250 questionnaires were returned. Their statistical and analytical investigations revealed interesting that we mention an instance while considering the confidentiality of information. For instance, gap analysis of the fifth gap by using the triangular model of gap analysis revealed an average state, while applying data mining techniques revealed that managers and some chairs with master or more education or individuals with low work records believed that using KM mechanisms are necessary for all staff of the company and also they have developed a self-learning spirit among this personnel.

Finally, by considering the findings out of the provided framework, it may be claimed that it assists with utilizing KM solutions in the company with more precisions which in turn will cause performance improvement and cost reduce in KM implementation in this company.

Appendix 1

	Questionnaire of KM mechanisms	Questionnaire of IT	Questionnaire of OC
Gap 1	It is possible to solve key issues of the company by social and organizational tools or through a human based structure	It is possible to solve key problem of the company by IT.	Are cultures manners clear to you?
	KM mechanisms are effective for facilitating knowledge processes in the company	In your opinion, to what extent there is enough perception and understanding of IT in the company	Can you feel OC environment in your company?
	In your opinion, to what degree enough perception and understanding of KM mechanisms prevail in the company?	IT cannot easily assist the company with value added knowledge	Do you think that the company is sensitive (susceptible) to changes in external environment?
	KM mechanisms cannot assist the company with value added knowledge by creating knowledge	IT can assist the mangers to have a deeper understanding of the current problem of the company	
	KM mechanisms cannot easily assist the company with value added knowledge by knowledge sharing	IT can assist the company with identification and application of key knowledge.	
	KM mechanisms can assist the managers to have a more deeper understanding of the current problems of the company	IT cannot improve decision taking processes for the managers.	
	KM mechanisms can assist the company with identification and application of key knowledge	Appropriate identification and application of IT tools can improve the effectiveness of KMS	
	KM mechanisms cannot improve decision taking processes for managers	In your opinion, to what extent smart tools (competitive intelligence, business intelligence) or customer relations system can assist the managers with achieving competitive advantage and taking strategic decisions?	
	Appropriate identification and application of KM mechanisms can improve the degree of effectiveness of KMS.		
	In your opinion, can mechanisms like knowledge map, registration of the best practices and registration of the instructions assist the managers with achieving competitive advantage and taking strategic decisions?		
Gap2	There is a place, obviously, for KM mechanisms in implementation of KM in the company.	In your opinion, are the goals for implementation of KM in the company explained clearly?	
	Mechanisms like knowledge map, registration of the best practices and registration of the instructions can assist KM mechanisms in the company	Storage systems of the company can be used in KM plan on an integrated basis.	
	KM mechanisms can facilitate knowledge sharing and optimal use of tacit and explicit knowledge resources and organizational creativity.	Continuous updating of knowledge can easily lead to knowledge sharing and optimal use of knowledge resource and organizational creativity.	
	KM mechanisms are effective in improving organizational memory.	Information categorization is available in the company as a systematic process.	
	Using KM mechanisms like exit interview can be effective in capturing tacit knowledge of individuals		In your opinion, is it possible to acquire knowledge from external environments in addition to acquire it from in house staff?
	KM mechanisms make it possible to acquire knowledge not only form in house staff but also for m external environments		Is there any full mechanism in the company that assists with all decision making processes?
	There are lots of opportunities for group discussions, specially unofficial ones, leading to promotion and exchange of tacit knowledge of the staff		
Gap 3	Registration of knowledge carriers in knowledge map is one way for identification and access to professional which is effective for better fulfillment of tasks.	In your view, do the IT tools available in the company allow you to directly access to the information?	
	In your opinion, can mechanisms like different meetings or knowledge café in the company encourage the staff to knowledge sharing?	In your view, can the IT tools available in the company encourage the staff to share knowledge with each other?	
	KM mechanisms can play the role of tools for exchange of tacit knowledge.	The IT tools available in the company can assist with control and monitoring of KM plan.	
	KM mechanisms can play the role of tools for converting tacit knowledge to explicit knowledge	IT can play role as a tool for making tacit knowledge explicit knowledge.	
	Existence of repository of lessons learned and also best practices are effective for facilitation of KM processes.		In your opinion, the values of staff are able to promote KM activities.
	Virtual space (internet, intranet etc.) has facilitated interaction of technology and intellectual capital.		Does the company hold special ceremonies for promotion of KM?
Gap 4	In your opinion, are current systems based on knowledge-management mechanisms available in the company y able of supporting KM implementation?		Are you satisfied with reward system for KM activities of the company?

	In your opinion, does the company utilize evaluation systems for measuring effectiveness of processes?		Is customer KM included in the strategic plan of the company?
	Certainly, an efficient KMS includes an appropriate knowledge map, repository of lessons learned and also repository of best practices.	In your opinion, do the IT based systems available in the company are able to support kl implementation.	
	In your view can, promotion of human based tools (based on KM mechanisms) is effective factors in implementation of KM processes.	In your view, the company utilizes evaluation systems for measuring effectiveness of the processes.	If one of your colleagues applies KM and arrives at satisfactory results out of it, does this encourage you to utilize KM?
	In your opinion, systems based on KM mechanisms available in the company are able to support KM implementation.	Certainly, an efficient KMS holds appropriate knowledge map, decision support software and work flow.	Does the company pay attention to individuals who deal with KM activities?
Gap 5	Despite lack of special information systems in the company, the company supports the possibility of communications and interactions in your department by promotion of human based tools (KM mechanisms).	In your view, facilitation of IT based systems can be an effective factor in implementation of KM processes.	
	Knowledge repository (lessons learned, best practices etc.) can reduce the probability of repetitive errors and parallel activities in the company.		
	There is no place for registration of lessons learned and best practices of projects in the information system of the company for others to access to.	Information systems of the company are able to share results out of projects for access of others.	
	To what extent, the activities of individuals taking part in KM processes of the company are evaluated?	Information systems of the company are able to support communications and interactions within your department.	To what extent, the current cultural environment of the organization allows every department to share knowledge more effectively
	In your view, the company supports the team works of the staff	Information system is able to reduce the probability of repetitive errors and parallel activities.	To what extent, the current OC assists with facilitation of knowledge exchange among departments.
	Using of KM processes leading to knowledge creation is only necessary for company ma agers.	To what extent, evaluation and measurement tools are utilized for supporting individuals active in KM activities.	
	There is the possibility of group learning in the company, so that self-motivated and unofficial learning groups can promote their activities through on-job negotiations and learning meetings.	In your view, the current IT systems provide the possibility of supporting team working of the staff.	
Gap 6	To what extent, do the KM mechanisms assist with knowledge exchanges among different departments of the company?	The current evaluation system of the company is effective for contribution and encouragement of the staff for implementation of KMS.	
	To what extent, can tools based on KM mechanisms effectively increase support of senior managers of the company for knowledge sharing activities of the staff.	There is a knowledge documentation system (lessons learned, best practices etc.) which deals with cooperation, decision support and information security.	To what extent, are communities-of-Practice utilized for facilitation of knowledge sharing by the staff.
	Systems like performance evaluation and suggestions system can play role in increasing staff motivation for participating in KM and knowledge sharing activities.	To what extent, IT based systems are utilized for facilitation of knowledge exchanges among departments.	In your view, to what extent, is the evaluation performance system effective for contribution and encouragement of staff for implementation of KMS?
	Communities-of –Practice have been used to a great extent by the staff for facilitation of knowledge sharing.	To what extent, can IT tools be effective in increasing the support of senior managers of the company form the staff for knowledge sharing?	
		Systems like performance evaluation and suggestions can play role in increasing staff motivation for taking part in KM and sharing activities.	

References

[1] P. Akhavan , "Semantic network of KM based on the key success factors", PhD thesis, university of Science and Technology,2007

[2] Berry m.,Lino G.,"Data Mining Techniques: for Marketing, Sales and Customer Support" New York: Jon Wiley and Son ,1997

[3] B. Minaei Bidgoli and M.. Nazaridoust, "Case Study: Data Mining of Associate Degree Accepted Candidates by Modular Method ",*Communications and Network* ,Vol. 4 No. 3, 2012, pp. 261-268. doi: 10.4236/cn.2012.43030

[4] C. Romero , S. Ventura, " Educational data mining: A survey from 1995 to 2005", Expert Systems with Applications, vol.33, pp. 135–146, 2007

[5] Ho. C. "The relationship between knowledge management enablers and performance" Industrial management & Data systems, Vol. 109, No. 1,pp:98-117, 2009

[6] Irma Becerra-Fernandez and Rajiv Sabherwal, "Knowledge management : systems and processes"

[7] J. Shahrabi and V. Shakourniaz , " Data Mining in SQL Server " Jahad Daneshgahi Published , Inc.2009

[8] J. Rezaienour and M. Nazaridoust , “Data Mining application in analysis of Knowledge management gaps” , Proceeding of The 2nd World Conference on Soft Computing , pages 551-557, 3-5 December 2012.

[9] Lin, Ch., Yeh Jong, M., Tseng Shu, M. " Case Study on knowledge- management gaps". Journal of Knowledge Management,9 (3): 36-50. 2005

[10] McBriar,I and Smith.C and Bain.G,Unsworth.P "Risk,gap, strength:key concepts in knowledge management" knowledge-Based systems Vol.16,: 29-36.2003

[11] Megdadi. M."knowledge management enablers & outcomes in the small & medium sized enterprises" Journal of Industrial Management & Data Systems, Emerald Group Publishing Limited, Vol. 109, No. 6,: 840-858, 2009

[12] A. Moteleb and M. Woodman "Notions of Knowledge Management Systems: a Gap Analysis" The Electronic Journal of Knowledge Management Volume 5 Issue 1, pp 55 - 62, 2007 , available online at www.ejkm.com

[13] M. Nazaridoust and B. Minaei Bidgoli , “Data Mining of associate degree accepted candidates by using unsupervised paradigm” , Proceeding of The 2nd World Conference on Soft Computing , pages 558-563, 3-5 December 2012.

[14] M. Qazanfari , S. Alizadeh, "Data mining and Knowledge Discovery ",University of Science and Technology, inc. 2008

[15] P.N. Tan, M. Steinbach, V. Kumar, "Introduction to Data Mining", 1th ed,Pearson:Addison Wesley, 2006

[16] R. RahmanSeresh, N.Symar Asl , "Study of Knowlegde gaps: Petrochemical Research and Technology research center in Tehran" ,61, Jornal of management Science, Third Year, No10 ,, pp3, 2008

[17] Spss Inc, "Clementine®12 Algorithms Guide",United State of America, pp.80-81, 2007

[18] Tseng.S and Lin, C., "Bridging the Implementation Gaps in the Knowledge Management System for Enhancing Corporate Performance", Expert systems with Applications, Vol.29, pp. 163-173, 2005

[19] Tseng.S and Lin.C, "The implementation gaps for the knowledge management systems", Industrial Management and Data systems, Vol. 105, No. 2, pp. 208-222. 2005

[20] Tseng.S,"The effects of information technology on knowledge management systems",Expert system with Application,Vol.35,pp.150-160, 2008

[21] Tseng. S et al, "Bridging knowledge management gaps by information technology and organizational culture",The university of Manchester,IEEE , 2009

Maryam Nazaridoust was born in Tehran, Iran in 1985. She received her B.S. degree with a first class Honors in Software engineering from UAST University. In 2010 she was selected for the Graduate studies without the entrance examination in the Department of Information Technology at University of Qom for M.S degree. She specializes in the field of Data Mining and Knowledge Management.

Dr. Behrouz Minaei-Bidgoli obtained his Ph.D. degree from Michigan State University, East Lansing, Michigan, USA, in the Field of Data Mining and Web-Based Educational Systems in Computer Science and Engineering Department. He is working as an assistant professor in Computer Engineering Department of Iran University of Science & Technology, Tehran, Iran. He is also leading at a Data and Text Mining research group in Computer Research Center of Islamic Sciences, NOOR co. Qom, Iran, developing large scale NLP and Text Mining projects for Farsi and Arabic languages.

Dr. Jalal Rezaeenoor obtained his Ph.D. degree from University of Science & Technology, Tehran, Iran. He is working as an assistant professor in industrial Engineering Department of University Qom, Qom, Iran. His research interests are in Performance Measurement, Knowledge Management, Information Technology, Decision Making and Multivariate Data Analysis. He is publishing two books and has more than 15 papers in different conferences and journals

Parallelization of Memetic Algorithms and Electromagnetism Metaheuristics for the Problem of Scheduling in the production Systems of HFS type

Kadda Zerrouki[1] and Khaled Belkadi[2]
[1]LAMOSI, Department of Computer, Faculty of Science, USTOran-Algeria

[2]LAMOSI, Department of Computer, Faculty of Science, USTOran-Algeria

Abstract

The metaheuristics are approximation methods which deal with difficult optimization problems. The Work that we present in this paper has primarily as an objective the adaptation and the implementation of two advanced metaheuristics which are the Memetic Algorithms (MA) and the Electromagnetism Metaheuristic (EM) applied in the production systems of Hybrid Flow Shop (HFS) type for the problem of scheduling. The Memetic Algorithms or hybrid genetic algorithms are advanced metaheuristic ones introduced by Moscato in 1989. Electromagnetism Metaheuristic (EM) draws its inspiration in the electromagnetic law of Coulomb on the particles charged. We will propose an adaptation of two methods to the discrete case on the problems of scheduling with the production systems (HFS). We present then a comparison between the Memetic Algorithms (MA), the Parallel Memetic Algorithms with Migration (PMA_MIG) and then we present a comparison between Electromagnetism Metaheuristic (EM) and Parallel Electromagnetism Metaheuristic with migration (PEM_MIG). Finally we give the results obtained by its algorithms applied to HFSs (HFS4: FH3 (P4, P2, P3) | | Cmax and HFS4: FH2 (P3, P2) | | Cmax) for the two problems: scheduling and assignment.

Keywords: *advanced Metaheuristics, Hybrid Flow Shop (HFS), Memetic Algorithms (MA), Electromagnetism Metaheuristic (EM), and Parallelism.*

1. Introduction

The objective of this work was to simulate two advanced metaheuristic called “MA and EM” and to study the contribution of parallelism in the improvement of the performances of these algorithms. The dealt problem is the scheduling of the production systems work of the type FSH. The complexity of this system lies in the nature of the problems itself.

It has a double complexity, on one hand it is necessary to find a sequence of work and on the other hand it is necessary to find an assignment as of this work to the resources in order to optimize a performance criterion such as Cmax (**Cmax: the total completion time of treatment or works**).

The general goal was the minimization of the time of completion of work (Optimization of a criterion of performance: Cmax).

Then we'll initially simulate MA and EM in manner Sequential and see if the time of simulation becomes more or less considerable. Finally, we'll be oriented to the Parallelization of two advanced metaheuristics.

2. Scheduling problem of production systems of HFS

2.1 Presentation of Hybrid Flow Shop

The model of Hybrid Flow Shop (HFS) is an extension of the model of Flow Shop.

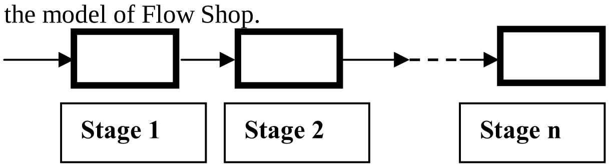


Fig. 1 Organization in Flow Shop [1, 2].

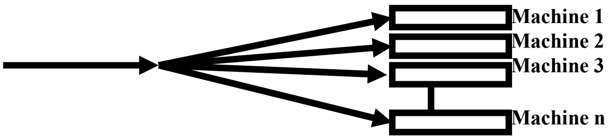


Fig. 2 Organization in parallel machines [2].

2.2 Organization in Hybrid Flow Shop (HFS)

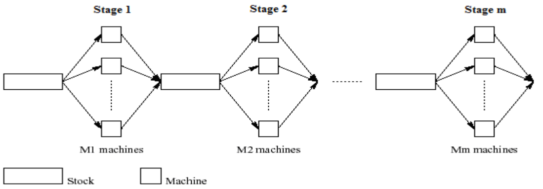


Fig. 3 Hydrid Flow Shop (HFS) [5].

2.3 Scheduling in the HFS

Scheduling in the HFS consists with:
- To find a sequence adequate of the jobs in entry.
- To find an assignment of the jobs on the various machines of the various stages.

3 Of which objective:

- Optimization of a criterion of performance.

Among the criteria one can quote Cmax, Fmax,… etc

4. Methods of resolution

It consists in finding a better solution who minimizes (or maximizes), according to the type of the problem, the selected criterion of performance. Among the methods, with the exact resolution and the approximate resolution (metaheuristics).

4.1 Approximate resolution

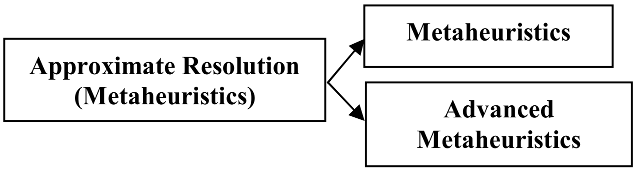


Fig. 4 Approximate Resolution.

5. Metaheuristics

It is a new generation of approximate methods. It allows the resolution of the problems of optimization of big size. Each metaheuristics is based on its own concepts and principles.

5.1 Objectives of the metaheuristics

The metaheuristics ones make it possible to guide research with an optimal solution and the effective exploration of the space of research. Those are mechanisms making it possible to avoid blocking.

5.2 Metaheuristics and Advanced Metaheuristics most known

We can quote among Metaheuristics and advanced Metaheuristics most known: research taboo [3], method of the kangaroo, the descent [4], Simulated Annealing [5], Genetic Algorithms [8], Particle Swarm Optimization method (PSO)) [6,7], the Memetic Algorithms (MA) [8], Scatter Search [8], Electromagnetism Metaheuristic (EM) [9,10].

6. Advanced Metaheuristics

The metaheuristics are very powerful for the resolution of a great number of problems and give better results, but one can note some defects like the limit to find a minimum total in a finished time, difficulties of adapting algorithms to certain problems, for certain problems they are not more powerful than the exact methods, some do not give success in connection with the intensification and diversification. To go further in the research solution, it is necessary for all to be able to detect new solutions. The algorithms containing populations present a private interest: parallelism necessary in the search for the solutions [11,12].

6.1 Memetic Algorithms

The Memetic Algorithms or hybrid Genetic Algorithms are the metaheuristic advanced introduced by MOSCATO in 1989 [5], the principal idea of this technique is to make an genetic algorithm more effective by the addition of a local research in addition to the change.
One of the general observations coming from the implementation of a basic genetic algorithm is often the low speed of convergence of the algorithm.
The idea of Moscato is thus to add a local research which can be a method of descent or a more advanced local research (reheated simulated or seeks taboo for example).
This local research will be applied to all new individual obtained during research.
It is obvious that this simple modification involves deep changes in the behavior of the algorithm.

A simple Memetic Algorithm (MA) is given below:

```
Initialization: to generate an initial population P of
  solutions with size = N
To apply a Local Research LR to each solution of P
Repeat
    To select two solution X and X' with a technique of
    selection
    To cross two parents X and X' to train children there
    For each child there
        To improve this solution with LR
        To apply a change to there
        To choose a solution to be replaced there y' and
         to replace it by their y in the population
    End For
Until criterion of stop.
```

The intensification in this algorithm is produced by the application of the local research and the operator of change ensures the diversification of the method [4, 5].

6.2 Electromagnetism Metaheuristic (EM)

6.2.1 Definition

The EM is a method with recent particles.
At introduced by Birbill (2003) [9, 10].
Used to optimize nonlinear continuous functions.

6.2.2 Principle

Its inspiration in the Coulomb law draws EM on the charged particles [9, 10].

6.2.3 The Coulomb law

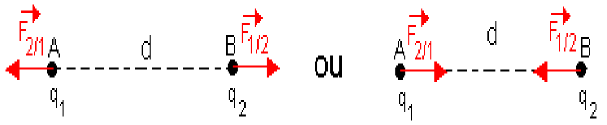


Fig. 5 Principle of the Coulomb law.

6.2.3.1 The formula of F:

$$\|F\| = K|q_1|.|q_2|/d^2$$

- q_i : charge of each particle i.
- $F_{i/j}$: particle i exerts a force (attraction/ repulsion) on particle j.
- d : the distance between two particles.
- k : constant.

6.2.3.2 Parameters of the Coulomb law

•There is a repulsion if $q_1.q_2 > 0$.

•There is an attraction if $q_1.q_2 < 0$.

•The common value $\|F\|$ of these forces is proportional to the loads q_1 and q_2 and inversely proportional to the square of the distance between A and B.
•the particles (or solutions) having ‘’bad properties‘’ (or negatively charged according to the law with Coulomb) exert a force of repulsion on the other particles.
•the particles having “good properties” (or positively charged according to the law with Coulomb) exert an attraction force on the other particles.

6.2.3.3 Example

The following figure illustrates the displacement of a particle located at the position x_4 which undergoes the forces of repulsion of the particles located at the positions x_1 and x_3 and the attraction force of the particle located at the position x_2.

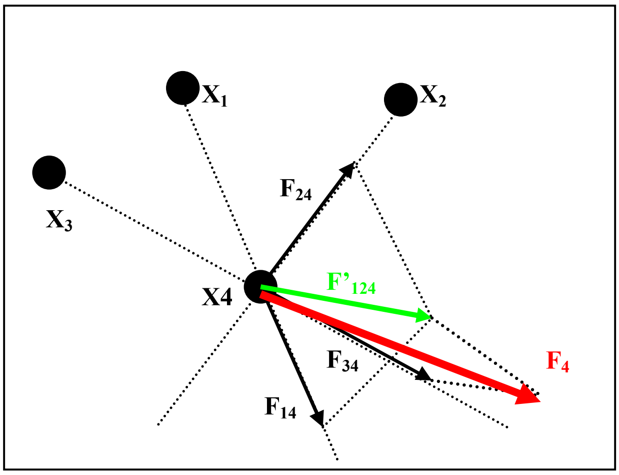


Fig. 6 Force exerted on particle 4 by particles 1, 2 and 3 [10].

Particles 1 and 3 exert a force of repulsion (F_{14}) and (F_{34}) on particle 4 and particle 2 exerts an attraction force (F_{24}) on particle 4.
The force F'_{124} it is the sum of the forces F_{14} and F_{34}. The force F_4 it is the sum of the forces F'_{124} and F_{24} i.e. the total force.

6.2.3.4 A particle i at the step k has the following characteristics

- Its current position $X_{i,k}$.

- The best position of its $X_{best,k}$.
- Its charge $q_{i,k}$.
- Its force F_i^k.
- $F(x_j,k)$ value of the function to optimize f at the x_j point, k where j is a particle.

6.2.3.5 General algorithm of EM (Algorithm. 1: General algorithm of EM [10])

```
BEGIN
    $k = 0$ ;
    For (each particle $i$) Then
        $X_{i,0}$ = GenerePosition () ;
    End For
        While ($k \le MaxIter$) Then
            For (each particle $i$) Then
                $Q_{i,k}$ = Calcul New Charge () ;
                $F_{i,k}$ = Calcul New Force () ;
                $X_{i,k+1}$ = Calcul New Position () ;
                $X_{i,k}$ = Update () ;
            End For
            $k := k + 1$ ;
        End While
END.
```

7. Parallelization of the Memetic Algorithms (PMA)

7.1 Weaknesses of the sequential memetic algorithms

The Sequential Memetic Algorithms (SMA) show their interest in the field of the scheduling of the workshops of the type Hybrid Flow Shop (HFS) but they present several weak points quote for examples the storage memory, the evaluation of the fitness, the centralized diagram of selection, the time of simulation…

1. Storage memory.
2. The evaluation of the fitness.
3. The centralized diagram of selection.
4. The time of simulation.

7.2 The standard model of Master /Slaves

There exist several models to implement the distributed model, the model completely distributed and the Master /Slaves model that is simplest. It consists in using the standard memetic algorithm applied to only one population but by carrying out stages of evaluation in parallel. The stage of selection cannot be carried out in parallel because it requires a total knowledge of the costs of all the individuals [6].

The principal processor (master) controls the population and distributes individuals to the processors slaves. These processors will carry out the operations of crossing, change and evaluate the children. After the evaluation the Master gathers the results and applies the replacement to produce the new generation to be managed.

7.3 Description of algorithm PMA of migration

The algorithm is implemented in mode Master /Slaves. Initially the Master generates an initial population and divides it into “P” under populations (“P” being the number of slaves). Each slave carries out a Sequential Memetic Algorithm (SMA) for a number “N” of iterations, with “N” the interval which separates two operations from migration.

The number of migrations to be carried out is determined by the frequency of the latter if it is frequent or not very frequent. Indeed, the number of migration to be carried out and cuts it “N” of the interval of migration are directly calculated starting from the frequency of migration and the full number of iterations to be carried out. At the point of migration, the Master recovers the results of the research carried out by the various slaves and lunch the operation of migration.

8. Parallelization of Electromagnetism Metaheuristic (PEM)

8.1 Description of algorithm PEM of migration

The algorithm is implemented in mode Master /Slaves. Initially the Master generates an initial whole of particles and divides it into “P” subset of particles (“P” being the number of slaves). Each slave carries out an algorithm of sequential (EM) for a number “N” of iterations. At the point of migration, the Master recovers the results of the research carried out by the various slaves and lances the operation of migration.

9. Numerical simulation and results

For better studying the problem of scheduling of the HFS4 (a stock in entry and inter stages) one will consider two problems of HFS4 (FH3 (P4, P2, P3) | | Cmax and FH2 (P3, P2) | | Cmax) and for each varied the number of parts to be scheduled what enables us to obtain four (4) problems for each one. First is FH3 (P4, P2, P3) | | Cmax (Fig. 7) and the second FH2 (P3, P2) | | Cmax (Fig. 8).

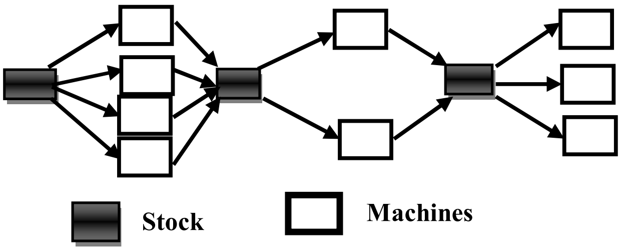


Fig. 7 Structure of a HFS4: FH3 (P4, P2, P3) | | Cmax.

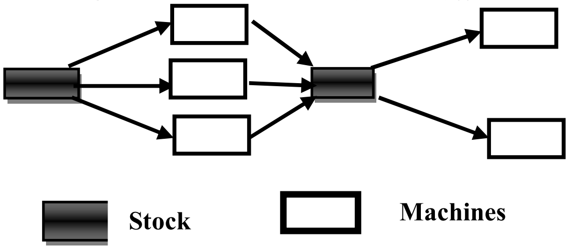


Fig. 8 Structure of a HFS4: FH2 (P3, P2) | | Cmax.

9.1 Application of the PMA_MIG and the PEM_MIG

The Parallel Memetic Algorithms with migration (PMA_MIG) and Parallel Electromagnetism Metaheuristic with migration (PEM_MIG) were implemented on a computer having architecture with shared memory, but this implementation depends on a number of parameters. The success of this strategy of Parallelization is bound by the influences of these parameters.

To study these parameters three values of the frequency were taken: 40%, 60% and 80%. This frequency determines the number of migrations to carry out during the execution of algorithms PMA_MIG and PEM_MIG as well as the interval which separates two operations from migration.

9.2 Application of the PMA_MIG on the HFS4: FH3 (P4, P2, P3) | | Cmax and on the HFS4: FH2 (P3, P2) | | Cmax

For simulation we fix the method of Local Research (LR) used by the descent, we take as number of threads 4 and we change for each number of parts the frequency of migration of 40%, 60% and 80%. One took the average of the ten tests for the four (4) problems by varying the number of parts (N) to be scheduled (5, 10, 20 and 50 parts), the results of Cmax are given by the graphs illustrated by the following Fig. 9 and Fig. 10:

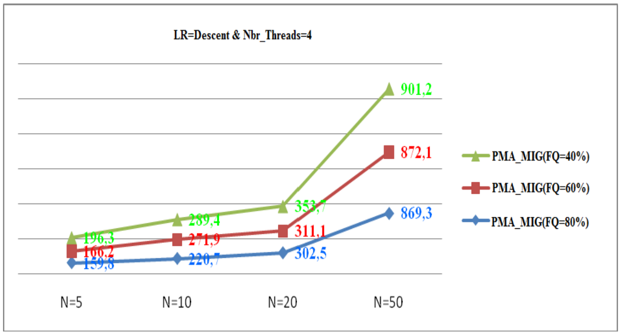


Fig. 9 Graph of variation of Cmax_Moy for the HFS4: FH3 (P4, P2, P3) | | Cmax with the application of the PMA_MIG by varying the frequency of migration.

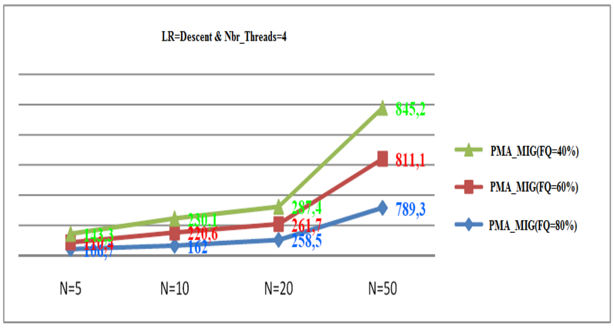


Fig. 10 Graph of variation of Cmax_Moy for the HFS4: FH2 (P3, P2) | | Cmax with the application of the PMA_MIG by varying the frequency of migration.

9.3 Application of PEM_MIG on the HFS4: FH3 (P4, P2, P3) | | Cmax and on the HFS4: FH2 (P3, P2) | | Cmax

For the simulation one taken as number of threads= 3 and one changed for each number of parts (N) the frequency of migration of 40%, 60% and 80% and took the average of the ten tests for the four (4) problems N = 5, 10, 20 and 50, the results of Cmax are given by the two graphs illustrated by the following Fig. 11 and Fig. 12:

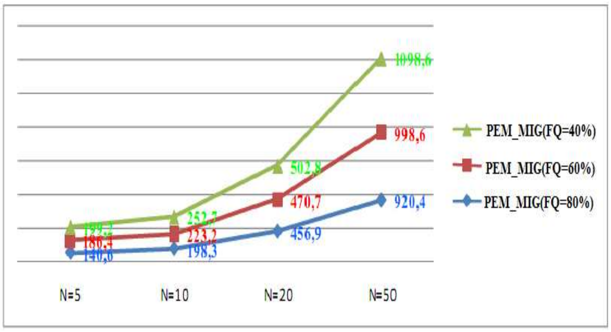


Fig. 11 Graph of variation of Cmax_Moy for the HFS4: FH3 (P4, P2, P3) | | Cmax with the application of the PEM_MIG by varying the frequency of migration.

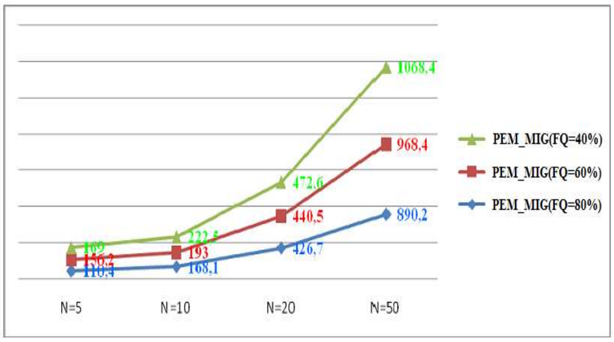


Fig. 12 Graph of variation of Cmax_Moy for the HFS4: FH2 (P3, P2) | | Cmax with the application of the PEM_MIG by varying the frequency of migration.

10. Comparison enters the PMA_MIG, MA, the PEM_MIG and EM on the HFS4: FH3 and HFS4: FH2 | | Cmax

For the two algorithms: MA and PMA_MIG one will fix the local research by the descent. For the PMA_MIG and the PEM_MIG the number of threads is fixed at 3 and the frequency with 80%.

The four (4) algorithms are applied to the HFS4: FH3 (P4, P2, P3) | | Cmax and for the four (4) problems of which the number of parts (N) varies as follows: N = 5, N = 10, N = 20 and N = 50.

The results of the comparison are given by the two graphs illustrated by the Fig. 13 and the Fig. 14.

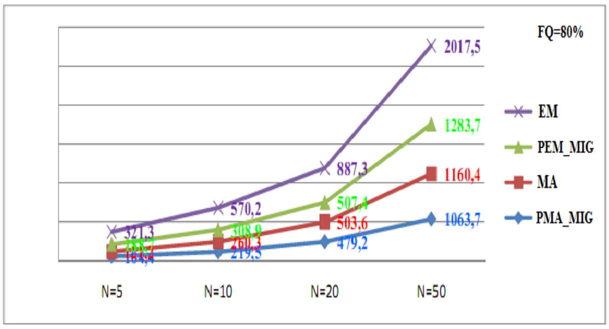


Fig. 13 Graph of variation of Cmax_Moy by the comparison enters the PMA_MIG, MA, the PEM_MIG and EM with the increase in the number of parts (N) on the HFS4: FH3 (P4, P2, P3) | | Cmax.

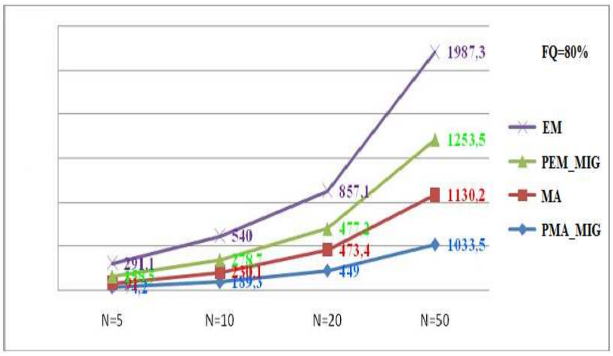


Fig. 14 Graph of variation of Cmax_Moy by the comparison enters the PMA_MIG, MA, the PEM_MIG and EM with the increase in the number of parts (N) on the HFS4: FH2 (P3, P2) | | Cmax.

According to the graphs below, we notice that the PMA_MIG presents a considerable improvement compared to MA sequential, the PEM_MIG and EM sequential with regard to the quality of the solutions obtained, and more the system is complex in addition, the algorithm manages to improve this quality.

11. Conclusion and Future Works

In this paper we gave practical details on two algorithms PMA_MIG and PEM_MIG and we implemented these algorithms on Hybrid Flow Shop (HFS) which were tested on plays of test generated in a random way. We analyzed the results obtained by the plays of test and the implementation of the algorithms. We approached a comparison between the Sequential Memetic Algorithms (SMA), the Parallels Memetics Algorithms with Migration (PMA_MIG), Electromagnetism Metaheuristic (EM) and Parallel Electromagnetism Metaheuristic with Migration (PEM_MIG).

Future Works:

To apply the Memetic Algorithms (MA) and the Electromagnetism Metaheuristic (EM) to other problems with other types of production systems such as Flow-Shop, Job Shop and Open Shop, and to apply other methods of Parallelization of MA and EM and that to grids computing and clusters.

12. References

[1] P. Delisle, "Parallélisation d'un Algorithme d'Optimisation par Colonie de Fourmis pour la Résolution d'un Problème d'Ordonnancement Industriel", Magister Thesis in Data processing, University of Quebec with Chicoutimi, 2002.

[2] F. Glover, "Genetic Algorithms and Scatter Search" Unsuspected Potentials, Statistics and Computing, 1994.

[3] M. R. Garey, and D. S. Johnson, "Computers and Intractability, A Guide to Theory of NP-Completeness", Freeman, San Francisco, 1979.

[4] M. Sevaux, "Métaheuristiques Stratégies pour l'Optimisation de la Production de Biens et de Services", Laboratory of Automatic, Mechanics of Industrial data processing and Human of CNRS, Equips of production Systems, 2004, pp.57-72.

[5] P. Moscato. "On Evolution Search Optimization, Genetic Algorithms and Martial Arts Towards Memetic Algorithms" Technical Report C3P 826, Caltech Concurrent Computation Program, 1989.

[6] A. Aribi. "Métaheuristiques Parallèles Appliquées à l'Ordonnancement dans les Systèmes de Production de type Flow Shop Hybrides", Magister Thesis, Framed by Belkadi.K, University of Sciences and the Technology of Oran USTO-Algeria, 2004.

[7] S. Hernane, and K. Belkadi, "Métaheuristiques Parallèles Inspirées du Vivant", Magister Thesis, Framed by K. Belkadi, University of Sciences and the Technology of Oran USTO-Algeria, 2006.

[8] M. R. Gourgand, N. R. Grangean, and S. Norre, "Problème d'Ordonnancement dans les Systèmes de type Flow Shop Hybride en Contexte Déterministe", University Blaise Pascal (Clermont Ferrand), 2003.

[9] S. I. Birbil, and S. Fang. "An Electromagnetism like Mechanism for Global Optimization", Journal of global optimization, Vol. 25, 2003, pp. 263-282.

[10] S. Kemmoé, M. Gourgand, A. Quilliot, and L. Deroussi, "Proposition de Métaheuristiques Hybrides Efficaces pour le Flow-Shop de Permutation", LIMOS CNRS UMR 6158-Campus of cézeaux, 63173 Aubière Cedex, 2006.

[11] I. H. Osman, and G. Laporte. "Metaheuristics: has Bibliography", Annales of Operations research, 1996.

[12] E. G. Talbi, "A Taxonomy of Hybrid Metaheuristics", Journal of Heuristics, Vol. 8, 2002, pp.541-64.

Refined Ontology Model for Content Anatomy and Topic Summarization

A. Mekala [1], Dr.C.Chandra Sekar [2]

[1] Manonmaniam Sundaranar University, Tirunelveli

[2] Department of computer science, Periyar University, Salem

Abstract

When the performance of any information processing system can be enhanced by the concept of ontologies, domain specific terms enclosing wealthy and defined semantics. Research has been accomplished with the help of variety of resources on automatic ontology construction. Each of these resources has different qualities that have need of special approaches to term and relationship extraction. On the consideration of terminological resources, semantic structure of ontology construction facilitates the NLP (Natural Language Processing) that extracts terms and relationships. Generally in this phase there can be a problem in that many relationships are incorrectly defined or applied excessively. For that reason, extracting ontological relationships from documents necessitates data cleaning and refinement of semantic relationships. In our research we provide the automatic term relationship and refinement ontology construction for the content anatomy and topic summarization. Where the automatic topic extraction mechanism will be done based on the significant score computation and the highest score will be the topics. Our proposed system supports effective joint inference approach, which simultaneously constructs the ISA (is a) and HASA (Has a)-tree, while mapping Topic models to WordNet, achieves the best performance. To end with, we estimate our ontology-based topic summarization results that formulate exploit of similarity-based metrics first enlarged for automatic term relationship findings and refinement of semantic relationships. The experimental result shows that the proposed system produce the better summarization result when compared with the existing methods.

Keywords: *Semantic Relationship Refinement, Noun Phrase Analysis, Semantic Web, Ontology, ISA (is a) &HASA (Has a)-tree, WordNet, Natural Language Processing, Automatic term relationship detection, Information Extraction.*

1. Introduction

In today’s world the quantity of information generation is increasing enormously by way of each day. In electronic form the amount of information being generated there has exactly been an ignition in and exposed through the World Wide Web. In the field of research and development this is extraordinarily spot-on where the frequency of the number of papers and articles being distributed is supplementing every day. This presents the requirements for users to be capable to browse through several different documents and rapidly discover the information. A précis submits to an abbreviated or a concentrated translation of a document. It is a brief and to the point depiction of the unique document exactness the most imperative positions enclosed surrounded by it thus removing the feel like to have to understand the full text. For the research purpose have to frequently undergo many research papers with the most spontaneous mode to refer that by reading the abstract and the conclusion together summing up the whole concept. But in various documents they do not automatically contain abstracts as part of it. The abstract is representing the overview of what the document is explaining about and does not essentially list out the most important ideas in a line. So there is a need of sorting out the documents can be achieved if only the summaries can be produced involuntarily to the user who is deciding the documents is useful before reading the whole thing that avoid the time complexity. While reading the news articles situation the summaries can be useful that help the readers can browse through the most important phase of the article as an alternative of reading the whole length article from this we can say that the summarization systems needed mostly in various situations.

That the summarization is compresses the large amounts of source information at the same time they maintain the aim of producing the consolidated main contents of the documents. Next to condensing information and eliminating redundancies this can be used to summative and collecting the information from different source documents with this highlighting the similarities and differences and produce the mostly concentrated information. These advantages are balanced through the fact that summaries can act as filter for irrelevant source information that is based on the user's need. These prospective benefits have motivated in the task of

growth of involuntarily reachable information formulates the automatic summarization for the effective information retrieval mechanism. When there is a lot superior quantity of controlled data obtainable to influence sophisticated applications the

because it is both inclusive and high eminence. With the aim of efficiently utilize extracted data using the spotless and reliable ontology.

Ontology enlargements along with inhabitants are assignments of dominant consequence in semantic web applications. The physical presentation of these responsibilities is labour-exhaustive and consequently cost-intensive, and yields from a maximum level of computerization. For this purpose, the recognition and taking out of terms which is a very important first step to facilitate obtain part in an important role in the field under concern. Essential model of lot of knowledge based applications are mainly depends on the Automatic term recognition which is also known as term extraction that applications such as automatic indexing, terminology mining, knowledge discovery and monitoring, knowledge management and biomedical domains. In this research which is mainly based on the automatic term extraction that is used to construct the refined ontology. Among the term recognition and information extraction there is a comparatively observable meaning moreover appear for other types of information than terms also it may not be always focused on specific domain. Conventionally, the term recognition mainly focused on statistical method; at the same time information extraction is based on machine learning methods.

In our research paper the input topics of the documents will be split as a paragraph where the association rule between each of the topics and paragraphs is achieved. As of the Web the input documents acquired which are normally noisy, necessitating deduplication and stemming and stop word removal is done before they could be organized into ontology. Let denote the input documents as $D = \{d_1, d_2 \dots . d_k\}$. We define the consecutive sentences of D as w blocks where let denote the topic as $T = \{t_1, t_2, t_3 \dots . . t_m\}$ are a set of stemmed terms removal of stop words. The topic is described as $m \times n$ which is called as term block association matrix B in that the columns represented blocks which can be $\{b_1, b_2, b_3 \dots . b_n\}$ decomposed from the topic documents. The association between each of T and D will be analysed and the related terms will be

automatic summarization through the quick

visualization of a Semantic Web only can be recognized. In this we can formulae the feasible automatic search scheme using the machine learning trained extractors which is logical source for extraction

considered by referring the WordNet and its structures makes it is a tool to Natural Language Processing for computational linguistics. In that the Noun phrase relationship will be analysed and forming the tree structure with ISA and HASA relationship by the way the ontology model is reconfigured. Through this the process of TSCAN model is achieved with automatic term extraction process to produce high accuracy in summarized result. The main contribution of the work is as follows:

1. First, we look for the automatic ontology construction. The association relationship is checked with each of the topics and the paragraph of the documents. In that the automatic topic extraction is based on the significant score calculation where the highest score of the terms will be extracted as topics.
2. Second, to attain the refined ontology it should contain a precise ISA tree and the HASA tree, where individual modules are semantically separate and accepted modules are well signified.
3. Third, each topic model should be defined with a well-off scheme, citation an inclusive list of instructive attributes. Topic models should be populated with numerous instances. We note that, while Topic models have rich schemata, many duplicate topic models and attributes exist. This is automatically removed.
4. At last, with the effective refined ontology model the topic summarization is examined and the experimental results are evaluated.

The rest of the paper is organized as follows: In section 1 the introduction about the paper is explained. In section 2 the related work is discussed. In section 3 the proposed work of refined ontology construction for the topic summarization is explained. The experimental results are placed in section 4. In section 5 the conclusion part is done.

2. Related Work

Ontology requires to be constantly modernized through new a perception which is to be a useful tool, relations and lexical resources. Consecutively we refer the several methods and techniques used to refine the domain ontology. It mainly spotlights on approaches appropriate to the researches on lexical attainment and linguistically aggravated mining. For the past, this comprises mounting algorithms and arithmetical methods for satisfying the gaps in obtainable machine legible dictionaries through seeming at the occasion models or declarations in huge content corpora. For the afterward this contains, review linguistic advances to text mining with more linguistic and semantic information useful to colonize ontology. Nicola Guarino [1] dealt with for reasons of information retrieval and extraction, the notional concerns associated to the devise and utilize of such ontologies. Later than a conversation on the character of semantic identical inside a model-theoretical structure, introduced the theme of reserved Ontology, viewing how the thinking of ingredient cover, reliability, uniqueness, and reliance can be of assist in perceptive, organizing and formalizing original ontological features where no background domain knowledge is offered. Marc Ehrig and Steffen Staab [2] believed QOM, Quick Ontology Mapping, as a technique to changeover between effectiveness and efficiency of the mapping creation algorithms also the QOM has subordinate run-time density than existing important advances.

Nenad Stojanovic [3] presented an application of the logic-based query refinement in the incisive for information in an information gateway. The refinement approach is supported on the detection of fundamental associations between queries concerning the addition relation between the responds of these queries. A formal model defined for the query-answering twosomes and employ techniques as of the inductive logic programming for the competent computation of a (lattice) sort between them. In a case revise demonstrated the reimbursement of with this approach in the conventional information retrieval missions. Chris WELTY [4] presented here OWL ontology on behalf of the essential OntoClean divisions, and a tool and method for concerning it to OWL ontologies. Here temporarily touched on the semantic problems oblique by using OWL Full syntax to distinguish the OntoClean meta-properties as properties of OWL Classes, and how that was explained to utilize an off-the-shelf OWL DL reasoner to ensure the OntoClean limitations on the classification. S’everin Lemaignan [5] presented an offer for a developed upper ontology, expected to summary an ordinary semantic mesh in mechanized domain. Convenience of ontologies for data formalization and distribution, particularly in an industrialized environment, are initially conversed. Particulars are given regarding the Web Ontology Language (OWL) and its adequation for ontologies in the modern systems.

Lina Zhou [6] provided an inclusive appraisal and conversation of foremost concerns, confronts, and opportunities in ontology learning. An innovative learning-oriented model proposed for ontology expansion and a structure for ontology learning. Furthermore, for classifying ontology learning advances identified and argued important proportions and techniques. In explanation of the conflict of pasture on choosing ontology erudition progress, recapped domain characteristics that can assist potential ontology knowledge endeavor. Yihong Ding [7] presents a generic architecture for automated ontology reuse. With our implementation of this architecture, we show the practicality of automating ontology generation through ontology reuse. Jens Dietrich [8] propose a novel approach to the formal definition of design patterns that is based on the idea that design patterns are knowledge that is shared across a community and that is by nature distributed and inconsistent. By using the web ontology language (OWL) we are able to properly define intend patterns and some connected notions such as pattern contestant, pattern refinement, and model occasion. Diana MAYNARD [9] described a method for word recognition using linguistic and statistical methods, construction exercise of appropriate information to bootstrap knowledge. Investigated afterward how term recognition techniques can be functional for the wider task of information extraction, production exploit of similarity metrics and background information. Two tools are described that have urbanized to formulate exploit of contextual information to assist the development of regulations for named entity recognition.

Achim Rettinger [10] the reasonable restrictions assumed from ontologies can be exploited to improve and manage the learning task by enforcing depiction logic satisfiability in a latent multi-relational graphical model. To show the possibility of the approach tests using real world social network data in form of SHOIN (D) ontology provided. Lei Liu [12] presented an iterative method extorting ISA relations as of large text for ontology learning. Routine acquisition of ISA relations is an essential trouble in knowledge acquisition from text. Initially, it determines a set of stretches using numerous special lexico-syntactic patterns from free text corpus. Secondly combine exterior coating elimination and in the interior coating gathering for acquiring concepts of constituting ISA relation. Asanee Kawtrakul [15] presented a hybrid advance for (semi-)automatically sensing the challenging relationships and for signifying extra specifically distinct ones. The system consists of three main components: Rule Acquisition, Detection and Suggestion, and Verification. The Refinement Rule Acquisition module is employed to acquire rules exacting by experts and throughout machine learning. The discovery and proposal constituent employs noun phrase analysis and WordNet position to intellect inaccurate relations and to propose more suitable ones supported on the application of the acquired rules. The confirmation module is a tool for proving the proposed associations. From the above mentioned methods there is always lacking in ontology reconfiguration model. These methods only considering the ISA relationship were also other relationships not considered and the automatic topic extraction is not done in all these methods. So there is need of ontology refinement with automatic topic extraction with that this will be summarized.

3. Proposed Methodology

3.1 Automatic Topic Extraction Model

In this phase of implementation model, initially the correlated keywords will be found out via WordNet tool. Calculating similarity values between the topics by means of WordNet. Commencing those topics recognize related topics, when finding document using swarm intelligence techniques those topics are exercised. In this technique applied TF-IDF as weighted values of terms and we measured relative terms during the penetrating time merely. In this approach the weight values of term is calculated with significance with other terms other than if apply the subsequent developments. This significance supports for score computation which is efficiently civilizing the tf-idf based computations. When first forthcoming an unidentified area or necessities document, to obtain a rapid clasp of what the important ideas and entities in the area are, it is frequently productive. This procedure is described as concept recognition, where the theme concept submits to a thing or notion that has an exacting consequence in the area. The most important reason of relating statistical techniques for concept recognition is to grade applicant abstractions based on an exacting measure that provides advanced attains to probable abstraction applicants.

The majority ordinary statistical technique is to suppose the implication of an applicant term from the number of periods it happens in the document. There is a meticulous dispute connected with multitopic terms because the majority methods, together with corpus-based frequency profiling, rely on recognizing individual topics, and count up these separately. There are collocation analysis techniques that can suppose lexical affinities; conversely, while mainly relationship procedures are distinct to compute the pair-wise devotion of topics (t_i, t_j) merely, they cannot be trained for measuring the relationship among more than two topics. In supplies construction, it is reasonably recognizable to assemble domain words, such as software requirements measurement, that encompass more than two topics. Suitably managing such successions is consequently an imperative dispute, as a number of researchers maintain that in particular value. Although multitopic terms can be recognized is key difficulty, in abstraction recognition desire to grade words in order of the significance of their signified abstractions.

In terms of pure frequency, it is common for significant multitopic terms to occur relatively occasionally in a document. Not as good as, no normative corpus of which we are conscious encloses large numbers of multitopic terms. This is for the reason that the majorities such terms are precise to meticulous areas and therefore are doubtful to discover their method into a corpus whose position is to provide as a direct to universal practice of a speech. Therefore, while the corpus-based frequency profiling technique described above works well for terms that are single topics, in practice it doesn't help with multitopic terms. To explain this problem, synthesize a significance value for all terms using a heuristic based on the number of topics of which the term is composed, and the LL (Log-Likelihood) value for each topic. In its simplest form, the significance value for a term $T = \{t_1, t_2, t_3 \ldots\ldots t_m\}$ is specified by the formula which is as follows: $S_T = \frac{\sum_i LL_{w_i}}{l}$

Basically computes the mean of the LL values for all the part topics comprising a multitopic term. Nevertheless, conjecture that not all the topics supply evenly to the implication assessment of the multitopic term of which they are a constituent. The suggestion is supported on supposition that such a term is normally created of a head topic and one or more modifiers. Presuppose that the head topic is the majority important constituent of the term; thus the term is extra important, and the LL value of term should bring more weight than the LL of term. To put up the assumption, the implication equation is modified to incorporate a weight, k_i which dispenses a weight to each topic that is a constituent of the term $S_T = \frac{\sum_i k_i LL_{w_i}}{l}$

It merges a number of obtainable natural language processing (NLP) methods in a work of fiction method to facilitate it to handle both single and multitopic terms, ranked in order of confidence. The evaluation method that employ for Relevance-driven Abstraction Identification (RAI) is one of the main offerings of the occupation, which circumvents the difficulties associated with using expert creature judgment for evaluating how fit the terms revisited against the trouble domain's fundamental abstractions. The significance based topic identification is realized by the subsequent practice:

1. Each topic in the domain document is interpreted with the use of Tree dragger (or) Stanford parser.

2. The set of topics is sorted to take away frequent topics doubtful to suggest concepts.

3. The left over topics are lemmatized to decrease them to their dictionary structure, to subside inflected shapes of topics to a support form.

4. Each theme is dispensed LL rate by concerning the approach described above, using the number of topics which is collected from Wikipedia.

5. Syntactic patterns are functional to the text to categorize multitopic terms.

6. A significance score is derivative for each term by applying the formula of:

$$S_T = \frac{\sum_i k_i LL_{w_i}}{l}$$

7. Recognized topics are sorted which depends on their significance score and the consequential list is revisited.

4. Refined Ontology Construction For Topic Summarization

Ontology refinement intends to fine adjust the ontology and is one of the reconfiguration method. In this step, all compositional sources are present to populate the ontology and refinement has to rely on amorphous sources like documents. Where the assessment procedure relies on information taken out from text for which IE (Information Extraction) is utilized. The subsequent segment specifies these IE tasks and presents the state of the art approaches. The specific necessary tasks are, if a specified word is a synonym of an existing perception occupies to find out the identification of the perception signified by the term and the aspirant synonym notions. If no specific perception is establish, a new impression is fashioned. To find the close relative of a given model we move onto the significant score calculation. If there is a relation between a concept and any other concept in the ontology we will move onto the ISA and HASA relationship. These tasks rely on recognizing portions of information applicable for a specified circumstance defined by the word we are meting out. An assortment of quotations is completed by IR (Information Retrieval). With this, the sentences belonging to the recovered documents will be graded by consequence each time an appropriate form can be created. This model will depend on the mining patterns that we propose to exercise. If a suitable approach for sentence ranking is not establish, a Boolean expression maintained on the words in the representation is employed to recover only on probable suitable sentences and speed up the extraction system.

In this model from the topic extraction phase the each topic will be associated with the set of paragraphs of the input document. There can be the associative block matrix will be formed. After that the related term of will be found out from the WordNet tool. From that the ontology (and the procedure utilized to generate it) will be formed and it must convince numerous criteria. First, we look for automatic ontology construction. Second, the ontology should contain a well-defined ISA hierarchy and HASA hierarchy, where individual classes are semantically separate and ordinary classes are well signified. Third, each class should be defined with a wealthy schema, listing a complete list of instructive attributes. Classes should be populated with numerous instances. Where the relationship and the frequency of each topic is found out within itself and the higher values of topic will be the parent and other will be formed based on that as a child. Then the ISA and HASA relationship tree structure is formed. In this way we are refining the ontology model. By computing a mapping from WordNet concept nodes, for illustration, if both c and d have completely equivalent nodes in WordNet and one WordNet node includes the other (say isaFT(c,d,isaWN)& say has aft(WN)), then this is probable to be extremely predictive for a beginner. Because computing the mapping to WordNet is involved. This procedure is given in Fig 1.

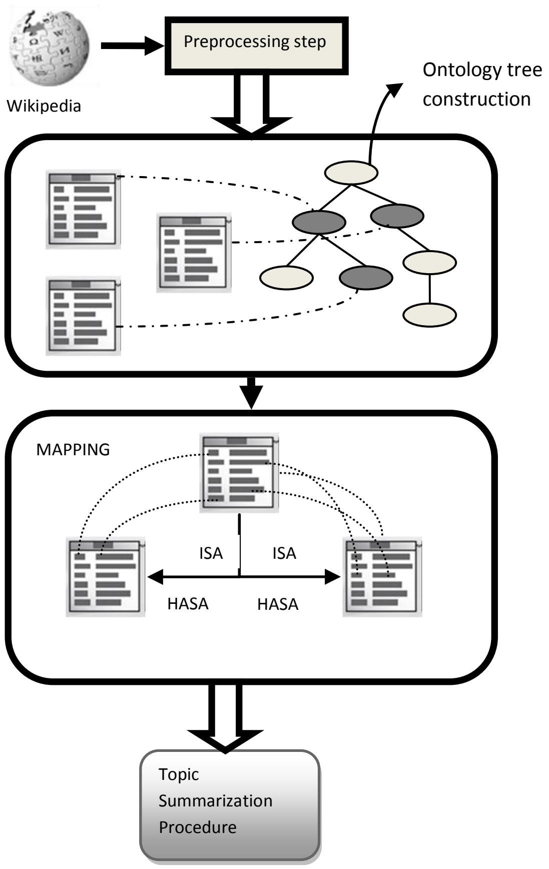


Fig 1: The Overall Architecture Diagram

The procedure is given as algorithm as follows:

```
$Refinement\ (t_1 \dots, t_m, Rel)$
$Input: \{t_1, t_2, t_3 \dots.. t_m\}, rel$
$Output: NewRel$
$For\ (i = 0; i < m; i + +)$
{
$If (Rel = isa\ \ |\ \ Rel = hasa)$
$Then\ if\ \ Agree\ except\ (t_i, Rel)$
$Return\ new\ refined\ relationship\ (NewRel)$
$else\ if\ parent\ is\ compatable(t_i)$
$then\ return\ parent\&child\ relationship$
$else\ if\ is\ wordnet\ hyername\ path(t_i)$
$then\ return\ \ parent\&child\ relationship$
$else\ if\ agree\ revision\ rules(t_i, Rel)$
$Then\ return\ new\ relationship$
}
$End\ the\ result$
```

5. Experimental Results

An experiment testing with the training rules technique using some examples for few semantic relationships. Using topics which are automatic extraction from the documents, done the noun phrase analysis, using WordNet for the experimental tests. Where the precision, recall rate and Fmeasure is measured and analyzed with existing methods. Each will be analyzed and described as follows.

5.1 Precision Rate

We analyze and compare the performance offered by Ontology & NonOntology with Automatic Topic Extraction and Refined ontology with Automatic Topic Extraction. Here if the number of document size increased the precision accuracy also increased linearly. The precision accuracy of the proposed method is high. Based on the comparison and the results from the experiment shows the proposed approach works better than the other existing systems with higher rate. The values are represented in the Table 1.

Table 1: Precision vs. Number of Documents

S. No	No. of Documents	Ontology & NonOntology with Automatic Topic Extraction	Refined ontology with Automatic Topic Extraction
1	10	0.75	0.85
2	20	0.69	0.75
3	30	0.58	0.64
4	40	0.45	0.56
5	50	0.39	0.49
6	60	0.33	0.47

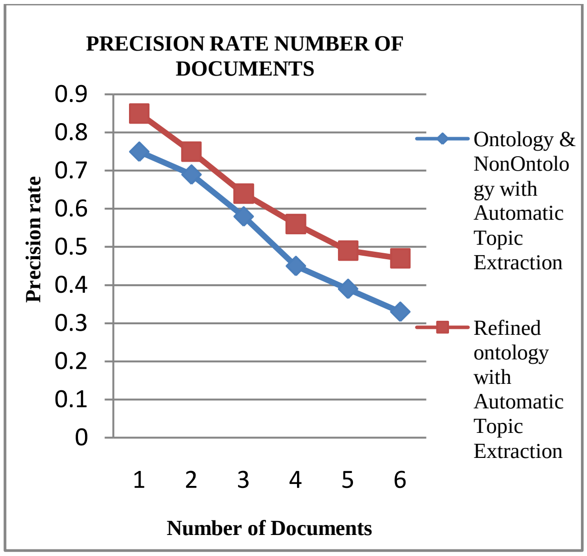


Fig 2: Precision Rate Number of Documents

In this graph we have chosen two parameters called number of Document and precision which is help to analyze the existing system and proposed systems. The precision parameter will be the Y axis and the number of document parameter will be the X axis. The blue line represents the proposed system and the red line represents the existingsystem. From this graph we see the precision of the proposed system is higher than the existing system. Through this we can conclude that the proposed system has the effective precision rate.

5.2 Recall Rate

This graph shows the recall rate of existing and proposed system based on two parameters of recall and number of Document. From the graph we can see that, when the number of number of Document is improved the recall rate also improved in proposed system but when the number of number of Document is improved the recall rate is reduced in existing system than the proposed system. From this graph we can say that the recall rate of proposed system is increased which will be the best one. The values of this recall rate are given below:

Table 2: Recall vs. Number of Documents

SNO	Number of Documents	Ontology & NonOntology with Automatic Topic Extraction	Refined ontology with Automatic Topic Extraction
1	10	0.75	0.87
2	20	0.67	0.77
3	30	0.58	0.69
4	40	0.48	0.58
5	50	0.41	0.51
6	60	0.38	0.48

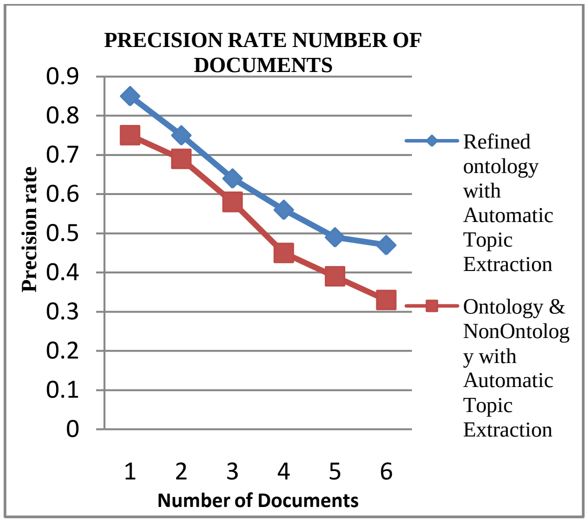


Fig 3: Recall Vs. Number Of Documents

In this graph we have chosen two parameters called number of Document and recall which is help to analyze the existing system and proposed systems on the basis of recall. In X axis the Number of document parameter has been taken and in Y axis recall parameter has been taken. From this graph we see the recall rate of the proposed system is in peak than the existing system. Through this we can conclude that the proposed system has the effective recall.

5.3 Fmeasure Rate

This graph shows the Fmeasure rate of existing and proposed system based on two parameters of Fmeasure and number of Document. From the graph we can see that, when the number of number of Document is improved the Fmeasure rate also improved in proposed system but when the number of number of Document is improved the Fmeasure rate is reduced in existing system than the proposed system. From this graph we can say that the Fmeasure rate of proposed system is increased which will be the best one. The values of this Fmeasure rate are given below:

Table 3: Fmeasure vs. Number of Documents

SNO	Number of Documents	Ontology & NonOntology with Automatic Topic Extraction	Refined ontology with Automatic Topic Extraction
1	10	0.67	0.75
2	20	0.56	0.69
3	30	0.48	0.62
4	40	0.4	0.53
5	50	0.34	0.45
6	60	0.23	0.39

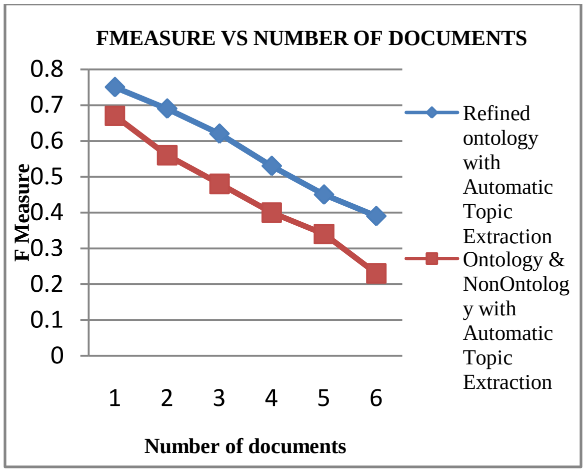


Fig 4: Fmeasure vs. Number of Documents

In this graph we have chosen two parameters called number of Document and recall which is help to analyze the existing system and proposed systems on the basis of Fmeasure. In X axis the Number of document parameter has been taken and in Y axis Fmeasure parameter has been taken. From this graph we see the Fmeasure of the proposed system is in peak than the existing system. Through this we can conclude that the proposed system has the effective Fmeasure.

6. Conclusion

This manuscript presents the three methodologies for data preprocessing and semantic relationship refinement and ontology refinement to resolve the difficulty of producing well-defined semantics from inadequately distinct or underspecified semantics in documents. The organization refines the semantic associations although noun phrase analysis, WordNet alignment, and semantic relationship topics, a number of produced by authorities and others produced from explained illustrations by an inductive statistical machine learning arrangement. Ontologies with accurate semantic are central for improving retrieval systems, for automating procedures from side to side machine reasoning, and for the Semantic Web. The refinement of ontology is very difficult to form and this methodology is efficient to make the configuration of ontology and automatic extraction of topics automatically through significant score calculation. The experimental results show that the proposed method is more efficient than the existing method.

REFERENCES

[1] Nicola Guarino," Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration", LADSEB-CNR, National Research Council, Corso Stati Uniti 4, I-35127 Padova-1997.

[2] Marc Ehrig and Steffen Staab," QOM - Quick Ontology Mapping" Institute AIFB, University of Karlsruhe, 2004.

[3] Nenad Stojanovic,"On the Query Refinement in Searching a Bibliographic Database", Internationale Tagung irtschaftsinformatik (WI2005) 23.-25. February 2005 in Bamberg.

[4] Chris WELTY,"OntOWLClean: Cleaning OWL ontologies with OWL", IBM Watson Research Center, NY, USA 2006.

[5] S´everin Lemaignan, Ali Siadat, Jean-Yves Dantan, Anatoli,"MASON: A Proposal For An Ontology Of Manufacturing Domain", Proceedings of the IEEE Workshop on Distributed Intelligent Systems: Collective Intelligence and Its Applications (DIS'06) 0-7695-2589-X/06 $20.00 © 2006 IEEE

[6] Lina Zhou, "Ontology learning: state of the art and open issues", Inf Technol Manage (2007) 8:241–252 DOI 10.1007/s10799-007-0019-5, Published online: 24 March 2007.

[7] Yihong Ding, Deryle Lonsdale, David W. Embley, Martin Hepp, and Li Xu,"Generating Ontologies via Language Components and Ontology Reuse", This work was partially funded under National Science Foundataion Information and Intelligent Systems grant IIS-0414644 2007.

[8] Jens Dietrich, Chris Elgar,"Towards a web of patterns", Massey University, Institute of Information Sciences and Technology, Palmerston North, New Zealand 2007.

[9] Diana MAYNARD, Yaoyong LI and Wim PETERS,"NLP Techniques for Term Extraction and Ontology Population", Corresponding Author: Diana

Maynard: Dept. of Computer Science, University of Sheffield, 211 Portobello St, Sheffield, UK; 2008.

[10] Achim Rettinger, Matthias Nickles, and Volker Tresp,”Statistical Relational Learning with Formal Ontologies’, Technische Universitat Munchen, Germany, University of Bath, United Kingdom, Siemens AG, CT, IC, Learning Systems, Germany,2009.

[11] Elena Beisswanger,”Exploiting Relation Extraction for Ontology Alignment”, Jena University Language and Information Engineering (JULIE) Lab and Friedrich-Schiller-Universitat Jena, 2010.

[12] Lei Liu, Sen Zhang, Luhong Diao, Cungen Cao.”An Iterative Method of Extracting Chinese ISA Relations for Ontology Learning”, Journal of Computers, College of Applied Sciences, Beijing University of Technology, Beijing, China, Vol. 5, NO. 6, June 2010.

[13] Javier Lacasta, Javier Nogueras-Iso, Jacques Teller, and Gilles Falquet,”Transformation of a keyword indexed collection into a semantic repository: applicability to the urban domain” International Conference on Theory and Practice of Digital Libraries 2011 - TPDL 2011.

[14] Farheen Siddiqui, M. Afshar Alam,”Web Ontology Language Design and Related Tools: A Survey”, Journal of Emerging Technologies in Web Intelligence, Vol. 3, No. 1, February 2011 Academy Publisher doi:10.4304/jetwi.3.1.47-59

[15] Asanee Kawtrakul , Aurawan Imsombut , Aree Thunyakijjanukit c, Dagobert Soergel, Anita Liang, Margherita Sini, Gudrun Johannsen g, and Johannes Keizer, “Automatic Term Relationship Cleaning and Refinement for AGROVOC”,March 2011.

[16] Ilaria Corda, Vania Dimitrova, Brandon Bennett,”An Ontological Approach to Unveiling Connections between Historical Events”, School of Computing, University of Leeds, United Kingdom ilaria,vania,brandon@comp.leeds.ac.uk, September 30, 2012

A weaving process to define requirements for Cooperative Information System

Mohamed Amroune[1], Jean Michel Inglebert[2], Nacereddine Zarour[1] and Pierre Jean Charrel[2]

[1] **University of Constantine, Algeria**

[2] **University of Toulouse II, France**

Abstract :

The development of a Cooperative Information System (CIS) becomes more and more complex, new challenges arise for managing this complexity. So, the aspect paradigm is regarded as a promising software development technique which can reduce the complexity and cost of developing large software systems. This opportunity can be used to develop a CIS able to support the interconnection of organizations information systems in order to ensure a common global service and to support the tempo of change in the business world that is increasing at an exponential level.

We previously proposed an approach named AspeCiS (An Aspect-oriented Approach to Develop a Cooperative Information System) to develop a Cooperative Information System from existing Information Systems by using their artifacts such as existing requirements, and design. In this approach we have studied how to elicit CIS Requirements called Cooperative Requirements in AspeCiS. In this paper we propose a weaving process to define these requirements by reusing existing requirements and new aspectual requirements that we define to modify these requirements in order to be reused.

Keywords: *Requirements engineering, Aspect, Cooperative Information System, Weaving*

1. Introduction

Today the organizations evolve in new environments characterized by changes in customer demands, the increased competition, communications performance, etc. In order to cope with these business conditions, enterprises migrate to inter-organizational relationships [4], [5], [6] as a way to adapt to their new environment, gain competitive advantage, and, increase their efficiency. So, the enterprise cooperation is not an easy task. It requires an effective Cooperative Information System (CIS) to support this inter-enterprise cooperation.

The Software Engineering discipline has emerged, to give response to the increasing demands for software development. So, it proposed structured processes and activities to facilitate the development of software. The initial phase of Software Engineering is Requirements Engineering. We suggest improving requirements engineering in CIS through the early identification of base concerns and crosscutting concerns (that affect several modularization units). In this strategy, managing of complexity is better supported than by traditional non aspect-oriented approaches. Thus, our research aims at developing a new approach called AspeCis, which ensures the effectiveness and efficiency of business cooperation based on the Aspect concept.

In AspeCiS, when a new requirement cannot be achieved directly by an existing Information System (IS), AspeCiS composes requirements issued from other ISs in order to fulfill this requirement. The main objectives of AspeCiS are: (i) to separate existing requirements from new requirements in the CIS; (ii) to provide a high degree of functional reuse, which helps to build again the same requirements on other existing ISs. AspeCiS includes three main phases (cf. 1): (i) elicitation and analysis of CRs, (ii) models weaving (conception of CRs models), and (iii) models to code (preparation of the implementation phase).

In our previous work we have exhibited a process to elicit requirements related to the CIS to be developed. In this paper we illustrate how existing requirements can be used by a weaving process that we propose in this paper to define requirements related to the CIS to be developed.

The remainder of the paper is organized as follows:
Section 2 presents an overview of AspeCiS. The AspeCiS weaving process is detailed in section 3. Section 4 draws some examples. Section 5 provides a summary of the paper and a brief overview of the continuation of this work.

2. AspeCiS approach: An overview

The CIS becomes more and more complex, new challenges arise for managing this complexity, for this reason general purpose methods no longer suffice. An important and particular software engineering phase we address in our research is requirements engineering (RE). The reason is that the development of a system is heavily determined by its requirement elicitation, specification and the design decision that derive from it. The later development phases will be based on the requirements elicitation and analysis.

AspeCiS propose to improve the requirements definition through the early identification of base concerns and crosscutting concerns. AspeCiS includes three main phases (cf. 1): (i) elicitation and analysis of CRs, (ii) models weaving (conception of CRs models), and (iii) models to code (preparation of the implementation phase).

Phase I: Elicitation and analysis of CRs. This phase is composed of four steps which are:
(1) the definition of CRs, (2) the refinement of CRs, (3) the formulation of CRs depending on the ERs and possibly with the definition of some aspectual requirements (ARs), (4) the selection of a set of Operators to be used to weave ERs and the ARs to define the CRs as can be seen in the figure 1.

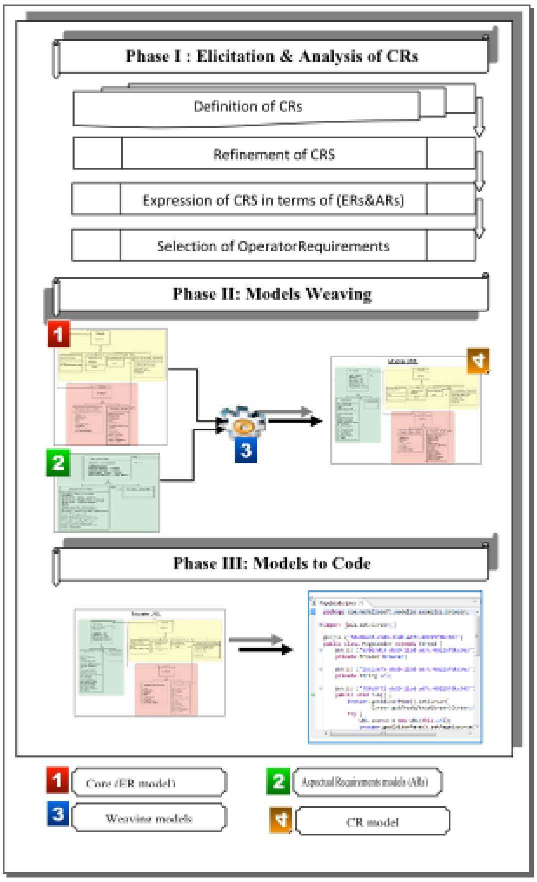

Phase I : Elicitation & Analysis of CRs
Definition of CRs
Refinement of CRS
Expression of CRS in terms of (ERs&ARs)
Selection of OperatorRequirements
Phase II: Models Weaving
1
2
3
4
Phase III: Models to Code
1
Core (ER model)
2
Aspectual Requirements models (ARs)
3
Weaving models
4
CR model

Fig. 1 Synopsis of AspeCiS

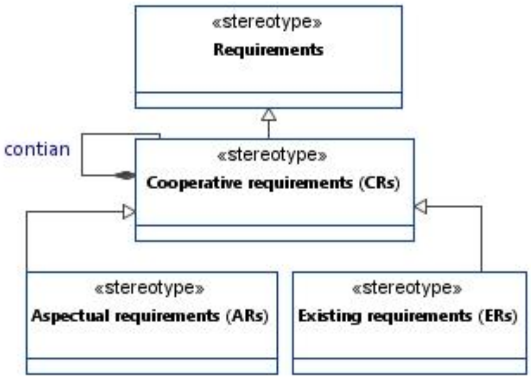

«stereotype»
Requirements
contian
«stereotype»
Cooperative requirements (CRs)
«stereotype»
Aspectual requirements (ARs)
«stereotype»
Existing requirements (ERs)

Fig. 2 Cooperative requirements metamodel

Phase II: Development of CRs models. The ARs should be composed with ERs models to produce CRs models. This phase includes a conflict resolution task that can appear during the requirements composition process. We proposed in [3] a conflict resolution process among aspectual requirements during the requirements engineering level: a priority value is computed for each AR, and it allows identifying a dominant AR on the basis of stakeholder priority. This process is more formal than those currently proposed, which requires a trade-off negotiation to resolve conflicts.

Phase III: Preparing the implementation phase. The purpose of this phase is to transform models into code templates.

2.1 The concept of Requirements in AspeCiS

Several definitions of requirement exist in the literature [7], but we adopt the following ones to differentiate between requirements in AspeCiS. So, in AspeCiS tree kind of requirements are defined.

Existing Requirements (ERs). They are statements of services or constraints provided by an existing system, which define how the system should react to particular inputs and how the system should behave in particular situations as shown in the figure 2.

Aspectual Requirements (ARs). They are concerns that cut across other existing requirements by a weaving operation in order to modify ERs to be reused to define CRs.
Cooperative Requirements (CRs). They are goal requirements that will be refined to relate on ERs and eventually ARs, exhibiting what parts of existing systems requirements will be reused.

In our previous work we have developed the first phase of AspeCiS that consists of a definition of cooperative requirements.

In the next section we present a weaving process that we used to develop a Cooperative Requirements (CRE) models.

3. A weaving process to define requirements for CIS

The weaving concept is used to support such a decoupling among models. The weaving concept is not new and the definition of model weaving considered in this paper is an extension of the generic metamodel weaving proposed by Didonet Del Fabro et al. in [8]. The general operational context of this generic metamodel weaving is depicted in Figure 3. It consists of the production of a weaving model *WM* representing the mapping between two metamodels: a left meatamodel *LeftMM* and a right metamodel *RightMM*. The *WM* model should be conform to a specific weaving metamodel *WMM*.

The generic metamodel weaving proposed by Didonet Del Fabro et al must be extended, to be used in our context. This extension (cf. figure 3), consists of the definition of a Core metamodel (*ALeftmm*), an Aspectual Requirements metamodel (*ARightmm*) and a weaving metamodel called *AWM* (for AspeCiS Weaving model) specific to our approach. So, the Core metamodel represents an ERs metamodel. It is conformed to the UML metamodel. We present in this paper the weaving model specific to AspeCiS (*AWM*).

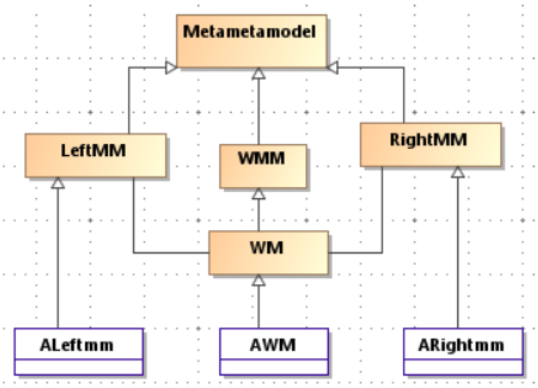


Fig. 3 Atlas model weaver AMW extension

The weaving models (*AWM*) is used to capture different kinds of links between input model elements (*ALeftmm* & *ARightmm*). The links have different semantics, depending on the application scenario. For instance (Attribute, Class) is a kind of link. It means that an attribute from Core model is added to a class from Aspectual Requirements model. The semantic of links is not in the scope of this paper.

Before presenting the *AWM* (AspeCiS metamodel), we briefly present the weaving metamodel (WM) proposed by Didonet Del Fabro et al. in [8]. This metamodel is illustrated in the Figure 4. This metamodel is composed by the following elements:

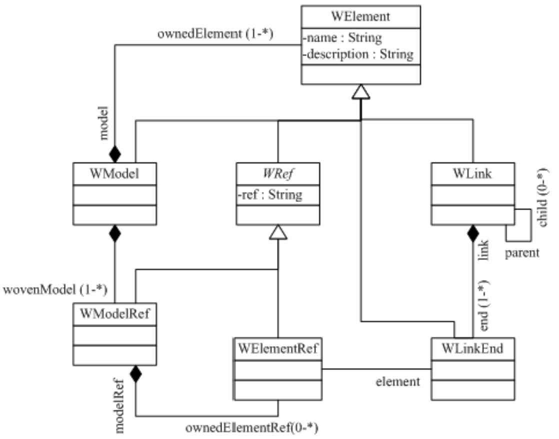


Fig. 4 Atlas Model Weaver (AMW) metamodel

- *WElement* is the base element from which all other elements inherit. It has a name and a description.
- *WModel* represents the root element that contains all model elements. It is composed by the weaving elements and the references to woven models.
- *WLink* express a link between model elements, i.e., it has a simple linking semantics. To be able to express different link types and semantics, this element is extended with different metamodels.
- *WLinkEnd* represents a linked model element.
- *WElementRef* is associated with an identification function over the related elements. The function takes as parameter the model element to be linked and returns a unique identifier for this element.
- *WModelRef* is similar to *WElementRef* element, but it references an entire model.

The *AWM* is produced and depicted in Figure 5. So, we define a model weaving which is the Weaving-Core_Aspect. It is composed of two models (Core & Aspect), as shown in the extract of KM3 following code. KM3 is a simple textual language to define metamodels [9].

```
class Weaving-Core_Aspect extends WModel {
reference Core container :
WModelRef;
reference Aspect container : WModelRef;
}
```

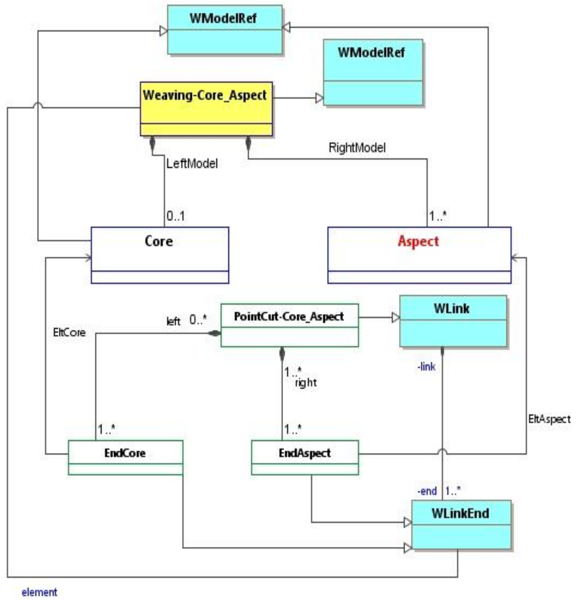


Fig. 5 AspeCiS Weaving Metamodel (AWM)

With respect to this metamodel, a Weaving-Core_Aspect consists of two models (Core model & Aspect model) related through weaving links (Pointcut-Core_Aspect). The Core model (LeftModel) is an extension of WModelRef, it represents a model of ERs. The Aspect model (RightModel) represents a model of a aspectual requirements. This model is also an extension of WModelRef.

The Weaving-Core_Aspect is also composed of the Pointcut (PointcutCoreAspect), which is an extension of WLink. The Pointcut-Core_Aspect is composed by two elements (EndAspect, EndCore), these elements are an extension of WLinkEnd. The EndCore element represents an artefact of Core model, and the EndAspect element represents an artefact of the Aspectual requirements.

4. The AWM Usage

In the previous sections, we have presented a weaving process to produce CRs models. It is now convenient to better illustrate the use of this process.

This example illustrates a part of the university students management system. It consists of the management of the student's subscription in the High Graduate School in Algerian universities. The Hight Graduate School composes of several universities; it aims to assure a high formation of students. After completion of courses of study, the student receives a doctor's degree.

We intend to build a CIS able to manage a cooperative project involving several universities to provide High Graduate School. Each university is supported by its existing IS. The new CIS is built on the basis of existing ISs (more details of this example are presented in [2]).

In this section, we illustrate how to weave two models (M1, M2) using AWM, in order to produce a model of CRs. At the requirements level of the existing ISs, the student subscription requirement is defined as:
ER1*= "Every student may have a second subscription in the same university"*. However, in the CIS, the **CR** is defined as:
CR1*= "Every student can have a second subscription in the same university provided that the number of hours of the second speciality does not exceed 50% of the number of hours of the first one".*

The existing ISs allow a second subscription in the same university. So, in order to participate in the Height Graduate School, each university must respect the constraint, of the number of hours for the second subscription, imposed by the Height Graduate School's regulation. This constraint is defined in the *CR1* cited previously. Furthermore, the CIS to be developed, to support the management of this Height Graduate School, will be developed by reusing the existing ISs after some modifications.

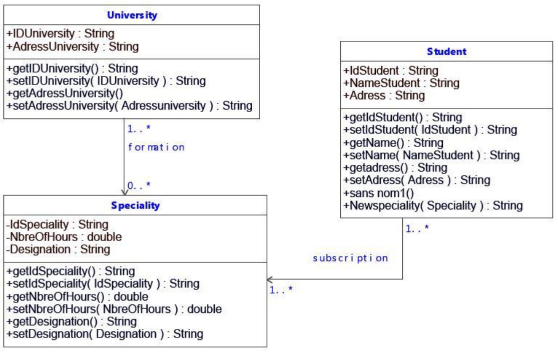

University
+IDUniversity : String
+AdressUniversity : String
+getIDUniversity() : String
+setIDUniversity(IDUniversity) : String
+getAdressUniversity()
+setAdressUniversity(Adressuniversity) : String
1..*
formation
0..*
Speciality
-IdSpeciality : String
-NbreOfHours : double
-Designation : String
+getIdSpeciality() : String
+setIdSpeciality(IdSpeciality) : String
+getNbreOfHours() : double
+setNbreOfHours(NbreOfHours) : double
+getDesignation() : String
+setDesignation(Designation) : String
Student
+IdStudent : String
+NameStudent : String
+Adress : String
+getIdStudent() : String
+setIdStudent(IdStudent) : String
+getName() : String
+setName(NameStudent) : String
+getadress() : String
+setAdress(Adress) : String
+sans nom1()
+Newspeciality(Speciality) : String
1..*
subscription
1..*

Fig. 6 A Core model

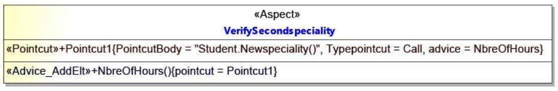

«Aspect»
VerifySecondspeciality
«Pointcut»+Pointcut1{PointcutBody = "Student.Newspeciality()", Typepointcut = Call, advice = NbreOfHours}
«Advice_AddElt»+NbreOfHours(){pointcut = Pointcut1}

Fig 7 An Aspect model

At the model level of ISs, M1 models the *ER1* (see figure 6). It represents the core model that models the student subscription in the Height Graduate School. M1 is a class diagram conforms to UML class diagram metamodel. M1 composes of three entities which are: University, Student and Speciality. M2 (see figure 7). Represents aspet, it conforms to the AspectOperator-Requirement metamodel that we presented in [1].

In this example, the aspect model contains the advice *advice_addElt* witch role is to verify the number of hours before the call of the function *Student.NewSpeciality()*. These informations are defined in the Pointcut *Pointcut1* through the (BodyAdvice) and the (*Typepointcut= "call")* (see figure 6).

In this example, we use the *Weaving-Core_Aspect* to add two operations to M1, especially to the *Student* class. These operations are called before the call of the *Student.NewSubscription()* operation, in order to add a second subscription. The first operation *VerifySpecialty.NbreOfHours(IdSpecialty)* consists of the computation of the number of hours for the new subscription, the result of this operation is used by the second operation get *SecondSpecialty()* to verify the constraint imposed to authorize or not a second subscription.

5. Conclusion

In a previous work [2], we proposed an approach named AspeCiS to develop a Cooperative IS (CIS) from existing ISs by using their artifacts such as requirements, and design. We developed a process to elicit CRs.

In the present work, we proposed a Model driven engineering weaving process to be used to develop a cooperative e requirements models. This process is based on the use of input models which are *Core* and *Aspect* models. The *Core* models represent existing requirements, and the aspect models represent Crosscutting Requirements. These crosscutting requirements are considered as aspectual requirements and must be woven with existing requirements in order to define cooperative requirements related to the CIS to be developed. The proposed weaving metamodel is used to capture different kinds of links between model elements. These links have different semantics. We will define these semantics in our future work.

6. Acknowledgements

This research is partially supported by the PHC TASSILI project under the number 10MDU817. We thank the anonymous reviewers for providing valuable comments.

References

[1] M. Amroune, N. Zarour, P-J. Charrel, and J-M. Inglebert, “An UML profile to design aspects in ASPECIS approach”, in IEEE second Int. workshop on advanced information systems for enterprises Constantine, Algeria. 2012, pp. 34-39.

[2] M. Amroune, J.-M. Inglebert, N. Zarour, and P.-J. Charrel, “Aspecis: An aspect-oriented approach to develop a cooperative information system,” in Model and Data Engineering - First International Conference, MEDI 2011, O´ bidos, Portugal, ser. Lecture Notes in Computer Science, vol. 6918. Springer, 2011, pp. 122–132.

[3] M. Amroune, J. M. Inglebert, N. Zarour, and P. J. Charrel, “Article: A conflict resolution process in aspecis approach,” International Journal of Computer Applications, vol. 44, no. 10, pp. 14–21, April 2012, published by Foundation of Computer Science, New York, USA.

[4] K. V. Andersen, J. K. Debenham, and R. Wagner, Eds., “How to design a Loose Inter-Organizational Workflow: An illustrative case study”, ser. Lecture Notes in Computer Science, vol. 3588. Springer, 2005.

[5] P. W. P. J. Grefen, H. Ludwig, A. Dan, and S. Angelov, “An analysis of web services support for dynamic business process outsourcing,” Information & Software Technology, vol. 48, no. 11, pp. 1115–1134, 2006.

[6] P. W. P. J. Grefen, N. Mehandjiev, G. Kouvas, G. Weichhart, and R. Eshuis, “Dynamic business network process management in instant virtual enterprises,” Computers in Industry, vol. 60, no. 2, pp. 86–103, 2009.

[7] 9B. Nuseibeh and S. Easterbrook, “Requirements engineering: a roadmap,” in Proceedings of the Conference on The Future of Software Engineering, ser. ICSE ’00. New York, NY, USA: ACM, 2000, pp. 35–46. [Online]. Available: http://doi.acm.org/10.1145/336512.336523

[8] 12M. D. Del Fabro, J. B´ezivin, F. Jouault, E. Breton, and G. Gueltas, “AMW: a Generic Model Weaver,” in Procs. of IDM05, 2005.

[9] A. Y. Halevy, Z. G. Ives, and A. Doan, Eds., “Rondo: a programming platform for generic model management”. ACM, 2003.

Mohamed Amroune is currently a Ph.D. Student, at the University of Toulouse II, Mirail, France and University of Tebessa, Algeria. He received his Engineer and Magister degrees in Software Engineering and Artificial Intelligence & Data Bases from the USTHB University of Algiers , Algeria, and the University of Tebessa, Algeria, in 1993 and 2007, respectively. His research interests include Information System, Requirements Engineering, Cooperation and Aspect oriented software development.

Jean Michel inglebert is a Doctor at the Computer Sciences Department of University of Toulouse, France. His current research activities are conducted at the IRIT laboratory, University of Toulouse.

.

Nacereddine zarour is a Professor at the Computer Sciences Department of University Mentouri, Constantine, Algeria. His current research activities are conducted at the LIRE laboratory, University of Constantine. He heads the project of PHC CMEP Tassili with IRIT laboratory of Toulouse 2 University. His research interests include advanced information systems, particularly cooperative information systems, architectures (based on SOA, SMA, ..), and requirements engineering.

Pierre Jean Charrel is a Professor at the Computer Sciences Department of University of Toulouse, France. His current research activities are conducted at the IRIT laboratory, University of Toulouse. He currently heads PHC CMEP Tassili project nr 10MDU817 with LIRE Laboratory of Mentouri University of Constantine, Algeria. His research interests include requirements engineering and knowledge engineering, in the context of cooperative information systems.

The Number of Terms and Documents for Pseudo-Relevant Feedback for Ad-hoc Information Retrieval

Mohammed El Amine Abderrahim[1], Saïd Benameur[2], Mohammed Alaeddine Abderrahim[3]

**[1] University of Tlemcen,
Laboratory of Arabic Natural Language Processing
BP 230 Chetouane, Algeria**

**[2] University of Tlemcen,
Laboratory of Arabic Natural Language Processing
BP 230 Chetouane, Algeria**

**[3] University of Tlemcen,
Laboratory of Arabic Natural Language Processing
BP 230 Chetouane, Algeria**

Abstract

In Information Retrieval System (IRS), the Automatic Relevance Feedback (ARF) is a query reformulation technique that modifies the initial one without the user intervention. It is applied mainly through the addition of terms coming from the external resources such as the ontologies and or the results of the current research.

In this context we are mainly interested in the local analysis technique for the ARF in ad-hoc IRS on Arabic documents. In this article, we have examined the impact of the variation of the two parameters implied in this technique, that is to say, the number of the documents «D» and the number of terms «T», on an Arabic IRS performance.

The experimentation, carried out on an Arabic corpus text, enables us to deduce that there are queries which are not easily improvable with the query reformulation. In addition, the success of the ARF is due mainly to the selection of a sufficient number of documents D and to the extraction of a very reduced set of relevant terms T for retrieval.

Keywords: Arabic Information Retrieval, Pseudo Relevance Feedback, Local Analysis, Query Reformulation.

1. Introduction

In order to reduce the distance between the system's relevance and that of the user, an IRS can lead the user to a useful formulation of his needs. The suggested solutions can appear in various approaches i.e.: the Query Reformulation (QR), the reestablishment of documents, the combining in between all the results set of different IRS and or the integration of the user's profile in the retrieval information process.

In this article, we are basically interested in the QR. The approaches for the latter are numerous and can be classified according to the resources used in three great classes. (see Fig. 1):

- The use of the external resources: this consists of using the external resources as the ontologies or the thesaurus to find other terms similar to those in the initial query.
- The global analysis: this approach aims at analyzing the set of documents collection, so that we can extract the pertinent terms to be added to the initial query. So, we get two techniques that are: similarity thesaurus and the statistical thesaurus [5].
- The local analysis: the documents coming back to a query after demand are analyzed so that we can extract other pertinent terms that are able to extend the query. The studies applied in [3, 5, 11, 15, 16, 19, 21, 22, 27, 28, 29, 31] show that conversely to the global analysis, the local one is simpler to be realized and allows an improvement of IRS performances. Two techniques are proposed for the local analysis in the literature [5]:
  - The local clustering: this consists of constructing a matrix of association that quantifies the correlation relations of terms got from the set of documents that were returned in response to the initial query. According to the method of construction of the correlation relations, we notice three types of clusters: association clusters, metric clusters and scalar clusters. We are interested to this technique. So, we are going to develop this technique and we implement the first type of clusters i.e.: association clusters.
  - The local context analysis: it consists of using the concepts instead of the keywords to represent the document [16].

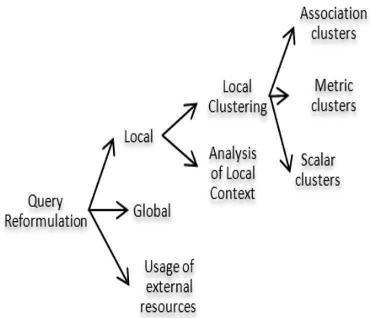


Fig. 1 The approaches for the query reformulation.

We distinguish two manners for the QR, the first one is based on an automatic process, it occurs without the intervention of the user. On the contrary, the second one is based on the interactive process between the IRS and the user. This is an interactive query reformulation (IQR). The experiments applied on behalf of this second manner have shown that it can allow the improvement of the results precision, all the same its efficiency is linked strongly to the attitudes of the users and their way of judging the documents pertinence [5, 9, 14, 20, 22, 26].

In order to avoid the heaviness of the judgment step of the documents relevance, this task has been automatized, and so the IRS considers the «D» first documents retrieved initially as pertinent. By these documents «D» the system establishes a list of «T» terms for the query expansion. This new form of pertinence reinjection is called blind, pseudo or ad-hoc (PRF). According to [9] and [12], this approach allowed to improve the results in comparison to the approach of global analysis. In fact, the attempts applied in the area of the campaign of evaluation TREC4 for the system SMART [12] and the work done by [11] assure that the technique of PRF leads to a gain in precision of responses of about 10%. In this article and in the area of the PRF, we suggest to examine the influence of the variation of «D» and «T» on the performances of an Arabic IRS. After this introduction, we are going to describe the PRF technique and we exhibit our experiment and we will discuss about the result.

2. The Relevance Feedback (RF)

The RF is a technique that consists of modifying the initial query of the user by adding some terms got from the list of documents retrieved in the IRS. It is based on three steps (see Fig. 2):

- The samples: it consists of selecting a set of «D» documents (samples) among the returned ones by the IRS and judged as pertinent.
- The extraction of evidences: it consists of establishing the list of «T» terms judged pertinent for the expansion of the query.
- The rewriting of the query: It consists of enriching the query with terms found in the previous step.

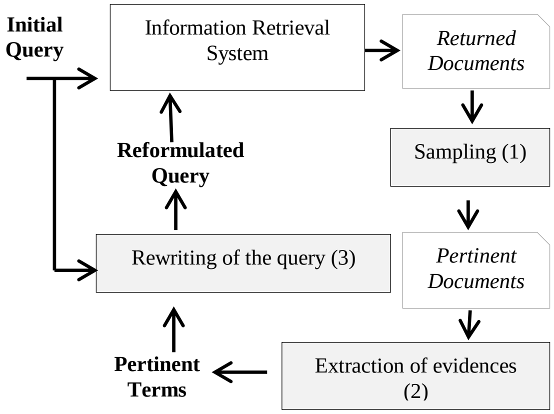


Fig. 2 The three phases of the relevance feedback.

The PRF is an approach for the RF that lies on the automatic sampling, otherwise, instead of judging explicitly the documents, we suppose that the «D» first documents are relevant.

The main problem with the PRF technique is summed up in the determination of two parameters «D» and «T». In the majority part of the work, these two parameters are chosen arbitrary, for instance:

- [23] has used: D=80 and T=10.
- [25] has used: D=20 and T=10.
- [21] has used: D= (5, 10, 25, 50, 75 and 100) and T= (10, 25, 50, 75 and 100).
- [16] has used: D= (5, 10, 20, 30, 50 and 100) and T= 70.
- [10] and [13] have used: D= (6, 10, 20, 25 et 30) and T= (1, 2, 3, … and 100).
- [12] proposed to use a «D» between 5 and 10 and recommend «D»=5 as an optimum value.
- [27] have used D=10 and T=30.
- [30] have used D=20.
- [4] and [2] have used: D=5.

On the other hand all the works that point at the effect analysis of the two parameters «D» and «T» on the IRS performance are not numerous. We find as examples:

- The experiment of [7, 8, 9] consists of analyzing the average precision of the IRS for all the combinations of the values: «D» and «T» between 1 and 100 (10 000 combinations). A list of 50 queries of the TREC8 was used with a collection of documents composed of newspapers articles. The results of the experiment have shown that there was an improvement in the average precision of the IRS for the values of «D» between 8 and 16 and the values of «T» between 7 and 42. It should be noted that for the same experiment but with a different collection of documents (TREC9 WT) composed of web pages, [9] has not obtained an improvement in the average precision. To explain that we find out that the web pages do not contain a pure content. They contain mainly some different objects that have nothing to deal with the content of the page such as: hyperlinks, decorations (images, animations, logos… etc.), interaction and other information (copyrights, information of contacts…). As for the matching of the queries, all these informations are considered as a noise and consequently degrading the performances of the IRS. This kind of content requires some approaches as the one used by [23] who has obtained some gains of performances at the order of 27%. The work set by [7, 8, 9] have led to two important results:

- The values D=15 and T=13 give the best value of average precision of the IRS.
- The estimated parameters (D, T) of one query are not obligatory suitable for another.

- In [10] many factors on which an IRS depends were studied. In the context of the PRF the variation of «D» and «T» was also examined. These experiments have not allowed to extract a conclusion concerning the estimated values for both «D» and «T». The final objective of the experiment was rather to analyze and try to understand the interaction between the IRS and the topics.
- [30] shows that the use of 10 terms leads to an improvement of more than half of the queries. However, there is no optimal and fixed number of terms for all queries.

The works concerning the RF evaluation for the Arabic texts are not enough, we find as examples:

- The researches done in [18] have shown that the manual RF by weighting the query terms have led to an improvement of the Arabic IRS performances (recall and precision). For his experiment, [18] has used a corpus of 242 documents and a list of 9 queries.
- The researches done by [6] linked to the query expansion with terms got from a thesaurus showed an improvement in the recall of the Arabic IRS. We notice that the corpus used was the Quran.
- The work achieved by [17] showed that the use of a thesaurus ameliorates considerably (18%) the performances of an Arabic IRS. It has also shown that the use of the roots is more effective than the use of schemes for the Arabic texts indexation.
- The work realized by [24] on the query expansion using an ontology in the field of law, and WordNet showed an improvement in the performances of the IRS.
- The work achieved by [1] on the query expansion using Arabic WordNet and a morphological analyzer showed an improvement in the recall but not the precision of the Arabic IRS.
- The work achieved by [2] on the strategy evaluation of the local PRF for the Arabic texts showed an improvement in the performances of the IRS.

In the context of this article, we are interested in the evaluation of the PRF technique for Arabic texts and particularly in the study of the two parameters (D, T) variation.

3. Experimentation

As for us, we are going to investigate the influence of the variation of the two parameters «D» and «T» involved in the PRF technique in Arabic IRS. For such a deal, we are going to vary «D» and «T» from 1 to 20 and for each combination (D,T) (with D=1, 20 and T=1, 20; 400 combinations) we are going to compare the different results obtained by the different systems (runs).

As for our experiment, we have used a corpus of Arabic texts in different fields. This corpus has been not used in an official evaluation campaign. The table 1 sums up the principal characteristics of this corpus.

Table 1: The main features of the corpus used

Number of text files	22 000
Fields	Health, sports, politics, sciences, religion, astronomy, nutrition, law, tales, family
Size	180 MB
Number of words	17 000 000
Number of different words	612 650

All the operations concerning the indexation and the interrogation of the documents collection are realized using the API Lucene version 3.0 (http : //lucene. apache. org). On the other hand, the process of PRF is developed in Java language. It implements the local clustering technique (see Fig. 3) to extract the most pertinent «T» terms that serve in the reformulation of the initial query. The table 2 shows examples of queries before and after reformulation.

Table 2: Examples of queries before and after RF.

N°	Query before RF	Query after RF
1	أضرار التدخين (damage of smoking)	طب (Medicine)، مدخن (Smoker)، عام (year)، موقع (location)، مواضيع (topics)، هنا (here)، سلب (robbed)، اثار (traces)، عادة (habit)، علاج (treatment)، علم (science)، جديد (new)، مقال (article)، امراض (disease)، اخر (another)، زوار (visitor)، عين (eye)، جمع (collect)، دراسة (study)، قائمة (list)، مصاب(infected)
2	سهم مالي (monetary action)	اسهم(Shares) ، تداول(trading) ، بنك(bank) ، سوق(market) ، مال(money) ، عام(Land)
3	مدار الارض (Earth's orbit)	ارض(earth) ، قمر(moon) ، شمس(sun) ، مدار(orbit) ، فلك (orbit)، رئيس(Chairman) ، موقع(location)

```
// Algorithm of the experimentation
Begin
 For each **D** (D=1, 20) do
     For each **T** (T=1, 20) do
         For each query qi (i=1, 50) do
             1- Interrogation of the collection of documents
             2- Sampling : select the «D» first returned documents: $D_F$
             3- Extraction of evidences
                 - Construct the matrix of local association (term - term) from the set of distinct terms of $D_F$ : $\vec{S}$ with each element
                       $S_{u,v} = \sum_{d_j \in D_F} f_{S_{u,j}} \times f_{S_{v,j}}$
                   $f_{S_{u,j}}$ : represents the term frequency
                   $S_u$ in the document $d_j$
                   $S_{u,v}$ : expresses the correlation
                   between $u$ and $v$
                 - For each term $t$ of $q_i$ extract its local association clustering $C_i$ set from the « T » highest values $S_{u,v}$ ($v \neq u$) of the
                   $u^{th}$ line of $\vec{S}$
             4- Rewriting of the query : construct the new query
                       $q_{new} = q_i \cup C_i$
             5- End
```

Fig. 3 Algorithm evaluation of the PRF (local clustering technique).

For each combination (Di, Tj) (i = 1, 20, j = 1, 20), the results of different queries are written in different files. For example: the number of documents found, the number of relevant documents found, the precision at 5, 10, 20, 100 and 1000 documents (P@5, P@10, P@20, P@100, P@1000) and the average precision.

4. Analysis and discussion of the results

To understand the effect of the variation of D and T on each query, we have established various measures that are mainly based on the comparison of results before and after enrichment.

For a given query, three cases can arise:

- Improvement (+): All precisions (at 11 points of recall) before are lower than those after. In other words, the curve (recall / precision) after is over before.
- No improvement (-): is the inverse of the previous case. The curve (recall / precision) before is over after.
- No decision (X): for some precisions, there is an improvement but for others there is no improvement. In other words, there is an intersection of the two curves (recall / precision).

4.1 Comparison based on the number of improved queries

We recorded in Table 3 for each value of D (respectively T) the average number of queries improved.

Table 3: (a) the average number of improved queries in accordance with D. (b) the average number of improved queries in accordance with T.

D	Average Number of queries (+)	D	Average Number of queries (+)
1	2,30	11	4,25
2	1,70	12	3,90
3	4,40	13	3,90
4	3,45	14	3,90
5	3,70	15	4,05
6	3,55	16	3,30
7	2,80	17	2,95
8	2,75	18	2,90
9	3,35	19	2,90
10	3,55	20	2,80

(a)

T	Average Number of queries (+)	T	Average Number of queries (+)
1	3,65	11	2,80
2	3,90	12	2,95
3	3,70	13	3,10
4	3,55	14	3,00
5	4,25	15	3,20
6	3,60	16	3,15
7	3,75	17	3,05
8	2,85	18	3,10
9	3,15	19	3,20
10	3,25	20	3,20

(b)

We note in Table 3 (a) that the average number of improved queries tends to increase when the value of D increases and vice versa in Table 3 (b) it decreases when the value of T increases.

4.2 Comparison based on the number of improved tests (runs)

We recorded in Table 4 the number of tests that led to an improvement (+) of the number of queries with the values of the corresponding D and T.

Table 4: The number of tests that led to an improvement (+) depending on the number of improved queries with the values of D and T.

Number of queries (+)	Number of tests	D	T
0 (0%)	2 (0.5%)	2	1,4
1 (2%)	10 (2.5%)	1	16,17,18,19,20
2 (4%)	86 (21.5%)	17,18,19, 20	8,9,10,11,12,13,14
3 (6%)	108 (27%)	5,6,7,8,9, 10	11,12,13,14,15,16
4 (8%)	151 (37.75%)	10,11,12,13, 14, 15	1,2,3,4,5,6, 7
5 (10%)	40 (10%)	3,4,11,13,15, 16	1,4,5,7
6 (12%)	3 (0.75%)	14,15,19	5,2

The analysis of the results of table 4 enables us to release the following remarks:
- In only two tests (D = 2, T = 1) and (D = 2, T = 4), no reformulated query has improves the IRS performance. This percentage is very low and we can confirm that the query reformulation improves the IRS performance.
- In three tests among 400 (0.75%), we achieved a maximal improvement of 12% (6 queries from 50).
- In 259 tests (151 108), about 65%, an improvement of 7% was recorded.
- In 191 tests (151 +40), about 48%, we achieved an improvement of 9%.
- Approximately 50% of the 151 systems (having shown an improvement of 8%) have a value of D between 10 and 15, and T between 1 and 7. This fact confirms the results obtained in the previous section (4.1), i.e., improving the performance of an Arabic IRS can be achieved by high values of D and low values of T.
- Approximately 50% of the 108 systems (having shown an improvement of 6%) have a value of D between 5 and 10, and T between 11 and 16.
- Approximately 50% of the 86 systems (having shown an improvement of 4%) have a value of D between 17 and 20, and T between 8 and 14.
In conclusion, we can say that the values {14, 15, 19} for D and {5, 2} for T can enable a maximum improvement of the IRS performance. Remember that these values are only valid for the corpus used in our experiments. We not pretend, in any case, a generalization of these results.

4.3 Comparison based on the number of improved queries

The table 5 shows for each query used in the experiment the number of systems with improvement indicator (+), no improvement (-) or no decision (X).

Table 5: The number of systems according to the three cases of query comparison (Improvement (+) No improvement (-), no decision (X)).

N° Query	(+)	(-)	(X)	N° Query	(+)	(-)	(X)
1	0	0	400	26	243	17	140
2	0	0	400	27	0	0	400
3	1	390	9	28	0	381	19
4	0	0	400	29	0	40	360
5	0	321	79	30	14	0	386
6	2	4	394	31	1	266	133
7	0	0	400	32	0	385	15
8	0	16	384	33	20	0	380
9	0	384	16	34	0	0	400
10	69	46	285	35	0	1	399
11	9	20	371	36	109	27	264
12	228	10	162	37	0	250	150
13	20	208	172	38	27	87	286
14	136	3	261	39	0	0	400
15	12	39	349	40	0	0	400
16	0	99	301	41	0	24	376
17	0	55	345	42	34	45	321
18	0	0	400	43	0	0	400
19	0	47	353	44	29	74	297
20	1	29	370	45	0	132	268
21	0	0	400	46	0	368	32
22	0	39	361	47	0	45	355
23	348	0	52	48	9	0	391
24	13	19	368	49	0	0	400
25	0	169	231	50	3	315	82

The analysis of the results of table 5 enables us to release the following remarks:
- Only 21 queries (42%) are concerned with the improvement (+).
- In 29 queries (58%), more than half, we did not obtain improvement. In other words, these queries are very difficult to improve with the query reformulation.

- In 12 queries (24%) we obtained indecision (X) in all tests (400 tests).
- From the point of view of the query domain, and knowing that we have worked with a text corpus of ten (10) different areas, we recorded at least one query improved by domain and two queries improved for seven domains (70%). Moreover, all queries have been improved for one domain only (law domain). Based on these results, we can conclude that there is no direct relationship between the query domain and the query reformulation.

4.4 Comparison on the basis of the analysis of the precision at 5 documents (P@5)

From the point of view of the P@5, the analysis of the results obtained has enabled us to deduce that:
- For 166 of the 400 systems (41.5%) the P@5 after is greater than the P@5 before, so we can confirm that the query reformulation improves the IRS performance. In addition, 202 of the 400 systems (50.5%) the P@5after is less than the P@5 before, so there is no performance improvement for the IRS. Table 6 shows the combinations of D and T for the best eight (8) P@5 obtained.

Table 6: T and D for the best P@5 obtained

D	T	P@5
3	3	0,660
3	4	0,656
3	2	0,644
18	2	0,644
5	2	0,644
17	2	0,640
18	8	0,640
18	13	0,640

The examination of the various values of D and T that allows for an improvement in P@5 are generally between 3 and 20 for D and between 2 and 20 for T. Also, we found that the increase in D (or T) does not necessarily imply an improvement in P@5. Moreover, if we fix the value of D, the increase in T does not consequently increase (improve) the P@5. The opposite is also true, that is to say, if we fix the value of T, the increase of D does not necessarily imply an increase (improvement) in P@5.

5. Conclusion

In IRS, automatic relevance feedback is a technique for reformulating the user query. It is based on a process composed of three steps: the sampling, the terms extraction and the rewriting of the initial query.
Two parameters are to be taken into account during the two first steps which are: the size of the sample D and the number of terms T to be extracted for the query expansion.
In this paper, an experiment was conducted on a corpus of Arabic text in order to study the effect of the variation of the two parameters in question (D and T) on the overall performance of the Arabic IRS.
The aim is not to propose optimum values for D and T and even less, to generalize the results on Arabic SRI.
The experiment allows us to deduce that there are queries that are hardly improvable with the query reformulation. Moreover, the success of the PRF is mainly due to the selection of a sufficient number of documents and the extraction of a reduced set of relevant terms for the search.
The results obtained allow us on the one hand, to study the correlation between D and T. On the other hand, opening the way to test the same technique with different types of corpora in order to provide the best possible settings for each query.

References

[1] Abderrahim M. A., Abderrahim Med Alaeddine (2010) Using Arabic WordNet for query expansion in information retrieval system ; IEEE,The Third International Conference on Web and Information Technologies, 16-19 June, 2010, Marrakech – Morocco.
[2] Abderrahim M. A., Abderrahim Med Alaeddine (2012) Réinjection Automatique de la pertinence pour la Recherche d’Informations dans les textes Arabes ; IEEE, 4th International Conference on Arabic Language Processing, May 2–3, 2012, Rabat, Morocco; pp 77-81.
[3] Attar R. and A. S. Fraenkel (1977) Local feedback in full-text retrieval systems. Journal of the ACM, 24(3):397-417, 1977.
[4] Aurélien Saint Requier, Gérard Dupont, Sébastien Adam et Yves Lecourtier (2010) Évaluation d'outils de reformulation interactive de requêtes. COnférence en Recherche d'Infomations et Applications - CORIA, 7th French Information Retrieval Conference, Sousse, Tunisia, March 18-20.
[5] Baeza-Yates, Ricardo and Berthier Ribeiro-Neto (1999) Modern Information Retrieval. Addison-Wesley, New York City, NY, ACM Press.
[6] Bassam Hammo, Azzam Sleit , Mahmoud El-Haj (2007) Effectiveness of Query Expansion in Searching the Holy Quran. Colloque internationale Traitement automatique de la langue Arabe, CITALA'07, 18-19 juin.
[7] Bodo Billerbeck, Justin Zobel (2004) Questioning Query Expansion: An Examination of Behaviour and Parameters. ADC: pp. 69-76.
[8] Bodo Billerbeck, Justin Zobel (2004) Techniques for Efficient Query Expansion. 30-42 ; String Processing and Information Retrieval, 11th International

Conference, SPIRE 2004, Padova, Italy, October 5-8, Proceedings. Lecture Notes in Computer Science 3246 Springer.

[9] Bodo Billerbeck (2005) Efficient Query Expansion. PHD Thesis, RMIT University, Melbourne, Australia.

[10] Buckley, C. and Harman, D. (2004) Reliable information access final workshop report.

[11] Claudio Carpineto, Giovanni Romano (2012) A Survey of Automatic Query Expansion in Information Retrieval ; Computing Surveys (CSUR), Volume 44 Issue 1; January.

[12] Cristopher D. M., Prabhaker R., Hinrich S. (2009) An Introduction to Information Retrieval. Cambridge University Press.

[13] Harman, D., Buckley, C. (2004) RIA and "Where can IR go from here? SIGIR Workshop, www.sigir.org/.../2004D/harman_sigirforum_2004d.pdf.

[14] Hlaoua L. (2007) Reformulation de Requêtes par Réinjection de pertinence dans les Documents Semi-Structurés. PHD Thesis, université Paul Sabatier.

a. http://nrrc.mitre.org/NRRC/publications.htm

[15] IJsbrand Jan Aalbersberg, (1992) Incremental relevance feedback, Proceeding SIGIR '92 Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 11 – 22, ACM New York, NY, USA.

[16] Jinxi Xu, W. Bruce Croft (2000) Improving the effectiveness of information retrieval with local context analysis. Transactions on Information Systems (TOIS), Volume 18 Issue 1, ACM, January.

[17] Jinxi Xu, Alexander Fraser , Ralph Weischedel (2002) Empirical Studies in Strategies for Arabic Retrieval.

[18] Kanaan, G., Al-Shalabi, R., Abu-Alrub, M., and Rawashdeh, M. (2005) Relevance Feedback: Experimenting with a Simple Arabic Information Retrieval System with Evaluation. International Journal of Applied Science & Computations,Vol. 12, No.2, USA.

[19] Kyung Soon Lee W. Bruce Croft James Allan (2008) A Cluster-Based Resampling Method for Pseudo-Relevance Feedback. SIGIR'08 , July 20-24, Singapore. pp. 235-242 ACM.

[20] Marie-France BRUANDE, Jean-Pierre CHEVALLET (2003) Assistance intelligente à la recherche d'information. Lavoisier ; pp. 99-129.

[21] Salton Gerard (1986) Recent trends in automatic information retrieval. Proceeding SIGIR '86 Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 1 – 10, ACM New York, NY, USA.

[22] Salton, G., and C. BUCKLEY (1990) Improving Retrieval Performance by Relevance Feedback. Journal of the American Society for Information Science, 41(4), 288-97.

[23] Shipeng Yu, Deng Cai, Ji-Rong Wen, Wei-Ying Ma (2003) Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In WWW '03: Proceedings of the 12th international conference on World Wide Web (2003), pp. 11-18, doi:10.1145/775152.775155.

[24] Soraya, Zaidi, M-T. Laskri (2007) Expansion de la requête Arabe dans le web. Barmajiat (CSLA): Les applications logicielles en arabe: Pas vers le e-gouvernement, 9-10 décembre Alger.

[25] Stefan Büttcher, Charles L. A. Clarke, Gordon V. Cormack (2010) Information Retrieval Implementing and Evaluating Search Engines. MIT press, Cambridge, Massachusetts London, England.

[26] W. Nesrine ZEMIRLI (2008) Modèle d'accès personnalisé à l'information basé sur les diagrammes d'influence intégrant un profil utilisateur évolutif. PHD Thesis ; université Paul Sabatier Toulouse III.

[27] Hazra Imran and Aditi Sharan (2010) Improving Effectiveness of Query Expansion Using Information Theoretic Approach; IEA/AIE 2010, Part II, LNAI 6097, pp. 1–11, Springer 2010.

[28] Karthik Raman and Raghavendra Udupa (2010) On Improvin g Pseudo-Relevan ce Feedback Usin g Pseudo-Irrelevant Do cuments; Springer, ECIR 2010, LNCS 5993, pp. 573–576, 2010.

[29] Guihong Cao, Jian-Yun Nie and Guihong Cao, Jian-Yun Nie and Guihong Cao, Jian-Yun Nie (2008) Selecting Good Expansion Terms for Pseudo-Relevance Feedback; SIGIR'08, July 20–24, 2008, Singapore.

[30] Paul Ogilvie and Ellen Voorhees and Jamie Callan (2009) On the number of terms used in automatic query expansion; Inf Retrieval (2009) 12:666–679; Springer; doi: 10.1007/s10791-009-9104-1.

[31] Johannes Leveling and Gareth J. F. Jones (2010) Classifying and Filtering Blind Feedback Terms to Improve Information Retrieval Effectiveness; RIAO'10, 2010, Paris, France.

Mohammed El Amine Abderrahim is a research teacher at the University of Tlemcen, Algeria. His research interests are natural language processing, information retrieval, information extraction, databases and data mining. Med El Amine has a Magister in computer science from UST Oran, Algeria, and a Doctorate in computer science from the University of Tlemcen, Algeria. He is a member of the Laboratory of Arabic Natural Language Processing, university of Tlemcen.

Saïd Benameur is a research teacher at the University of Tlemcen, Algeria. His research interests are natural language processing, applied linguistics and translation. Saïd has a Magister in linguistics from the University of Tlemcen, Algeria. He is a Doctorate candidate and a member of the Laboratory of Arabic Natural Language Processing in the University of Tlemcen.

Mohammed Alaeddine Abderrahim is a research teacher at the University of Tiaret, Algeria. His research interests are natural language processing, information retrieval, information extraction, data mining and ontology. Med Alaeddine has a Magister in computer science from the University of Tlemcen, Algeria. He is a Doctorate candidate and a member of the Laboratory of Arabic Natural Language Processing in the University of Tlemcen.

A Formalization of the End User Service Development Approach

Meriem Benhaddi[1], Karim Baïna[2] and El Hassan Abdelwahed[3]

**[1] Faculty of Sciences Semlalia, Cadi Ayyad University
PO.Box 2390, Marrakesh, Morocco**

**[2] ENSIAS, Mohammed V Souissi University
PO.Box 713 Agdal-Rabat, Morocco**

**[3] Faculty of Sciences Semlalia, Cadi Ayyad University
PO.Box 2390, Marrakesh, Morocco**

Abstract

The end user service development known as the user-centric SOA emerged as a new approach that allows giving the end user the ability to create on the fly his own applications that meet a situational need. In fact, the classical SOA was designed for developers and is characterized by a heavy technical stack which is out of reach of end users. Lightweight Web 2.0 technologies such as Mashup appeared to bridge this gap and provide a new agile and quick way to compose and integrate different resources in a dynamic and on the fly manner. However, Mashups are emerging applications, and thus consist of immature, non intuitive and non formalized area. In this paper, we formalize the user-centric SOA development by proposing a new cloud-based architecture for user-centric SOA platforms, and by introducing a new rich integration language based on the advanced Enterprise Integration Patterns (EIPS). We also propose a new intuitive and self-explanatory semantic process for end users services integration.

Keywords: *SOA, Mashup, integration patterns, end user development, end user satisfaction, intuitiveness, Cloud Computing.*

1. Introduction

The text must be in English. Authors whose English language is not their own are certainly requested to have their manuscripts checked (or co-authored) by an English native speaker, for linguistic correctness before submission and in its final version, if changes had been made to the initial version. The submitted typeset scripts of each contribution must be in their final form and of good appearance because they will be printed directly. The document you are reading is written in the format that should be used in your paper.

1.1 Problems and limitations of SOA

The concepts behind the Service Oriented Architecture has proved that it is the best way to urbanize the enterprise information system by modulating applications as interoperable services; in fact SOA promotes the modulating applications as fine or coarse grained services, the reuse of services to build more complexes ones, the interoperability between different heterogeneous system, and the standardized languages and protocols (WSDL, SOAP, BPEL). SOA's goal is to lower costs and make information systems more flexible. Nevertheless, enterprises that applied SOA didn't get the great promised added value, which has prevented the installation of the global SOA, and has lowered the percentage of companies planning the SOA [9].

In this section, we introduce the concept of "End User", to signify the non-computer user, who has very little computer knowledge. We will give a further definition of this concept in the next section.

- The limitations of SOA could be summarized as:
- Exclusion of the end user from the hierarchy of the SOA actors: users kept away and out of the loop. In fact, the SOA technologies (WSDL, SOAP, SCA, BPEL, etc) are hard to master and require advanced knowledge [22] [28].
- Rigidity, heaviness and incompatibility of SOA implementations with the real constraints of end users:
  - Lack of accessibility: UDDI registries are dedicated to expert; therefore, end users have to browse different web sites in order to use services. [17] states that SOA was originally designed as an architecture focused fundamentally on the B2B context, and does not offer support for B2C interactions.
  - Lack of interoperability and openness: The implementation of SOA has been limited to

the use of WS* technologies (WSDL, SOAP, UDDI), which prevents the development of SOA specifications, that are independent of any technology.

- Lack of flexibility and scalability: SOA technologies cannot support the services composition on the fly: After composition design, implementation, testing and deployment, it becomes very difficult to change the composition logic according to the changing needs of users, as it involves a long life cycle [18].
- Lack of mobility: SOA implementation and integration technologies are very heavy for devices with limited capabilities. WSDL and SOAP are instances of complicated XML documents, which makes the WS* services very demanding in terms of computing power, bandwidth and storage [10].

1.2 End users: Who are they? What do they need?

Definition of end user: A software end user is a person who interacts with information systems solely as a final information consumer. It’s a user with minimal technical knowledge, and who uses the software in the context of daily life or daily work for personal (business or leisure) purposes, without having any intentions to produce other systems. He is not interested in computers per se, and do not worry about system technologies as long as he can get what he needs quickly [8] [1].

End users have many requirements that should be respected by system designers and developers in order to deliver systems satisfying end users. Based on the work of [20] and [15], we have grouped into four criteria the end users requirements, which are listed in table 1.

Table 1. Criteria of a user-centric solution

Criteria	***Description***	***Problem of criteria lacking***
Functional richness	Features requested to execute different tasks.	Limited set of offered features.
Usability & intuitiveness	User interfaces, interaction and dialogue mode.	Lack of visibility, feedback, consistency, non-destructive operations, discoverability, scalability, reliability [23].
Efficiency, reliability, maintainability and portability (ERMP)	Difficulties that do not refer directly to system features.	Lack of documentation, performance, security, supportability.
Personnalizability, customizability	Capability of end user to tailor themselves their systems.	Useless systems that lack many important features.

Based on this section, we define the user-centric SOA as the expectation of end users, their future hope, and the promise for better information systems. A user-centric SOA offers:

- Empowerment of the end user: Easy and flexible composition on the fly of services by all end users that can design and create new services through the combination and composition of existing services, made possible by reduction of the complexity of services composition techniques.
- Openness of the Information System to the public: the democratization of SOA and the installation of the global SOA. In this context, [26] speaks about the Internet of Services where every user use and access to services.
- More independence of SOA: the adoption of a variety of interoperable technologies in order to meet the great variety of the web.
- Lightweight SOA technologies: the support of SOA technologies by all mobile devices.

2 State of the art

2.1 Mashup frameworks limitations

Mashup is a new paradigm of the Web 2.0 [24] – the new generation of the web - that enables the user generation of services by allowing end users to personalize and customize their applications [13][19][6]. Today, there are a big number of Mashup frameworks on the web, which allow end users to mix visually different heterogeneous resources and thus create new applications called mashups. Mashup frameworks have helped to bridge the gap between end users and software development, but they are still some technical gaps [4]:

- Mashup frameworks use lightweight resources (RSS, ATOM, REST services, etc): [25] affirms that existent Mashup frameworks focus on the integration of lightweight Web Services, and do not take into consideration enterprise-class and complex services, that may use any SOA technology and not only Web Services. [21] says that the conversion between inputs and outputs parameters is limited to simple data types, and do not consider complex parameters.Moreover, the Mashup tools do not allow diversity of the output type; an example is Yahoo Pipes [27] that provide only RSS as output of the Mashup. Besides, Mashup tools require ready-to-use sources, which prevent the flexibility of these tools. Thus, existent Mashup tools cannot support the Web Services Mashup and more generally the SOA Mashup. In this context, [25] underline the need of the enablement of Web Services Mashup.

- Mashup frameworks do not allow the creation of business process mashups: the existent Mashup frameworks do not provide ways to design and create complicated use case. In fact, the resources composition and the interaction are based only on the data flow. Moreover, the event-handling concerns only the events from sources and doesn't satisfy the user interaction level [21].
- Mashup tools do not provide stable applications: [2] asserts that the solutions provided by Mashup tools are fragile, neither stable nor robust.
- Mashup frameworks are still outside the scope of end users: Mashup frameworks still lack simplicity for the end user. In fact, the Mashup tools often use technical concepts like port or wires. For the simple end user, handling these technical concepts is not easy and requires a learning time [22].

These critics show that the Mashup is at an early stage and needs more research. In fact, there is a lack of a powerful language for describing Mashup and realizing advanced Mashup applications. Hence, in order to achieve the user-centric SOA, there is a need to introduce new elements consisting of patterns and models to enhance the development of Mashup applications.

The next section introduces the Enterprise Integration Patterns, and shows their contribution to any integration solution.

2.2 Contribution of the Enterprise Integration Patterns

The Enterprise Integration Patterns (EIPs) collected by [12] describe a number of design patterns for enterprise application integration and message oriented middleware. The EIPs are implemented by a set of sophisticated mediation bus, such as Camel, Mule and Apache, in order to achieve very complex integration scenarios. Enterprise Integration Patterns propose the best and common solutions to integration problems. Therefore, when EIPs are used, they enhance the quality of the integrated applications. EIPs consist of six groups of patterns: messaging channels, message construction, message routing, message transformation, messaging endpoints and system management. Based on the book of [12], we categorize these patterns groups according to the four end user satisfaction criteria that we defined and presented in section 1.2.

As it can be seen from table 2, the Enterprise Integration Patterns, when used together, help achieving a high level of system quality by ensuring four of the end user satisfaction criteria. The use of EIPs is therefore considered as a proof of the system quality. Hence, we had the idea of studying different Mashup frameworks based on the EIPs. The next section gives the result of this study and positions the Mashup frameworks against the user-centric SOA.

Table 2. Categorization of EIPs according to end user SOA criteria

		Non-functional		
Patterns/ Criteria	Functional Richness	Efficiency	Reliabili ty	Mainta-inability
Messaging Channels	X		X	
Message construction	X			
Message routing	X			
Message transformation	X			
Endpoint patterns	X	X		
System management			X	X

2.3 Study: Mashup frameworks and the user-centric SOA

As we announced in the previous section, we studied three Mashup frameworks according to the EIPs. The Mashup frameworks considered are: Yahoo! Pipes [27], Jackbe Presto Wires [16] and IBM Mashup Center [14]. As the latter two groups of the EIPs – endpoint and system management patterns - are related to the internal implementation of the solution, we could study the Mashup frameworks only according to the first four groups which are: messaging channels, message construction, message routing and message transformation.

Table 3 shows the number of patterns used among all the existing patterns. The quotient x/y means that x patterns are used among y existing patterns.

Table 3. Study of three Mashup frameworks according to EIPs

Patterns/ Mashup Frameworks	***Yahoo! Pipes***	***Jackbe Presto Wires***	***IBM Mashup Center***
Messaging Channels	3/7	3/7	3/7
Message construction	2/9	2/9	2/9
Message routing	4/12	4/12	3/12
Message transformation	3/7	4/7	4/7

Table 3 shows that the three Mashup frameworks implement a limited set of the integration patterns. Our study showed also that the used patterns are very basic and simple; the three Mashup frameworks fail to implement advanced and sophisticated integration patterns. According to this study and to table 2, we deduced that the three Mashup frameworks fail to totally ensure the criteria of “Functional richness”, “Efficiency”, “Reliability” and “Maintainability”.

We also studied the three Mashup frameworks according to the other end user satisfaction criteria, which are “Usability & intuitiveness” and “Portability”, and the study showed also that they are not completely ensured. Unfortunately we could not introduce this study in this paper because of the restricted number of pages.

All the study that we conducted showed that the Mashup frameworks are not user-centric SOA solutions. To enhance Mashup, we propose the idea of using the EIPs within Mashup frameworks to improve their acceptance by different end users. The next section gives a brief description of our proposed new SOA-Mashuped language based on the Enterprise Integration Patterns.

Table 4. Mashup frameworks and user-centric SOA criteria

UCSOA criteria/Mashup Frameworks	*Yahoo! Pipes*	*Jackbe Presto Wires*	*IBM Mashup Center*
Functional Richness	2	2	2
Personnalizability	3	3	3
Usability & Intuitiveness	2	2	2
Efficiency, Reliability, Maintainability and Portability	3	2	2

3=High, 2=Medium, 1=Low

In this section, we presented and criticized existing Mashup frameworks. Mashup development is still immature and at an early stage and thus needs more research. In particular, there is no significant formalization of Mashup integration. For this reason, we conducted a study of three Mashup frameworks regarding to the end user satisfaction criteria defined in section 1.2. The conclusion drawn from this study led us to the need for new patterns and methodologies to improve Mashup development. The next section is dedicated to the proposal of a new Cloud-based Mashup architecture, that uses a new EIPs-based integration language, while allowing the end user service creation through a new intuitive and self-explanatory creation process. The last requirement – non functional requirement – is out of the scope of this paper.

3 User-centric SOA proposal

3.1 Cloud-Based Architecture

We presented the technical architecture of the user-centric SOA in [5]. This Architecture includes six vertical layers – Web or non Web resources, Resources access, Gadget or Mashup component development, Integration or Mashup components assembly and Visualization or consumption – and two cross layers – Enterprise infrastructure and Web 2.0 collaborative community –. Each layer relies on several services; usability is a very important dimension that should be considered in Gadgets layer, Integration layer and Visualization layer in order to provide end users with intuitive and self-explanatory creation process.

The different services used by Mashup platforms can be homemade (developed internally), or accessible through the Cloud Computing. Indeed, the Cloud Computing can be considered as a novel way to retrieve and use IT-enabled services by customers. The new Software-as-a-Service (SaaS) paradigm allows the supply of services through the internet. According to [7], the Cloud Computing is an emerging paradigm that is based on compute and storage virtualization to deliver reliable services to customers. Customers can access data and applications anywhere in the world on demand.

This way, Mashup platforms can rely on the Cloud Computing services to ensure the operation of each layer of the technical architecture. For example, Enterprise Service Buses could be used for their routing and translation capabilities, BPEL engines could be used for their orchestration capability and the CRUD services offer different services such as identity management, persistent storage, resources access, routing and translation.

As stated before in this paper, end users have four requirements: functional richness, usability && intuitiveness, infrastructure requirements such as reliability, efficiency, maintainability and portability, and Personalizability. As Mashup platforms were created to let end users personalize their applications, we consider that the fourth requirement is ensured. The third requirement is out of the scope of this paper. We focus our work on the first two requirements. The next section is dedicated to the study of the first requirement -functional richness – and provides a solution based on the Enterprise Integration Patterns (EIPs).

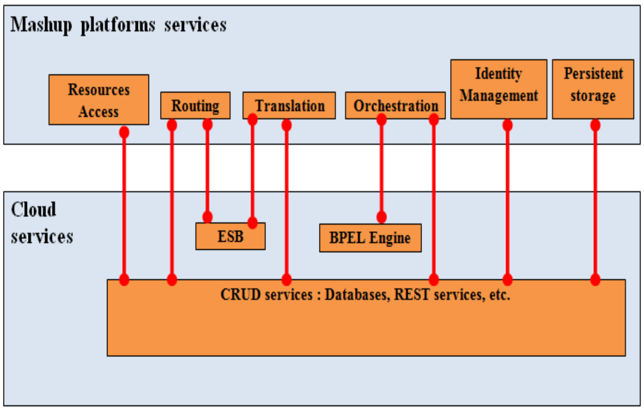


Fig. 1 Application architecture of our proposal

3.2 The system point of view: Functional Richness

As it was showed in section 2.2, the Enterprise Integration Patterns help enhancing the system quality in terms of the functional richness. Therefore, our proposal is based on the Enterprise Integration Patterns.

In the following, we give an example of use case to help defining the different entities that will form our language.

Our example is taken from the world of physical Mashup. According to [11], physical Mashup is a concept that allows to link and combine real-world objects. So let's take the example of an end user whose goal is to model and customize his Mini Cooper car. The Mini Cooper is considered as an object with features and services. In addition, end users can add various accessories to personalize the car and develop new services. Accessories are considered as objects to be integrated with the car. Examples are integrating the car with an object that displays the temperature, the state of the seatbelt and some advices, with dataflow from central system to the accessory; or integrating the car with a car burglary detector, with an event as a message between the two objects.

To summarize, in order to achieve his task, the end user needs a platform that encapsulates the following elements:

- Objects/resources to integrate: Mini Cooper car, accessories.
- Fields on interface allowing the entry of intermediate data.
- Communication channels that allow binding and forwarding the results between different objects.
- Messages of different types which will be carried by channels and sent by one object to another. A message can be of different types: a message representing a document, a message representing an order, etc.
- Routing components whose role is to route the results of an object to another.
- Translation components that transform the results of an object before sending them to another object.
- A view showing a graphical interface that displays the final result of the integration.

From this simple illustrative example, we have identified the different basic elements that will form our future language that we named SOA4EU (SOA for End User). Table 2 lists these elements.

Table 5. Constructs of SOA4EU language

Construct	***Description***
Task	is the goal of the end user performing the integration. Each task can have a frequency of execution.
Tag	key words used to describe a task
Mashup	A Mashup application represents the realization of a task and includes a set of integration taking place between several resources.
Process	Is the composition process of the Mashp application resources and consists of parallel or sequential integration flows.
Step	Is a step in the integration process and consists of a link between two or several components.
Component	Is the integration process node: resource, input of the end user, router or translator.
Partner	represents the external partner of the Mashup: resource or end users.
EndUser	Represents the interaction with end users during the integration process.
Resource	Represents the applications to integrate by the Mashup. A resource is described by its type, address and exchange format.
Expose Resource	Represents an exposed resource with input and output variables. The same resource can be exposed many times within the integration process.
Channel	Allows communication between two components and supports the single atomic integration step.
Message	is the entity transferring in a channel between two components.
Router	Is a node forwarding messages between resources, end user fields or translators.
Translator	Is the messages translation node.
Data	Represents any data type handled by the Mashup application.
View	Is the view or graphical interface displaying the final result of the integration.
Transaction	End users may want to synchronize actions of components to realize a transaction.

The formalization of UCSOA language was done using Backus-Naur Form (BNF). Because of the pages number restriction, we present only the main part of the formalization, and it is as follow:

```
<Task>::= {<Tag>} {<Frequency>} <Mashup>
<Tag>::= [a-zA-Z][0123456789]
<Mashup>::= {<Resource>}+ {<Expose_Resource>}+ {<EndUser>} {<Router>} {<Translator>} <Process> <View>
<Process>::= <Sequence> | <Flow>
<Sequence>:=sequence({<Step>} {<Flow>} {<Step>})
<Flow>::= flow({<Step>} {<Sequence>} {<Step>})
<Step>::= <FromComponant> <ToComponants> <Channel> <Message>
<FromComponant>::= <Componant>
<ToComponants>::= {<Componant>}+
<Componant>::= <Partner> | <IntegrationService>
<Partner>::= <Expose_Resource> | <EndUSer>
<IntegrationService>::= <Router> | <Translator>
<EndUser>::= <Input>
<Resource>::= <Type> <DataFormat> <URL>
<Expose_Resource>::= <Resource> <ExpectVariable> <ReturnVariable> {<Transaction>}
<Data>::= <Input> | <Content> | <Event> | <Address> | <Identifier> | <Time> | <Version>| <Key> | <Schema> | <Datatype>
```

The formalization of “Channel”, “Message”, “Router” and “Translator” elements is done based on the Enterprise Integration patterns that define five patterns for channels, nine patterns for messages, twelve patterns for routers and six patterns for translators.

The next section focuses on the second requirement – usability & intuitiveness – and presents a methodology helping end users to easily compose services.

3.3 The end user point of view: Usability and Intuitiveness

3.3.1 Goals Composition vs Services Composition

When creating new applications, end users try to achieve a new goal by composing existing sub-goals. Each sub-goal is represented by a service. In this way, when composing services, end users try to resolve a problem whose solution does not exist yet on the web. In fact, the answer exists in the form of many subparts – services – dispersed on the web. Therefore, the inexperienced end user faces many challenges when trying to compose services in response to a new goal:

- Determine the types of resources: what to do?
- Find resources that meet the end user criteria (quality, price, etc.).
- Determine necessary actions for the use of interfaces (selection problems): what and how to use interfaces?
- Determine how to arrange and coordinate resources (integration): how to coordinate the elements?
- Determine the final interface of the integrated resources.

The system has the role of helping end user to answer these different questions, by suggesting resources, providing guidelines for the coordination of resources and providing feedback and documentation for each selected action.

Faced with these design problems, the end user will use the knowledge he possesses that describe his goal and which consists of:

- The objective or set of operations that the goal task must accomplish,
- The final result of the goal task (output of the process),
- The frequency of the goal task execution,
- The degree of importance of the goal task,
- The duration of the goal task.

This end-user knowledge represents the semantic which, alone, should be involved in the interaction between the end user and the user-centric SOA platform. Indeed, the service-to-service interaction, which is based on the syntax, is not valid at the interface level. The interface provides gadgets that represent a sub-goal, which is an abstraction of services; therefore, the interaction and communication way at the interface level should also be an abstraction of the communication way between services (Figure 2). Being the abstraction of the syntax, the semantic should be defined as the only way of interaction at the interface level. The semantic is what should be offered to the end user so that he could compose services.

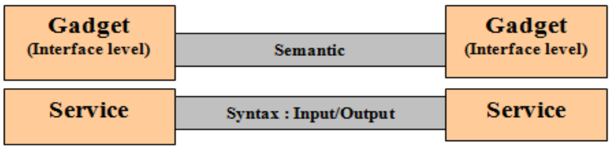


Fig. 2 Interaction way on the service level and the interface level

As the knowledge of the end user is limited to the semantic - goal, output, frequency, importance and duration of the task -, the end user should not and cannot manipulate the syntax. Therefore, the end user knowledge is insufficient to enable the integration of resources and the creation of new applications. The user centric SOA platform has to allow to end users to link the various resources in a very intuitive and self-explanatory way, requiring no knowledge of how to map an output of a resource to an input of another. The interface has then the role of intermediary between the end user and the services and should translate the end user interactions from semantic to syntax or code, as shown in Figure 3.

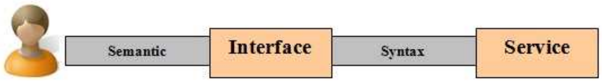


Fig. 3 Interaction between end users and the user centric SOA platform

To achieve this, the user centric SOA platform has to provide the end user with a set of goal prototypes or goals patterns. These goals patterns have the role of guiding the end user through the goals composition process. The next section presents our goals patterns-based suggestion system.

3.3.2 Goals Patterns-Based Suggestion System:
1) What are the goals patterns? In the world of software development, design patterns are solutions or best practices in response to common problems in software design. For example, the "Model-View-Controller" pattern help organizing an application by splitting it into a data model, an interface or a presentation and a controller (control logic, event management and synchronization). Goals patterns represent common and repetitive use cases, and can also be called end users experience patterns. They provide answers to questions like "How to automate the execution of two consecutive tasks - eg. Turn on the light on the entrance of the house and turn on the heating - in response to a triggered event? - ex. presence of a person detected by the sensor.
The following are examples of goals patterns:

- Booking airline ticket, hotel room and car for a destination.
- Purchase order for a product whose quantity reached a limit value.
- Turning on the room light and the coffeemaker when the alarm clock goes off.

While software design patterns are derived from the experience of the software developers, goals patterns are created, improved and enriched by end users themselves.
Our objective is to create a relational database of end users goals that end users will feed and develop as they create new applications. This database can also be automatically enriched by systems such as systems for smart homes patterns discovering.
2) Suggestion system: The usefulness of the goals patterns is the suggestion system. In fact, end users will be guided in the process of services composition through the database of goals patterns that contains the possible links between the various gadgets. As gadgets represent sub-goals, the database links represent also relations between sub-goals. The system will utilize this goals patterns database to suggest to the end user links and components in order to build new applications.
The suggestion system should be based on the semantic information, as it is explained in section 3.3.1. In fact, the different links between components should be represented by semantic information as input/output matching.
The database of goals patterns being built through the experience of end users, the system will score the various components, depending on the frequency of use, and thus offer to the end user the best one - which has the highest score.
Our suggestion model is similar to e-mail interfaces - ex. Gmail. When writing a new message, and when the first recipient address is entered by the user, other addresses are proposed and suggested at the basis of the previous messages sent by this user.
The goals patterns database elements that constitute also the components of the services composition interface are managed by the following description:

- An end-user profile is described by the age, the types of goals (work, leisure or both) the end user is interested in, the areas of interest, the physical environment.
- A profile is a set of goals.
- A goal is described by its type, its physical environment of execution, its objective, its frequency and its degree of importance.
- The realization of a goal involves several composition steps. A step represents a link from a component to one or several components (one-to-one or one-to-many).
- A component can be another application participating in the composition as sub-goal or an operator (translator or router).
- In order to suggest to the end user the appropriate actions, the database must store the various possible relationships between components. Thus, each composition step possesses a relation.
- Each link between two components (composition step) is described by a semantic data that corresponds to the output of the message transmitter and the input of the message receiver.
- The semantic data of a component can be information, event, interface or nothing.
- The participating applications or sub-goals can be synchronized in order to realize a transaction.

The object model of the goals patterns database is represented by Figure 4.

3.4 Linking the end user point of view with the system point of view

The end user point of view allows representing the end user services composition in terms of goals, relation, semantic data and other operators. To be able to be executed, the services composition application has to be represented using the technical system elements such as mashup, service, channel, message, etc. Thus, it is necessary to translate the services composition application from the end user point of view to the system point of view. As described earlier in this paper, the system point of view elements are based on the Integration Patterns which

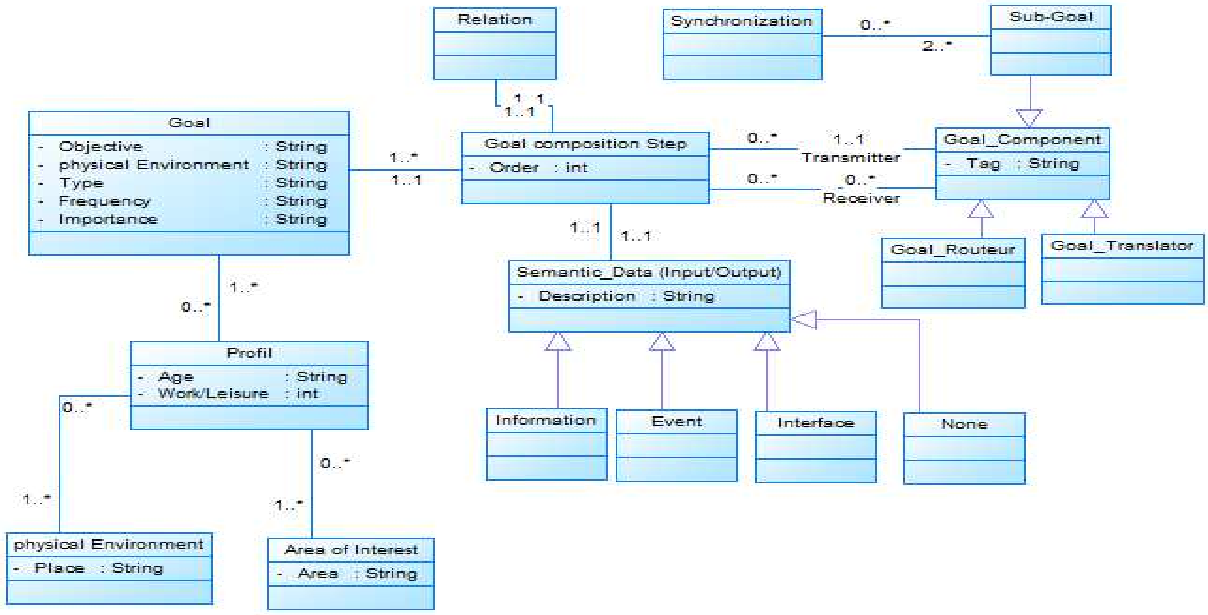


Figure 4.The object model of the goals patterns database

represent solutions to integration problems whose purpose is to achieve a goal.

Table 6 summarize the main elements of correspondence between the two set of elements: the end user point of view elements and the system point of view elements.

Table 6. Correspondences between end user and system points of view

End user point of view element	***System point of view element***
Goal	Mashup
Sub-Goal	Service
Relation	Channel
Semantic Data	Message
Goal Translator	Service Translator
Goal Router	Service Router
Goals Synchronization	Services Synchronization (Transaction)
Goal composition Step	Service composition Step

The formalization of this correspondence with Backus-Naur Form (BNF) is as follow :

<Goal>::= <Mashup> <Profil> {<Tag>}
<Goal Composition Step> ::= <Service composition Step>
<Relation>::= <Canal>
<Semantic data>::= <Message>
<Goal Component>::= <Service Component>
<Sub Goal>::= <Partner>
<Goal Router>::= <Service Router>
<Goal Translator>::= <Service Translator>
<Goal Synchronization>::= <Services Transaction>

4 Illustrative Example

To illustrate our new proposal, we choose an example from the WebOfThings world [10][3] which allows physical objects – called smart objects – to belong to a network and to be linked trough what is called the physical Mashup.

Our end user, Alice, wants to schedule a task to be executed every day at 7:00 in the morning - when the alarm goes off. The task, that represents Alice’s goal, consists of turning on the light on the bedroom and the coffee maker in the kitchen. When Alice is in the kitchen, the light must be lit. After Alice had opened the fridge and eaten food, the refrigerator recalculates the food quantities and displays them. If a food quantity reaches a minimum limit, a grocery order is automatically sent (Figure 5).

In the goals patterns database, there is a set of gadgets that Alice could use and that the platform could suggest to her. The gadgets are represented in four sub-directories depending on their output type (information, event, interface, none).

The steps followed by Alice to perform her task are as follows:

- Alice launches the platform, looks in the different sub-directories of smart objects she owns in her home and which are the resources of the applications she will creates with the user centric SOA platform. She selects the first object - alarm - from the sub-directory of event objects that she adds to the interface. The alarm requires the time as input and returns an event presented by ringing. Alice sees on her interface the gadget "Alarm" with the tag "Time" on its left and the red tag "Ring" on its right. At this level, since the output of the "Alarm" gadget is an event, any resource can be added without any constraint on the compatibility of input / output. In fact, an event role is to trigger a sequence of sub-tasks and not to deliver inputs for a sub-task. Thus, the platform does not make any suggestion at this level.

- Alice looks a second time in the sub-directory gathering objects that return nothing (the fourth category) and chooses the "Room Light" object which does not require input data and returns no result. Alice adds the “Room Light” object in sequence to the previous object. Alice also adds – in the same manner – the "Coffee Maker" object in sequence to the previous objects.
- From the event object sub-directory, Alice selects and adds – in sequence – the "I am in the kitchen" object, whose role is to notify the presence of a person in the kitchen.
- Alice adds in sequence the "Kitchen Light" object from the fourth sub-directory (objects that return nothing).
- Alice also adds the "Refrigerator" object which displays the different foods quantities. This object is represented by a gadget with different blue tags on its right representing the amounts of different foods (milk, eggs, cheese, butter, etc). As Alice has already used a filter with the "Refrigerator" object, the platform stored this link in the goals patterns database. At this level and based on the goals patterns database, the platform suggests to Alice, by displaying a button at the top of the window, to add a filter in order to show only foods with a specified limit amount.
- The platform suggests a second time to Alice, based on the goals patterns database, to add the “Grocery" object in order to make purchases for foods with small quantities.

At this level, the role of our end user is finished. In order to be run, Alice’s new application which is made of visual objects and links between these objects should be translated into services and links between these services. Those services links should be built based on the Enterprise Integration patterns presented in section 2.2.

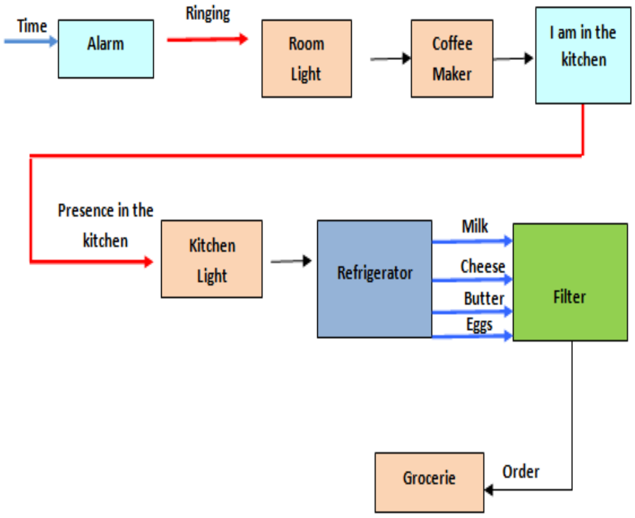


Fig. 5. Illustrative example for our intuitive creation process

The translation of visual objects and links into code (services and EIP links) is the translation of the goals composition – the end user point of view – into the services composition – the system point of view. This translation is realized based on the correspondences already established between the two points of view (section 3.4).

5 Conclusion and future work

In this paper, we presented the limitations of the Service Oriented Architecture that prevent it to be widely accepted in the web by inexperienced end users. We gave a definition of the end user and the end user satisfaction criteria. At a second time, we introduced the Mashup as a new web 2.0 paradigm and discussed its limitations resulting from its immaturity and its need to new patterns. We studied three Mashup platforms against the end users satisfaction criteria (based on the Enterprise Integration Patterns for the functional richness criteria) and we concluded that the Mashup frameworks fail to be user-centric SOA solutions. Our contribution aims at the formalization of the end user service creation. It consists of the proposal of a new Cloud-based architecture, a new EIPs-based integration language and a new intuitive and self-explanatory service creation methodology. Our future work consists of the completion and the implementation of our model in an intuitive graphical environment using AJAX technology, and its testing by real end users to guarantee the end users satisfaction. Our objective is to prove that our proposal prevails over the classical SOA and the existing Mashup solutions.

References

[1] Allison, H and R. Kelly, R. (1992) 'The Influence of Individual Differences on Skill in End-User Computing'. Journal of Management Information Systems I Summer 1992, Vol. 9, No. 1, pp. 93-111. (1992).

[2] Anjomshoaa, A., Tjoa, A.M. and Hubmer, A. (2010) 'Combining and integrating advanced IT-concepts with semantic web technology, Mashup architecture case study'. Paper presented at The 2nd Asian Conference on Intelligent Information and Database Systems, ACIIDS 2010, 24–26 March 2010, pp.13–22, Hue City, Vietnam, Part I, LNAI 5990. (2010).

[3] Avilés-López, E. and García-Macías, J.A. (2009) 'UbiSOA Dashboard: Integrating the Physical and Digital Domains through Mashups'.Paper presented at The Human Interface and the Management of Information Conference. Designing Information Environments.San Diego, CA, USA, July 19-24, 2009.

[4] Benhaddi, M., Baïna, K. and Abdelwahed, E. (2010) 'Towards an approach for a user centric SOA'. Paper presented at The third International Conference on Web & Information Technologies, Marrakech, Morocco, April 2010. ISBN: 978-9954-9083-0-3. Pages: 91-104.

[5] Benhaddi, M., Baïna, K. and Abdelwahed, E. (2012) 'A user centric Mashuped SOA'. Int. Journal of Web Science. Vol. 1, Issue 3. DOI: 10.1504/IJWS.2012.045812

[6] Bradley, A. (2007) Reference Architecture for Enterprise Mashups, Gartner Research.

[7] Buyya, R., Yeo, C. and Venugopal, S. (2008) 'Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities'. Paper presented at The 10th IEEE International Conference on High Performance Computing and Communications (HPCC-08), pages 25{27, Los Alamitos, CA, USA, 2008. IEEE

[8] Cypher, A.(1993) Watch What I Do: Programming by Demonstration. The MIT Press, Cambridge.

[9] Gartner. (2005) Gartner Newsroom http://www.gartner.com/it/page.jsp?id=790717. (2008). (Accessed 10 June 2012).

[10] Guinard, D. and Trifa, V. (2009) 'Towards the Web of Things: Web Mashups for Embedded Devices'. Paper presented at The 18th Int World Wide Web Conference, April, 2009, Madrid, Spain.

[11] Guinard, D., Trifa, V., Pham, T. and Liechti, O. (2009) 'Towards Physical Mashups in the Web of Things'. Paper presented at The 6th international conference on Networked sensing systems, INSS'09. 17-19 June 2009. Pittsburgh, PA, USA.

[12] Hohpe, G. and Woolf, B. (2003) Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions, Addison-Wesley Professional.

[13] Hoyer, V., Janner, T., Schroth, C., Delchev, I. and Urmetzer, F. (2009) 'FAST Platform: A Concept for user-centric, enterprise class Mashups'. Paper presented at The 5th Conference of Professional Knowledge Management, Poster Session, Solothurn, Switzerland, 25-3-2009, pp.5-8.

[14] IBM Mashup Center. [Online] http://www-01.ibm.com/software/info/mashup-center/ (Accessed 04 March 2012).

[15] ISO/IEC 9126-1. (2001) Software engineering – Product quality - Part 1: Quality model. ISO.

[16] Jackbe Presto Wire. [Online]. www.jackbe.com/ (Accessed 04 March 2012).

[17] J. Hierro, J., Janner, T., Lizcano,D., Reyes,M., Schroth,C. and Soriano,J.(2008) 'Enhancing User-Service Interaction Through a Global User-Centric Approach to SOA'. Paper presented at The Fourth International Conference on Networking and Services IEEE Computer Society, ICNS '08. Washington, DC, USA (2008).

[18] Liu, X., Hui, Y., Sun, W. and Liang, H. (2007) 'Towards service composition based on Mashup'. Paper presented at The IEEE Congress on Services, 9–13 July 2007, pp.332–339, Salt Lake City, Utah, USA.

[19] López, J., Pan, A., Bellas, F., and Montoto, P. (2008) 'Towards a Reference Architecture for Enterprise Mashups'. Paper presented at The Jornadas de Ingeniería del Software y Bases de Datos, 7-10 October 2008. Gijón, Spain.

[20] McCall, J.A., Richards, P.K., and Walters, G.F. (1977) Factors in Software Quality, RADC TR-77-369, 1977, Vols I, II, III, US Rome Air Development Center Reports. Italie. (1977).

[21] Nestler, T. (2008) 'Towards a Mashup-driven end-user programming of SOA-based applications'. Paper presented at The 10th International Conference on Information Integration and Web-based Applications & Services, iiWAS 2008, 24–26 November 2008, pp.551–554, Linz, Austria.

[22] Nestler, T., Dannecker, L. and Pursche, A. (2009) 'User-centric composition of service front-ends at the presentation layer'. Paper presented at The 2009 International Conference on Service-oriented Computing, ICSOC/ServiceWave, 24–27 November 2009. Stockholm, Sweden.

[23] Norman, D. and Nielsen, J.. (2010) 'Gestural Interfaces: A Step Backward In Usability'. Interactions' magazine, Vol. 17 Issue 5, September + October 2010 ACM New York, NY, USA.

[24] O'Reilly, T. (2005). 'What is Web 2.0 – design patterns and business models for the next generation of software', O'Reilly [Online] 30 September. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html. (Accessed 10 November 2011).

[25] Roy, M. (2010) 'Towards end-user enabled web service consumption for Mashups. International conference on software engineering'. Paper presented at The 32nd ACM/IEEE International Conference on Software Engineering, ICSE 2010, Vol. 2, pp.413–416, Cape Town, South Africa.

[26] Schroth, C. and Janner, T. (2007) 'Web 2.0 and SOA: converging concepts enabling the internet of services'. Journal of IT Professional, Vol. 9, No. 3, pp.36–41. (2007).

[27] Yahoo! Pipes [Online]. http://pipes.yahoo.com/pipes/. (Accessed 04 March 2012).

[28] Zhao, Z., Laga, N. and Crespi, N. (2009) 'The Incoming Trends of End-user driven Service Creation'. Paper presented at Digital Business : the first Iternational ICST Conference, DigiBiz, London, UK, June 17-19, 2009 Springer (Ed.) (2010) 98-108.

Palm Vein Verification Using Gabor Filter

Ali Mohsin Al-juboori[1,2], Wei Bu[1,3] , Xingqian Wu[1] and Qiushi Zhao[1]

**[1] School of Computer Science and Technology, Harbin Institute of Technology
Harbin, 150001, China**

**[2] College of Computer Science and Mathematics, University of al-Qadisiyah
,Iraq**

**[3]Department of New Media, Harbin Institute of Technology,
Harbin, 150001, China**

Abstract

Palm vein authentication is one of the modern biometric techniques, which employs the vein pattern in the human palm to verify the person. The merits of palm vein on classical biometric (e.g. fingerprint, iris, face) are a low risk of falsification, difficulty of duplicated and stability. In this research, a new method is proposed for personal verification based on palm vein features. In the propose method, the palm vein images are firstly enhanced and then the features are extracted by using bank of Gabor filters. Then Fisher Discriminated Analysis (FDA) is used to reduce the dimension of the features vectors. For vein pattern verification, this work uses Nearest Neighbors method. The EER of the proposed method is 0.2335%.

Keywords: Palm vein, Gabor Filter, EigenVein, FisherVein .

1. Introduction

Biometric technology refers to a pattern recognition system which depends on physical or behavioral features for the person identification. Many biometric systems exist today by using fingerprint, face, iris, etc. Palm vein is a new member of biometric family. Palm vein is defined as vascular patterns under the skin of the palm [1]. Like the fingerprint, the pattern of vain very state in the life and different in each part in same body. Because the vein pattern is hidden underneath the skin and invisible directly by the eye, the vein pattern is difficult to copy compared with other biometric types [2]. Besides, the palm vein is impossible to fake [1].The researcher and the communities are increasingly interested in vein pattern recognition. In [3] the researchers take the shape and texture of the hand vein for person authentication. They used Hausdorff distance and like edge mapping for shape authentication and Gabor filter for feature vein extraction. However, the researchers work on a database of 1600 images and get recognition rate is 80%, which makes this system have not a good result. In [2] the researchers analyzed the infrared back hand image. They used the minutiae features extracted from hand vein pattern for recognition. This pattern includes bifurcation point and ending point as fingerprint. However, they evaluated the method using small database (141 images), making it hard to draw strong conclusions. In [4] the researchers built a multimodal identification system based on fusion of the palm print and palm vein on image level. By using the novel integrated line preserving and contrast enhancement fusion method the palm print and palm vein are fused. The modified multiscale edge of palm vein and palm print images are combined additionally the image contrast and interaction point (IP) of palm vein line and palm print are enhanced. By using the IP, the feature vectors of the combine images are extracted. However, they implement the image acquisition using two separated cameras and requires a time consuming registration procedure, which makes it difficult to use in real time. In [14], the researchers worked on the same database (PolyU) that is used in the proposed method. They combined the palm print and palm vein. The method that is used to extract the vein is matching filter. The EER to the system is 0.3091%. However, they fused the palm print with palm vein features to evaluate the system. In [19], the researchers consider the palm vein as a piece of texture and apply texture based feature extraction techniques to a palm vein authentication. A 2D Gabor filter is applied for extracting the local features in the palm vein. The researcher proposed a directional code technique to encode the palm vein features in bit string representation called vein code. The similarity between two vein codes is measure by normalized Hamming distance. All the above studies they implemented using fusing multimodal or used a small database or the accuracy are low.

In this paper, we proposed a new method for palm vein extraction and features reduction dimensional and matching to get a less EER to make the used method is more secure. The remainder of this paper is organized as follows. Section 2, perform the palm vein features extraction using Gabor Filter. Section 3, used the fisher discriminated analysis to dimensional reduction and remove the redundancy in the features vector. Section 3, verify the test data by using Nearest Neighbors method.

2.The framework of the propose method

As show in figure 1, the propose method is consisted of three parts, preprocessing, feature extraction and matching. Then based on the matching method, can verify the user.

2.1 Preprocessing

Palm vein images are preprocessing by enhancement vein pattern before feature extraction. In the propose method, the method that is used for vein image enhancement is histogram equalization as show in figure 2.

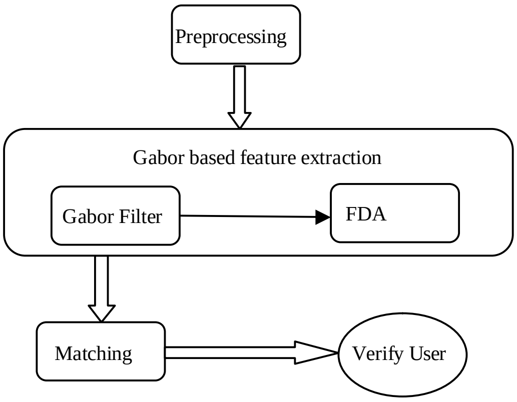


Fig. 1. Framework of our method

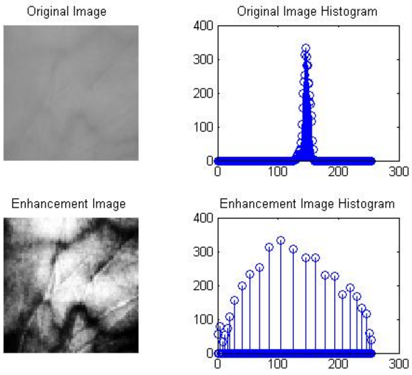


Fig. 2. Palm Vein Image Enhancement

3. Feature Extraction Based on Gabor Filter

The feature extraction is implement using Gabor filter. Gabor filter is a band pass filter which have orientation-selective and frequency-selective features and optimal joint resolution in both spatial and frequency domain [5, 6]. Gabor filters have been extensively used for extracting texture information, that they were powerful in capturing some specific local features in the texture image. A two-dimensional Gabor filter is a combine function with two components: a complex plane wave and a Gaussian-shaped function. It is defined as following:

$$G(x,y)=k\exp\left\{-\frac{1}{2}\left(\frac{x_\circ^2}{\sigma_x^2}+\frac{y_\circ^2}{\sigma_y^2}\right)+\hat{j}\,2\pi f_\circ x_\circ\right. \qquad (1)$$

$$x_\circ = x\cos\varnothing + y\sin\varnothing \qquad (2)$$

$$y_\circ = -x\sin\varnothing + y\cos\varnothing \qquad (3)$$

Where $k = {1}/{(2\pi\sigma_x\sigma_{y)}}$, $\hat{j}=\sqrt{-1},\theta$ is the orientation of Gabor filter, $f_\circ$ represent the filter center frequency, σ_x and σ_y are the scale of the Gaussian shape, $x_\circ$ and $y_\circ$ are the two vertical Gaussian axes. The most important parameters $f_\circ,\sigma_x$ and σ_y in Gabor function that make the filter appropriate for some specific application. The Gabor filter can be split to imaginary part and real part. The imaginary part (odd symmetric) Gabor filter is used for edge detection. The real part (even symmetric) Gabor filter is used for detecting the ridge in the image [7, 8, 16]. To analysis the Gabor filter in terms of real part and imaginary part, can be represented as following:

$$G_{mk}^{e}(x,y)=k\exp\left\{-\frac{1}{2}\left(\frac{x_\circ^2}{\sigma_x^2}+\frac{y_\circ^2}{\sigma_y^2}\right)\cos(2\pi f_{mk}x_{\circ k})\right. \qquad (4)$$

$$G_{mk}^{o}(x,y)=k\exp\left\{-\frac{1}{2}\left(\frac{x_\circ^2}{\sigma_x^2}+\frac{y_\circ^2}{\sigma_y^2}\right)\sin(2\pi f_{mk}x_{\circ k})\right. \qquad (5)$$

where m is the scale index, k is the channel index and f_{mk} is represent the center frequency of the real part and imaginary part of Gabor filter at the kth channel. After create a bank of Gabor filter, the enhanced palm image is convolved with the Gabor filter bank. The best output to the Gabor filter is depend on its parameters ($f_\circ,\sigma_x,\sigma_y$ *and* $\varnothing$).In the propose method, θ is begin from 0 to π by increment is equal to $\pi/8$ and the center frequency f_{mk} is change with the orientation. In [7] propose a method to determine the relation between σ and f_{mk} and we used it in the research which is defined as following

$$\sigma f_{mk} = \frac{1}{\pi}\sqrt{\frac{\ln 2}{2}}\,\frac{2^{\Delta\varnothing_{mk}}+1}{2^{\Delta\varnothing_{mk}}-1} \tag{6}$$

where $\Delta\varnothing_{mk}\,(\in[0.5, 2.5])$ is represent the spatial frequency bandwidth to the Gabor filter in the kth channel and m scale. $\Delta\varnothing$ are put as $\Delta\varnothing_1 < \Delta\varnothing_5 < \Delta\varnothing_2 < \Delta\varnothing_3 < \Delta\varnothing_4, \Delta\varnothing_2 = \Delta\varnothing_8, \Delta\varnothing_3 = \Delta\varnothing_7, \Delta\varnothing_4 = \Delta\varnothing_6$.

In the propose method we built a bank of Gabor filter with 8 channels and 8 orientations and the central frequency is change depending on Eq. (6). Figure 3 shows some sample of the bank of Gabor filter. Assume that I(x; y) denote a palm-vein image, F(x; y) denotes a filtered I(x; y), we can obtain

$$F(x,y) = \sqrt{(G^{e}_{mk}(x,y) * I(x,y))^2 + (G^{o}_{mk}(x,y) * I(x,y))^2} \tag{7}$$

where * denotes convolution in two dimensions. Thus, for a palm-vein image, 64 filtered images are generated by a bank of Gabor filters.

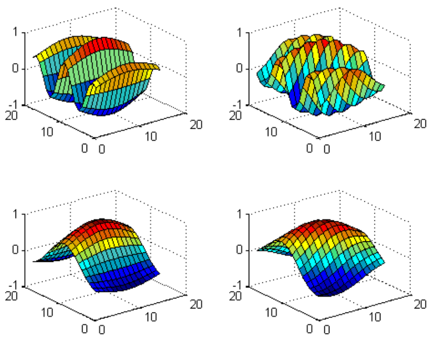


Fig. 3: A bank of Gabor filter

After creating the filter bank, the convolution operation is performed with the enhanced image in figure 2 with all the Gabor filters. The some sample of the results are shown in figure 4.

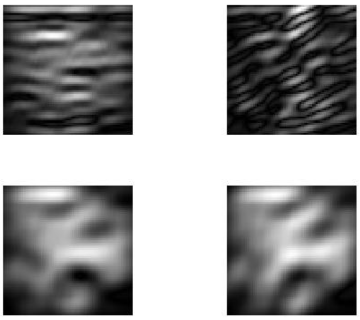


Fig. 4: The output of convolution operation

4. Dimensional Reduction

Dimensionality reduction of the feature set is a common preprocessing step used for pattern recognition and classification applications. Feature selection is effective in the data decreasing and increasing accuracy and improving the result of the pattern recognition system[9]. In many applications such as data mining, pattern recognition and information retrieval the data reduction is very important. LDA is the most popular dimensionality reduction. By implement the eigen-decomposition on the scatter features matrix of the training data can get an optimal projection of the LDA. In the propose method, when implement the Gabor filter and used all filtered images pixels value as a features vector, the number of features is more than the number of sample, that lead a non-stable solution of LDA. To solve this problem the scatter features matrix must be non-singular. The preprocessing steps as PCA (Principle Component Analysis) and SVD (Singular Value Decomposition) must be implemented to ensure the features scatter matrix is a non-singularity. PCA techniques, also known as Karhunen-Loeve methods, choose a dimensionality reducing linear projection that maximizes the scatter of all projected samples [10].

4.1 EigenVein

PCA aims to find a subspace whose basis vectors correspond to the maximum-variance directions in the original space. The features extracted by PCA are the best description of the data, but not the best discriminated features [20]. Assume a set of *N* sample images $\{x_1, x_2, \ldots, x_N\}$ taking values in an *n*-dimensional image space, and assume that each image belongs to one of *c* classes $\{X_1, X_2, \ldots, X_c\}$. Let us also consider a linear transformation mapping the original *n*-dimensional image space into an *m*-dimensional feature space, where *m* <*n*. The new feature vectors $y_k \in R^m$ are defined by the following linear transformation [11].

$$y_k = W^T x_k \qquad k = 1, 2, \ldots, N \tag{8}$$

where $W \in \mathbb{R}^{n\times m}$ is a matrix with orthonormal columns. If the total scatter matrix S_T is define as following:

$$S_T = \sum_{k=1}^{N}(x_k - \mu)(x_k - \mu)^T \tag{9}$$

where N is the number of sample images, and $\mu \epsilon \mathbb{R}^n$ is the mean image of all samples, then after applying the linear transformation W^T, the scatter of the transformed feature vector $\{y_1, y_2, \ldots, y_N\}$ is $W^T S_T W$. In PCA the projection W_{opt} is chosen to maximize the determinant of the total scatter matrix of the projected samples.

$$\begin{aligned} W_{opt} &= \arg\max_{W} \left| W^T S_T W \right| \\ &= \left[W_1, W_2, \ldots, W_m \right] \end{aligned} \tag{10}$$

where $\{W_i \,|\, i = 1, 2, \ldots, m\}$ is the set of n-dimensional eigen-vector of S_T corresponding to the m largest eigen-vector. Thus if PCA is implemented with images of vein, the projection matrix W_{opt} will contain principle components, we will refer it's as EigenVein.

4.2 Fisher Discriminated Analysis

Fisher Discriminated Analysis (FDA) is a well-known approach for feature extraction and dimension reduction. It computes a linear transformation by maximizing the ratio of between-class distance to within-class distance, thereby achieving maximal discrimination [21]. FDA finds the set of the most discriminated projection vectors that can map high dimensional samples onto a low-dimensional space. Using the set of projection vectors determined by FDA as the projection axes, all projected samples will form the maximum between-class scatter and the minimum within-class scatter simultaneously in the projective feature space [20].The FDA is find the set of basis vectors which maximizes the ratio between class scatter and within class scatter [11-13]. Let the between class scatter is define as following

$$S_B = \sum_{i=1}^{c} N_i (\mu_i - \mu)(\mu_i - \mu)^T \tag{11}$$

The within class scatter matrix be define as following:

$$S_w = \sum_{i=1}^{c} \sum_{x_k \in X_i} \left(x_k - \mu_i \right)\left(x_k - \mu_i \right)^T \tag{12}$$

Where μ_i is the mean image of class X_i , and N_i is the number of samples in class X_i. If S_w is a non-singular, the optimal projection W_{opt} is chosen as the matrix with orthonormal columns which maximizes the ratio of the determinant of the between class scatter matrix of the projection samples to determinant of the within class scatter matrix of the projection samples, i.e.:

$$\begin{aligned} W_{opt} &= \arg\max_{w} \frac{\left| W^T S_B W \right|}{\left| W^T S_w W \right|} \\ &= [W_1, W_2, ..W_m] \end{aligned} \tag{13}$$

where $\{W_i \,|\, i = 1, 2, \ldots, m\}$ is the set of generalized eigen-vector of S_B and S_W corresponding to the m largest generalized eigen-vector $\{\lambda_i \,|\, i = 1, 2, \ldots, m\}$ .i.e.

$$S_B w_i = \lambda_i S_w w_i, \qquad i = 1, 2, ..., m \tag{14}$$

To avoid the difficulties of a singular S_w, substitute the principle in Eq.(13). This method, which we call FisherVein, avoids this problem by projecting the images set to a lower dimensional space so that the resulting within class scatter matrix S_w is non-singular. This is implement by using PCA to reduce the dimension of the feature space to $N - c$ and then applying the standard FLD defined by Eq. (13) to reduce the dimension to $c - 1$.The W_{opt} will become [11]:

$$W_{opt}^T = W_{fld}^T W_{pca}^T \tag{15}$$

where

$$W_{pca} = \arg\max_{W} \left| W^T S_T W \right|$$

$$W_{fld} = \arg\max_{W} \left| \frac{W^T W_{pca}^T S_B W_{pca} W}{W^T W_{pca}^T S_W W_{pca} W} \right|$$

In this paper, the used database contains 500 different person palms (12 images to each person). Split the database into two sets and used one of the set images to obtain the eigen basis vectors. Then the remaining set is projected into those vectors. After implement PCA, the FDA finds a set of basis vector which maximizes the ratio between class scatter and within the class scatter.

5. Palm Vein matching

The nearest neighbor method is used to compute the matching between the train set and test set. To measure the similarity between two biometric feature vectors, we used Euclidean distance as a similarity measures. Let y denoted the test feature vector and $x_i^k, i = 1, .., C_k, k = 1, .., C$ represent the i^{th} gallery image of subject ID_k ,where C_k is the number of images of subject ID_k and C is the totally numbers of the images in the train set. The smallest Euclidean distance [17].

$$ID_y = \arg\min\nolimits_k \left\| y - x_i^k \right\|^2 \tag{16}$$

6. Experimental results

The Biometric Research Centre (UGC/CRC) at The Hong Kong Polytechnic University has developed a real time multispectral palm print capture device which can capture palm print images under blue, green, red and near-infrared (NIR) illuminations, and has used it to construct a large-scale multispectral palmprint database. Multispectral palmprint images were collected from 250 volunteers, including 195 males and 55 females. The age distribution is from 17 to 60 years old. The samples are collected in two separate sessions. In each session, the subject was asked to provide 6 images for each palm. Therefore, 24 images of each illumination from 2 palms were collected from each subject. In total, the database contains 6,000 images from 500 different palms for one illumination. The average time interval between the first and the second sessions was about 9 days. The proposed method used the near-infrared (NIR) illuminations images of PolyU multi-spectral palm print database [15].

To establish the sturdiness of the propose method in the experiment the total number of the palm vein images was 6000 images, which were collected from 500 person each with 12 images captured at two session. In verification, receiver operating characteristics (ROC) curve is used to show the behavior of the propose method. In the experimental randomly select 6 images from each person for training set and the other for testing set. The nearest neighbor method is used to verify the feature vector from test set with the train set feature vectors and take the minimum distance for verification. The distance distribution of genuine and impostor of the palm vein images is show in figure 5, and the ROC curve is show in figure 6. As show from figure 5 the EER is 0.2335% by using Euclidean distance. The Min–max normalization is used to normalize the matching scores. This normalization maps the raw matching scores to interval [0,1] and retains the original distribution of matching scores except for a scaling factor. Given that max(X) and min(X) are the maximum and minimum values of the raw matching scores, respectively, the normalized score is calculated as [18].

$$x' = \frac{x - \min(X)}{\max(X) - \min(X)} \qquad (17)$$

Two methods for palm vein authentication are proposed in [14] and [19] are also implemented for comparison. The method in Ref.[14] is tested on the same database. Table 1 show the comparison of our method and all above methods. Figure 5 show the distance distribution of the impostor and genuine and figure 6 show the ROC curve of the proposed method. From the result illustrate in table 1 and figure 5 and 6, we can find that the propose method has better performance from the methods that describe in [14] and [19].

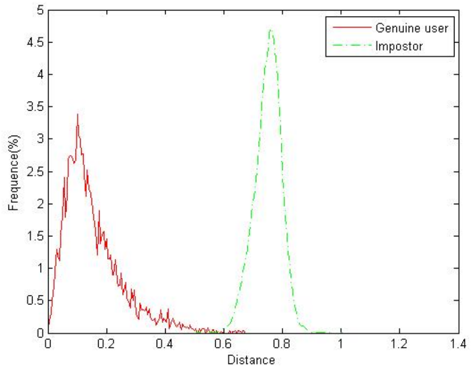


Fig. 5. Matching distance distribution of palm vein

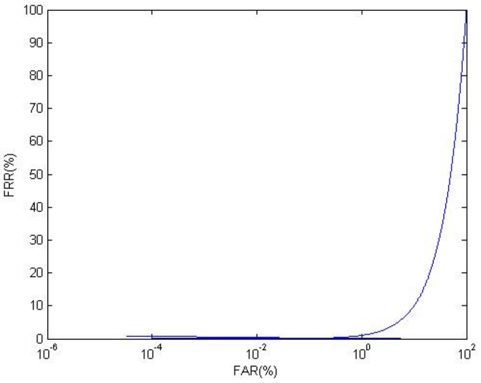


Fig. 6. ROC curve of palm vein

Table 1: Methods comparison on our palm vein images database.

Method	EER %
David [14]	0.3091 %
Lee [19]	1.6312%
Proposed method using EigenVein	6.5250%
Proposed method using FisherVein	0.2335%

The experimental results show that our method has better result than David [14] and Lee [19] and propose EigenVein methods. The main benefit of the proposed method, is that implement the Gabor filter with 8 scale and 8 direction and after that implement the dimensional reduction using FisherVein method that give best authentication features to reach to lowest EER value.

8. Conclusion

A new method of personal authentication based on palm vein has been discussed indetail. First, the palm vein images are enhancement using histogram equalization. Then a bank of Gabor filter is created and convolution on the enhanced images and used the convolution images as feature vectors. The dimensional reduction is implemented using FDA to get best features for verification. Finally, the palm vein verification was implemented using Nearest Neighbor classifier. In our used database of 6000 images to 500 person, we get an EER is 0.2335%.

Acknowledgement

This work was supported by the Natural Science Foundation of China (NSFC) (No. 60873140, 61073125 and 61071179), the Program for New Century Excellent Talents in University (No. NCET-08-0155 and NCET-08-0156), and the Fok Ying Tong Education Foundation (No. 122035).

References

[1] Ashok Rao, Mohammad Imran, Raghavendra R, Hemantha Kumar G," Multibiometrics: analysis and robustness of hand vein & palm print combination used for person verification", International Journal of Emerging Trends in Engineering and Technology ,Vol. I, No. 1, 2011,pp 11-20.

[2] LingyuWang, Graham Leedham, David Siu-Yeung Cho," Minutiae feature analysis for infrared hand vein pattern biometrics",Pattern Recognition, Vol. 41,2008, pp. 920 – 929.

[3] Zhonhli Wang, Baochang Zhang, Weiping Chen, Yongsheng Gao," A performance Evaluation of Shape and Texture based method for Vein Recognition", Congress on Image and Signal Processing, Vol. 2, 2008, pp. 659-661.

[4] Jian-Gang Wang, Wei Yun Yan, Andy Suwandy,Eric Sung, "Fusion of Palm print and Palm Vein Images for Person Recognition Based on Laplacianpalm Feature", IEEE conference on computer vision and image processing , 2007, pp. 1-8.

[5] Chao Ni, Qi Li, Liang Z. Xia,"A novel method of infrared image denoising and edge enhancement" , Signal Processing, Vol. 88, 2008, pp. 1606–1614

[6] Yi Hu, Xiaojun Jing, Bo Zhang, Xifu Zhu," Low Quality Fingerprint Image Enhancement Based on Gabor Filter ", International Conference on Advance Computer Control (ICACC), 2010, pp. 195-199.

[7] Jinfeng Yang , Yihua Shi, Jinli Yang," Personal identification based on finger-vein features", Computers in Human Behavior, Vol. 27, 2011,pp. 1565–1570.

[8] Jianwei Yang, Lifeng Liu, Tianzi Jiang, Yong Fan," A modified Gabor filter design method for fingerprint image enhancement", Pattern Recognition Latter,Vol. 24, No. 12, 2003, pp. 1805-1817.

[9] Lei Yu, Huan Liu "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", Proceedings of the Twentieth International Conference on Machine Learning, Washington DC, 2003, pp. 856-863.

[10] Deng Cai, Xiaofei He, and Jiawei Han "SRDA: An Efficient Algorithm for Large-Scale Discriminant Analysis", IEEE transaction on knowledge and data engineering, VOL. 20, NO. 1, 2008, pp. 1-12.

[11] Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", IEEE transactions on pattern analysis and machine intelligence, VOL. 19, NO. 7,1997, pp. 711-720.

[12] Deng Cai, Xiaofei He, Yuxiao Hu, Jiawei Han, Thomas Huang," Learning a Spatially Smooth Subspace for Face Recognition", IEEE computer society conference on computer vision and image processing , 2007, pp. 1-7.

[13] Hamid M. Hasan, Waleed A. AL.Jouhar , Majed A. Alwan," Face Recognition Using Improved FFT Based Radon by PSO and PCA Techniques", International Journal of Image Processing, Vol. 6, No. 1, 2012, pp. 26-37.

[14] David Zhang, Zhenhua Guo, Guangming Lu, Lei Zhang, Yahui Liu, Wangmeng Zuo," Online joint palmprint and palmvein verification", Expert Systems with Applications,Vol. 38, 2011,pp. 2621–2631.

[15] Multispectral PolyU database, www4.comp.polyu.edu.hk/~biometrics/.

[16] Jen-Chun Lee," A novel biometric system based on palm vein image ", Pattern Recognition Letters , Vol. 33, 2012, pp. 1520–1528.

[17] Muhammad Talal Ibrahim, YongjinWang, Ling Guan, Anastasios N. Venetsanopoulos," A Filter Bank Based Approach for Rotation Invariant Fingerprint Recognition ", J Sign Process Syst,Vol. 68,2012, pp. 401–414.

[18] Mingxing He, Shi-JinnHorng, PingzhiFan, Ray-ShineRun, Rong-JianChen, Jui-LinLai, MuhammadKhurramKhan, KevinOctaviusSentosa," Performance evaluation of score level fusion in multimodal biometric systems ", Pattern Recognition, Vol. 43, 2010, pp. 1789–1800.

[19] Jen-Chun Lee "A novel biometric system based on palm vein image", pattern recognition letter, Vol. 33,2012, pp. 1520-1528.

[20] Lei Wang, Hongbing Ji, Ya Shi," Face recognition using maximum local fisher discriminated analysis",18th IEEE International Conference on Image Processing, 2011,pp.1737-1740.

[21] Jing Liu, Yue Zhang,"Palm-Dorsa Vein Recognition Based on Two-Dimensional Fiher Linear Discriminated ", Proceeding

of International Conference on Image Analysis and Signal Processing, 2011,pp. 550-552.

Ali Mohsin Al-juboori received his B. S. degree in 2002 and M. Sc. degree in 2005 in Computer Science both from the college of Science, Al-Nahrain University, Baghdad, Iraq. He is now a Ph. D. candidate at the School of Computer Science and Technology of Harbin Institute of Technology. His research interests include Biometrics, Pattern Recognition, and Image Processing.

Wei Bu received her B.Sc., M.Sc. and Ph.D. from Harbin Institute of Technology (HIT), Harbin, China in 2000, 2006 and 2010, respectively. Now she is a post doctor in the School of Computer Science and Technology, and a lecturer in the Department of New Media, HIT. Her current research interests include pattern recognition, image analysis, biometrics and digital art design, etc.

Xinagqian Wureceived his B.Sc., M.Sc. and Ph.D. in computer science from Harbin Institute of Technology (HIT), Harbin, China in 1997, 1999 and 2004, respectively. Now he is a professor in the School of Computer Science and Technology, HIT. His current research interests include pattern recognition, image analysis and biometrics, etc.

Qiushi Zhao received his B. S. degree and M. S. degree in Computer Science and Technology from the School of Computer Science and Information Technology, Northeast Normal University, Changchun, P. R. China. He is now a Ph. D. candidate at the School of Computer Science and Technology of Harbin Institute of Technology. His research interests include Biometrics, Pattern Recognition, etc.

Knowledge Management System as Enabler for Knowledge Management Practices in Virtual Communities

Setiawan Assegaff[1], Ab Razak Che Hussin[2], Halina Mohamed Dahlan[3]

**[1] Program Magister Sistem Informasi, STIKOM Dinamika Bangsa
Jambi, 36138, Indonesia**

**[23] Faculty Computer Science and Information System, Universiti Teknologi Malaysia
Skudai, Johor 183000, Malaysia**

Abstract

Knowledge Management System was recognized as one of the key enablers in a Knowledge Management initiative. This is because KMS have been prove bring value for Knowledge Management initiatives such as eliminate distance and time barriers. Furthermore KMS also made KM more effective for the organization. As general KMS have two main functions, first managing people interaction and managing information/knowledge. Because knowledge creates from the interaction of the people, than KMS has a vital function in knowledge creation with managing people interaction. In KM on an organization, they conduct interaction by developing some activities such as: Communities of Practices, Communities of Interest, Peer Assist and Share Learning as a method to create and leverage the knowledge. The activities could be conducted manually or virtually. In support that activities done in virtual, using IT could bring potential value. In this study we would like to propose the framework for organizations on how to implement KMS as a powerful enabler for KM in virtual communities.

Keywords: *Information Technology, Knowledge Management System, Knowledge Management Practices, Knowledge Management, Virtual Communities*

1. Introduction

Information Technology (IT) has been used for a long time in support activities in organizations. IT use in the organization to make numerous contribution such as reducing time, cost, support better services for customers. In the knowledge era, practitioners also consider IT to support KM. IT use in Knowledge Management (KM) in the various methods [1].

Many applications have been developed and used to support KM. Social network software, video/teleconference, organization directories, e-mail, e-learning, repositories were potential tools in support KM [1].

IT founded very potential in support KM. The main function of IT in KM is to support and enabler KM process. IT use in KM known as Knowledge Management System (KMS) [2].

Implementing the KMS has been considered as an important part of the KM project. It is believed that KMS give huge opportunities to break down barriers by making the information presented at every level and units in organization hence it will help to enhance organization becomes more effective [3]. However the vital function of KMS is about managing people interaction. This is because knowledge creates from people interaction. In this study we would propose the framework that can use for organizations to develop and implement KMS especially focus on support people interaction in virtual communities. When developing KMS organizations should pay attention in social such as people interaction and technological aspect such as IT itself [4].

This paper consists of six sections. In section two KM will be defined and discussed. Section three will explains about KM practices and virtual communities. Section four explains our overview in KMS as socio-technical system. Section Five would discuss ours propose framework on how KMS can be used to support KM practices in organizations. The last sections propose a summarization and highlight some factors for attention in using KMS as enabler in KM.

2. Knowledge Management in Organization

Major Scholars still have not had an agreement about what knowledge management is [5]. However some of KM definitions have been used widely by KM communities. One of famous KM definitions was by Karl and Erik

Sveiby. They defined KM as: Knowledge management is creating value by leveraging intangible assets.

We highlight there are three important components in KM is about; first, is about managing people, it related with at networking; collaboration; the second is about managing knowledge/information, it is related to accessibility, searching, validating, taxonomy, up-to date, knowledge flooding and Managing Information technology is related with information/ knowledge security, speed, and reliability

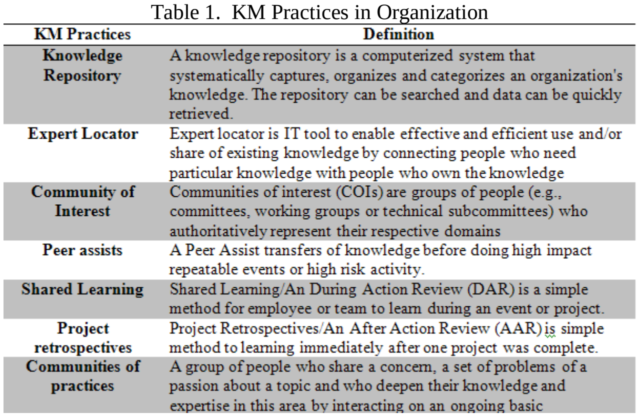
Knowledge is created when interaction among people in an organization occurs. Nonaka believed that an organization can create and utilizes knowledge through converting tacit knowledge, and vice versa. He proposes SECI model to figure four modes of knowledge conversion [6], which can be described follows; four modes of knowledge conversion were identified:

1. **Tacit to Tacit (Socialization)** - This dimension explains Social interaction as tacit to tacit knowledge transfer, sharing tacit knowledge through face-to-face or share knowledge through experiences.
2. **Tacit to Explicit (Externalization)** - Between tacit and explicit knowledge by Externalization (publishing, articulating knowledge), which embed the combined tacit knowledge which enable its communication.
3. **Explicit to Explicit (Combination)** - Explicit to explicit by Combination (organizing, integrating knowledge), combining different types of explicit knowledge, for example building prototypes.
4. **Explicit to Tacit (Internalization)** - Explicit to tacit by Internalization (knowledge receiving an application by an individual), enclosed by learning by doing; on the other hand, explicit knowledge becomes part of an individual's knowledge and will be assets for an organization.

3. Knowledge Management Practices and Virtual Communities in Organization

KM has been recognized by organizations for decades. On many KM implementations, most of the organizations focus their KM on knowledge creation and how it can leverage in their organization [6]. To achieve that, some methods and roles have been developed. One of successful methods was by creating interaction among people in communities. Communities had proven as powerful entities could use by organizations for knowledge creation and knowledge leverage [7]. Some of communities interaction could develop in a KM initiative in an organization such as: Community of interest (CoI), Community of Practices (CoP), Share Learning (SL), Project Retrospective (PR), and Peer Assist (PA). Beside that an organization could implement knowledge repositories and expert locator to support the communities. Table 1 bellow explain the activities interaction among communities.

Table 1. KM Practices in Organization



KM Practices	Definition
Knowledge Repository	A knowledge repository is a computerized system that systematically captures, organizes and categorizes an organization's knowledge. The repository can be searched and data can be quickly retrieved.
Expert Locator	Expert locator is IT tool to enable effective and efficient use and/or share of existing knowledge by connecting people who need particular knowledge with people who own the knowledge
Community of Interest	Communities of interest (COIs) are groups of people (e.g., committees, working groups or technical subcommittees) who authoritatively represent their respective domains
Peer assists	A Peer Assist transfers of knowledge before doing high impact repeatable events or high risk activity.
Shared Learning	Shared Learning/An During Action Review (DAR) is a simple method for employee or team to learn during an event or project.
Project retrospectives	Project Retrospectives/An After Action Review (AAR) is simple method to learning immediately after one project was complete.
Communities of practices	A group of people who share a concern, a set of problems of a passion about a topic and who deepen their knowledge and expertise in this area by interacting on an ongoing basic

From that activity, knowledge have been created and leveraged. Each activities base on communities have their own benefits and characteristics. For example CoI is best for organization in getting idea from communities furthermore Peer assists give opportunity for a team project to learn from their senior. All of the activities above are about people interaction. It is important for organizations to clearly understand the benefit and values from each community activities and decide the best activities they could develop for their KM.

The concepts of community activities itself is relevant to the SECI model of Nonaka. SECI Model describes people interaction and how knowledge could create and leverage in some different activities. In SECI model is merely about socialization aspect.

Nowadays with support of technologies the activities could conduct in virtual way known as virtual communities [1]. Virtual communities consist of a large number of people, who is connecting each others in the internet and exchange views on specific subjects. Virtual communities are similar to virtual organizations (VO) in many aspects. [8]. Both, VOs as well as communities need support for information sharing, for communication and for sharing of resources across organizational borders [9]

Support for virtual communities by using collaboration and knowledge management application base on web technology has to meet several challenges. Firstly, communication between the community members has to be improved to reduce the geographical and cultural distances. Secondly, simplified and effective sharing of knowledge has to be enabled. A structured knowledge base is an important step in (re) as common knowledge. Thirdly, management of a community has to be simplified. The goal of virtual communities is about communication and collaboration [9].

4. Knowledge Management System

Information technology used in KM known as Knowledge Management System (KMS). In common, Knowledge Management Systems (KMS) are IT that enables organizations to manage effective and efficient knowledge. Some definition of KMS has been proposed by some researchers. One of widely use is the KMS definition of Alavi and Leidner [3] . They defined KMS as a class of information systems applied for managing organizational knowledge. Another perspective of KMS comes from Ericsson, F. & Avdic, A. (2004). They defined KMS as a system that increase organizatinal performance by increase the better decision by employee when they use knowledge in daily work activities [10].

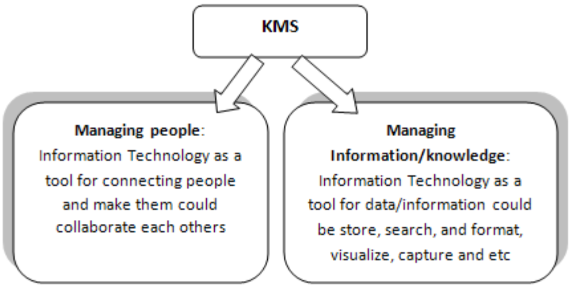

KMS
Managing people:
Information Technology as a tool for connecting people and make them could collaborate each others
Managing Information/knowledge:
Information Technology as a tool for data/information could be store, search, and format, visualize, capture and etc

Fig 1. KMS Function

From the definition of function KMS above, we have highlighted two of the elements that should exist in KMS. First KMS should have ability to connecting people, its mean by hardware and software KMS enable people to support interaction among people in communities, communicate with them and making collaboration. Another KMS function is to manage information/knowledge in order help people to reuse knowledge and make better decision with their knowledge [13]. Figure 2 above describes the function of KMS as general

5. KMS as Enabler

In KM. IT is enabler of human-based methodology. Most of KM reseachers agree that knowledge does not exist in technical elements-they only exist in human beings who are able to act upon the knowledge [4]. Therefore, no technology by itself couldl be exist for the creation and leveraging of knowledge assets without integration and tailored to the needs of specific communities of people. SECI model is very good to describe the proces in knowledge creation and how it could be leverage by people [6].

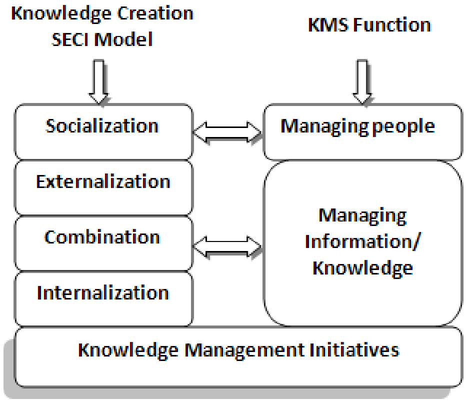


Fig 2. SECI link to KMS Function

Figure 2 above describes how KMS function has potential aspect to enable KM initiatives. KMS has In fact, the KMS have a great opportunity to support KM in organizations [14]. KMS has many opportunities to bring benefit in KM such as KMS could help people to connect with the right people, in example if an engineering face a problem in his/her job he/she can use IT tool in KMS to find expert easier compare if he/she do in manual.

However some aspect should identify carefully when organizations have a plane to decide what type of KMS to support their KM activities. If the organization focus on a specific kind of their knowledge, it is better for them to develop KMS in very specific function [5], otherwise they could develop KM in general function. In this case KMS would have all functions such as managing people interaction and managing knowledge/information.

To help organizations aim their goal in developing and implemented KMS. We propose a framework to guide them what aspects they should carefully attend and awareness in KMS. Figure 3 below describes our framework.

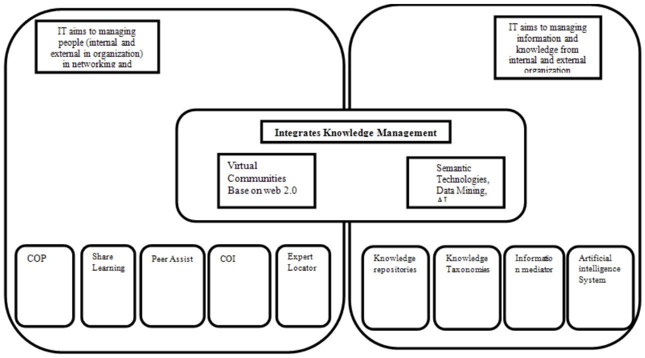

IT aims to managing people (internal and external in organization) in networking and
IT aims to managing information and knowledge from internal and external organization
Integrates Knowledge Management
Virtual Communities Base on web 2.0
Semantic Technologies, Data Mining, AI
COP
Share Learning
Peer Assist
COI
Expert Locator
Knowledge repositories
Knowledge Taxonomies
Information mediator
Artificial intelligence System

Fig 3. KMS as Enabler for KM Practices in Virtual Communities

The aspects that organizations should also consider when develop KMS such as: KMS today is about web technology and the environment. People have a chance to interact each other as virtual. Some trend in virtual communities in the KM base on web technology are Virtual CoP, Virtual CoI and Virtual Peer Assist. Some research has identified in virtual community factors like trust is becoming a very important issue [11]. When organization decides to conduct virtual communities they must careful attention to the issues beside technology part, In fact, the manual procedure to conduct virtual communities should develop [12].

6. Conclusions

KMS has great opportunity as enable for virtual communities in KM. The organization should pay attention both in social and technological aspect when develop and implement KMS. As general KMS have two functions: managing people interaction and managing knowledge/information. The organization should carefully decide what type of KMS they should adapt. Each of KMS has different treatments and factors for attention. The right decision of choosing tools for support and enabler KM practices in an organization is critical and would impact the benefit of KM process.

References

[1] I. Ajiferuke and A. Markus, ". . Potentials of Information Technology in Building Virtual Communities," in *Encyclopedia of Multimedia Technology and Networking*, I. M. P. (Ed.), Ed., Ed. Hershey: PA: Information Science Reference, 2005, pp. (Pp. 836-841).

[2] R. McDermott, "Why Information Technology Inspired But Cannot Deliver Knowledge Management," *California Management Review*, vol. 41, 1999.

[3] M. Alavi and D. E. Leidner, "Review: Knowledge Management and Knowledge Management Systems Conceptual Foundations and Research Issues," *MIS Quarterly* vol. 25 (1), pp. 107-136, 2001.

[4] M. Jelavic, "Socio-Technical Knowledge Management and Epistemological Paradigms: Theoretical Connections at the Individual and Organisational Level," *Interdisciplinary Journal of Information, Knowledge, and Management* vol. Volume 6, 2011.

[5] Hansen, *et al.*, "What's your strategy for managing knowledge? Harvard Business Review,". *Taylor \& Francis*, vol. 77(2), 106-116, 187, (1999).

[6] I. Nonaka, "A Dynamic Theory of Organizational Knowledge Creation," *Organization Science*, vol. Vol. 5, No. 1, pp. 14-37, 1994.

[7] E. Wenger, "Knowledge management as a doughnut: Shaping your knowledge strategy through communities of practice," *Ivey Business Journal Online*, vol. January/February, 2004.

[8] D. Langenberg, *et al.*, " Knowledge Management in Cloud Environments " *I-KNOW Conference* 2011.

[9] v. D. Langenberg and M. Welker, "Knowledge management in virtual communities," *Open Journal of Knowledge Management*, 16. Mai 2011.

[10] F. Ericsson and A. Avdic, "Knowledge Management Systems Acceptance," *Knowledge Management: Current Issues and Challenges*, pp. (Pp. 39-51), 2003.

[11] A. Alexander, *et al.*, "Motivation and barriers to participation in virtual knowledge-sharing communities of practice," *Journal of Knowledge Management*, vol. 7, pp. 64-77, 2003.

[12] J. Li, *et al.*, "Exploring the Contribution of Virtual Worlds to Learning in Organizations," *Human Resource Development Review*, vol. 10, pp. 264-285, September 1, 2011.

[13] D. V. Subramanian and A. Geetha, "Evaluation Strategy for Ranking and Rating of Knowledge Sharing Portal Usability," *IJCSI International Journal of Computer Science Issues*, Vol. 9, 2012.

[14] ARC Hussin and Setiawan Assegaff, "Knowledge Management System as Socio-Technical System", *IJCSI International Journal of Computer Science Issues*, Vol. 9, 2012.

Setiawan Assegaff is a PhD candidate at the Faculty of Computer Science and Information System, University Teknologi Malaysia. His education includes BS and MS in Information System, University of Gunadrma, Indonesia in 2000 and 2003. His research interests focus on Knowledge Management, Technology Adoption and Computer and Society.

Ab Razak Che Hussin is a senior lecturer in Faculty of Computer Science and Information System, Universiti Teknologi Malaysia. He received his PhD from University of Manchester in 2006 on the field of Trust in E-commerce. His research interests focus on information system, web application and trust and privacy in e-commerce.

Halina Mohamed Dahlan is a senior lecturer in Faculty of Computer Science and Information System, Universiti Teknologi Malaysia. She received his PhD from University of Manchester in 2008 on the field of Intelligent Decision Support. Her research interests focus on business intelligent, evolutionary computing, and fuzzy logic.

Forecasting Russian renewable, nuclear, and total energy consumption using improved nonlinear grey Bernoulli model

Hsiao-Tien Pao[1] , Hsin-Chia Fu[2], Hsiao-Cheng Yu[3]

**[1] Department of Management Science, National Chiao Tung University,
Taiwan, ROC**

**[2] College of Engineering, Huaqiao University,
Quanzhou, Fujian 362021, China**

**[3] Graduate Institute of Technology Management, National Chiao Tung University,
Taiwan, ROC**

Abstract

Forecasts of renewable, nuclear, and total primary energy consumption are essential for a green energy system and the understanding of climate change in a rapidly growing market such as Russia. In this paper, nonlinear grey Bernoulli with power *j* model (NGBM^{j}) is applied to predict these three different types of energy consumption. A numerical iterative method to optimize the powers of NGBM using mathematical software is also proposed. The NGBM with optimal power model is named NGBM^{op}. The forecasting ability of NGBM^{op} has remarkably improved, comparing with the grey model. For each time series, the best NGBM^{op} provides an accurate and reliable multi-step prediction with a MAPE value of less than 2.90 during the out-of-sample period of 2004-2009. The prediction results show that Russia's compound annual renewable, nuclear, and total energy consumption growth rates are set respectively at 1.95%, 2.44%, and 0.88% between 2010 and 2015.

Keywords: *Grey prediction model; Nonlinear grey Bernoulli model; Nuclear; Renewable; Russia.*

1. Introduction

A good forecasting technique is prerequisite for studies of green energy systems, not only for the cost-effectiveness of investment planning but also for the monitoring of environmental issues as well as demand side management. Forecasting studies for different types of energy consumption, e.g. renewable, nuclear, and total primary energy consumption, constitute a pivotal part of green energy policies, especially for an emerging market like Russia. Russia spreads out over a vast swath of land from the searing pre-Caspian deserts to the Arctic tundra and extends across 11 time zones (GMT+2 to GMT+12). The impact of climate change, including the adverse socio-economic consequences of natural hazards, plays a critical role in the spatial and economic development of the country [1]. Therefore, an accurate prediction model is necessary for the clean energy system in Russia.

The prediction method includes multivariate models, univariate time-series models, and nonlinear intelligent models. A limitation of multivariate models is that their predictive ability depends on the availability and reliability of the independent variables data. Univariate time-series models only need the historical data of the target variable to predict its future behavior; however, they require many observations in order to produce accurate forecasts. Due to the instability of energy consumption, nonlinear intelligent prediction models have been employed, such as artificial neural network (ANN) [2-4], fuzzy regression [5, 6], and some of the hybrid models [7, 8], in order to more efficiently forecast the demand for energy. However, the prediction accuracy of the above-mentioned nonlinear models also relies on the number of training data and its representation. In developing countries, the trend of energy consumption may change rapidly over time. Therefore, only the most recent sample data are adequate for the prediction of renewable, nuclear, and total primary energy consumption. Grey prediction models, on the other hand, are appropriate when dealing with rapidly changing data because of their low data requirements.

Grey theory was first proposed by Deng [9] in 1982 and has been widely used in forecasting studies. When compared with other forecasting techniques, advantages of grey prediction model include no statistical distribution of data, small sample requirements, and high prediction accuracy [10, 11]. One of the main characteristics of grey theory is the accumulated generating operation (AGO), and its aim is to reduce the source data to a monotonic increasing series. Some of the modified nonlinear grey hybrid models have been proposed, such as Pao et al. [12], Taguchi-grey [13], grey-Markon [14], trigonometric-grey [15], and gray-based learning model [16]. They are not only complex mathematical inference but also difficult to apply.

The nonlinear grey Bernoulli with power j model (NGBM^j) was named by Chen et al. [17] and was first mentioned in the book by Liu et al. [18]. NGBM is built based on the modification of Bernoulli differential equation in the GM model [19]. The power j in the Bernoulli differential equation can be adjusted to achieve the best prediction performances. Pao et al. [12] proposed a numerical iterative method to optimize power j in NGBM to improve model precision and the best power NGBM is named as NGBM^{op}.

The remainder of this paper is organized as follows. Section 2 outlines the GM and NGBM approaches. Section 3 presents the forecasting results and discussions. Finally, the last section concludes the paper.

2. Methodology

This section describes nonlinear grey prediction models GM (1, 1), NGBM^j (1, 1), and NGBM^{op} (1, 1). Both GM and NGBM^{op} are employed to forecast three different energy consumptions of Russia from 2009 to 2015, namely renewable, nuclear, and total primary energy consumption. NGBM^{op}'s abilities of multi-period forecasts are compared with the GM by using the out-of-sample during 2003-2008 for renewable and total energy consumption, and 2004-2009 for nuclear energy consumption, where the in-sample period is during 1997-2002 or 1998-2003.

One of the advantages of GM (1, 1), NGBM^j (1, 1), and NGBM^{op} (1, 1) grey prediction models is that they can utilize a limited amount of data to achieve accurate predictions. The value '1' in the first dimension for grey prediction models means that only one variable needs to be forecasted, and the other '1' represents the first order grey differential equation to build a grey model. Grey theory was proposed by Deng [9], the detail algorithm of GM (1, 1) was described by Pao [11], and the detail algorithms of NGBM^j (1, 1) and NGBM^{op} (1, 1) were described by Pao [12].

Based on the modification of Bernoulli differential equation in the GM model [18], the algorithm of NGBM^j (1, 1) can be summarized as follows.

Considering the non-negative time-series data:

$$v^{(0)}=[v^{(0)}(0),v^{(0)}(1),\cdots,v^{(0)}(i),\cdots,v^{(0)}(n)],\ \text{where } n\geq 3 \quad (1)$$

NGBM^j (1, 1) is as follows:

$$v^{(0)}(k)+\alpha W^{(1)}(k)=\beta\left[W^{(1)}(k)\right]^{j},\ j\in R, \quad (2)$$

where

$$W^{(1)}(k)=0.5[v^{(1)}(k)+v^{(1)}(k-1)],\ k=1,\ 2,\dots,n \quad (3),$$

$$v^{(1)}(k)=\sum_{i=0}^{k}v^{(0)}(i),\ k=0,1,\cdots,n \quad (4)$$

and $v^{(1)}(0)=v^{(0)}(0).$ (5)

$v^{(1)}(k)$ is obtained by accumulated generating operation (AGO). The optimal value of power j in Eq. (2) is determined by the minimum mean absolute percentage error. NGBM^j is reduced to GM when j=0, and it is reduced to grey Verhust model when j=2 [18]. The parameters α and β can be estimated as

$$[\alpha,\beta]^T=\left[D^TD\right]^{-1}D^Ty_n$$

where

$$D=\begin{bmatrix}-W^{(1)}(1) & \left[W^{(1)}(1)\right]^j\\ -W^{(1)}(2) & \left[W^{(1)}(2)\right]^j\\ \vdots & \vdots\\ -W^{(1)}(n) & \left[W^{(1)}(n)\right]^j\end{bmatrix} \text{ and } y_n=\begin{bmatrix}v^{(0)}(1)\\ v^{(0)}(2)\\ \\ v^{(0)}(n)\end{bmatrix},\ j\in R \quad (6)$$

Following is the response equation

$$\hat{v}^{(1)}(k)=\left[\left(v^{(0)}(0)^{(1-j)}-\frac{\beta}{\alpha}\right)e^{-\alpha(1-j)k}+\frac{\beta}{\alpha}\right]^{1/(1-j)},\ j\neq 1 \text{ and } k=0,1,\cdots \quad (7)$$

By performing inverse accumulated generating operation (IAGO) on $\hat{v}^{(1)}(k+1)$, the predicted value of $\hat{v}^{(0)}(k+1)$ is

$$\hat{v}^{(0)}(k+1)=\hat{v}^{(1)}(k+1)-\hat{v}^{(1)}(k),\ k=0,1,\dots \quad (8)$$

where $\hat{v}^{(0)}(1),\ \hat{v}^{(0)}(2),\dots,\ \hat{v}^{(0)}(n)$ is called a fitted sequence, and $\hat{v}^{(0)}(n+1),\ \hat{v}^{(0)}(n+2),\dots$ are prediction values.

Three different statistics: RMSE, MAE, and MAPE are employed to evaluate the accuracy of the forecasts using the out-of-sample period. Lewis [20] developed a scale to evaluate forecasting performance. In this scale, if the value of MAPE is lower than 10%, it is considered highly accurate. 10-20% is good, 20-50% reasonable, while greater than 50% is considered inaccurate. The power j in NGBM can be adjusted to minimize the MAPE value using a numerical iterative method. In the next section, the iterative results will demonstrate that parameter j is efficient in improving the model precision. Also, the prediction results of NGBM with optimal power j model

($NGBM^{OP}$) are compared with the results of GM (1, 1) models.

3. Forecasts and Discussion

3.1 Data analysis

In this research, we collected annual total data from EIA of renewable (R) and total energy consumption (TE) for the period from 1997 to 2008 as well as nuclear energy consumption (NE) and CO_2 emissions (CO2) from 1997 to 2009. Real GDP data between 1997 and 2009 were obtained from World Development Indicators (WDI), and are measured in 2000 US dollars. The three different types of energy consumption are measured in quadrillion Btu (British thermal unit). CO_2 emissions are measured in Million Metric Tons (MMT), produced by the burning of fossil fuels and the manufacture of cement.

Table 1 displays the summary statistics associated with the Russian energy consumption series. The trends of these series are shown in Figs. 1-3. In Table 1, nuclear energy consumption exhibits the largest related variation (defined by coefficient of variation (CV)) while renewable shows the smallest related variation. Table 2 shows the average growth rates of the three different types of energy consumption, emissions, and real GDP. Fifteen-year (1993~2008), ten-year (1998~2008), and five-year (2003~2008) growth rates are respectively calculated to demonstrate the long-term, medium-term, and short-term growth trends. For the short-term period, the Russian compound annual growth rate (CAGR) in real GDP is 7.09%, which is almost 2.1 times higher than the world CAGR of 3.41%; nuclear boasts a CAGR of 1.76%, almost 3 times higher than the world CAGR of 0.59%. But Russia has lower CAGRs in both renewable (0.47%) and total energy use (1.31%) than the world CAGRs of respectively 3.67% and 2.98%. Russia’s CAGR in emissions is 0.88%, almost 3.7 times lower than the world CAGR of 3.26%. In addition, the Russian long-term growth rates in both emissions (-0.72%) and energy use (-0.03%) are negative, while the world growth rates are positive in emission (2.30%) and energy use (2.42%). The results show that Russia is a booming market and that the government effectively conserved energy resources, controlled emission, developed clean energy, and responded to climate change in the past five years.

Table 1: Descriptive statistics for Russia from 1997 to 2008

Renewable (Quadrillion Btu)			Nuclear (Quadrillion Btu)			Total (Quadrillion Btu)		
Mean	S.D.	CV(%)	Mean	S.D.	CV(%)	Mean	S.D.	CV(%)
1.69	0.07	4.14	1.43	0.20	13.98	28.06	1.62	5.70

Table 2: Compound annual growth rates towards 2008 for each variable (in percentages)

	Russia					World				
	R	NE	TE	GDP	CO2	RE	NE	TE	GDP	CO2
15-year	-0.57	2.06	-0.03	2.79	-0.72	2.33	1.41	2.42	3.10	2.29
10-year	0.16	4.57	1.69	6.84	1.35	2.48	1.11	2.58	3.08	2.83
5-year	0.47	1.76	1.31	7.09	0.88	3.67	0.59	2.98	3.41	3.26

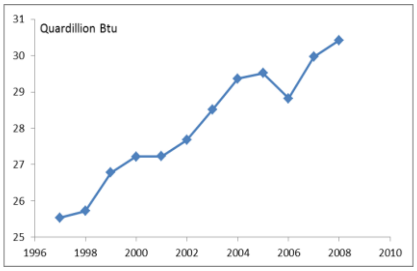


Fig. 1. Trend plot of total energy consumption from 1997 to 2008.

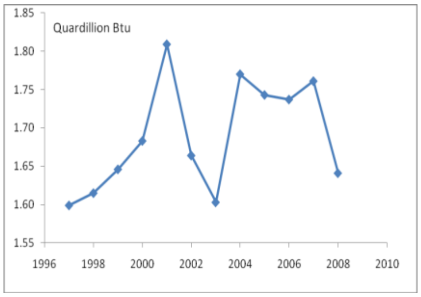


Fig. 2. Trend plot of renewable energy consumption from 1997 to 2008.

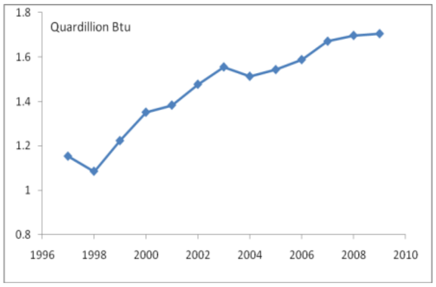


Fig. 3. Trend plot of nuclear energy consumption from 1997 to 2009.

3.2. Forecasting results

The multi-step forecasting performances of the $NGBM^{op}$ models are compared with the GM (1, 1) models by using out-of-sample actual data during the period 2003-2008 for renewable and total energy consumption and 2004-2009 for nuclear energy consumption. For each variable, the GM/NGBM^{op}-k (k=6, 5 or 4) models connect six-year (GM-6, 1997-2002 or 1998-2003), five-year (GM-5, 1998-2002 or 1999-2003), and four-year (GM-4, 1999-2002 or 2000-2003) data sets as the in-sample period. The in-sample data are employed to build models, and the out-of-sample data are used to evaluate the prediction accuracy by using RMSE, MAE, and MAPE statistics. The best prediction model enjoys the lowest value of MAPE. For NGBM^{i}, the proposed numerical iterative method with MAPE value is employed to determine the optimal power i. Figs. 4-6 show the impact on the MAPE values in NGBM when the powers i are set to -0.2 to 0.2 with 0.01 increments for the three different types of energy consumption. Figures show that the proposed iteration method is an effective optimization algorithm for the power selection of NGBM. In particular, for all tested i, alternatives 0.07, 0.14, and 0, each for renewable, nuclear, and total energy consumption, have the lowest MAPE in NGBM^{i} (1, 1).

Three observations can be thus made, the first of which is that the best GM (1, 1) models for renewable, nuclear, and total energy consumption are GM-4, GM-5, and GM-6 with MAPE values being respectively 4.01, 13.89, and 1.39, as shown in Table 3. The parameters a and b for the best GM models are shown in Table 4. The RMSE, MAE and MAPE statistics for all of the GM-k (k = 4, 5 and 6) models are shown in Table 3. As we can see, the ranges of the MAPE values are 4.01–7.27, 13.89–18.28, and 1.39–3.42 respectively for renewable, nuclear, and total energy consumption. According to Lewis's criteria [20], GM model presents a highly accurate forecast for renewable and total energy consumption and a good forecast for nuclear. Secondly, the best NGBM^{op} for renewable, nuclear, and total energy consumption, as shown in Table 3, are NGBM$^{0.07}$-6, NGBM$^{0.14}$-5, and NGBM^{0}-6 with respective MAPE values of 2.90%, 2.20%, and 1.39% where NGBM^{0}-6 is equal to GM-6. The parameters a and b for the best NGBM^{op} models are shown in Table 4. As shown in Table 3, we can see that the ranges of the MAPE values are 2.90–3.13%, 2.20–2.93%, and 1.39–1.58% respectively for renewable, nuclear, and total energy consumption, which are much lower than Lewis's criteria, 10%. Thus, NGBM^{op} model presents a highly accurate forecast for the three different types of energy consumption. Thirdly, Figs. 4-6 show that the proposed numerical iterative method is an effective optimization algorithm for choosing optimal powers *i* in the NGBM to improve the accuracy of the model. Finally, this study finishes by using the best NGBM^{op} model to forecast the three types of energy consumption for Russia from 2009 to 2015. The forecast values, together with the actual values, are presented in Tables 5. The prediction results show that Russia's renewable, nuclear, and total energy consumption will grow at compound annual growth rates (CAGRs) of 1.95%, 2.44%, and 0.88% respectively over the period of 2010-2015.

The results are compared with leading research in the field of energy forecasting, e.g., Azadeh et al. [2] for Iran, Pao [21, 7] and Lee & Shih [16] for Taiwan, and Kumar and Jain [14] for India. For one-period out-of-sample forecasting, Azadeh et al. used a simulated-based ANN univariate model to forecast monthly electricity consumption in Iran, and because of ANN's dynamic structure, the value of MAPE is lower than that of ANN. For multi-period out-of-sample forecasting, Pao [21] proposed a multivariate ECSTSP model to forecast electricity consumption in Taiwan with the value of MAPE at approximately 3.90%; Pao [7] proposed an ANN-based hybrid univariate model for energy consumption in Taiwan, and the value of MAPE is lower than 5%. Lee and Shih proposed a novel grey-based cost efficiency model (GCE) to improve short-term prediction of power generation cost for renewable energy technologies. Empirical results demonstrated that the GCE model has a highly accurate forecasting power. Additionally, Kumar and Jain applied three univariate models, namely Grey-Markov, Grey-Model with rolling mechanism, and singular spectrum analysis, in order to forecast the consumption of conventional energy (petroleum, coal, electricity, and natural gas) in India. As for the two out-of-sample forecasts (2006-2007), the MAPE values of Kumar and Jain's models ranged from 1.6% to 3.4%. In this paper, all of the MAPE values of the best NGBM^{op} for medium-term forecasting are lower than 3.20. Therefore, NGBM^{op} shows a highly accurate predictive model for green energy systems.

Table 3: Out-of-sample comparisons between GM and NGBM models from 2003 to 2009

	GM-4	GM-5	GM-6	NGBM-4	NGBM-5	NGBM-6
Forecasts of renewable energy consumption (2003-2008)				i=-0.04	i=0.04	i=0.07
RMSE	0.07	0.11	0.14	0.07	0.07	0.07
MAE	0.07	0.08	0.12	0.05	0.05	0.05
MAPE(%)	4.01	5.06	7.27	3.13	2.94	2.90
Forecasts of nuclear energy consumption (2004-2009)				i=0.17	i=0.14	i=0.19
RMSE	0.32	0.24	0.31	0.06	0.05	0.04
MAE	0.30	0.23	0.29	0.05	0.03	0.35
MAPE(%)	18.28	13.89	17.53	2.93	2.20	2.25
Forecasts of total energy consumption (2003-2008)				i=-0.03	i=-0.03	i=0
RMSE	1.09	0.91	0.48	0.56	0.52	0.48
MAE	1.01	0.83	0.41	0.46	0.45	0.41
MAPE(%)	3.42	2.79	1.39	1.58	1.52	1.39

Table 4: The parameters a and b in both GM and NGBM models for energy consumptions

Parameter	Renewable energy	Nuclear energy	Total energy	Renewable energy	Nuclear energy	Total energy
	GM-4	GM-5	GM-6	$NGBM^{0.07}$	$NGBM^{0.14}$	$NGBM^{0}$
a	0.005	-0.010	-0.016	0.008	0.006	-0.016
b	1.742	1.51e+03	25.447	1.541	1.218	25.447

Table 5: Forecasts of renewable, nuclear, and total energy consumption from 2009 to 2015

Year	Renewable		Nuclear		Total energy	
	Actual	NGBM	Actual	NGBM	Actual	NGBM
2003	1.603	1.603			28.512	28.512
2004	1.770	1.813	1.513	1.513	29.370	29.106
2005	1.743	1.730	1.543	1.554	29.520	29.361
2006	1.737	1.704	1.588	1.598	28.818	29.618
2007	1.761	1.700	1.671	1.641	29.969	29.878
2008	1.641	1.709	1.697	1.684	30.426	30.139
2009		1.727	1.705	1.726		30.403
2010		1.750		1.769		30.669
2011		1.779		1.813		30.938
2012		1.811		1.858		31.209
2013		1.847		1.903		31.482
2014		1.886		1.949		31.758
2015		1.928		1.997		32.036

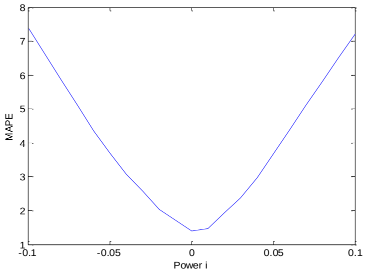


Fig. 4. MAPE values with different power *i* in NGBM^{i} (1, 1) for total energy consumption over the out-of-sample period of 2003-2008.

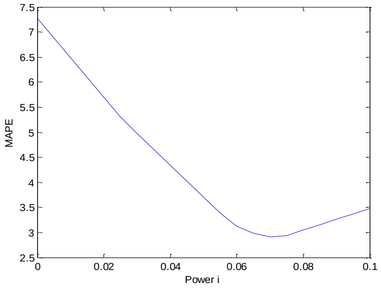


Fig. 5. MAPE values with different power *i* in NGBM^{i} (1, 1) for renewable energy consumption over the out-of-sample period of 2003-2008.

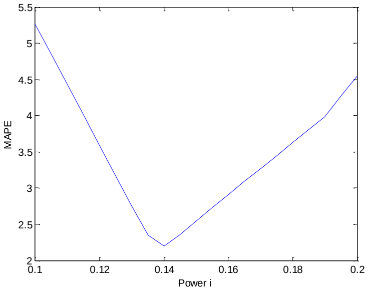


Fig. 6. MAPE values with different power *i* in NGBM^{i} (1, 1) for nuclear energy consumption over the out-of-sample period of 2003-2008.

4. Conclusions

Forecasts of renewable, nuclear, and total energy consumption are key requirements for a green energy system and understanding climate change in an emerging market such as Russia. This research uses recent four- to six-year historical data to construct univariate GM and NGBM^{op} models for forecasting these three indicators over the period of 2009-2015, while 1997-2003 is the in-sample period and 2004-2009 is the out-of-sample period. The multi-step forecasting ability of the best NGBM^{op} is compared with GM models over the out-of-sample period. The proposed numerical iterative method with the value of MAPE is an effective optimization algorithm for choosing optimal power of NGBM. NGBM$^{0.07}$-6 with a MAPE value of 2.90 for renewable and NGBM$^{0.14}$-5 with a MAPE value of 2.20 for nuclear are both better than GM models, whose value of MAPE is the lowest. For total energy consumption predictions, NGBM^{0}-6 and GM-6 are equally good. Performance evaluation results are clear and it is shown that NGBM^{op} can be used safely for future projection of these indicators in a green energy system. Future projections have also been carried out for these indicators using NGBM^{op} for the period between 2009 and 2015. The prediction results show that Russia’s renewable, nuclear, and total energy consumption will grow respectively at compound annual growth rates (CAGRs) of 1.95%, 2.44%, and 0.88% over the period of 2010-2015. The Russian government can apply these results for the dynamic adjustment of its green energy policy.

Because of the global economic uncertainty, high-tech progresses, and ever-changing domestic social structures, it is strictly recommended to revise the results every five years using NGBM^{op} to obtain more accurate outcomes. In the future, NGBM^{op} can be used to improve the accuracy of multi-step predictions of conventional or sectoral energy consumption in other fast-growing markets to effectively develop a clean energy economy.

References

[1] Perelet R, Pegov S, Yulkin M. Climate Change. Russia Country Paper 2007.

[2] Catalao JPS, Pousinho HMI, Mendes VMF. Short-term wind power forecastingin Portugal by neural networks and wavelet transform. Renewable Energy 2011;36:1245-1251.

[3] Kavoosi H et al. Forecast global carbon dioxide emission by use of genetic algorithm (GA). IJCSI 2012;9:418-427.

[4] Pao HT. Forecasting electricity market pricing using artificial neural networks. Energy Conversion and Management 2007;48:907-912.

[5] Blonbou R, Very short-term wind power forecasting with neural networks and adaptive Bayesia. Renewable Energy 2010;36:1118-1124.

[6] Badar UI Islam E. Comparison of conventional and modern load forecasting techniques based on artificial intelligence and expert systems. IJCSI 2011;8:504-513.

[7] Azadeh A, Khakestani M, Saberi M. A flexible fuzzy regression algorithm for forecasting oil consumption estimation. Energy Policy 2009;37:5567-79.

[8] Pao HT. Forecasting energy consumption in Taiwan using hybrid nonlinear models. Energy 2009;34:1438-1446.

[9] Li K, Su H. Forecasting building energy consumption with hybrid genetic algorithm-hierarchical adaptive network-based fuzzy inference system. Energy and Buildings 2010;42:2070-2076.

[10] Chaabene M, Ben Ammar M. Neuro-fuzzy dynamic model with Kalman filter to forecast irradiance and temperature for solar energy systems. Renewable Energy 2008;33:1435-1443.

[11] Monfared M, Rastegar H, Kojabadi HM. A new strategy for wind speed forecasting using artificial intelligent methods. Renewable Energy 2009;34:845-848.

[12] Cadenas E, Rivera W. Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA-ANN model. Renewable Energy 2010;35:2732-2738.

[13] Deng JL. Control problems of grey systems. System & Control Letter 1982;5:288-294.

[14] Akay D, Atak M. Grey prediction with rolling mechanism for electricity demand forecasting of Turkey. Energy 2007;32:1670–1675.

[15] Pao HT, Tsai CM. Modeling and forecasting the CO_2 emissions, energy consumption, and economic growth in Brazil. Energy 2011;36:2450-2458.

[16] Pao HT, Fu HC, Tseng CL. Forecasting of CO2 emissions, energy consumption and economic growth in China using an improved grey model. Energy 2012; 1:10.

[17] Yao AWL, Chi SC. Analysis and design of a taguchi-grey based electricity demand predictor for energy management systems. Energy Conversion and Management 2004;45:1205–1217.

[18] Kumar U, Jain VK. Time series models (Grey-Markov, GreyModel with rolling mechanism and singular spectrum analysis) to forecast energy consumption in India. Energy 2010;35:1709-1716.

[19] Zhou P, Ang BW, Poh KL. A trigonometric grey prediction approach to forecasting electricity demand. Energy 2006;31:2839–2847.

[20] Lee SC, Shih LH. Forecasting of electricity costs based on an enhanced gray-based learning model: A case study of renewable energy in Taiwan. Technological Forecasting and Social Change 2011;78(7):1242-1253.

[21] Chen, C.I., Chen, H.L., Chen, S.P. Forecasting of foreign exchange rates of Taiwan’s major trading partners by novel nonlinear grey Bernoulli model, NGBM(1,1). Communications in Nonlinear Science and Numerical Simulation 2008;13(6):1194–1204.

[22] Liu SF., Dang, Y.G., Fang, Z.G. The Theory of Grey System and its Applications. Science Press, Beijing (in Chinese); 2004.

[23] Chen, C.I. Application of the novel nonlinear grey Bernoulli model for forecasting unemployment rate. Chaos Solitons & Fractals 2008;37:278-287.

[24] Lewis, C.D. Industrial and Business Forecasting Method, London: Butterworth–Heinemann; 1982.

[25] Pao HT. Forecast of electricity consumption and economic growth in Taiwan by state space modeling. Energy 2009;34:1779-1791.

Hsiao-Tien Pao is a professor of the Department of Management Science at National Chiao Tung University, in Taiwan.

Hsin-Chia Fu received the B.S. degree from National Chiao-Tung University in Electrical and Communication engineering in 1972, and the M.S. and Ph.D. degrees from New Mexico State University, both in Electrical and Computer Engineering in 1975 and 1981, respectively. From 1981 to 1983 he was a Member of the Technical Staff at Bell Laboratories. From 1983 to 2012 he has been on the faculty of the Department of Computer science and Information engineering at National Chiao-Tung University, in Taiwan, ROC. Currently, he is a distinguished professor at the College of Engineering, Huaqiao University, QuanZhou, Fujian, China.

Hsiao-Cheng Yu is a professor of the Graduate Institute of Technology Management at National Chiao Tung University, in Taiwan. Currently, he is the deputy director of the National Communications Commission, in Taiwan.

Sustainability Criteria Model: A Field Study of ICT4D Project

Haslinda Sutan Ahmad Nawi[1,2], Nur Syufiza Ahmad Shukor[1,2], Suzana Basaruddin[1], Siti Fatimah Omar[1], Azizah Abdul Rahman[2], Rohaya Abu Hassan[1], and Mohammad Ashri Abu Hassan[1]

**[1] Faculty of Computer Science and Information Technology, Universiti Selangor
Bestari Jaya, Selangor, Malaysia**

**[2] Faculty of Computing, Universiti Teknologi Malaysia
Johor Bahru, Johor, Malaysia**

Abstract

Community ICT hubs provision in rural areas has been recognised as a promising tool to improve ICT literacy especially in developing regions. However, there are particular challenges in sustainable community ICT hubs provision that lead to low success rates and consequently derive economical, institutional, social, and cultural aspects consideration. The purpose of this study is to identify and understand the sustainability criteria of community ICT hubs implemented at 9 districts in one of the most progressive states in Malaysia. This study uses case study as a strategy to collect its qualitative data through document review, observation, and interview involving 92 respondents. There are 8 sustainability criteria discovered, grouped within 3 sustainability dimensions: social/cultural; economical; and institutional.

Keywords: *Community ICT Hubs, ICT Sustainability, ICT4D, Sustainability Dimensions.*

1. Introduction

Continually dropping prices of information and communications technology (ICT), and continually increasing ubiquitous ICT power, have created an environment where ICT literacy is vital for competing effectively in a globalised world. In achieving the Millennium Development Goals (MDGs), most governments around the world have proposed rural environments ICTs to ensure that the benefits of new technologies especially information and communication technologies are available to all. Several initiatives born within civil organisations, universities, and research institutes have developed specific low-cost computers, wireless communication infrastructures, and open-source software to be used in such environments. However, ICT initiatives in developing countries have shown low success rates in terms of sustainability [1]. *Pusat Siber Rakyat* (PSR) was introduced by one of the most progressive states in Malaysia, to support community development and to bridge the digital divide through the use of communication and computing technology. This PSR is another community ICT hub for citizen. However, after its 10 years in operations, only 19% were sustained, while the other 81% failed to sustain as they were, among other reasons, not fully utilised, abandoned or totally closed. This study is derived from [2] where its aim is to identify and understand the sustainability criteria of community ICT hubs in developing countries with reference to Malaysia.

1.1 ICT for Development

There is growing optimism that technology, particularly information and communication technologies (ICTs), can help achieve development goals and spur progress in developing countries. It is agreed by many researchers that there is a consensus that ICTs can play an important role in development, for examples by connecting people to more accurate and up-to-date information, equipping them with new skills, and connecting them to an international market. ICT for development (ICT4D) project is similar to conventional projects in that they are transitory undertakings that use resources, incur costs, expected to produce deliverables over a period of time, and typically have a high rate of failure. In ICT4D project, ICT-based solutions are developed to meet needs or to address a problem. These projects improve processes and methodologies in the scope of ICT. Moreover, these projects also introduce technological changes in organisations that are intended to be beneficial to the organisations and their target groups.

The evolution of ICT4D has three phases: ICT4D 0.0, 1.0, and 2.0 [3]. In the first phase, until about 1990, computers were used in government administration and by multinationals to foster economic growth. From the mid-1990s onwards, ICT4D 1.0 started as development actors such as the World Bank called for the adoption of ICTs as a tool for development – a call which was in response to the growth of the Internet and the adoption of the Millennium Development Goals (MDGs). Due to the need for a rapid response to the plight of poor, rural communities, a popular choice was the deployment of telecentres or community ICT hubs to deliver information, communication, and various services.

1.2 The Concept of Sustainability

Sustainability is a concept and strategy for integrating and balancing three bottom lines (TBL) namely, economic, environmental, and social dimensions [4] into a specific domain, while sustainable development is defined as a process of achieving sustainability. However, this definition is rather broad and difficult for organisations to understand and apply. As a result, much of the focuses on sustainable development tend towards an ecological perspective without explicit incorporation of the social aspects of sustainability [5]. The broad definition of sustainability has been reinterpreted in the domain of information systems to address challenges in the design and implementation of sustainable IT solutions [6-9]. According to [8], sustainable IT is a technology that is capable of being maintained over a long span of time independent of shifts in both hardware and software. Other than that, [10] defined a sustained programme or project as a set of durable activities and resources aimed at program-related objectives. This current research worked around the definition of sustainability by [8] and [10].

1.3 Sustainability Dimensions

Sustainability is not only about being better environmental stewards, but it should also include giving a comprehensive response to both the internal and external impacts of social, cultural, and resource trends [11]. As managers internalise strategic approaches and responses that encapsulate these four sustainability factors, only then will that the organisation, whether public or private, be able to adequately sustain its existence in the future. In other discussion, most researches consider sustainability to be closely linked to the ability of a project to be financially sustainable, in that a project must be capable of cost recovery in order to be continuously operative and dynamic in the services they provide [12]. However, sustainability encompasses more than just the financial or economic aspect of the project; it also considers other significant facets such as rootedness in local communities, cultural and political acceptance, and value to rural individuals [13]. Therefore, ICT projects have to take these aspects into account. Most ICT projects have been proven to be unsustainable in the long run because they have not been accompanied by, or failed to, generate the broader economic and social changes [14], which would consequently lead to unsustainable demand for ICT resources in rural development. Five main dimensions of sustainability in the ICT4D literature have been identified namely financial sustainability, social sustainability, institutional sustainability, technological sustainability, and environmental sustainability [15]. There is a need to integrate all these five dimensions as they are vital elements in the planning and operation of ICT projects [16]. Many studies refer to sustainability as a key to long-term development outcomes for ICT projects. Other than that, there are three dimensions of sustainability as follows: (1) economic sustainability (achieved when a given level of expenditure can be maintained over time); (2) social and cultural sustainability (achieved when social exclusion is minimised and social equity is maximised); (3) and institutional sustainability (achieved when prevailing structures and processes have the capacity to perform their functions over the long term) [17]. This current research adopts these three dimensions of sustainability. The dimensions are as follows:

Social and Cultural Sustainability: This dimension considers the social and cultural context in which a project operates, and the impacts of the ICT project to this context. As the ICT project takes into account the social and cultural aspects of the community, people in the community feel empowered by the project and hence they become active in seeking ways to keep the project running as it is in their own vital self-interest [16]. According to [17], as a consequence to considering social and cultural sustainability, social exclusion is therefore minimised, and social equity is continuously built on and not undermined.

Economic Sustainability: This dimension could be associated with the level of expenditure that can be sustained in the long term [17]. ICT projects in rural areas are initially funded by development organisations; however, in the long term, the ICT services provided will need to develop cost recovery mechanisms to generate enough income to keep the project sustainable. [18] indicated that the ability for ICT services to be financially self-sustainable is a key concern; hence, there is a need to promote a spirit of entrepreneurship to market ICT services rendered and to secure grant contributions.

Institutional Sustainability: Institutional sustainability is achieved when structures and processes of an organisation are able to perform their functions over the long term [17]. Aspects of institutional sustainability that need to be put in place include well-defined ICT laws, participatory policy-making processes, and effective public and private sector organisations that develop a framework in which the livelihoods of the community can be continuously improved.

2. Case Study: *Pusat Siber Rakyat* (PSR)

Malaysia consists of thirteen states and three federal territories, and is separated into two similar sized regions, Peninsular Malaysia and Malaysian Borneo. The Malaysian capital city is Kuala Lumpur, while Putrajaya is the seat of the federal government. The state, which is situated at the west of Malaysia, is one of the richest states in Malaysia. This state is also the most developed in Malaysia with good infrastructure and infostructure. This state also has the largest population in Malaysia. There are nine districts and all nine districts involved in the PSR project.

The cyber community centre or locally known as *Pusat Siber Rakyat* (PSR) was established in 1999 to serve the needs of ICT usage for the local community in the state. Formerly, it was known as *Pusat Komuniti IT* (PKIT) or information technology community centre with the vision to increase ICT literacy among the state's community. There are 39 PSRs in the nine districts under responsibilities of the State Public Library and State Federal Office. Each PSR is equipped with minimum six computers for users, one computer for administration, one scanner, one printer, and a dial-up internet connection for all computers. For the past ten years of their operation, the facilities have benefited the users, especially those in rural and sub-urban areas of the state. Nevertheless, these PSRs have many management and operational problems, which have lead to discontinued services of a few PSR facilities. The State Federal Office through their ICT Department has noticed the PSR operational problems and is looking forward to overcoming all the problems in sustaining the PSR. Table 1 shows the number of PSR by district.

Table 1: PSR by District and the Responsible Agencies

Location	Number of PSR	Responsibility of	
		State Federal Office	State Public Library
District 1	9	3	6
District 2	6	5	1
District 3	5	5	0
District 4	5	4	1
District 5	4	3	1
District 6	3	2	1
District 7	3	3	0
District 8	2	2	0
District 9	2	1	1

Source: State's Economic Planning Unit (UPEN), 2010

This current research was conducted at the 9 districts involving 26 PSRs. Data from semi-structured interview, observation, and document reviews were collected from the State Federal Office's officers and the users at each PSR.

3. Methodology

The overall research investigation examined the implementation of ICT hubs. The study aimed to draw general lesson where an exploration of the criteria of sustainability formed a part of the overall research investigation. A survey was conducted, where a case study qualitative research methodology was adopted to assess the ICT hubs project in a real-life environment [19]. Techniques employed in data collection were semi-structured interviews, non-participant observations, and document reviews. Data were collected from 26 ICT hubs. The main instrument used in this research was semi-structured interviews questionnaire.

Semi-structured interviews: Semi-structured interviews were done with two groups of respondents. The first group represented people who were responsible in managing and running the ICT hubs. The officers were contacted and appointments were made prior to the interview date. A total of 28 officers responsible for the operation of the hubs were interviewed. Each of the interview session lasted between 45 to 90 minutes and data gathered were transcribed. The respondents were asked on their routine operation activities of providing services to the customers, challenges and problems in administrating the hubs, their perception on how the hubs could be sustained, experience in operating the hub, and the history of the hub from the first day it was opened to date.

The second group interviewed was the community. 64 respondents from the community who had been using the facilities in the ICT hubs were interviewed separately from the first group of the respondent. The second interview was done with the intention of exploring the users' perspectives on the uses and challenges of ICT hubs, and the project's approach to promote sustainability.

Non-Participant Observations: Apart from the semi-structured interview, researchers were also exposed directly to the operations associated with the hub through observation of the administrators' and users' activities at the hubs for two weeks. 26 hubs were visited within the three weeks period and observation of how the hubs operate was documented. The observation focused on how the operators handle their daily operations of the hub where the users involved particularly on how the service was served to the users at the hub. The observation included the process of registering the users' attendance, recording the users' activities (i.e., printing, browsing the

Internet, scanning, etc.), and recording the maintenance of the software and hardware supplied to the hubs. There were hubs that were still in operation but did not have many users. The hubs were in a good condition but the hardware and software were obsolete. There were abandoned hubs with hardware broken and not fixed.

Document Reviews: To help the researchers to understand the processes observed better, documentations of the PSR operation activities were examined. These included the attendance log, the usage log, and the maintenance log. Data from the documents were reviewed and analysed, conforming what was previously observed at the hubs. Attendance for the first 2 years was high with an average of 30 users per day. After 2 years, there was a decline in the number of visitors as recorded in the log book. In some cases or hubs, the hubs were closed from operation.

4. Result Analysis and Discussion

4.1 ICT Hub Sustainability Criteria Model

From the result analysis, there are eight significant criteria that directly influence the sustainability of the ICT hubs, namely community development, ethics, social network, financial support, people, infrastructure, policy and strategy, and political influences. These criteria are then categorised into three dimensions as mentioned in the earlier section accordingly. Figure 1 depicts the eight criteria identified and categorisation made to the criteria into three sustainability dimensions.

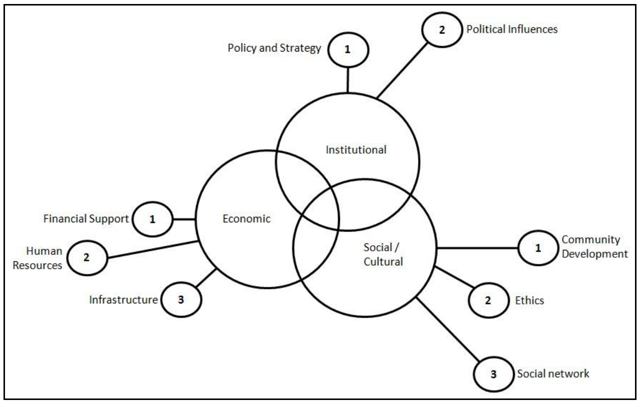


Fig. 1 Proposed ICT Hubs Sustainability Criteria Model

As discussed earlier, the social and cultural dimension takes into account the social and cultural aspects of the community and how people in the community feel empowered by the project and become active in seeking ways to keep the project running as it is in their own vital self-interest. There are three criteria that fit into this dimension: community development, ethics, and social network.

Community Development: The hub could be used as a medium for personal development opportunities for the community via learning environment. The role of the Internet connectivity is to create the learning environment and to enable linkages between local community and outside world. It also helps to encourage local ownership, to establish the community of practice (CoP), to facilitate local content development, and to create technically literate user. This hub also serves as a platform for communication between the government and citizen through the e-government application. This is especially significant for the low-income earners as it helps them to communicate directly to the Government without discrimination and bureaucracy barriers.

"...I know how to use computer and Internet. I learned it at school but my mother and her friends attended the computer classes held here, and had made them an expert on how to use the computer and Internet..."
(Respondent29- user)

"...at first the majority of users who came to use the facilities were teenagers and schoolchildren. But when we started to introduce our computer classes for the community, more adults and senior citizens came to the centre. Some of them came to get more information from the government agencies' official websites, some came to download tax forms, and there were also people who paid their bills. I, for an instance, could use the Internet facility to email my child who's now studying in Australia..."
(Respondent10- PSR Officer)

Ethics: ICT hub should also be administered with some guidelines or code of conduct to ensure its sustainability. It is important for the users to be able to differentiate between what is right and what is wrong. Thus, ethical practices among the CoP are mostly important especially as it involves users from all ages.

"...I hope the operator could monitor the PSR usage by the children and teenagers. We, parents are worried if the children here use the Internet unethically. If there is no monitoring, it would be difficult for me and my husband to allow our children to come here and use the facilities..."
(Respondent32- user)

"...In my opinion, any public ICT centre should have clear ethics guideline on the use of the facilities. If it is not in place, then this centre no longer provides value and positive impact towards community development. At this centre, my staff and I will always ensure the users especially the teenagers do not abuse the Internet facilities

provided. They are monitored and prohibited from surfing some websites that could give negative impacts in shaping their personality…”

(Respondent6- PSR Officer)

Social Network: Another contribution of the Internet is that it helps people from different location to communicate easily. People can connect to each other by many ways that they feel comfortable with. People want to stay connected, and with the help of e-mails and other applications, the social networking is easily done and maintained. Having the hub with reliable Internet facility is now a mandatory.

“…I come here to use FB (Facebook). I could chat with my friends who are far away from here and we share stories. We even share recipes …”

(Respondent17- user)

“…I had resigned from my work. So, with this PSR, I could come and use the Internet to communicate with my ex-officemate…”

(Respondent32- user)

“…from my observation, there are users who always come here just to open and use their Friendster and Facebook accounts…”

(Respondent10- PSR Officer)

The second sustainability dimension is economic dimension. Three criteria are categorised into this dimension namely financial support, human resources, and infrastructure. Like the first dimension, these three criteria are the most cited and mentioned criteria by the interviewees. Based on the observation and documentation reviewed, it was obvious that strong financial support, adequate human resources, and good infrastructure will ensure the hub’s life-long usage. Alongside with strong financial plan, spirit of entrepreneurship to market ICT services is also important to enable the hub to be self-sustained. On top of that, human resources factor, which is the centre of the operation of any ICT hubs, is also taken into account.

Financial Support: It is clear that the hubs need to be funded for a long period to provide the hub with enough time and effort to grow and nurture. A thorough and detailed execution plan of financing strategy is needed. A schedule and sufficient financial allocation or support for each PSR is required, especially for seed money. A financial strategy that empowers the PSR operators to manage their own financial affairs (including generating and distributing income) is also seen as important by the respondents. This is an investment for producing future champions – a life-long working ICT hub.

“…there are 2 critical and important factors to maintain and sustain this centre. First, a scheduled financial support, in terms of money and allocation, and the second factor is a fulltime qualified ICT staff …”

(Respondent6- PSR Officer)

“… if we don’t have yearly budget from the Government, it’s hard for us to maintain this PSR. I’ve heard that a centre has been shut down due to this reason …”

(Respondent10- PSR Officer)

“…the government did not have a proper yearly budget for this PSR. We don’t have enough allocation to maintain this centre. The state ICT centre’s officer just comes for the maintenance only twice a year. But, you know, when dealing with computers for public use, we need to service it more often…”

(Respondent12- PSR Officer)

“…Here in my centre, we need a budget allocation to maintain it. Our office is the Chief Office (Pejabat Penghulu). We don’t have budget for computers, and we need special allocation for this purpose…”

(Respondent22- PSR Officer)

“…Money is the most critical factor to maintain and sustain this centre... Why I say this? Every year, we need to service the computers, clean them from virus, sometimes format the hard disk and change the faulty devices especially the mouse... state’s officer just comes twice a year, but our computers always have problem. Now, at my PSR, there are 3 computers not functioning, we couldn’t turn on the computers. I don’t know why they are so, so we just put those computers aside; we don’t have enough money to send them for service …”

(Respondent26- PSR Officer)

Human Resources: Evidently, the issues on staff competency and qualifications, staff incentive, and dedicated and permanent staff appointment dominate the human criteria. The hubs are desperately in need of qualified permanent staffs who could provide and deliver better service needed by the community, for examples, in conducting computer classes and doing simple maintenance tasks on the computers when necessary.

“…there are 2 critical and important factors to maintain and sustain this centre. First, a schedule financial support in term of money and allocation, and the second factor are a fulltime qualified ICT staff …”

(Respondent6- PSR Officer)

“We need a fulltime qualified staff to operate this centre. With this fulltime staff, other than making sure the centre can be opened daily, the centre could also give more services to the community such as scheduling and

> *conducting a computer workshop, scan service, and other computer-related services."*
>
> *(Respondent26- PSR Officer)*

Infrastructure: To provide an ICT centre that offers an added value to its users, it requires updated technology, reliable Internet access, comfortable ICT services centre, and continuous and scheduled maintenance of ICT devices. Accessible and reliable, if not a state-of-the-art, infrastructure is very important and this will ensure user satisfaction and continuous ICT centre services.

> *"…other than that, a comfortable ICT service centre is another criteria to making sure its sustainability."*
>
> *(Respondent6- PSR Officer)*

> *"…during the transition of state government reign, the centre was broken in. A few computers that were in good condition were stolen, but there was no sign of forced entry; however, it is known that this centre is not really secured and needs to be improved..."*
>
> *(Respondent22- PSR Officer)*

> *"…one more, in my opinion, the Internet access here needs to be improved. I always have to wait for a few minutes just to open my email, it's too slow…"*
>
> *(Respondent32- user)*

> *"…other than improving the Internet access, the operator should carry out regular maintenance works. If there is any computer infected by viruses, it must be cleaned up immediately. Sometimes it was found that the (computer) mouse could not be used, it needs to be serviced…"*
>
> *(Respondent33- user)*

Finally, institutional sustainability does play a own role in sustaining the ICT4D project. Institutional sustainability ensures well-defined ICT laws and policy are in place by the authoritative party(s) and it also promotes for the public and private organisations to work hand-in-hand in developing, owning, and maintaining the hubs through their active participations. Based on the transcribed interviews, two criteria have been identified and categorised under institutional sustainability: policy and strategy, and political influences.

Policy and Strategy: In previous study, [2] mentioned about policy and strategy as influencing factors in sustaining community ICT hubs. Policy makers are responsible to derive a set of policy and strategy and then disseminate them to the related parties. When there is clear policy and strategy, these will help in identifying the roles and responsibilities all the related parties should play. This leads to a proper organisational structure that helps to identify the roles and responsibilities of each party involved. Following that, a continuous monitoring, control, and evaluation that are compliant with government rules and regulations will easily be executed. Thus, it is highly important for any ICT4D project to have project ownership so that the project champion would have full authority to delegate work and provide continuous support and commitment. The project champion is the responsible party(s) that focuses on the following: (1) strategising the existing public facilities for each ICT service centre at strategic location; (2) making sure systematic documentation on facilities and operations, and hand them over when there is change in management; (3) implementing affordable membership fees; and (4) providing value-added services to users with minimum charge, for examples, computer classes, and scanning and printing services.

> *"…to me, we should have a proper SOP and a clear structure of the people involved in managing and maintaining this centre…"*
>
> *(Respondent6- PSR Officer)*

> *"The Government should give full support and commitment especially in terms of monetary, allocating a qualified staff to operate all PSR, and in making it a successful and sustained project."*
>
> *(Respondent22- PSR Officer)*

> *"At our centre, we have do a systematic documentation on the centre's operation and maintenance. We have a schedule on the community development, for example the computer workshop. We do have a strategic location of our PSR and people do come to utilise our service. But, I strongly believe if we would like to sustain the success of this PSR, we need sufficiently allocated money and dedicated staff to do it."*
>
> *(Respondent6- PSR Officer)*

Political Influences: A well-defined project ownership regardless of political changes in management is relevant in any projects. The project champion that is free from any political influences will enable a clear workflow to be exercised whenever there is a need of handover. A clear defined roles and responsibilities of all parties involved in the handover will then make a smooth transition process that will be almost unnoticed by the users. This is highly applicable especially when there is a transition of the state government reign.

> *"…there was a handover problem between the old operator and new operator during the transition of state government reign that happened not long ago. It happened because there was a new appointment (operator) and both operators were representing two*

conflicting political parties and they have different views on how the PSR should be managed."

(Respondent22- PSR Officer)

"...during the transition of state government reign, the centre was broken in. A few computers that were in good condition were stolen, but there was no sign of forced entry; however, it is known that this centre is not really secured and needs to be improved There were a few allegations made that some fanatic supporters of the political turmoil had intentionally made the computers "disappeared" to blame certain parties.."

(Respondent22- PSR Officer)

4.2 Hub Lesson Learned

The presented case allows us to draw some important lessons and provide recommendations for policymakers in other similar ICT4D future projects. The ICT hubs at all 26 sites have great potential to be sustained if the project champion that is free from any political influences is identified. Before the project could even start, the project champion must be equipped with a comprehensive policy and strategy to ensure smooth administration. Continuous financial support to run the ICT hubs with adequate resources and good infrastructure for the community development has been identified to be crucial. Ethical use of the facilities and their applications are also important to ensure a life-long usage as it helps parents to be rest assured of the quality of services provided by the ICT hubs. Moreover, ICT hubs also help the community to reach out to the Government, or to their social acquaintances and friends through social network. In order for the whole model to work, the human factor is utterly important. Good people resources will ensure all the other mentioned criteria are synergised and exploited to their maximum potential in delivering sustainable ICT4D project.

5. Conclusions

For community ICT hubs provision to contribute to mainstream development, there is a need for evidence-based interpretation from the real project perspective. Therefore, this study concludes eight sustainability criteria of PSR project according to the three sustainability dimensions adopted from [17]. The three criteria classified under social and cultural dimension are community development, ethics, and social network. Next, another three criteria grouped into economic dimension are financial support, human resources, and infrastructure. Finally, the criteria policy and strategy, and political influences are categorised under institutional dimension, which are also significant to the sustainability of community ICT hubs. It is recommended to use these ICT hubs sustainability criteria model as a guideline for future development of community ICT hubs to ensure they will be fully utilised and give high impact to the society. Hence, by applying this model, it is expected that the failure to sustain the community ICT hubs will be decreased.

The sustainability criteria model proposed in this research open for further research to investigate the implementation of community ICT hubs in developing countries. Future research on the measurement of the most significant criteria towards community ICT hubs sustainability, among other examples, could also be conducted. Study to provide evidence whether the ICT hubs sustainability criteria model is an adequately valid and reliable instrument to measure the implementation of community ICT hubs is also relevant. Further investigation is still needed and it might focus on using confirmatory research approach to revise and improve the model.

Acknowledgments

The research team would like to express their sincere thank to Centre for Research & Innovation, Universiti Selangor (Project No.: G-SUK/ST-001) for the financial support of the research work.

References

[1] R. Heeks, "Information systems and developing countries: Failure, success, and local improvisations," *The Information Society*, vol. 18, pp. 101-112, 2002.

[2] S. Basaruddin, *et al.*, "Influencing Factors for Effective Community ICT Hubs," *World Applied Sciences Journal*, vol. 11, pp. 114-117, 2010.

[3] R. Heeks, "ICT4D 2.0: The Next Phase of Applying ICT for International Development.," *Computer*, vol. 41, pp. 26-33, 2008.

[4] Global Reporting Initiative. (2006, November 2010). RG Sustainability Reporting Guidelines Version 3.0. Available: www.globalreporting.org/ReportingFramework/G3Guidelines

[5] C. R. Carter and D. S. Rogers, "A framework of sustainable supply chain management: moving toward new theory," *International Journal of Physical Distribution & Logistics Management*, vol. 38, pp. 360-387, 2008.

[6] J. Reynolds and W. Stinson, "Sustainability analysis," presented at the Primary Healthcare Management Advancement Programme, Bangkok, 1993.

[7] M. Korpela, *et al.*, "Blueprint for an African Systems development methodology: an action research project in the health sector," in *Avgerou, C. (Ed.), Implementation and Evaluation of Information Systems in Developing Countries, International Federation for Information Processing*, Vienna 1998, pp. 173-286.

[8] G. Misund and J. Hoiberg. (2003, February, 2012). Sustainable information technology for global sustainability.Digital Earth. *Information Resources for*

Global Sustainability Symposium. Available: http://www.ia.hiof.no/~gunnarmi/omd/dig_earth_03/

[9] G. Z. Oyomno, "Sustainability of governmental use of microcomputer-based information technology in Kenya," in *Mayuri Odedra-Straub (Ed.), Global IT & socio-economic development*, Marietta, GA, 1996.

[10] M. Scheirer, "Are the levels of institutionalization scales ready for prime time? A commentary on development of level of institutionalization scales for health promotion programs," *Health Education Quarterly*, vol. 20, pp. 179-183, 1993.

[11] A. Werbach, *Strategy for Sustainability: A Business Manifesto*: Harvard Business Press, 2009.

[12] C. Pade, *et al.*, "An Exploration of the Categories Associated with ICT Project Sustainability in Rural Areas of Developing Countries: A Case Study of the Dwesa Project," in *Proceedings of SAICSIT 2006*, 2006, pp. 100-106.

[13] K. Keniston. (2005, January 2011). Notes on Sustainability. Available: http://web.mit.edu/~kken/Public/PAPERS/on_sustainability.html

[14] K. S. McNamara, "Information and Communication technologies, Poverty and Development: Learning from Experience," The World Bank, Washington DC2003.

[15] M. Ali and S. Bailur, "The challenge of “sustainability in ICT4D – Is bricolage the answer?," presented at the 9th International Conference on Social Implications of Computers in Developing Countries, Sao Paulo, Brazil, 2007.

[16] K. Stoll. (2003, December 2011). Telecentres Sustainability: What Does it Mean? Available: http://topics.developmentgateway.org/ict/sdm/previewDocument.do~activeDocumentId==442773

[17] S. Batchelor and P. Norrish. (2003, September 2010). Sustainable Information Communication Technologies (ICT). Available: http://www.sustainableicts.org/Sustainable.htm

[18] The World Bank, "ICT for Development Contributing to the Millennium Development Goals: Lessons learnt from Seventeen InfoDev Projects.," The World Bank, Washington DC2003.

[19] R. K. Yin, *Case Study Research: Design and Methods* vol. 5. Thousand Oaks, California: SAGE, 2009.

Haslinda Sutan Ahmad Nawi, a doctoral candidate from Faculty of Computing, UTM. She is an academician from UNISEL and her research focuses on Information Systems and corresponding IT projects sustainability. Currently she holds a Master of Information Technology degree awarded by UiTM, Malaysia and possesses over eight years of relevant industrial experience in IT projects management, implementation and operation. Haslinda is also a member of PMI and AIS.

Nur Syufiza Ahmad Shukor, a senior lecturer in the Department of Information Systems, UNISEL and currently she is pursuing her Ph.D at UTM, Malaysia. She has implemented many ICT projects specifically software development projects during her previous work in the industry. Her current research interests focus on the area of Information Systems and Knowledge Management. She is also a member of AIS and iKMS.

Suzana Basaruddin, a graduate in the field of IT, and currently pursuing her Ph.D at UiTM. She is a member of AICSIT and a lecturer at UNISEL. Her research interests include ICT and Knowledge Management.

Siti Fatimah Omar CTFL, graduated in Bachelor of Science in Information Technology in 2007 and Master in Computer Science (Software Engineering) in 2010 from Universiti Selangor (UNISEL). Currently, she is a lecturer at UNISEL and her research interests include software development and information security.

Azizah Abdul Rahman, **Ph.D**, a member of the IACSIT and IEEE. She is currently an Associate Professor at UTM. Her research interests focus on the area of Information Systems and Knowledge Management.

Rohaya Abu Hassan, a Lecturer in Computer Science Department at Universiti Selangor (UNISEL). She graduated from Universiti Teknologi MARA with Master in Information Technology. Her areas of interest focus on systems analysis and design, and object-oriented programming. She is a member of AIS.

Mohammad Ashri Abu Hassan, graduated in Master of IT by Research, UiTM. He is a member of Computer Science Department in FCSIT, UNISEL. His areas of interest in research are ICT and imaging. He is also interested in software development and algorithm analysis.

Fuzzy Logic System for Opportunistic Spectrum Access using Cognitive Radio

Kaniezhil. R[1], Daniel Nesa Kumar. C[2] and Prakash. A[3]

[1] Research Scholar, Department of Computer Science, Periyar University, Salem, TamilNadu-636011, India

[2] Asst Professor, Dept Of Computer Applications, Hindusthan college of arts and science, Coimbatore, India

**[3] Asst Professor,
Dept Of Computer Applications,
Hindusthan college of arts and science, Coimbatore.**

Abstract

Opportunistic spectrum access approach is enabled by cognitive radios which are able to sense the unused spectrum and adapt their operating characteristics to the real-time environment. However, a naive spectrum access for secondary users can make spectrum utilization inefficient and increase interference to adjacent users. In this paper, we propose a novel approach using Fuzzy Logic System (FLS) to control the spectrum access and spectrum for secondary user is selected by possibility of the spectrum access instead of considering the antecedents.

Keywords: *Antecedent, FLS, Interference, Spectrum access, spectrum utilization.*

1. Introduction

In recent studies, the spectrum allocated by the traditional approach shows that the spectrum allocated to the primary user is under-utilized and the demand for accessing the limited spectrum is growing increasingly. Spectrum is no longer sufficiently available, because it has been assigned to primary users that own the privileges to their assigned spectrum.

The concept of opportunistic spectrum access is used in order to efficiently utilize the spectrum which is under-utilized. The opportunistic spectrum access improves the spectrum utilization by cognitive radio adopting the secondary user to use the unused spectrum of the primary user. This opportunistic spectrum access, avoids the spectrum scarcity.

The idea of cognitive radio was first presented officially in an article by Joseph Mitola III and Gerald Q. Maguire, Jr in 1999. Regulatory bodies in various countries found that most of the radio frequency spectrum was inefficiently utilized. For example, cellular network bands are overloaded in most parts of the world, but amateur radio and paging frequencies are not. This can be eradicated using the dynamic spectrum access.

The key enabling technology of dynamic spectrum access techniques is cognitive radio (CR) technology, which provides the capability to share the wireless channel with licensed users in an opportunistic manner. From this definition, two main characteristics of the cognitive radio can be defined as follows:

- Cognitive capability: It refers to the ability of the radio technology to capture or sense the information from its radio environment. Through this capability, the portions of the spectrum that are unused at a specific time or location can be identified. Consequently, the best spectrum and appropriate operating parameters can be selected.
- Reconfigurability: The cognitive capability provides spectrum awareness whereas reconfigurability enables the radio to be dynamically programmed according to the radio environment.

More specifically, the cognitive radio technology will enable the users to (1) determine which portions of the spectrum is available and detect the presence of licensed users when a user operates in a licensed band (spectrum sensing), (2) select the best available channel (spectrum management), (3) coordinate access to this channel with other users (spectrum sharing), and (4) vacate the channel when a licensed user is detected (spectrum mobility).

The paper is organized as follows; Section 2 and 3 provide the related work and fuzzy logic system for its implementation. In section 4 and 5, opportunistic spectrum access by Fuzzy logic system to improve the spectrum efficiency and performance analysis. Finally, conclusions are presented in Section 6.

2. Related Work

In the research literature on the opportunistic spectrum access, different methods are applied using fuzzy logic to access spectrum efficiently. In [7], a novel approach using Fuzzy logic system provides the possibility of accessing spectrum band for secondary users and the user with the greatest possibility has to be assigned the available spectrum band. For enhancing the performance of Cognitive radio fuzzy logic based scheme is developed by Anilesh Dey et al., [8], where efficient spectrum utilization depends upon the link loss, hold time and interference temperature. With fuzzy controller, Cognitive radio opportunistically adjust its transmit power in response to the changes of the interference level to the primary user is discussed in [9].

A Fuzzy logic based algorithm [10] is used to reduce the spectrum handoff and improves the probability of PU's occupancy at a certain channel. In [11], Fuzzy rules are used to optimize the bandwidth allocation based on three antecedents as: arrival rate of both licensed and unlicensed users and the availability of unoccupied number of channels within the system. A decentralized method has been developed using fuzzy based techniques for both channel estimation and channel selection in [12].

The efficient decision making in the cognitive radio by fuzzy logic is also discussed by Matinmikko et al [13], which presented an overview of application of fuzzy logic to telecommunications. Our work is different from the above said works available on Cognitive radio based on fuzzy logic system.

This paper presents a novel approach using Fuzzy logic system to utilize the available spectrum by the secondary users without interference with the primary user. The secondary users access the spectrum based on the highest possibility of the secondary users.

3. Fuzzy Logic

3.1 Fuzzy sets

Fuzzy sets theory is an excellent mathematical tool to handle the uncertainty arising due to vagueness. Fuzzy sets may be viewed as an extension and generalization of the basic concepts of crisp sets. An important property of fuzzy set is that it allows partial membership. Fuzzy set is the description of fuzzy events and conceptions. The fuzzy events means some events that there are no strict boundary and cannot be characterize easily by exact measurement. Fuzzy logic theory can model the vagueness of the real world.

A Fuzzy Logic System (FLS) is unique in that it is able to simultaneously handle numerical data and linguistic knowledge. It is a nonlinear mapping of an input data (feature) vector into a scalar output, i.e., it maps numbers into numbers. FL provides a simple way to arrive at a definite conclusion based upon vague, ambiguous, imprecise, noisy, or missing input information. FL's approach to control problems mimics how a person would make decisions, only much faster. Figure 1 shows the structure of a fuzzy logic system.

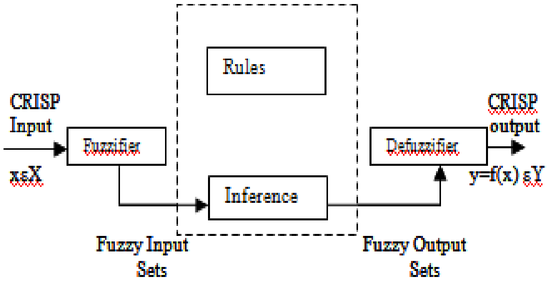


Fig. 1. The Structure of the Fuzzy Logic System

Since there is a need to "fuzzify" the fuzzy results we generate through a fuzzy system analysis i.e., we may eventually find a need to convert the fuzzy results to crisp results. Here, we may want to transform a fuzzy partition or pattern into a crisp partition or pattern; in control we may want to give a single-valued input instead of a fuzzy input command. The "dufuzzification" has the result of reducing a fuzzy set to a crisp single-valued quantity, or to a crisp set.

Generally, Fuzzy Logic and Fuzzy decision making is divided into three consecutive phases namely Fuzzification, Fuzzy reasoning and Dufuzzification [17].

1. Fuzzification: The input variables are fuzzified using predefined membership functions (MF). Unlike in binary logic where only 0 and 1 are accepted, also numbers between 0 and 1 are used in fuzzy logic. This is accomplished with MF to which the input variables are compared. The output of fuzzification is a set of fuzzy numbers.

2. Fuzzy reasoning: Fuzzy numbers are fed into the predefined rule base that presents the relations of the input and output variables with IF – THEN Clauses. The output of the fuzzy reasoning is a fuzzy variable that is composed of the THEN clauses.

3. Dufuzzification: The output of the fuzzy reasoning is changed into a non-fuzzy number that represents the actual output of the system.

3.2 Membership Functions

Consider a fuzzy logic system with a rule base of M rules, and let the lth rule be denoted by R_l. Let each rule have p antecedents and one consequent (as is well known, a rule with q consequents can be decomposed into rules, each having the same antecedents and one different consequent), i.e., it is of the general form

$$R_l : \text{IF } u_1 \text{ is } F_l^1 \text{ and } u_2 \text{ is } F_l^2 \text{ and } \ldots \text{ and } u_p \text{ is } F_l^p ,$$

$$\text{THEN } v \text{ is } G^l .$$

where $u_k, K = 1,\ldots.p$ and v are the input and output linguistic variables, respectively.

While applying a singleton fuzzification, when an input $X^{'} = \{x_1^{'}, x_2^{'}, x_3^{'}, \ldots x_p^{'}\}$ is applied, the degree of firing corresponding to the lth rule is given by

$$x^{*} = \frac{\sum_{i=1}^{n} x_i . \mu(x_i)}{\sum_{i=1}^{n} \mu(x_i)} \tag{1}$$

Where * denotes a T-norm, n represents the number of elements, x_i 's are the elements and $\mu(x_i)$ is its membership function.

There are many kinds of dufuzzification methods, but we have chosen the centre of sets method for illustrative purpose. It computes a crisp output for the FLS by first computing the centroid, C_{G^l} of every consequent set G^l and, then computing weighted average of these centroids. The weight corresponding to the lth rule consequent centroid is the degree of firing associated with the lth rule, $T_{i=1}^{p} \mu_{F_1^l}(x_1^{'})$, so that

$$y_{cos}(x^{'}) = \frac{\sum_{l=1}^{M} C_{G^l} T_{i=1}^{p} \mu_{F_1^l}(x_1^{'})}{\sum_{l=1}^{M} T_{i=1}^{p} \mu_{F_1^l}(x_1^{'})} \tag{2}$$

where M is the number of rules in the FLS.

We use trapezoidal membership functions (MFs) to represent near, low, far, high, very low and very high, and triangle MFs to represent moderate, low, medium and high. MFs are shown in Fig. 2, 3, 4. Since we have 3 antecedents and fuzzy subsets, we need setup $3^3 = 27$ rules for this FLS.

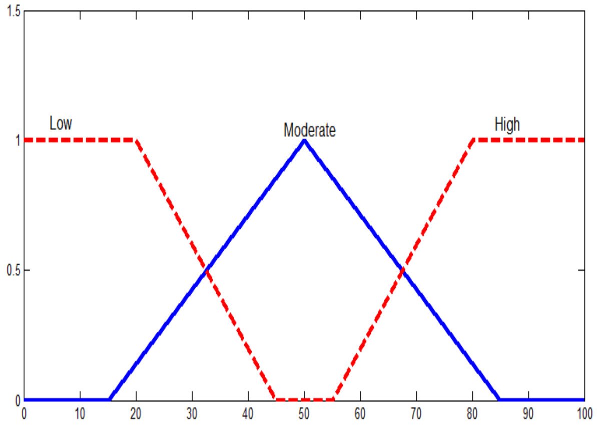


Fig. 2. Membership Function (MF) used to represent the antecedent1

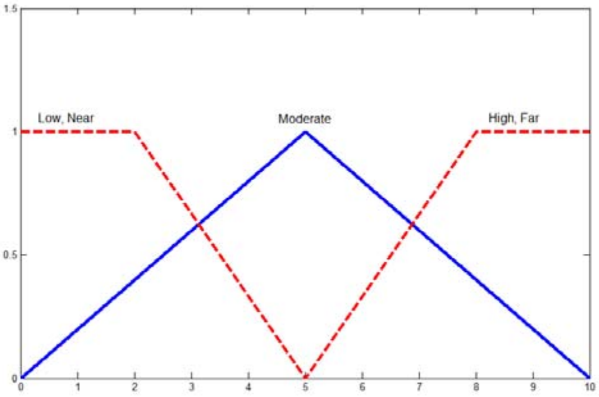


Fig. 3. Membership Function (MF) used to represent the antecedent2

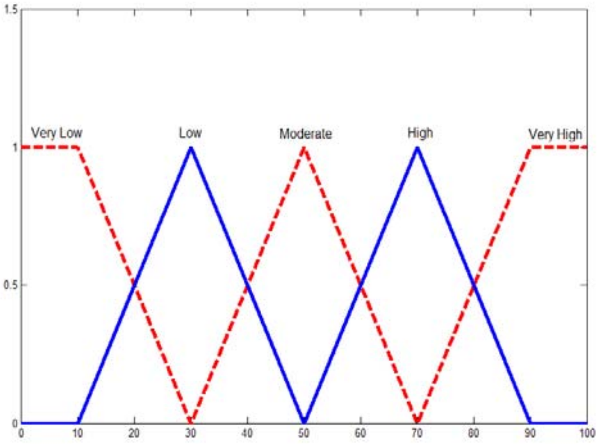


Fig. 4. Membership Function (MF) used to represent the antecedent3

Linguistic variables are used to represent the spectrum utilization efficiency; distance and degree of mobility are divided into three levels: low, moderate, and high. We use 3 levels to represent the distance, i.e., near, moderate.

The consequence, i.e., the possibility that the secondary user is chosen to access the spectrum is divided into five levels which are very low, low, medium, high and very high.

4. Fuzzy Logic for Opportunistic Spectrum Access

We design the fuzzy logic for opportunistic spectrum access using cognitive radio. In this paper, we are selecting the best suitable secondary users to access the available users without any interference with the primary users. This is collected based on the following three antecedents i.e., descriptors. They are

Ant 1 : Spectrum Utilization Efficiency
Ant 2 : Degree of Mobility
Ant 3 : Distance of Secondary user to the PU.

Fuzzy logic is used because it is a multi-valued logic and many input parameters can be considered to take the decision. Generally, the secondary user with the furthest distance to the primary user or the secondary user with maximum spectrum utilization efficiency can be chosen to access spectrum under the constraint that no interference is created for the primary user.

In our approach, we combine the three antecedents to allocate spectrum opportunistically inorder to find out the optimal solutions using the fuzzy logic system.

Mobility of the secondary user plays a vital role in the proposed work. Wireless systems also differ in the amount of mobility that they have to allow for the users. Spectrum Mobility is defined as the process when a cognitive radio user exchanges its frequency of operation. The movement of the secondary user leads to a shift of the received frequency, called the Doppler shift.

We apply different available spectrum inorder to find out the spectrum efficiency which is the main purpose of the opportunistic spectrum access strategy. Hence, we calculate the spectrum efficiency as the ratio of average busy spectrum over total available spectrum owned by secondary users.

Since we chose a single consequent for each rule to form a rule base, we averaged the centroids of all the responses for each rule and used this average in place of the rule consequent centroid. Doing this leads to rules that have the following form:

$R^{'}$: If Degree of mobility (x_1) is F_l^1 ; and its distance between primary user and the secondary users (x_2) is F_l^2 ; and the spectrum utilization efficiency of the secondary user (x_3) is F_l^3 , then the Possibility (y) choosing the available spectrum is c_{avg}^l , where l = 1,2,..27 and c_{avg}^l is defined as follows:

$$c_{avg}^{l} = \frac{\sum_{i=1}^{5} w_i^l c^i}{\sum_{i=1}^{5} w_i^l} \tag{3}$$

which w_i^l is the number of choosing linguistic label i for the consequence of rule l and c^i is the centroid of the ith consequence set (i: 1; 2; ...; 5; l: 1; 2; ...; 27). Table 2 provides c^i for each rule. For every input (x_1, x_2, x_3) , the output y(x_1, x_2, x_3) of the designed FLS is computed as

$$y(x_1, x_2, x_3) = \frac{\sum_{i=1}^{27} \mu_{F_1^l(x_1)} \mu_{F_2^l(x_2)} \mu_{F_3^l(x_3)} c_{avg}^l}{\sum_{i=1}^{27} \mu_{F_1^l(x_1)} \mu_{F_2^l(x_2)} \mu_{F_3^l(x_3)}} \qquad (4)$$

which gives the possibility that a secondary user is selected to access the available spectrum.

In Figure 5, shows that the Spectrum user (SU4) having the highest possibility of accessing the spectrum than the other users. Even though the other secondary users have the furthest distance from primary user to the secondary users, highest spectrum utilization and highest mobility but we prefer SU4 to access the spectrum because it has the highest possibility i.e., 58.62.

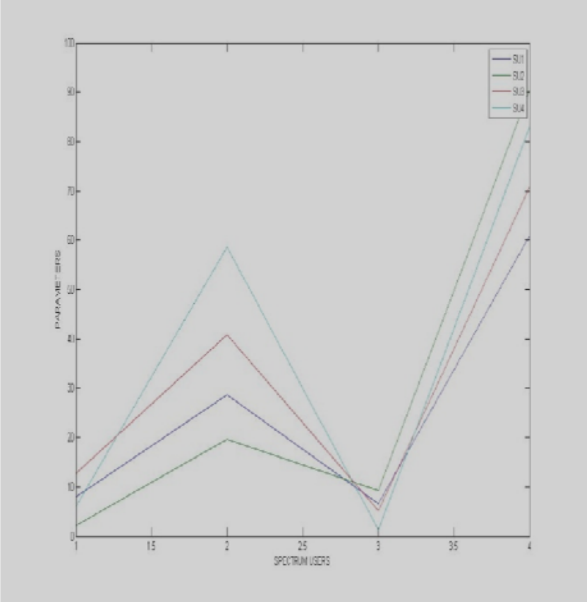


Fig. 5. Possibility of choosing spectrum for opportunistic spectrum access

Thus, the secondary user will select the spectrum for accessing based on the highest possibility rather than the highest spectrum utilization and the furthest distance from the primary user.

4. Performance Analysis

In this section, we present simulation results on the performance of our proposed work based on Fuzzy logic System. In the proposed work, we are choosing the available channel with the high possibility and high spectrum utilization efficiency.

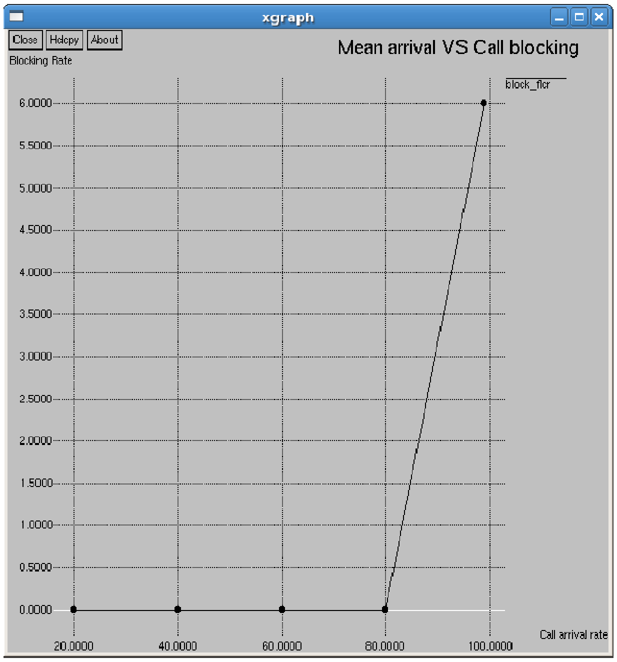


Fig. 6. Mean Arrival VS Call blocking rate

Figure 6 shows the call blocking of the service provider using the Fuzzy logic system. As the call arrival rate increases the blocking rate gets decreased. Traffic rate increases along with the call blocking rate.

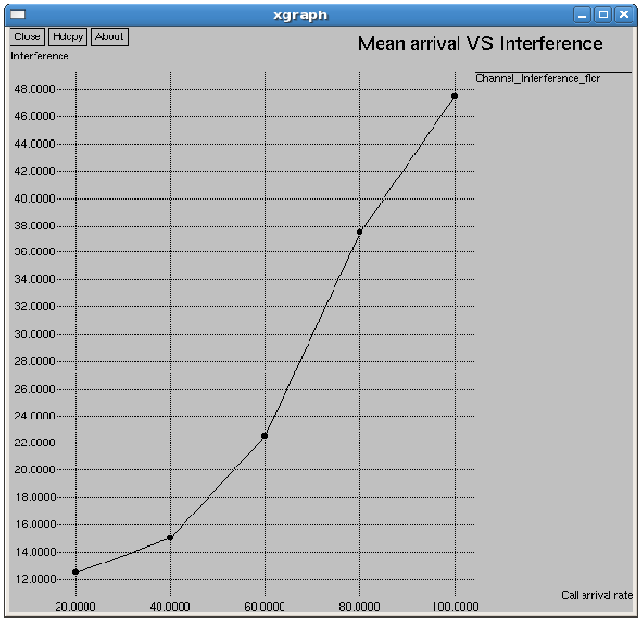


Fig. 7. Mean Arrival VS Interference using FLS

When interference increases spectrum utilization will decrease. The Figure 7 shows that there is a decrease in the interference which provides an increase spectrum utilization using FLS. This leads to efficient spectrum utilization.

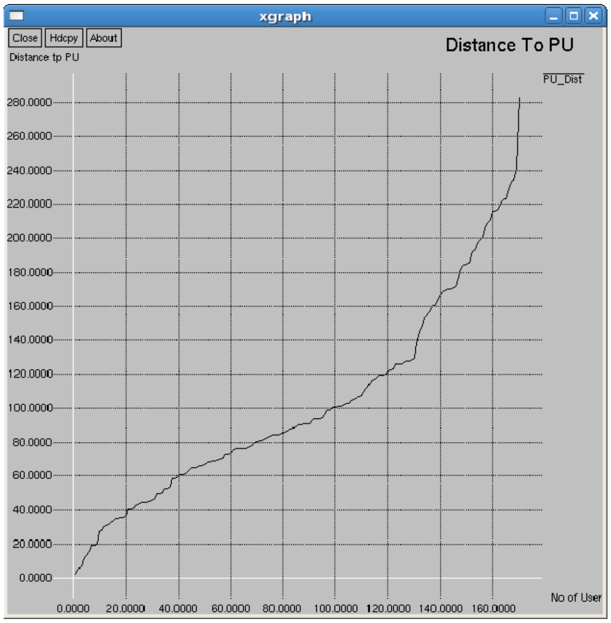


Fig. 8. Distance of Secondary Users to the Primary users using FLS

As illustrated in the Figure 8, the distance from primary user to the secondary users is shown. The distance between primary user and the secondary users helps us for calculating the possibility for accessing the spectrum opportunistically.

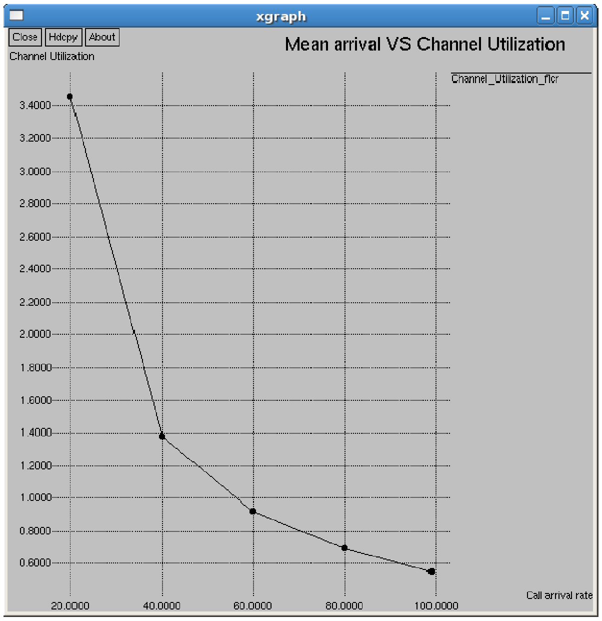


Fig. 9. Mean Arrival VS Channel Utilization using FLS

The Spectrum Efficiency (Channel Utilization) is defined as the ratio of average busy channels over total channels owned by service providers. The Figure 9 shows that there is an increase in channel utilization with decrease in the call arrival rate.

4. Conclusions

The proposed approach using a Fuzzy Logic System detects the effective spectrum access for secondary users via cognitive radio. The secondary users are selected on the basis of spectrum utilization, degree of mobility and distance from secondary users to the primary user. Our designed FLS is used to control the spectrum assignment and access procedures in order to prevent multiple users from colliding in overlapping spectrum portions. Therefore, the secondary user with the highest possibility is guaranteed to access the spectrum.

Acknowledgments

The First Author extends her gratitude to UGC as this work was supported in part by Basic Scientist Research (BSR) Non SAP Scheme, under grant reference no.11-142/2008(BSR) UGC XI Plan.

References

[1] H.J. Zimmermann, "Fuzzy Set Theory and its Applications," ISBN 81-8128-519-0, Fourth Edition, Springer International Edition, 2006.

[2] Timothy J. Ross, "Fuzzy Logic with Engineering Applications," ISBN 9812-53-180-7, Second Edition, Wiley Student Edition, 2005.

[3] James J. Buckley, Esfandiar Eslami, "An Introduction to Fuzzy Logic and Fuzzy Sets," ISBN 3-7908-1447-4, Physica Verlag, 2002.

[4] Andreas F. Molisch, "Wireless Communications," ISBN: 978-0-470-74187-0, Second Edition, John Wiley & Sons Ltd , 2011.

[5] R. Kaniezhil, Dr. C. Chandrasekar, S. Nithya Rekha, "Dynamic Spectrum Sharing for Heterogeneous Wireless Network via Cognitive Radio," Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, March 21-23, 2012, pp. 156-162.

[6] R. Kaniezhil, Dr. C. Chandrasekar, "Multiple Service providers sharing Spectrum using Cognitive Radio in Wireless Communication Networks", International Journal of Scientific & Engineering Research (IJSER), Volume 3, Issue 3, Feb 2012, pp.1-6.

[7] R. Kaniezhil, Dr. C. Chandrasekar, "Spectrum Sharing in a Long Term Spectrum Strategy via Cognitive Radio for Heterogeneous Wireless Networks", International Journal on Computer Science and Engineering (IJCSE), Volume 4, No.6, June 2012, pp 982-995.

[8] R. Kaniezhil, Dr. C. Chandrasekar, S. Nithya Rekha, "Performance Evaluation of QoS Parameters in Spectrum Sharing using SBAC Algorithm", IEEE International Conference on Advances in engineering, Science and Management (IEEE-ICAESM 2012), Nagapattinam, March 30-31, 2012, pp 755-760.

[9] R. Kaniezhil, Dr. C. Chandrasekar, "An Efficient Spectrum Utilization via Cognitive Radio using Fuzzy Logic System

for Heterogeneous Wireless Networks", INCOSET 2012, Dec 13-14, 2012.

[10] R. Kaniezhil, Dr. C. Chandrasekar, "Comparing Spectrum Utilization using Fuzzy Logic System for Heterogeneous Wireless Networks via Cognitive Radio", International Journal of Scientific & Engineering Research, Volume 3, Issue 7,July 2012, pp.1 - 10.

[11] Jerry M. Mendel, "Fuzzy Logic Systems for engineering: A Tutorial," IEEE Proc., March 1995, vol. 83, no. 2, pp. 345-377.

[12] Ila Sharma, G. Singh, "A Novel Approach for Spectrum Access Using Fuzzy Logic in Cognitive Radio," I.J. Information Technology and Computer Science, 8, 2012, pp. 1-9.

[13] Anilesh Dey, Susovan Biswas, Saradindu Panda, and Santanu Mondal, "A New Fuzzy Rule Based Spectrum Utilization and Spectrum Management for Cognitive Radio," National Conference on Electronics, Communication and Signal Processing, NCECS-2011, 19th September 2011, pp. 41-44.

[14] Hong-Sam T. Le and Qilian Liang, "An Efficient Power Control Scheme for Cognitive Radios," IEEE conference on Wireless Communications and Networking, 2007, pp. 2559-2563.

[15] Tang Wanbin, Peng Dong, "SPECTRUM HANDOFF IN COGNITIVE RADIO WITH FUZZY LOGIC CONTROL," JOURNAL OF ELECTRONICS (CHINA), Vol.27 No.5 , September 2010, pp. 708-714.

[16] Prabhjot Kaur, Moin Uddin, Arun Khosla," Adaptive Bandwidth Allocation Scheme for Cognitive Radios," International Journal of Advancements in Computing Technology, Volume 2, Number 2, June, 2010, pp. 35-41.

[17] Ala Al-Fuqaha, Bilal Khan, Ammar Rayes, Mohsen Guizani, Osama Awwad, Ghassen Ben Brahim, "Opportunistic Channel Selection Strategy for Better QoS in Cooperative Networks with Cognitive Radio Capabilities," IEEE Journal on Selected Areas in Communications, vol. 26, no. 1, January 2008, pp. 156 – 167.

[18] Marja Matinmikko, Tapio Rauma, Miia Mustonen,Ilkka Harjula, HeliS Arvanko and Aarne Mammela,―Application of fuzzy logic to cognitive radiosystems," IEICE Trans. Communication, vol. E92B, no.12, December 2009, pp. 3572-3580.

R. Kaniezhil is a member of the IEEE and she received her B.Sc Degree from University of Madras in 1998. She received her MCA and M.Phil degrees from Periyar University and Annamalai University in 2001 and 2007, respectively. Her research interests include mobile computing, Cognitive Radio and wireless communication.

C. Daniel Nesa Kumar completed his MCA in Bishop Heber College (Autonomous), Bharathidasan University, and Trichy. He is currently working as an Assistant Professor in Hindusthan College of Arts & Science. His Research area is Networking.

A. Prakash completed his M.Sc (CT), MCA, M.Phil., Ph.D. Currently he is working as an Assistant professor in Department of Computer Science & Applications, Hindustan College of arts and Science, Bharathiar University, Coimbatore. His Area of Interest is Data Mining.

Parallel K-Means Algorithm on Agricultural Databases

V.Ramesh[1], K.Ramar[2], S.Babu[3]

**[1,3]Assistant Professor,
Department of CSA, SCSVMV University, Kanchipuram, India**

[2]Principal, Einstein College of Engineering, Tirunelveli, India

Abstract

A cluster is a collection of data objects that are similar to each other and dissimilar to the data objects in other clusters. K-means algorithm has been used in many clustering work because of the ease of the algorithm. But time complexity of algorithm remains expensive when it applied on large datasets. To improve the time complexity, we implemented parallel k-means algorithm for cluster large dataset. For our study we take agricultural datasets because of limited researches are done in agricultural field.

***Keywords**: Clustering,k-means,parallel k-means, agriculture*

1. Introduction

Clustering is grouping input data sets into subsets called ‘clusters’ within which the elements are somewhat similar. In general, clustering is an unsupervised learning task as very little or no prior knowledge is given except input data sets. The tasks have been used in many fields and therefore various clustering algorithms have been developed. Clustering task is however, computationally expensive as many of the algorithms require iterative or recursive procedures and most of real-life data is high dimensional. Therefore, the parallelization of clustering algorithms is inevitable and various parallel clustering algorithms have been implemented and applied to many applications.

Parallel computing is simultaneous use of multiple compute resources to solve a computational problem. In parallel computing a problem is broken into discrete parts that can be solved concurrently and each part is further broken down to a series of instructions. The instructions from each part execute simultaneously on different CPUs.

There are different ways to classify parallel computers. One of the more widely used classifications in use since 1972 is called Flynn’s taxonomy. Flynn’s taxonomy distinguishes multi-processor computer architectures according to how they can be classified along the two independent dimensions of Instruction and Data. Each of these dimension have only one of two possible states, single or multiple. According to Flynn, the matrix given below defines the four possible classifications.

SISD (Single Instruction, Single Data)	**SIMD** (Single Instruction Multiple Data)
MISD (Multiple Instruction, Single Data)	**MIMD** (Multiple instruction, Multiple Data)

A Single Instruction, Single Data (SISD) machine is a traditional sequential computer that provides no parallelism in hardware. Instructions are executed in a serial fashion. One only data stream is processed by the CPU during a given clock cycle. A Multiple Instruction, Single Data (MISD) machine is capable of processing a single data stream using multiple instruction streams simultaneously. A Single Instruction, Multiple Data (SIMD) machine is one in which a single instruction stream has the ability to process multiple data streams simultaneously. A Multiple Instruction, Multiple Data(MIMD) machine is capable of is executing multiple instruction streams, while working on a separate and independent data stream. Given that modern computing machines are either the SIMD or MIMD machines, software developers have the ability to exploit data-level and task level parallelism in software.

Talia[1] identified three main strategies in the parallelism used in data mining algorithms as the following: (1) Independent parallelism where each processor accesses to the whole data to operate but do not communicate each other. (2) Task parallelism where each processor operate different algorithms on the partitioned or on the whole data set. (3) SPMD (Single Program Multiple Data) parallelism where multiple processors execute the same algorithm on different subsets and exchange the partial results to co-operate each other. Most of the parallel clustering algorithms follow the combinations of task and SPMD parallelism with Master – Slave Architecture.

Our study is based on the Single Program Multiple Data (SPMD) model using message-passing which is currently the most prevalent model for computing on distributed memory multiprocessors. Many applications for clustering algorithms, particularly applications in data mining, usually require the algorithms to work on massive data sets with an acceptable speed. For instance, in [2] NASA launches satellites for studying the earth’s ecosystem. The Earth Observing System (EOS) is capable of generating about a terabyte of data per day. These terabytes of data will then be used to identify anomalies on earth by a visualization program. A grouping of such data sets could be done by clustering algorithms. The k-means and Apriori data mining algorithm was implemented with the GPUMiner system [3]. The CPUMiner system is a one of the post useful system for data mining and data clustering. The researchers have used three modules of the

GPUMininer. The performances of these algorithms were improved significantly, and the computational speedup is also improved significantly. Judd *et al*.[4] designed and implemented a parallel clustering algorithm for a network of workstations. They used a client-server approach where a block of work assignments were sent to each client process, which then calculates the block partial sum of each cluster and send the results back to the server. The server collects the partial sums from all clients, calculates the new centroids and returns the new centroids to all clients to begin a new iteration. They used parallel virtual machine and message Passing Interface for their implementation.

The main objective of this paper is compare the performance of commonly used classical k-means clustering procedures as well as parallel k-means clustering to realize the advantage of the parallelism of algorithm on agricultural data sets. The present investigation has been taken up to achieve the above mentioned goal.

2. Parallel k-means Algorithm

2.1. k-Means Algorithm

The k-means method has been shown to be effective in producing good clustering results for many practical applications. K-means method is well known for its relatively simple implementation and decent results. However, a direct algorithm of k-means method requires time proportional to the product of number of documents (vectors) and number of clusters per iteration. This is computationally very expensive especially for large datasets.

The algorithm is an iterative procedure and requires that the number of clusters k be given a priory. Suppose that the k initial cluster centers are given, and then the algorithm follows the steps as below :

1. Compute the Euclidean distance from each of the documents to each cluster center. A document is associated with a cluster such that its distance from that cluster is the smallest among all such distances.
2. After this assignment or association of all the documents with one of k clusters is done, each cluster center is recomputed so as to reflect the true mean of its constituent documents.
3. Steps 1 and 2 are repeated until the convergence is achieved.

Suppose there are n data points or documents, $X_1, X_2, \dots X_n$ are given such that each one of them belongs to R^d . The problem of finding the minimum variance clustering of this dataset into k clusters is that of finding the k points $\{m_j\}^k{}_{j=1}$ in R^d such that

$$(1/n) * \Sigma(\min_j d^2(X_i , m_j)), \text{ for } i = 1 \text{ to } n$$

is minimized, where $d(X_i , m_j)$ denotes the Euclidean distance between X_i and m_j. The reason for popularity of k-means algorithm is ease of interpretation, simplicity of implementation, scalability, speed of convergence and adaptability to sparse data. The k-means algorithm can simply be explained as follows :

1. Phase Initialization

Select a set of k starting points $(m_j\}^k{}_{j=1}$ in R^d . This selection may be done in a random manner or making use of some heuristic.

2. Phase Distance Calculation

For each data point X_i $1 \le i \le n$, compute its Euclidean distance to each cluster m_j $1 \le j \le k$ and then find the closest cluster centroid.

3. Phase Centroid Recalculation

For each $1 \le j \le k$, recompute cluster centroid m_j as the average of data points assigned to it.

4. Convergence Condition

Repeat steps 2 and 3 until convergence.

2.2. parallel k-means Algorithm

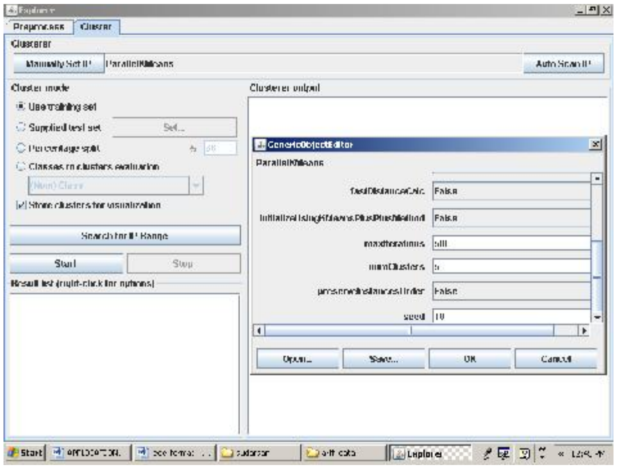
In contrast, the parallel k-means algorithm was developed for cluster analysis of very large data sets. The parallel algorithm can scale a problem size upto O(K) times the size of the problem on a single machine[6]. The implementation was done in the Java using Message Passing Interface(MPI) for distributed memory parallelism. The dataset to be clustered is divided among the available system. The initial centroid is selected by one system and is broad cast to all the system. Every process operates only on its chunk of the dataset, carries out the distance calculation of points from the centroids and assigns it to the closest centroid. Each process also calculates partial sum along each dimension of the points in each cluster for its chunk for the centroid calculation. At the end of the iteration, a global reduction operation is carried out, after each process obtains the information, to calculate the new cluster centroids. Iterations are carried out until convergence, after which the cluster assignments are written to an output file. A simple life cycle of the program can be listed as :

1. Master reads the data and divides it into N portions and sends one of them to each slave.
2. Master randomly initializes the centroids.
3. Master sends all of the centroids to all of the slaves.
4. Each slave receives a portion of dataset along with the centroids from master.
5. Each slave calculates the cluster membership of the data points assigned to it, and sends the results back to the master.
6. Master calculates new cluster centers by taking the means.
7. If converged, the process terminates, otherwise the master once again send the new centroids to the slaves for to calculate cluster membership for the new centroids.

3. Applications

Performance of the parallel k-means algorithm and implementation was evaluated using soil datasets which was collected from the department of Agriculture, Government of Tamil Nadu. By using parallel k-means clustering, the soil of Tamil Nadu was clustered according to their soil characteristics. The soil characteristics such as pH, EC, available nutrients like N, P, K, Zn, Fe, Mn, Cu, B and other features like Soil textures, and Lime status are considered as attributes for clustering.

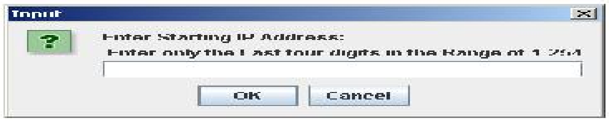
A tool was developed using Java for the implementation of parallel k-means clustering. ServerKmeans is a jar file to do the master process. ClientKmeans is a jar file to run in client/slave systems. (Fig.1). ClientKmeans is installed and in running condition in all clients.



fastDistanceCalc False
initializeUsingKMeansPlusPlusMethod False
maxIterations 500
numClusters 5
preserveInstancesOrder False
seed 10

Fig.-1 ServerKmeans accepts Number of cluster points

ServerKmeans accepts the IP numbers of nodes connected with the master(Fig. 2).



Enter Starting IP Address:
Enter only the Last four digits in the Range of 1-254

Fig.-2 Accepts IP number of clients

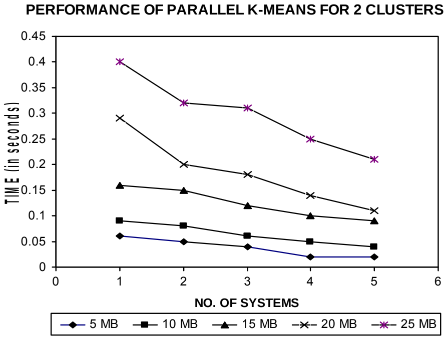
After getting IP address of the each slaves, the master sends the partitioned data and centroids to the clients.

4. Results and Discussion

We run the program in Pentium I5 with 2 GB ROM available in SCSVMV University, Kanchipuram Computer Centre. We used maximum of ten systems for our study. We tested our tool with some UCI datasets also. The performance of the parallel k-means algorithms was tested proved having a better performance compared with sequential k-means. Initially the number of clusters required is set as two. The time taken to cluster the different sizes of data is given in Table 1.

Table 1. Time complexity for the two clusters

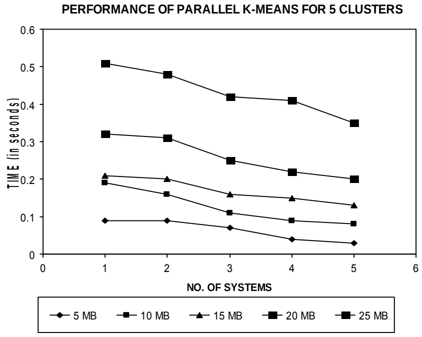
No. of Slaves / Data Size	1	2	3	4	5
5 MB	0.06	0.05	0.04	0.02	0.02
10 MB	0.09	0.08	0.06	0.05	0.04
15 MB	0.16	0.15	0.12	0.1	0.09
20 MB	0.29	0.2	0.18	0.14	.011
25 MB	0.4	0.32	0.31	0.25	0.21



The number of required clusters is also increased to five. The time taken to cluster the different sizes of data for the given cluster=5 is given in Table 2.

Table 2 Time complexity for the five clusters

No. of Slaves / Data Size	1	2	3	4	5
5 MB	0.09	0.09	0.07	0.04	0.03
10 MB	0.19	0.16	0.11	0.09	0.08
15 MB	0.21	0.2	0.16	0.15	0.13
20 MB	0.32	0.31	0.25	0.22	0.2
25 MB	0.51	0.48	0.42	0.41	0.35



We can conclude that our parallel k-means algorithm achieves more efficiency and time complexity than the normal sequential k-means algorithm. The processing speed is also disturbed with the bandwidth of the network available in the centre. However, our tool can be used to cluster the very large data sets in any field.

Acknowledgements

We would like to thank the Department of Computer Science & Engineering and Computer Science & Applications at SCSVMV University, Kanchipuram for allowing us to use the computer facilities for our experiments.

References

[1] Domenico Talia, Parallelism in Knowledge Discovery Techniques, PARA'02 Proceedings of the 6th International Conference on Applied Parallel Computing and Advanced Scientific Computing, pages 127-138, London,U.K

[2] K.Assiter, K.P. Lentz, A. Couch, and C.Currey, "Locating Anomalies in Large Data Sets", Society for Computer Simulation Military, Government, and Aerospace Simulation, April 5, 1998, pp. 218-223

[3] Wenbin Fang, Ka Keung Lau, Mian Lu, Xiangye Xiao, Chi Kit Lam, Philip Yang Yang, Bingsheng He, Quong Luo, Pedro V.Sander, and Ke Yang, "Parallel Data Mining on Graphics Processors", Technical Report HKUSTCS0807 (Oct.2008).

[4] D.Judd, P.McKinley, A. Jain, Larger-scale parallel data clustering, Pattern Analysis and Machine Intelligenc,e IEEE Transactions on 20 (8) (1998) 871-876.

[5] Sanpawat Kantabutra and Alva L.Couch, Parallel k-means Clustering Algorithm on NOWs, Technical Journal, Vol.1, No.6, January – Feburary 2000.

[6] Wooyoung Kim – Parallel Clustering Algorithms : Survey, Spring, 2009.

A Novel Particle Swarm Optimization-based Algorithm for the Optimal Centralized Wireless Access Network

Dac-Nhuong Le[1]**, and Gia-Nhu Nguyen**[2]

[1] **Faculty of Information Technology, Haiphong University**
Haiphong, Vietnam

[2] **Duytan University**
Danang, Vietnam

Abstract

The wireless access networks design problem is formulated as a constrained optimization problem, where the goal is to find a network topology such that an objective function is optimized, subject to a set of constraints. The objective function may be the total cost, or some performance measure like utilization, call blocking or throughput. The constraints may be bounds on link capacities, cost elements, or some network performance measure. However, the optimization problem is too complex. In this paper, we propose a novel Particle Swarm Optimization (PSO) algorithm to finding the total cost of connecting the BSs to the MSCs, and connecting the MSCs to the LE called by the optimal centralized wireless access network. Numerical results show that performance of our proposed algorithm is much better than previous studies.

Keywords: *Wireless Access Network, Base Station, Mobile Switching Center, Particle Swarm Optimization.*

1. Introduction

The wireless access network of a cellular telephone system consists four interacting layers. These layers are the mobile station or user equipment layers, the base transceiver stations layer, the mobile switching centers layer, and lastly local exchanges of the public switched telephone network (PSTN). Each cell in the hexagonal cell grid contains one base station (BS) and mobile station (MS). A set of BS's are physically connected to and served by a mobile switching center (MSC). In turn, a set of MSC's are physically connected to and served by a local exchange (LE). Fig.1 depicts the general configuration of a cellular access network. Each BS is typically assigned a group of radio channels (*frequency carriers*) to support a number of mobile stations in its cell. BS's at adjacent cells are assigned different sets of frequencies. The antennas of a BS are designed to achieve coverage only within the particular cell. By limiting coverage of a BS to its cell area, the set of frequencies assigned to this BS can be reused at other BS's that are distant enough to keep co-channel interference within acceptable limits.

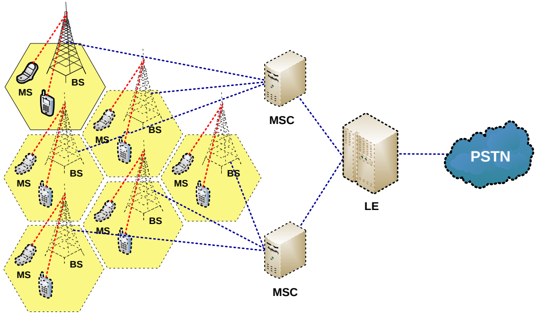


Fig.1. A cellular access network

A BS contains the radio transceivers defined for its cell, and handles the radio-link protocols with the user's wireless device (cell phone). In addition, it may house a controller that handles radio-channel setup, frequency hopping and handovers. In a large metro area, a potentially large number of BS's are deployed at pre-determined locations. The BS controllers are connected by land-wires to nearby MSC's in the area. The MSC provides all the functionality needed to handle a mobile subscriber, such as registration, authentication, location updating, handovers, and call routing to a roaming subscriber. To switch calls from/to local mobile users to/from remote users, MSCs are connected by land-cables to nearby LEs of the PSTN. The potential locations of MSCs are judiciously determined with respect to the BS locations and to the LEs in the region. Typically, the locations of the LEs are fixed, and a single LE serves an area with many BS's and multiple MSCs [1-2].

In the latest paper [3], we have proposed a novel Particle Swarm Optimization (PSO) [4] algorithm based on Ford-Fulkerson algorithm find maximum flow in networks for the optimal location of controllers in a mobile communication network. In [5], the authors have presented the topological design of the network connecting the BSs to the MSCs and the MSCs to the LEs in a typical region of the cellular system. The access network has a centralized tree topology. That is, a single LE facility controls a set of MSCs, and a single MSC controls, in turn, a set of BS's. Finally, a BS supports a group of mobile stations through wireless connections. A tree topology of the wireless access network, consisting of 1 LE, 2 MSCs, 4 BSs, and 18 MSs is shown in Fig.2.

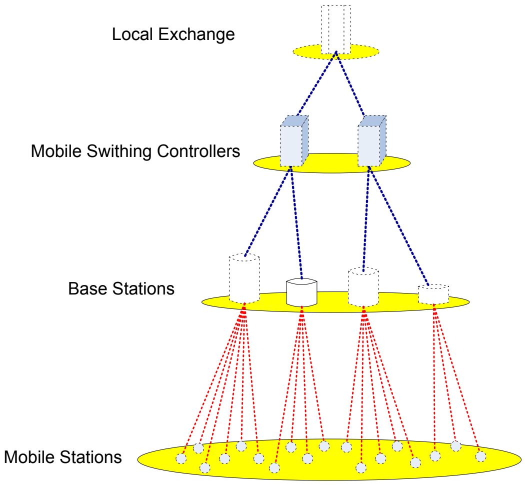


Fig.2. A Centralized access network

The topological design of access networks has been very important part of cellular network research in recent years. Recent studies are given in references [5-9]. Generally, the design problem is formulated as a constrained optimization problem, where the goal is to find a network topology such that an objective function is optimized, subject to a set of constraints. The objective function may be the total cost, or some performance measure like utilization, call blocking or throughput. The constraints may be bounds on link capacities, cost elements, or some network performance measure. However, the optimization problem is too complex, or it's computationally impractical to search for the optimal solution. So, one usually resorts to heuristic methods that enable one to determine a near-optimum network topology more easily.

The simple design of access network has one single LE. The objective function is the total cost of connecting the BSs to the MSCs, and connecting the MSCs to the LE. Authors in [6] proposed an exhaustive search algorithm to generating all the possible matrices and searches for the matrix that yields the minimum cost. In [7-9], the authors presented a heuristic algorithm to finding the best solution is the topology with the smallest cost across all the iterations.

In this paper, we propose a novel Particle Swarm Optimization (PSO) algorithm [10] to finding the total cost of connecting the BSs to the MSCs, and connecting the MSCs to the LE. Numerical results show that our proposed algorithm is much better than previous studies. The rest of this paper is organized as follows: Section 2 presents the problem formulation the simple centralized access network. Section 3 presents our new algorithm for optimization wireless access network based on PSO algorithm. Section 4 presents our simulation and analysis results, and finally, section 5 concludes the paper.

2. Problem Formulation

The simple centralized access network can be defined as follows [5]:

Let *N* be the number of BSs ($T_1, T_2, ..., T_N$). The locations of the *N* terminals are assumed known and fixed. Let *M* be the number of potential sites ($S_1, S_2, ..., S_M$), where up to *M* MSCs can be placed. In one extreme situation, none of the *M* sites is used, and all the *N* BSs are linked directly to the central LE, S_0.

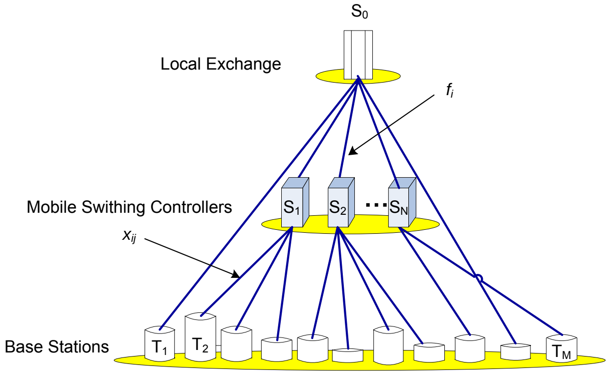


Fig.3. Simple centralized access network

In the other extreme, all the *M* MSC sites are used, each serving a subset of BS's. The principal constraint is that the MSC at site S_j can handle up to a maximum of P_j BSs ($j = 1..M$). This can be a hardware limitation, or a capacity constraint of the land-cable connecting the MSC to the LE. The central site is assumed to have no such constraint.

2.1 The Simple Centralized Wireless Access Network

We want to formulate the network design problem as an optimization problem. Let c_{ij} be the cost of connecting base station T_i to MSC site S_j or to the central site S_0. The cost

c_{ij} is measured in some unit (e.g., *dollar/month*), and represents the overall BS-MSC connection cost (e.g., *transmission cabling, interfacing, maintenance, leasing*). Note that a base station may be located at the site of an MSC, in which case the corresponding c_{ij} cost is zero.

These cost elements c_{ij} can be written in the form of a matrix, as follows:

$$C = \left(c_{ij}\right)_{N\times M+1} = \begin{pmatrix} c_{10} & c_{11} & \cdots & c_{1M} \\ c_{20} & c_{21} & \cdots & c_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ c_{N0} & c_{N1} & \cdots & c_{NM} \end{pmatrix} \qquad \textbf{(1)}$$

If an MSC at site S_j is utilized, the MSC capital cost and its connection cost to the LE are also incurred. Let f_j be the cost of connecting an MSC at S_j to the central LE S_0, and b_j the capital cost of the MSC at S_j.

We can write these 2 costs as row vectors, as follows:

$$\begin{cases} F = \left(f_0, f_1, ..., f_M\right) \\ B = \left(b_0, b_1, ..., b_M\right) \end{cases} \qquad \textbf{(2)}$$

We assumed that the capital cost of the central LE is not counted. That is, b_0=0, and clearly f_0=0. Similarly, we can write the MSC constraints as the following row vector:

$$P = \left(p_0, p_1, ..., p_M\right) \qquad \textbf{(3)}$$

where, p_j is the maximum number of BSs that MSC at site S_j can handle $(j = 1..M)$, with p_0=N (i.e., the central LE can handle all the N base stations).

A network design can be defined by the following matrix variable:

$$X = \left(x_{ij}\right)_{N\times M+1} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1M} \\ x_{20} & x_{21} & \cdots & x_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N0} & x_{N1} & \cdots & x_{NM} \end{pmatrix} \qquad \textbf{(4)}$$

where, the element variable x_{ij} $(i = 1..N, j = 0..M)$ is defined as:

$$x_{ij} = \begin{cases} 1 \ \ if \ \ T_i \ \ is\, connected \ \ to\, S_j \\ 0 \ \ if \ \ T_i \ \ is \ \ not\, connected \ \ to\, S_j \end{cases} \qquad \textbf{(5)}$$

Note that since a BS may be connected to at most one of the M MSC sites or to the central LE site, there must be only one "1" in each row of matrix X. In addition, note that the number of 1's in column j is the number of BSs connected to the MSC at site S_j $(j = 0..M)$. Thus, an all-zero column of matrix X corresponds to an MSC site that is not used.

Displayed equations or formulas are centered and set on a separate line (with an extra line or half line space above and below). Displayed expressions should be numbered for reference. The numbers should be consecutive within each section or within the contribution, with numbers enclosed in parentheses and set on the right margin. From matrix X, we extract the following MSC-usage vector:

$$Y = \left(y_0, y_1, ..., y_M\right) \qquad \textbf{(6)}$$

where, the element variable y_j $(j = 0..M)$ is defined as:

$$y_j = \begin{cases} 1 \ \ if \ \ S_i \ \ used, \ if \ \sum\limits_{i=1}^{N} x_{ij} > 0 \\ 1 \ \ if \ \ S_i \ \ not\, used, \ if \ \sum\limits_{i=1}^{N} x_{ij} = 0 \end{cases} \qquad \textbf{(7)}$$

The cost of a network design (defined by matrix X and vector Y) is thus expressed as follows:

$$\begin{aligned} & Z = \sum_{i=1}^{N}\sum_{j=0}^{M}\left(c_{ij}\times x_{ij}\right) + \sum_{j=0}^{M}\left(f_j \times y_j\right) + \sum_{j=0}^{M}\left(b_j \times y_j\right) \\ & \Leftrightarrow Z = sumdiag(C\times X^T) + F\times Y^T + B\times Y^T \end{aligned} \qquad \textbf{(8)}$$

In expression (8), the superscript T means transpose of matrix or vector, and *sumdiag*(A) is a function that sums up the diagonal elements of matrix A. The first term in the cost function Z is the cost of connecting the N BSs to the M MSCs used or to the central LE, the second term is the cost of connecting the MSCs to the LE, and the third term is the hardware cost of the MSCs used.

2.2 The Optimal Centralized Wireless Access Network

The optimal centralized wireless access network (OCWAN) in network design problem can thus be stated as the following optimization problem.

Problem instance:

- A set of BSs at known locations: $T_1, T_2, ..., T_N$.
- A set of possible MSC sites: $S_1, S_2, ..., S_M$.
- BS-connection cost matrix: $C = \left(c_{ij}\right)_{N\times M+1}$
- The cost of connecting an MSC at S_j to the central LE S_0: $F = \left(f_0, f_1, ..., f_M\right)$
- The capital cost of the MSC at S_j:
  $$B = \left(b_0, b_1, ..., b_M\right)$$
- The mux capacity constraint vector:
  $$P = \left(p_0, p_1, ..., p_M\right)$$

Objective function: Find the matrix *X* (thus the vector *Y*) that minimizes the network cost *Z*:

$$Z = sumdiag(C \times X^T) + F \times Y^T + B \times Y^T \to \min \qquad \textbf{(9)}$$

Subject to the following 2 constraints:

- The first constraint indicates that the sum of the elements in row *i* of matrix *X* must be 1 (*i*=1,2,…,*N*). *E* is the column vector of all 1's.

$$X \times E = E \qquad \textbf{(10)}$$

- The second constraint indicates that the sum of elements in column *j* of matrix *X* must be less than or equal to p_j $(j = 0..M)$.

$$E^T \times X \le P \qquad \textbf{(11)}$$

In the matrix inequality of equation (11), the inequality relation is defined element by element.

3. Particle Swarm Optimization for OCWAN

3.1 Particle Swarm Optimization

Particle swarm optimization (PSO) is a stochastic optimization technique developed by Dr. Eberhart and Dr. Kennedy [4],[10], inspired by social behavior of bird flocking or fish schooling. It shares many similarities with other evolutionary computation techniques such as genetic algorithms (GA). The algorithm is initialized with a population of random solutions and searches for optima by updating generations. However, unlike the GA, the PSO algorithm has no evolution operators such as the crossover and the mutation operator.

In the PSO algorithm, the potential solutions, called particles, fly through the problem space by following the current optimum particle. By observing bird flocking or fish schooling, we found that their searching progress has three important properties. First, each particle tries to move away from its neighbors if they are too close. Second, each particle steers towards the average heading of its neighbors. And the third, each particle tries to go towards the average position of its neighbors. Kennedy and Eberhart generalized these properties to be an optimization technique as below.
Consider the optimization problem *P*. First, we randomly initiate a set of feasible solutions; each of single solution is a *"bird"* in search space and called *"particle"*. All of particles have *fitness values* which are evaluated by the *fitness function* to be optimized, and have *velocities* which direct the flying of the particles. The particles fly through the problem space by following the current optimum particles. The better solutions are found by updating particle's *position*.

In iterations, each particle is updated by following two "best" values. The first one is the best solution (*fitness*) it has achieved so far. (The fitness value is also stored.) This value is called *pbest*. Another "best" value that is tracked by the particle swarm optimizer is the best value, obtained so far by any particle in the population. This best value is a global best and called *gbest*. When a particle takes part of the population as its topological neighbors, the best value is a local best and is called *lbest.*

```
PARTICLE SWARM OPTIMIZATION ALGORITHM
{ FOR each particle
        Initialize particle
  ENDFOR
  DO
     FOR each particle
        Calculate fitness value
        IF the fitness value is better than
        the best fitness value (pBest) in history
           Set current value as the new pBest
        ENDIF
     ENDFOR
        Choose the particle with the best fitness
value of all the particles as the gBest (or Choose the
particle with the best fitness value of all the
neighbors particles as the lBest)
        FOR each particle
            Calculate particle velocity according
                 to(12)or (13))
            Update particle position according to 14)
        ENDFOR
   WHILE (STOP CONDITION IS TRUE)}
```

Fig.4. Particle Swarm Optimization Algorithm

After finding the two best values, the particle updates its velocity and positions with following equation (12) (which use global best gbest) or (13) (which use local best *lbest*) and (14).

$$\begin{aligned} v[] = v[] + c_1 * rand() * (pbest[] - present[]) \\ + c_2 * rand() * (pbest[] - present[]) \end{aligned} \qquad \textbf{(12)}$$

$$\begin{aligned} v[] = v[] + c_1 * rand() * (pbest[] - present[]) \\ + c_2 * rand() * (lbest[] - present[]) \end{aligned} \qquad \textbf{(13)}$$

$$present[] = present[] + v[] \qquad \textbf{(14)}$$

In those above equation, *rand()* is a random number between *0* and *1*; c_1 and c_2 are cognitive parameter and social parameter respectively. The stop condition mentioned in the above algorithm can be the maximum number of interaction is not reached or the minimum error criteria are not attained.

3.2 Solving the OCWAN based on PSO algorithm

In this section, we present application of PSO technique for the OCWAN problem. Our novel algorithm is described as follows. We consider that configurations in our algorithm are sets of *N* BSs and set of *M* MCSs.

1) Represent and decode a particle: The encoding of the configuration is by means of matrix *x*, say

$$x=(x_{ij})_{N\times M+1},\ (i=1..N,\ j=0..M)$$

where $x_{ij}=1$ means that the corresponding BS T_i has been connected to MSC site S_j, and otherwise $x_{ij}=0$ means that the corresponding BS T_i has been not connected to MSC site S_j. We use fully random initialization in order to initialize the population.

2) Initiate population: We use fully random initialization in order to initialize the population. We present *Particle_Repair Algorithm* to ensure that the particle *x* satisfies constraints in (10) and (11) show in Fig.5.

```
PARTICLE_REPAIR ALGORITHM ( $x=(x_{ij})_{N\times M+1}$ )
Input: The particle $x$
Output: The particle $x$ will satisfies constraints
in (10) and (11)
{
FOR $i$=1..$N$
   $r_i$ = Number of 1s in row $i$
     IF $r_i$>1
             Select ($r_i$ -1) 1s randomly and
             removes them from row $i$
     ELSE IF Count$_i$<1
      Adds 1 1s in random positions in row $i$.
   ENDIF
ENDFOR
FOR j=0..M
     $c_j$ = Number of 1s in column $j$
     IF ( $c_j > p_j$ )
             Select ($c_j$ - $p_j$) 1s randomly and
             removes them from column$j$
     ELSE IF ( $c_j < p_j$ )
      Adds ($p_j$ -$c_j$) 1s in random positions in
column $j$.
   ENDIF
ENDFOR}
```

Fig.5. Particle_Repair algorithm

After that, the particle *x* will have the sum of the elements in row *i* of matrix *x* must be 1 (*i*=1,2,…,*N*) and the sum of elements in column *j* of matrix *x* must be less than or equal to $p_j\ (j=0..M)$.

3) Fitness function: The cost function of the particle *x* given by:

$$f_k = sumdiag(C\times k^T)+F\times Y^T+B\times Y^T \qquad \textbf{(15)}$$

In which, *Y* defined in (7) and (8).

4) Stop condition: The stop condition used in this paper is defined as the maximum number of interaction N_{gen} (N_{gen} is also a designated parameter).

4. Experiments and Results

For the experiments, we have tackled several OCWAN instances of different difficulty levels. There are 8 OCWAN instances with different values for *N* and *M*, and BS-connection cost matrix show in Table 1.

Table 1. The experimental of the problems tackled

Problem #	***Number of MSCs***	***Number of BSs***
#1	4	10
#2	5	20
#3	8	40
#4	10	80
#5	20	100
#6	40	150
#7	50	200
#8	60	250

We have already defined parameters for the PSO algorithm in Table 2 below:

Table 2. The PSO algorithm specifications

Population size	*P* = 1000
Maximum number of interaction	N_{gen} = 500
Cognitive parameter	c_1 = 1
Social parameter	c_2 = 1
Update population according to	(13) and (14)
Number of neighbor	*K* = 3

The experiment was conducted on Genuine Intel® CPU DuoCore 3.0 GHz, 2 GB of RAM machine. We ran experiment PSO algorithm, Exhaustive Search algorithm [5] and Heuristic algorithm [8] implemented using C language.

The experimental results of our algorithm was finally compared with others algorithm shown in Fig.6.

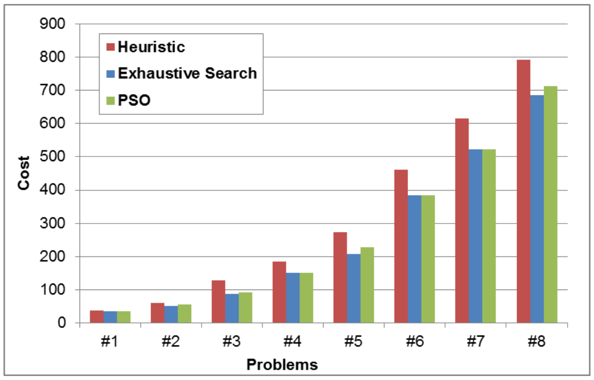


Fig.6. The results obtained in the OCWAN instances tackle

The results show that the objective function values of our algorithm has achieved a much better than a Heuristic algorithm and approximate good solutions of Exhaustive Search algorithm. But, the performance of our proposed algorithm is better than other algorithm.

The comparison of time processing shows in Fig.7.

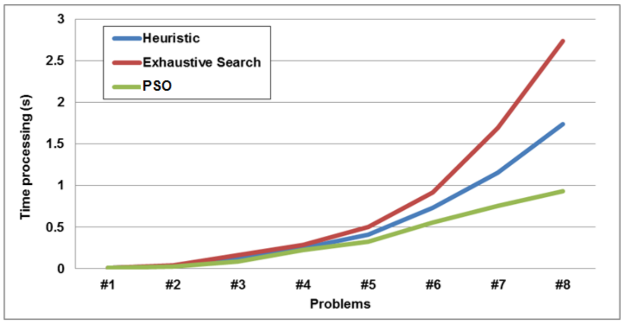


Fig.7. Comparison of time processing OCWAN instances tackle

5. Conclusion

In this paper, we have proposed a novel Particle Swarm Optimization (PSO) algorithm to finding the total cost of connecting the BSs to the MSCs, and connecting the MSCs to the LE called by the optimal centralized wireless access network. Numerical results show that performance of our proposed algorithm is much better than previous studies.

With a growing need for anywhere and anytime access to information and transaction, optimal capacity expansion of wireless networks to accommodate next-generation wireless service is our next research goal.

References

[1]. Mathar, R., Niessen, T., *Optimum positioning of base stations for cellular radio networks*, Wireless Networks, No.6, 2000.

[2]. Schwartz, M., *Computer-communication Networks*, Prentice-Hall, 1977.

[3]. Dac-Nhuong Le, Nhu Gia Nguyen, and Trong Vinh Le, *A Novel PSO-Based Algorithm for the Optimal Location of Controllers in Wireless Networks*, International Journal of Computer Science and Network Security, Vol.12 No.8, pp.23-27, 2012.

[4]. J. Kennedy and R. Eberhart, *Swarm Intelligence*, Morgan Kaufmann Publisher Inc, 2001.

[5]. K. Kraimeche, B. Kraimeche, K. Chiang, *Optimization of a wireless access network*, 2005.

[6]. Glaber, C., S. Reith, and H. Vollmer., *The Complexity of Base Station Positioning in Cellular Networks. Approximation and Randomized Algorithms in Communication Networks*, In Proceedings of ICALP, pp.167-177, 2000.

[7]. Galota, M., Glaber, C., Reith, S. Vollmer, H., *A Polynomial-time approximation scheme for base station positioning in UMTS networks*, ACM 2001.

[8]. B. Krishnamachari, S.Wicker, *Base station location optimization in cellular wireless networks using heuristic search algorithms*, Wang, L.(Ed.), Soft Computing in Communications, Springer.

[9]. Raisane, L., Whitaker, R., Hurley, S., *A comparison of randomized and evolutionary approaches for optimizing base station site selection*, ACM Symposium on Applied Computing, 2004.

[10].James Kennedy and Russell Eberhart, Particle Swarm Optimization, *in Proceedings of IEEE International Conference on Neural Networks*, pp.1942-1948, Piscataway, NJ, USA 1995.

[11].F. Houeto, S. Pierre, Assigning cells to switches in cellular mobile networks using taboo search, Systems, Man and Cybernetics, Part B, IEEE Transactions, vol. 32 No.3, pp.351-356, 2002.

[12].S. Salcedo-Sanz, and Xin Yao, A hybrid Hopfield network-genetic algorithm for the terminal assignment problem, IEEE Transaction on Systems, Man and Cybernetics. B. v34 i6, pp.2343-2353, 2006.

[13].S. Salcedo-Sanz, J. A. Portilla-Figueras, E. G. Ortiz-García, A. M. Pérez-Bellido, C. Thraves, A. Fernández-Anta, and X. Yao, *Optimal switch location in mobile communication networks using hybrid genetic algorithms*, Applied Soft Computing, pp.1486-1497, 2008.

[14].Dac-Nhuong Le, *Optimizing Resource Allocation to Support QoS Requirements in Next Generation Networks using ACO Algorithm*, International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol.2, No.5, pp.931-938, 2012.

[15].Dac-Nhuong Le, and Nhu Gia Nguyen, *A New Evolutionary Approach for Gateway Placement in Wireless Mesh Networks*, International Journal of Computer Networks and Wireless Communications (IJCNWC), Vol.2, No.5, pp.550-555, 2012

Dac-Nhuong Le received the BSc degree in computer science and the MSc degree in information technology from College of Technology, Vietnam National University, Vietnam, in 2005 and 2009, respectively. He is a lecturer at the Faculty of information technology in Haiphong University, Vietnam. He is currently a Ph.D student at Hanoi University of Science, Vietnam National

University. His research interests include algorithm theory, computer network and networks security.

Gia Nhu Nguyen received the BSc degree in computer science and the MSc degree in information technology from Dannang University, Vietnam, in 1998 and 2006, respectively. He currently works in Duy Tan University, Danang, Vietnam. His research interests include algorithm theory, network and wireless security.

Preserving Privacy Using Gradient Descent Methods Applied for Neural Network with Distributed Datasets

Mr.Sachin P.Yadav, Mr.Amit B.Chougule

[1] **D.Y.Patil College of Engineering**
Kolhapur,Maharashtra,India

[2] **Bharati Vidyapeet's College of Engineering,**
Kolhapur, Maharashtra, India

Abstract

The learning problems have to be concerned about distributed input data, because of gradual expansion of distributed computing environment. It is important to address the privacy concern of each data holder by extending the privacy preservation concept to original learning algorithms, to enhance co-operations in learning. In this project, focus is on protecting the privacy in significant learning model i.e. Multilayer Back Propagation Neural Network using Gradient Descent Methods. For protecting the privacy of the data items (concentration is towards Vertically Partitioned Data and Horizontally Partitioned Data), semi honest model and underlying security of El Gamal Scheme is referred [7].

Keywords: *Cryptography Techniques, Distributed Datasets, Gradient Descent Methods, Neural Network.*

1. Introduction

Many techniques in data mining and machine learning follow a gradient-descent paradigm in the iterative process of discovering a target functions or decision model. For instance, neural networks generally perform a series of iterations to converge the weight coefficients of edges in the network; thus, settling into a decision model. Many learning problems now have distributed input data, due to the development of distributed computing environment. In such distributed scenarios, privacy concerns often become a big concern. For example, if medical researchers want to apply machine learning to study health care problems, they need to collect the raw data from hospitals and the follow-up information from patients. Then, the privacy of the patients must be protected, according to the privacy rules in Health Insurance Portability and Accountability Act (HIPAA) [1], which establishes the regulations for the use and disclosure of Protected Health Information. Why the researchers would want to build a learning model (e.g. Neural networks) without first collecting all the training data on one computer is a natural question.

If there is a learner trusted by all the data holders, then the trusted learner can accumulate data first and build a learning model. However, in many real-world cases, it is rather difficult to find such a trusted learner, since some data holders will always have concerns like "What will you do to my data?" and "Will you discover private information beyond the scope of research?" On the other hand, given the distributed and networked computing environments at present, alliances will greatly benefit the scientific advances [2].

The researchers have the interest to obtain the result of cooperative learning even before they see the data from other parties. As a concrete example, the progress in neuroscience could be boosted by making links between data from labs around the world, but some researchers are reluctant to release their data to be exploited by others because of privacy and security concerns.

2. Related Work

2.1 "Privacy-Preserving Data Mining"

D. Agrawal and R. Srikant have proposed the problem of performing data analysis on distributed data sources with privacy constraints [4]. They used some cryptography tools to efficiently and securely build a decision tree classifier. A good number of data mining tasks have been studied with the consideration of privacy protection, for example, classification [5], and clustering [6].

In particular, privacy-preserving solutions have been proposed for the following classification algorithms (to name a few): decision trees, naive Bayes classifier [8], and support vector machine (SVM) [9] Generally speaking, the existing works have taken either randomization-based approaches or cryptography- based approaches[7] Randomization-based approaches, by perturbing data, only guarantee a limited degree of privacy.

2.2 "Protocols for Secure Computations"

A.C.Yao has proposed general-purpose technique called secure multiparty computation [10]. The works of secure multiparty computation originate from the solution to the millionaire problem proposed by Yao, in which two millionaires can find out who is richer without revealing the amount of their wealth. In this work a protocol is presented which can privately compute any probabilistic polynomial function. Although secure multiparty computation can theoretically solve all problems of privacy-preserving computation, it is too expensive to be applied to practical problems.

Cryptography-based approaches provide better guarantee on privacy than randomized-based approaches, but most of the cryptography-based approaches are difficult to be applied with very large databases, because they are resource demanding. For example, although Laur et al. proposed an elegant solution for privacy-preserving SVM in [9], their Protocols are based on circuit evaluation, which is considered very costly in practice.

2.3 "Privacy-Preserving Gradient-Descent Methods"

L.Wan, W. K. Ng, S. Han, and V. C. S. Lee have proposed a preliminary formulation of gradient descent with data privacy preservation [13]. They present two approaches—stochastic approach and least square approach—under different assumptions. Four protocols are proposed for the two approaches incorporating various secure building blocks for both horizontally and vertically partitioned data.

Major headings are to be column centered in a bold font without underline. They need be numbered. "2. Headings and Footnotes" at the top of this paragraph is a major heading.

3. Gradient Decent Method

Gradient descent is a general paradigm that underlies many algorithms in machine learning and knowledge discovery. In neural networks, updating the weight value of the output and hidden nodes is a form of gradient descent. There are two approaches of Gradient Decent Method-Stochastic Approach and the Least Square Approach.

For proposed system Least Square Approach is suitable for Back Propagation Neural Network.

3.1 Stochastic Approach:

For neural networks, the two component functions for the prediction function f in unipolar sigmoid activation function in the multilayer perceptron are: $\mathbf{hj(xj, wj) = wjxj}$ and $\mathbf{gh = \frac{1}{1 + e^{-\alpha h}}}$ where α is a positive constant.

Correspondingly, the bipolar sigmoid activation function is $\mathbf{g(h) = \frac{1 - e^{-\alpha h}}{1 + e^{-\alpha h}}}$ and $\mathbf{hj(xj, wj) = wjxj}$.

3.2 Least Square Approach:

The objective of the gradient descent is to determine w to best fit the training data that minimize the error function. In the following, we introduce a simpler case that can be computed based on the least square approach. Here, we assume that the global error function is the Residual Sum of Squared (RSS) values in (1) for the training data.

$$\mathbf{RSS = \sum_{i=1}^{n} \left(yi - f(xi)\right)^2} \tag{1}$$

Besides, we define the prediction function f as a composition of two functions g and h where 1) function g is an invertible function (such as the inverse function of $\mathbf{y = \frac{1}{1 + e^{-\alpha x}}}$ is $\mathbf{x = -\frac{1}{\alpha} \ln\left(\frac{1}{y} - 1\right)}$ and 2) function h is a linearly separable function:

$$\mathbf{h(xi) = \sum_{j=1}^{m} xi, jwj} \tag{2}$$

That is applied by many gradient-descent methods.

4. Comments and Need of Work

From the above survey it can be comment that:

1. The Gradient Descent methods are used for solving optimization problems. This method gives general formulation for preserving privacy of distributed datasets to solve optimization problems.

2. The privacy preserving Gradient Descent Method can be applied for some specific areas like Neural Network, Linear Regression, and Bayesian Network.

3. There is no any privacy preserving mechanism for distributed datasets (Vertically partitioned and horizontally partitioned) using Gradient Descent Method in neural network for solving classification problem.

The work carried out in above Literature Survey is based on privacy preserving Data Mining for performing data analysis on distributed datasets. Further the solution proposed in the Privacy Preserving Gradient Descent Methods gives general formulation for preserving privacy of distributed datasets to solve optimization problems. It is therefore required that this solution for the privacy preserving gradient descent method needs to be converted to the Neural Network for solving classification problem.

5. Proposed Work

Here focus is to implement the privacy-preserving distributed algorithm to securely compute the piecewise linear function for the neural network training process to obtain the desired output.

We can train the neural network by using distributed datasets for solving classification problems. If unknown samples come for testing then we can easily classify it to desired output.

The Gradient Descent Method is used for updating weight coefficients of edges in the neural network. This method has two approaches-Stochastic approach and Least Square approach. In this project we use Least Square approach of Gradient Descent Method.
Briefly the work can be summarized into following proposed system architecture.

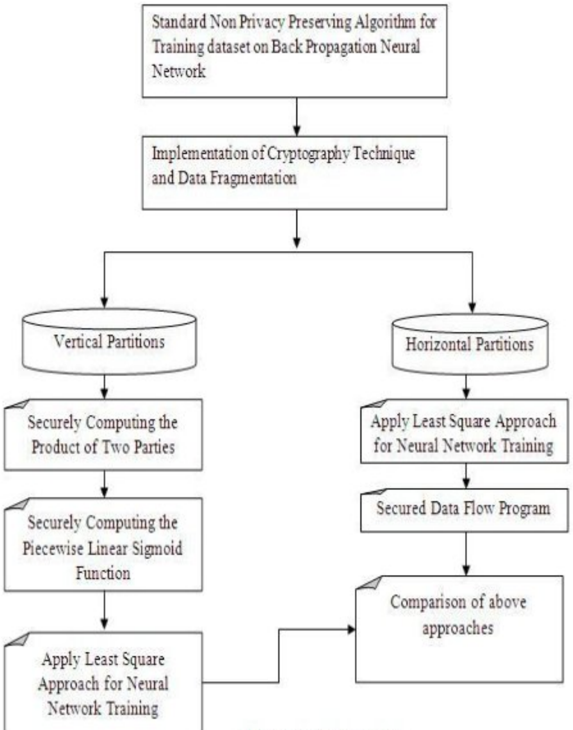

Standard Non Privacy Preserving Algorithm for Training dataset on Back Propagation Neural Network
Implementation of Cryptography Technique and Data Fragmentation
Vertical Partitions
Horizontal Partitions
Securely Computing the Product of Two Parties
Apply Least Square Approach for Neural Network Training
Secured Data Flow Program
Securely Computing the Piecewise Linear Sigmoid Function
Comparison of above approaches
Apply Least Square Approach for Neural Network Training

Fig: System Architecture

The Proposed system will contain following modules.

5.1 Implementing a standard non privacy preserving algorithm on Neural Network

For better understanding, the back propagation learning algorithm can be divided into two phases: propagation and weight update.

Phase 1: Propagation

Each propagation involves the following steps:

1) Forward propagation of a training pattern's input through the neural network in order to generate the propagation's output activations.

2) Backward propagation of the propagation's output activations through the neural net-work using the training pattern's target in order to generate the deltas of all output and hidden neurons.

Phase 2: Weight update

For each weight-synapse follow the following steps:

1) Multiply its output delta and input activation to get the gradient of the weight.

2) Bring the weight in the opposite direction of the gradient by subtracting a ratio of it from the weight.

3) This ratio influences the speed and quality of learning; it is called the learning rate. The sign of the gradient of a weight indicates where the error is increasing; this is why the weight must be updated in the opposite direction.

4) Repeat phase 1 and 2 until the performance of the network is satisfactory.

There are two modes of learning to choose from, one is on-line (incremental) learning and the other is batch learning. In on-line (incremental) learning, each propagation is followed immediately by a weight update. In batch learning, much propagation occurs before weight updating occurs. Batch learning requires more memory capacity, but on-line learning requires more updates.

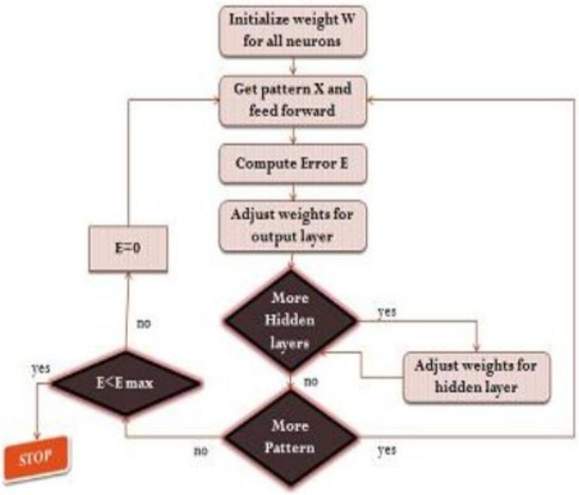

Initialize weight W for all neurons
Get pattern X and feed forward
Compute Error E
Adjust weights for output layer
E=0
no
More Hidden layers
yes
Adjust weights for hidden layer
yes
E<E max
no
STOP
no
More Pattern
yes

Fig: Data Flow of a standard non privacy preserving algorithm on Neural Network

5.2 Implementing cryptography technique for preserving privacy of owner's dataset and fragmentation of data

For ease of presentation, in this paper, we consider a neural network of three layers, where the hidden-layer activation function is sigmoid and the output layer is linear. Note that it is trivial to extend our work to more layers.

a. Semi honest Model:

As many existing privacy-preserving data mining algorithms, here in this work semi honest model is adopted. Semi honest model is a standard adversary model in cryptography. In this work, the security of algorithm is guaranteed in this model. When computing function in a distributed fashion, semi honest model requires that each party that participates in the computation follow the algorithm, but a party may try to learn additional information by analyzing the messages that she receives during the execution. In order to guarantee the security of distributed algorithm of computing, it must be ensured that each party can learn nothing beyond what can be implied by her own input and output.

Semi honest model is a right t for this work's setting, because normally participants want to learn the neural network learning results and thus they are willing to follow the algorithm to guarantee the results correctness. The security guaranteed in semi honest model can relieve the concerns about their data privacy. Of course, in reality, there may be scenarios in which there are malicious adversaries. It has been shown that a distributed algorithm that is secure in the semi honest model can be converted to one that is secure in the malicious model, with some additional costs in computation and communications for zero knowledge proofs.

We use here El-Gamal Cryptography technique.

- El Gamal Encryption Scheme:
- El Gamal is a public-key encryption scheme.
- Setup
  - Choose Large Prime p
  - Choose primitive element $\alpha \varepsilon Zp^*$
  - Choose secret key a ε {2,3,…..,p-2}.
  - Compute $\beta = \alpha a \bmod p$.
  - Public Key = Kpub = (p, α, β).
  - Private Key = Kpr = (a).
- Encryption
  - Choose k ε {2,3,…..,p-2}.
  - Y1= αk mod p.
  - Y2= x.βk mod p
  - Encryption :Ekpub (x, k) = (Y1, Y2).
- Decryption
  - x= Dkpr(Y1, Y2) = Y2 (Y1a)-1 mod p

1. Homomorphic Property: For two messages m1 and m2, an encryption m1m2 of can be obtained by an operation on E(m1,r) and E(m2,r) without decrypting any of the two encrypted messages.

2. Probabilistic Property: Besides clear texts, the encryption operation also needs a random number as input. There exist many encryptions for each message. One encrypted message as input and outputs another encrypted

message of the same clear message. This is called re-randomization operation.

5.3 Implementation of Securely Computing the Piecewise Linear Sigmoid Function for Vertically partitioned dataset.

In this section, we present a privacy-preserving distributed algorithm for training the neural networks with back propagation algorithm. A privacy-preserving testing algorithm can be easily derived from the feed forward part of the privacy-preserving training algorithm. Our algorithm is composed of many smaller private computations. We will look into them in detail after first giving an overview

Algorithm No 1:

Securely Computing the Product of Two Integers.

Assume M= Integer hold by Party A.

N= Integer hold by Party B.

Party A:

1) Generates a Random Number R
2) Computes M.i – R for each I, s.t –n<i<n Mi = M.i – R.
3) Encrypts each Mi using ElGamal Scheme using new random number for each Mi
4) Sends each E(Mi,ri) to Party B in increasing order of i

Party B:

1) B picks E(MN,RN), randomizes it and sends back to A E(MN,r'), r' = rN+S where S is known to Party B

Party A:

1) Party A partially decrypts E(Mn,r') and sends to B

Party B:

1) Finally decrypts to get Mn = M.N-R

Algorithm No 2:

Securely Computing Piecewise Linear Sigmoid Function

Assume M= Integer hold by Party A.

N= Integer hold by Party B.

Party A:

1) Generates a Random Number R
2) Computes y (X1+i)-R for each I, s.t –n<i<n Mi = y (X1+i)-R
3) Encrypts each Mi using ElGamal Scheme using new random number for each Mi
4) Sends each E(Mi,ri) to Party B in increasing order of i

Party B:

1) B picks E(MN,RN), randomizes it and sends back to A E(MN,r'), r' = rN+S where S is known to Party B

Party A:

1) Party A partially decrypts E(Mn,r') and sends to B

Party B:

1) Finally decrypts to get Mn = y(x1+x2)-R

5.4 Implementation of Privacy preserving Least Square Approach Algorithm on Vertically Partitioned dataset for Back propagation Neural Network Training.

Algorithm: Least Square Approach for Vertically Partitioned case for Two Parties.

Initialization to Random weight values and making them known to both parties
And for all Training Sample Repeat below steps
Step1: Feed Forward Stage

1) For each hidden layer node hj, Party A computes weight * input for Ma attributes
2) Party B computes weight * input for Mb attributes
3) Using Algorithm 2, Party A and B jointly compute Sigmoid Function for each hidden layer node hj obtaining their random shares hj1 and hj2 respectively.
4) For each output layer oi , Party A computes oi1as
$$o_{i1} = \sum_i w_{ij}{}^{o} h_{j1}$$
5) For each output layer oi , Party B computes oi2 as
$$o_{i2} = \sum_i w_{ij}{}^{o} h_{j2}$$

Step 2: Back Propagation Stage

1) For each Output Layer weight , parties A and B apply Algorithm 1 to securely compute product $h_{j1}o_{i2}$ obtaining random shares r11 and r12. Similarly they compute $h_{j2}o_{i1}$, to get r21 and r22 as shares
2) Party A Computes Δ1 wij as (oi1-ti)hj1+r11+r21
3) Party B Computes Δ2 wij as (oi2-ti)hj2+r12+r22
4) Similarly step is repeated to Hidden Layers to calculate the delta values back propagating from output layer to hidden layer

Step 3:

1) A sends Δ1 of output and hidden layers to B and B sends Δ2 of output and hidden layers to A.
2) A and B compute new weight vector values accordingly also considering the learning rate.(At hidden and Output Layers)
3) This rate is kept same at both parties.
4) Finally, repeat above three steps until terminating condition for error threshold occurs or after predefined number of iterations.

5.5 We can apply Least Square approach algorithm for horizontal partitioned dataset for effectively obtaining exact output for classification of data.

Algorithm: Least Square Approach for Horizontally partitioned case for Two Party

Two rounds of BPA algorithm should be called and for a single round of algorithm logic is same as Non privacy preserving algorithm for BPA training. This training suits the Least Square Approach.

6. Conclusions

Using Privacy preserving gradient decent method applied for Back Propagation Neural network on distributed datasets. We can preserve privacy of dataset holders and no one can get others private data then also our neural network is gets trained for distributed datasets.

We can extend this work for other types of neural network in future. And also generalize the neural network.

Acknowledgments

I thanks to Prof. G.A.Patil Head of Department Computer Science and Engineering, D Y Patil College of Engineering, Kolhapur. And I also thanks to Prof. Amit B.Chougule for supporting me.

References

[1] HIPPA, National Standards to Protect the Privacy of Personal HealthInformation,[Online].Available:http://www.hhs.gov/ocr/hipaa/finalreg.html

[2] M. Chicurel, “Data basing the brain,” Nature, vol. 406, pp. 822–825, Aug. 2000.

[3] D. Agrawal and R. Srikant, “Privacy-preserving data mining,” in Proc. ACM SIGMOD

[4] Y. Lindell and B. Pinkas, “Privacy preserving data mining,” in Lecture Notes in Computer Science. Berlin, Germany: Springer-Verlag, 2000, vol. 1880, pp. 36–44.

[5] N. Zhang, S. Wang, and W. Zhao, “A new scheme on privacy-preserving data classification,” in Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining, 2005.

[6] G. Jagannathan and R. N. Wright, “Privacy-preserving distributed k-means clustering over arbitrarily partitioned data,” in Proc. ACM

[7] O. Goldreich, Foundations of Cryptography. Cambridge Univ. Press, 2001.

[8] R. Wright and Z. Yang, “Privacy-preserving Bayesian network structure computation on distributed heterogeneous data,” in Proc. 10th ACM SIGKDD.

[9] H. Yu, X. Jiang, and J. Vaidya, “Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data,” in Proc. Annu. ACM Symp. Appl. Comput., 2006.

[10] A. C. Yao, “Protocols for secure computations,” in Proc. 23rd Annu. Symp. Found. Comput. Sci., Chicago, IL, Nov. 1982.

[11] M. Barni, C. Orlandi, and A. Piva, “A privacy-preserving protocol for neural-network-based computation,” in Proc. 8th Workshop Multimedia Security, New York, 2006.

[12] A. Yao, “How to generate and exchange secrets,” in Proc. 27th IEEE Sym p. Found. Comput. Sci., 1986, pp. 162–167.

[13] L.Wan, W. K. Ng, S. Han, and V. C. S. Lee, “Privacy-preservation for gradient descent methods,” in Proc. IEEE Transactions on Knowlede and Data Engineering, 2010.

First Author: Mr. S. P. Yadav, BE degree in Information Technology from Shivaji University Kolhpur,Maharashtra,India.Currently he is pursuing his ME in Computer Science and Engineering in D.Y.Patil College of Engineering and Technolgy,Kolhapur,Maharashtra and working as a Assistant Professor at Annasaheb Dange college of engineering,Ashta,Tal:Walwa,Dist Sangli.

Second Author: Prof. A. B. Chougule, M.Tech.Working as Professor at Bharati Vidyapeet's College of Engineering, Kolhapur, Maharashtra, India.

Pair versus Solo Programming – An Experience Report from a Course on Design of Experiments in Software Engineering

Omar S. Gómez, José L. Batún and Raúl A. Aguilar[1]

[1]Faculty of Mathematics, Autonomous University of Yucatan
Merida, Yucatan 97119, Mexico

Abstract

This paper presents an experience report about an experiment that evaluates duration and effort of pair and solo programming. The experiment was performed as part of a course on Design of Experiments (DOE) in Software Engineering (SE) at Autonomous University of Yucatan (UADY). A total of 21 junior student subjects enrolled in the bachelor's degree program in SE participated in the experiment. During the experiment, subjects (7 pairs and 7 solos) wrote two small programs in two sessions. Results show a significant difference (at α=0.1) in favor of pair programming regarding duration (28% decrease), and a significant difference (at α=0.1) in favor of solo programming with respect to effort (30% decrease). With only a difference of 1%, our results regarding duration and effort are practically the same as those reported by Nosek in 1998.

Keywords: *Software Engineering, Pair Programming, Design of Experiments, Latin Square Design, Experimentation, Experience Report.*

1. Introduction

Since the seminal work of Fisher on principles of experimental design [13], the design of experiments (DOE) for obtaining information has been widely used in natural sciences, social sciences and engineering.

When a researcher is designing an experiment, (s)he is interested in analyzing the effect produced in a treatment or intervention that is applied on certain objects or experimental units such as: Persons, plants, animals, etc. SE experiments use to employ persons acting as experimental units, where persons are asked to perform certain tasks that usually constitute a treatment or intervention.

The SE degree program at Autonomous University of Yucatan offers a course on DOE. In this course, students learn to analyze the effect produced in a treatment or intervention by using different types of experimental designs.

As part of this course, during the summer semester 2012 we decided to carry out an experiment; this with the aim of students learn to collect and analyze measures given an experimental design. The experiment selected for the course consisted in analyzing a couple of pair programming aspects.

One of the twelve main practices of extreme programming created by Kent Beck in the late 90s [3, 4] is pair programming. In this practice, two programmers work together on the same task using a computer. One of the programmers (the driver) writes the program whereas the other (the observer) reviews actively the work done by the controller. The observer reviews against possible defects, writes down annotations, or defines strategies for solving any issue that can rise over the task they are working on.

Some experiments have been conducted to study the effect of pair programming [24, 28, 19, 21, 22, 7, 20]. In a general way, these experiments report beneficial effects of applying this practice. Some beneficial effects reported are that it helps to produce shorter programs and helps to implement better designs; programs contain less defects than those written individually, and pairs usually require less time to complete a task than programmers working individually.

Under an academic context, the experiment proposed for the DOE course analyzes the duration and effort needed to write small programs in pairs and individually. The rest of the paper is organized as follows: Section 2 presents the experiment definition. Section 3 describes the design and conduction of the experiment. Section 4 presents the analysis. Section 5 discusses some experiment limitations. In section 6 we discuss the results we found. Finally, in section 7 we present the conclusions and further work.

2. Experiment Definition

We use the Goal-Question-Metric approach [2] for defining the experiment. This approach facilitates to identify the object of study, purpose, quality focus, perspective and context of an experiment. We define the experiment as follows:

Study pair and solo programming with the purpose of evaluating possible differences between these two programming types with respect to duration and effort. This study is conducted from the point of view of the researcher under an academic context. This context is composed by juniors students enrolled in a course of DOE where they will write, by pairs or individually, two small programs.

From the experiment definition we derive the following hypotheses:

$H_0$1: The time required to write a program in pair is equal to the time required to write it individually or: Pair programming = Solo programming regarding time duration.

H_a1: The time required to write a program in pair is different to the time required to write it individually or: Pair programming ≠ Solo programming with respect to time duration.

$H_0$2: The effort required to write a program in pair is equivalent to the effort required to write it individually or: Pair programming = Solo programming regarding effort.

H_a2: The effort required to write a program in pair is different to the effort required to write it individually or: Pair programming ≠ Solo programing with respect to effort.

3. Experiment Design and Conduct

The previous hypotheses will be tested through different measures that we will collect from subjects during the experiment. In a general way, measures belong to two subject groups: Those who perform a task in pairs and those who perform it individually. With these measures, we will perform statistical analyses given an experimental design.

At the beginning of the DOE course, we decided to conduct the experiment at the midterm (semester) in order to students had certain knowledge of DOE and that they had sufficient time to write a report before the semester ended.

The experimental design to use was selected according to the designs listed in the DOE course syllabus. Specifically, we chose the Latin square design because it was scheduled in the course syllabus at midterm, just a few days before the experiment was conducted.

3.1 Latin Square Designs

The main features of Latin square designs are that there are two blocking factors. Each treatment is present at each level of the first blocking factor and is also present at each level of the second blocking factor. This design is arranged with an equal number of rows (factor one) and columns (factor two). Treatments are represented by Latin characters symbols where each symbol is present exactly once in each row, and exactly once in each column. An example of the arrangement of this design is shown in Table 1.

Table 1: Latin square design with three treatments

A	B	C
B	C	A
C	A	B

In a Latin square design, blocking is used to systematically isolate the undesired source of variation in the comparison among treatments. In this case, pair versus solo programming. As a teaching purpose, we decided to block treatments by program and by tool support. Table 2 shows the arrangement used for the experiment.

Table 2: Latin square design arrangement

Program / Tool Support	***IDE***	***Text Editor***
Calculator	Solo	Pair
Encoder	Pair	Solo

The program block has two levels: Calculator an encoder whereas tool support block has the levels: IDE (Integrated Development Environment) and text editor. The treatments to examine are: Pair and solo programming.

3.2 Subjects, Tasks and Objects

Junior students enrolled in the DOE course participated as subjects in the experiment; in total, for this experiment there were 21 subjects. Most of the subjects were in their third year of the program's degree in SE; the rest of them (three subjects) were in their four year. According of Dreyfus and Dreyfus programming expertise classification [12], we categorized subjects as advanced beginners; subjects have working knowledge of key aspects of Java programming practice.

Subjects were randomized and allocated into two groups: Pair and solo programmers. The experiment was split into two sessions, where in each session subjects wrote a different program. In both sessions we employed the same subjects, so we collected 14 measures with respect to solo programmers (7 solos per session) and 14 measures regarding pair programmers (7 pairs per session). In the first session, subjects that worked individually used NetBeans IDE (as tool support) to write the first program, whereas subjects that worked in pairs used only a text editor. In the second session the tool support was changed, so subjects that before worked individually with the NetBeans IDE, in the second session they worked with a text editor and conversely (See Latin square design arrangement in Table 2).

Before the experiment was conducted, we gave a talk to the students about pair programing. In this talk we explained the main concepts of this programming practice and how it can be used in practice. We also explained how to compile a Java program using only a text editor. Finally, we explained to students how to collect the measures during the experiment sessions. The collection procedure consisted in writing down the time duration that students spent writing a program. They recorded the start and finish time and computed the difference (in minutes).

We selected to small programs that subjects could write, compile, run and test in each session. In the first program (identified as calculator) we asked the subjects to write a calculator that evaluates expressions with decimal numbers, and the operators: Plus (+), minus (-), times (×), divide (/), and prints the result on the screen. In the second program (identified as encoder) we asked the subjects to write a simple encoding-decoding program. Given a specified letter switch the program must be able to encode or decode a line of text.

3.3 Conduct

The allotted time for each session was 90 minutes. Both sessions were carried out in one of the computer classroom of the faculty. The first session started almost 30 minutes late because we were waiting for some students to arrive. Once students were complete, we started the session. We gave to subjects some directions and projected on the screen the specification of the program to be written (program calculator). Due to we did not start on time, some subjects did not complete the assignment, so we asked them to pause their work and record the time. Subjects that were working individually we asked them to finish the program at home. At the other hand, subjects that were working in pairs and did not complete the program, we programmed them an extra session on the next day. In this extra session all the remaining pairs completed the program.

The second session started on time; again, we gave to subjects some directions and projected on the screen the second specification (program encoder). In this session all the subjects finished on time. In both sessions programs were verified according to its specification.

3.4 Measures

We used the time records of subjects to define the following measures:

Duration: It is the elapsed time in minutes to write the program. Before starting the program assignment, subjects wrote down the current time. When they completed the program, they registered the finish time; then we calculate the difference in minutes between start and finish time.

Effort: It measures the amount of labor spent to perform a task. It is the total programming effort in person-minutes to write a program. Total effort for a pair is the duration multiplied by two. Tables 3 and 4 show the measures (in minutes) collected for the experiment.

Table 3: Measures collected for duration

Program / Tool Support	*IDE*	*Text Editor*
Calculator	Solo: 110, 136, 281, 239, 126, 69, 205	Pair: 256, 184, 114, 59, 37, 89, 135
Encoder	Pair: 70, 48, 88, 85, 43, 39, 56	Solo: 66, 102, 128, 107, 106, 76, 64

Table 4: Measures collected for effort

Program / Tool Support	*IDE*	*Text Editor*
Calculator	Solo: 110, 136, 281, 239, 126, 69, 205	Pair: 512, 368, 228, 118, 74, 178, 270
Encoder	Pair: 140, 96, 176, 170, 86, 78, 112	Solo: 66, 102, 128, 107, 106, 76, 64

4. Data Analysis

Once we have the measures, we are able to test the hypotheses through statistical inferences. The statistical model associated with a Latin square design is shown in equation (1).

$$y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + \epsilon_{ijk} \qquad (1)$$

Where μ is the overall mean, α_i is the block effect common to row i, β_j is the block effect common to column j, τ_k is the *k* th treatment effect, and ϵ_{ijk} is a random error which is assumed to be $N(0, \sigma^2)$.

This design uses analysis of variance (ANOVA) to assess the components (overall mean, blocks, treatment and random error) of the model. ANOVA is based on looking at the total variability of the collected measures and the variability partition according to different components. ANOVA provides a statistical test of whether or not the means of several groups are all equal. The null hypothesis is that all groups are simply random samples of the same population. This implies that all treatments have the same effect (perhaps none). Rejecting the null hypothesis implies that different treatments result in altered effects. In this experiment, we have two groups of means (Pair and Solo programming), which are blocked by program and tool support.

4.1 Model Assumptions

Before we start to draw any conclusion, we must assess the following model assumptions:

1. All observations are independent (independence)
2. The variance is the same for all observations (homogeneity)
3. The observations within each treatment group have a normal distribution (normality)

The first assumption is addressed by the principle of randomization used in this experimental design; all the measures of one sample are not related to those of the other sample. The second and third assumptions are assessed by using the estimated residuals [6, 16]. To assess homogeneity of variances we use a plot to show a scatter plot of the standardized residuals against the estimated mean values (sometimes called fitted values). We also use the Levene test for homogeneity of variances [17]. The third assumption (normality) is evaluated by using a normal probability plot, and applying the Kolmogorov-Smirnov test for normality [15, 26].

Selecting the duration measure, Fig. 1 shows a scatter plot of the standardized residuals versus fitted values. Violations to the homogeneity variance assumption can be detected with either plot by noting that the variation in the vertical direction seems to differ at different points along the horizontal axis. In this case, Fig. 1 shows a different pattern between the vertical points. Applying the Levene test [17] we get a p-value of 0.0594. Setting an alpha level of 0.05 this test is significant (selecting only two decimal of the p-value with no rounding off), so the assumption of homogeneity is violated.

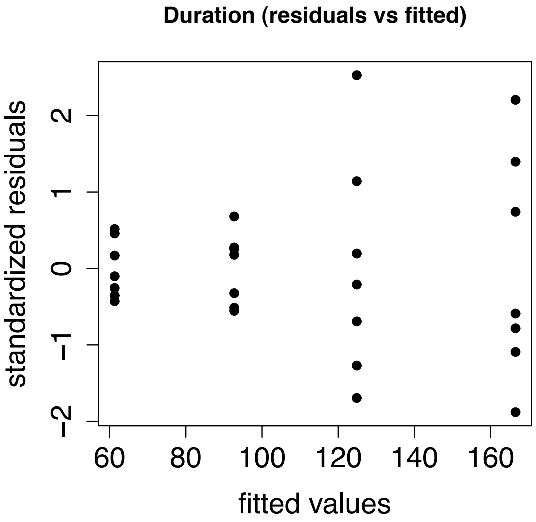


Fig. 1 Scatter plot of standardized residuals vs. fitted values.

Taking a further analysis, we found that the time duration to write the second program was less than the first one. In Fig. 1, the first and second vertical data points correspond to the second program (encoder). Fig. 2 shows the mean time duration to write both programs. To fulfill this assumption, in future experiments we will select programs with similar complexity.

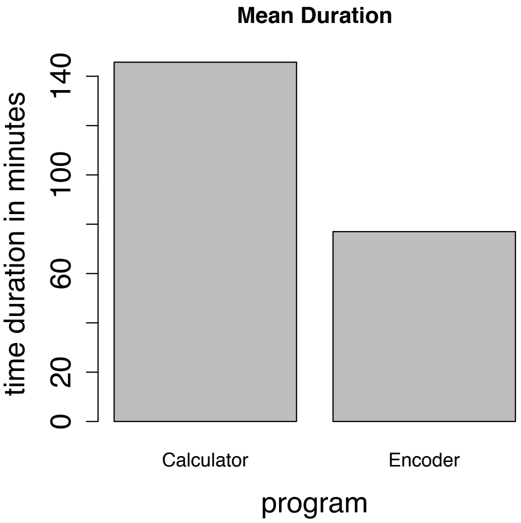


Fig. 2 Mean duration to write a program.

Continuing with the next assumption assessment, Fig. 3 shows a normal probability plot. If points (in this case standardized residuals) fall close to a straight line pattern then residuals are approximately normal. Points that are above the straight line pattern correspond to residuals that are bigger than we might expect for normal data. Points that are below the straight line pattern correspond to residuals that are smaller than we might expect for normal data. Applying the Kolmogorov-Smirnov test for normality [15, 26] we get a p-value of 0.8806; it means that we accept the null hypothesis in favor of normality.

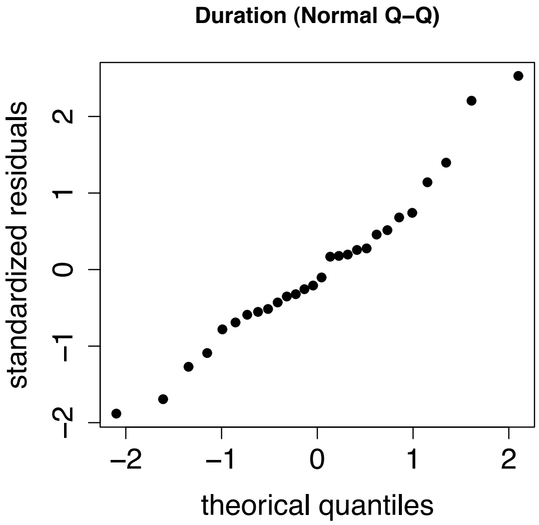


Fig. 3 Normal probability plot.

With respect to the assumptions assessment for effort, we get similar results to those we report regarding duration; performing the Levene test for homogeneity of variances [17] we get a p-value of 0.0241. Setting an alpha level of 0.05 this test is significant. It means that variances are not equal due to differences between programs duration. The Kolmogorov-Smirnov test for normality [15, 26] gives a p-value 0.8059. It means that we accept the null hypothesis in favor of normality.

Due to the experimental design used, another assumption that is worth to assess is the additivity. Experiment designs that implement blocking assume that there is no interaction between the treatment and the block. Under this situation it is told that treatment and block effects are additive [16]. We test this assumption by using the Tukey test for nonadditivity [27]. Table 5 shows the results of this test for the Latin square design used in the experiment.

Table 5: Nonadditivity test results

Measure	*Block*	*F-value*	*p-value*
Duration	Program	0.0084	0.9277
Duration	Tool support	1.0936	0.3061
Effort	Program	0.0899	0.7669
Effort	Tool support	0.9861	0.3306

Setting an alpha level of 0.1 (or less), p-values are not significant. It means that experiment results satisfy the assumption of additivity in lack of interaction between treatment and blocks.

4.2 Analysis of Variance (ANOVA)

Once model assumptions were assessed, we proceed to perform the ANOVA. Table 6 shows the ANOVA for the duration measure whereas Table 7 shows the ANOVA for effort.

Table 6: ANOVA for duration measure

Source	*Df*	*SS (Type I)*	*MS*	*F-value*	*p-value*
ProgramBlock	1	33,052	33,052		
ToolSupport Block	1	185	185		
Treatment	1	9,362	9,362	2.9843	0.0969
Residuals	24	75,293	3,137		

If we set an alpha level of 0.05 neither treatment (both ANOVA tests) are significant. However setting an alpha level of 0.1 which represents a confidence level of 90% we get significant differences in both treatments. For the first treatment (Table 6) we get a p-value = 0.0969 with respect to duration, whereas we get a p-value = 0.1017 for the second treatment (Table 7). Although this second p-value is slightly greater than 0.1, we also consider it significant.

Table 7: ANOVA for effort measure

Source	*Df*	*SS (Type I)*	*MS*	*F-value*	*p-value*
ProgramBlock	1	70,702	70,702		
ToolSupport Block	1	4,969	4,969		
Treatment	1	22,346	22,346	2.8953	0.1017
Residuals	24	185,232	7,718		

4.3 Treatment Comparisons

Taking this alpha level (α=0.1) into account, we perform a treatment comparison test (also referred as contrast test) for each measure. Table 8 shows the treatment means, standard error and replications for duration measure whereas Table 9 shows the same information for effort.

Table 8: Treatment means, standard error and replications for duration

Treatment	*Duration (minutes)*	*Std. err*	*Replication*
Solo	129.6428	17.8114	14
Pair	93.0714	16.7054	14

Table 9: Treatment means, standard error and replications for effort

Treatment	*Effort (minutes)*	*Std. err*	*Replication*
Solo	129.6428	17.8114	14
Pair	186.1429	33.4108	14

There are several tests for performing treatment comparisons. These tests help us to analyze pairs of means to assess possible differences between means. Using Scheffé test [21] for treatment comparisons, Table 10 shows the treatment comparison with respect to duration.

Table 10: Comparison with respect to duration

Comparison	*Difference*	*p-value*	*LCL (95%)*	*UCL (95%)*
Solo-Pair	36.5714	0.0969	6.1578	66.9850

As shown in Table 10, there is a significant difference (at α=0.1) of 36 minutes in favor of pair programming (28% decrease in time). At a confidence interval of 95% this difference ranges between 6 and 66 minutes (4% to 51% decrease in time).

Table 11 shows the treatment comparison with respect to effort. As we see, there is a significant difference (at α=0.1) of 56 minutes in favor of solo programming (30% decrease in effort). At a confidence interval of 95% this difference ranges between 8 and 104 minutes (4% to 55% decrease in effort).

Table 11: Comparison with respect to effort

Comparison	*Difference*	*p-value*	*LCL (95%)*	*UCL (95%)*
Pair-Solo	56.5	0.1017	8.7967	104.2032

4.4 Effect Size and Power Analysis

Effect size is a measure for quantifying the difference between two data groups. Usually, it is used to indicate the magnitude of a treatment effect. Using the function defined in equation (2) [5], we calculate Cohen's *d* coefficient [10]. This coefficient is used as an effect size estimate for the comparison between two means (in this case Solo and Pair programming). According to Cohen [10], a *d* value between 0.2 and 0.3 represents a small effect size, if it is around 0.5 it is a medium effect size, and an effect size bigger than 0.8 is a large one.

$$d = \sqrt{\frac{F\left(n_1 + n_2\right)}{n_1 n_2}} \qquad (2)$$

Using the F-value 2.9843 of the first ANOVA (Table 6) we get an effect size *d* of 0.6529 and an effect size *d* of 0.6431 for the F-value 2.8953 regarding second ANOVA (Table 7). According to Cohen's classification, both effect sizes are considered medium effects. The first effect size is against of solo programming (with respect to duration) whereas the second is against of pair programming (with respect to effort).

Once we have calculated effect sizes, we carry out a power analysis. The power of a statistical test is the probability of rejecting the null hypothesis when it is false. In other words, the power indicates how sensitive is a test to detect an effect in the treatment examined.

Power is equal to 1–β where β is the probability of committing a Type II error [10]. Power analysis can be conducted before or after the experiment is run. When it is performed before, a sample size is estimated with the aim of achieving an adequate power in the statistical test used in the experiment. On the other hand, when the experiment is run, power analysis is used to determine what the power was in the experiment test. We use this second approach to perform power analysis.

Once we know the effect size it is possible to compute the power of a test. In order to determine the power, we use the function *pwr.t.test()* of the R environment [9] which implements power analysis as outlined by Cohen [10]. Given an effect size of 0.6529 (related to duration) and a sample size of n=14 (number of measures in each group; pair and solo programming), and setting a significance level α=0.1; we get a power of 0.51 (51%). Similarly, a power of 0.5 (50%) was obtained with the same sample size and significance level, but replacing the effect size for the value 0.6431 (related to effort).

5. Experiment Limitations

Experiments are subject to concerns regarding validity. In this section we discuss experiment limitations based on the four categories of threats to validity described in [11]. Each category has several threats that can negatively impact on the experiment results. We list, both, threats that can impact on this experiment and suggestions for improvements in future versions of this experiment.

5.1 Threats to the Conclusion Validity

These threats concern with issues that affect the ability to draw a correct conclusion about the existence of a relationship between the treatment and the outcome. Next, we describe threats in this category that may have affected our experiment.

Although the experiment results show a moderate power of 50%, results may have been affected by low statistical power. With the aiming of increase the power at 80%, we will perform a power analysis to estimate the sample needed before we conduct replications of this experiment.

Regarding to assumptions of statistical tests, although experiment results satisfy the principle of independence and normality, results may have been affected by lack of variance homogeneity. We have identified the program as a source of variation. With the aiming of reduce variance heterogeneity, in future replications we will use programs with similar complexity.

Another threat that might have affected conclusion validity is with respect to reliability of measures. Although all measures were collected during second session, some measures regarding solo programmers were not collected during first session; it was due to time constraint. In this session subjects that did not finish on time were told to record the time at home. To avoid this threat in future replications we will be careful with managing the time of sessions.

5.2 Threats to Internal Validity

These threats concern whether the observed outcomes were due to other factors and not due to the treatment. To avoid these threats, subjects were randomly assigned to the treatments. Latin square design eliminated possible problems with learning effects, boredom or fatigue as the subjects tried different program and tool support. Subjects (pairs and solos) were in the same classroom with equal working conditions, and sitting apart with no interaction.

A possible threat that might have exposed this validity is that subjects knew the experiment, so a competition between pairs and solos could have happened.

5.3 Threats to Construct Validity

Construct validity threats concern the relationship between theory and observation. An issue in our experiment that might have affected this validity is that subjects had little or no previous experience with pair programming and they had not programmed with their partners before. These experiment results might be conservative with respect to the effect of pair programming. In subsequent experiment replications, we will reinforce this validity by assigning training programs to pairs.

5.4 Threats to External Validity

These threats concern with issues that may limit our ability to generalize the results of the experiment to other contexts, for example generalize it to industry practice. The use of students as subjects instead of practitioners might have exposed this validity. However, as pointed in [8] the use of students as subjects enable us to obtain preliminary evidence to confirm or refute hypotheses that can be tested later in industrial settings.

6. Discussion

In this section we discuss some results of other experiments and we contrast them with our results regarding duration and effort.

6.1 Duration

The experiment run by Nosek [24] employed 15 practitioners grouped in 5 pairs and 5 solos. Subjects wrote a database script. Results show a decrease of 29% in time duration in favor of pair programming.

Williams et al. [28] used 41 students grouped in 14 pairs and 13 solos. During the experiment, subjects completed four assignments. Authors reported that pairs completed the assignments 40 to 50 percentage faster.

Nawrocki and Wojciechowski [23] employed 16 student subjects (5 pairs and 6 solos). Subjects wrote four programs. Authors did not find differences between pairs and solos.

Lui and Chan [19] used 15 practitioners grouped in 5 pairs and 5 solos. Authors reported 52% decrease in time in favor of pair programming.

Müller [22] used 38 students (14 pairs and 13 solos). Students worked on four programming assignments where tasks were decomposed into implementation, quality assurance and the whole task. Author reported that pairs spent 7% more time working on the whole task, however this difference is not significant.

Arisholm et al. [1] used 295 practitioners grouped in 98 pairs and 99 solos. Subjects performed several change tasks on two alternative systems with different degrees of complexity. Authors reported 8% decrease in favor of pairs.

In contrast, the results reported in this paper infer a significant (at α=0.1) 28% decrease in time (in favor of pairs) and an effect size d=0.65. With respect to duration, our results reinforce those reported in [24].

6.2 Effort

This measure is not present in all of the experiments previously discussed, so we compute it (doubling the time duration of pairs) only in the cases where data is available.

According to Nosek data [24] we observe a decrease in effort of 29% in favor of solo programming. Conversely, data of Lui and Chan [19] indicate a decrease of 4% in favor of pairs. Finally, Arisholm et al. [1] Report an increase in effort of 84% (against of pairs).

In contrast, the results reported in this paper infer a significant (at α=0.1) 30% decrease in effort (in favor of solos), and an effect size d=0.64. Our results, again, reinforce those calculated in [24].

7. Conclusions and Further Work

This paper presented a controlled experiment that was run as part of a university course in DOE. The aim of the experiment was to evaluate pair versus solo programming with respect to duration and effort. Subjects who jointly wrote the program assignments took less time (28%) than subjects who worked individually. Conversely subjects grouped in pairs spent more effort (30%) than those who worked individually. These results are very close to those reported in [24].

With the aiming of striving towards better research practices in SE [18] we reported all the collected measures. This data will help other researchers to verify or re-analyze [14] the experiment results presented in this work. This data can also be used to accumulate and consolidate a body of knowledge about pair programing.

We are planning to conduct future replications of this experiment to get more insight about the effect of pair programming. Although we did not observe interactions between treatment and blocks, we plan to use another experimental design to assess possible interactions.

References

[1] E. Arisholm, H. Gallis, T. Dybå, and D. I. Sjøberg. Evaluating pair programming with respect to system complexity and programmer expertise. IEEE Transactions on Software Engineering, 33(2):65–86, 2007.

[2] V. Basili, G. Caldiera, and H. Rombach. Goal question metric paradigm. Encyclopedia of Software Eng, pages 528–532, 1994. John Wiley & Sons.

[3] K. Beck. Embracing change with extreme programming. Computer, 32(10):70–77, 1999.

[4] K. Beck. Extreme programming explained: embrace change . Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2000.

[5] M. Borenstein. The handbook of research synthesis and meta analysis. Chapter: Effect sizes for continuous data, pages 279–293. Russell Sage Foundation, New York, USA, 2009.

[6] G. E. P. Box, W. G. Hunter, J. S. Hunter, and W. G. Hunter. Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. John Wiley & Sons, June 1978.

[7] G. Canfora, A. Cimitile, F. Garcia, M. Piattini, and C. A. Visaggio. Evaluating performances of pair designing in industry. Journal of Systems and Software, 80(8):1317 – 1327, 2007.

[8] J. Carver, L. Jaccheri, S. Morasca, and F. Shull. Issues in using students in empirical studies in software engineering education. In METRICS '03: Proceedings of the 9th International Symposium on Software Metrics, page 239, Washington, DC, USA, 2003. IEEE Computer Society.

[9] S. Champely. pwr: Basic functions for power analysis , 2012. R package version 1.1.1.

[10] J. Cohen. Statistical power analysis for the behavioral sciences . L. Erlbaum Associates, Hillsdale, NJ, 1988.

[11] T. Cook and D. Campbell. The design and conduct of quasi-experiments and true experiments in field settings. Rand McNally, Chicago, 1976.

[12] H. L. Dreyfus and S. Dreyfus. Mind over Machine. The Power of Human Intuition and Expertise in the Era of the Computer . Basil Blackwell, New York, 1986.

[13] R. A. Fisher. The Design of Experiments. Oliver & Boyd, Edimburgh, 1935.

[14] O. S. Gómez, N. Juristo, and S. Vegas. Replication, reproduction and re-analysis: Three ways for verifying experimental findings. In International Workshop on Replication in Empirical Software Engineering Research (RESER'2010) , Cape Town, South Africa, May 2010.

[15] A. N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. Giornale dell'Istituto Italiano degli Attuari, 4:83–91, 1933.

[16] R. Kuehl. Design of Experiments: Statistical Principles of Research Design and Analysis. Duxbury Thomson Learning, California, USA. second ed. edition, 2000.

[17] H. Levene. Robust tests for equality of variances. In I. Olkin, editor, Contributions to probability and statistics . Stanford Univ. Press. Palo Alto, CA, 1960.

[18] P. Louridas and G. Gousios. A note on rigour and replicability. SIGSOFT Softw. Eng. Notes, 37(5):1–4, Sept. 2012.

[19] K. M. Lui and K. C. C. Chan. When does a pair outperform two individuals? In Proceedings of the 4th international conference on Extreme programming and agile processes in software engineering, XP'03, pages 225–233, Berlin, Heidelberg, 2003. Springer-Verlag.

[20] K. M. Lui, K. C. C. Chan, and J. Nosek. The effect of pairs in program design tasks. IEEE Trans. Softw. Eng., 34(2):197–211, Mar. 2008.

[21] C. McDowell, L. Werner, H. E. Bullock, and J. Fernald. The impact of pair programming on student performance, perception and persistence. In Proceedings of the 25th International Conference on Software Engineering , ICSE '03, pages 602–607, Washington, DC, USA, 2003. IEEE Computer Society.

[22] M. M. Müller. Two controlled experiments concerning the comparison of pair programming to peer review. Journal of Systems and Software, 78(2):166 – 179, 2005.

[23] J. Nawrocki and A. Wojciechowski. Experimental evaluation of pair programming. In Proceedings of the 12th European Software Control and Metrics Conference, pages 269–276, London, April 2001.

[24] J. T. Nosek. The case for collaborative programming. Commun. ACM , 41(3):105–108, Mar. 1998.

[25] H. Scheffé. A method for judging all contrasts in the analysis of variance. Biometrika , 40(1/2):87–104, 1953.

[26] N. V. Smirnov. Table for estimating the goodness of fit of empirical distributions. Ann. Math. Stat., 19:279–281, 1948.

[27] J. W. Tukey. One degree of freedom for non-additivity. Biometrics, 5(3):pp. 232–242, 1949.

[28] L. Williams, R. Kessler, W. Cunningham, and R. Jeffries. Strengthening the case for pair programming. Software, IEEE, 17(4):19 –25, jul/aug 2000.

Omar S. Gómez received a BS degree in Computing from the University of Guadalajara (UdG), and a MS degree in Software Engineering from the Center for Mathematical Research (CIMAT), both in Mexico. Recently, he received a PhD degree in Software and Systems from the Technical University of Madrid (UPM). Currently he is a full time professor of Software Engineering at Mathematics Faculty of the Autonomous University of Yucatan (UADY). His main research interests include: Experimentation in software engineering, software process improvement and software architectures.

José L. Batún received a BS degree in Mathematics from the Autonomous University of Yucatan (UADY). He received a MS degree and a PhD degree in Probability and Statistics, both, from the Center for Mathematical Research (CIMAT) in Guanajuato, Mexico. He is currently full time professor of Statistics at Mathematics Faculty of the Autonomous University of Yucatan (UADY). His research interests include: Multivariate statistical models, copulas, survival analysis, time series and their applications.

Raúl A. Aguilar was born in Telchac Pueblo, Mexico, in 1971. He received the BS degree in Computer Science from the Autonomous University of Yucatan (UADY) and a PhD degree (PhD European mention) at the Technical University of Madrid (UPM), Spain. Currently he is full time professor of software engineering at Mathematics Faculty of the Autonomous University of Yucatan (UADY). His main research interests include: Software engineering and computer science applied to education.

Photovoltaic potential assessment on direct and diffusive radiation conditions in Albania

Anuela Prifti

Immovable Property Registration Central Office

Department of Cartography & First Registration

"Jordan Misja", Tirana, Albania

Abstract

By data extracted from survey stations installed throughout the Albanian territory and time series for 30 years, which are dispersed throughout the country, is clearly expressed solar energy property all year, despite growing differences between daily, monthly and annual average values of the energy. In setting survey points is generally observed such a spatial diffusion, which creates an opportunity for a more realistic assessment of solar radiation. In these conditions will be treated the annual and monthly average values of direct and distribute radiation in the fixed plan and in real weather. Generally, our country has a tremendous potential solar radiation, because in one year are recorded over 280 days of sunshine, which provides more than 3700Wh/m^2 at a 40º angle from the south, according to data at the point of observation of Tirana.

Keywords: *Station, Energy, Diffusion, Radiation, Observation*

Introduction

Data extracted from survey stations throughout Albania territory are used to see the difference between daily, monthly and yearly average values of solar energy: Calculation of annual and monthly average values. There's made a comparison with direct and distribute radiation in fixed plan and a real weather. Also is made a comparison with direct and distribute radiation in 2-axis plan and a real weather.

1. The Average of annual solar radiation in two axes plan

With very particular importance are presented average values of solar radiation in mobile plan with two axes, which are generally 23% higher than the average in fixed plan, regardless of Western Lowland they build up to 25% of this value. In these circumstances our country takes the average 4800W/m² every year from solar radiation, characterized both by an apparent local distribution. Larger quantities of radiation takes the annual average in Western Lowlands (Shkodra 5287W/m², Lezha 5123W/m², Durres 5388W/m², Lushnja 5270W/m², Fier 5329W/m², Vlora 5228W/m²). Then comes the part of the SE province of Central Highlands (Korça 5245W/m², Bilisht 5114W/m², Erseka 4984W/m², Pogradec 4935W/m². Lesser amount of the country takes part in Kukes VL 4631W/m² and B. Curri with 4491W/m². In general, such a distribution of average solar radiation, appears to seasonal values, which shows up the summer season with 1762W/m², representing 35% of the annual amount, then 27% in spring, autumn and winter 24% to 14% of this amount. So monthly changes are presented with different values, in particular the relatively small values of the transition between seasons (3%) from fall to spring, rise to 10% between autumn and winter, spring and summer have a change of 8 % for the account of the summer. These changes are closely related two predominant types of weather in our country, which determine both the number of clear days and those cloudy Difference between summer and winter reaches on average 38 % of annual amount for the account of the summer. Another phenomenon of the annual average radiation in a mobile plan with two axes is the local distribution of its values, so large amounts of this radiation takes the Western Lowlands, as Shkodra 1847W/m², Lezha 1831W/m², Durres 1885W/m², Lushnja 1856W/m², Fier 1874W/m², Vlora 1872W/m², Kuçova 1807W/m², Tirana 1829W/m². It is followed by the SE part of the Central Highlands province, which takes over 1700W/m², such as Erseka 1745W/m², Korça 1778W/m², Bilishti 1781W/m², Pogradec 1733W/m², while the southern hill and NE part of the North take on the 1600W/m², as Gjirokastra 1687W/m², Saranda 1637W/m², Berat 1681W/m², Permeti 1744W/m², Kukes 1616W/m² and B. Curri 1507W/m². The same zonality appears in the distribution of values of solar radiation in three other seasons, so larger quantities of distributet radiation are the regions mentioned above. There's presented a particularly important quantity of monthly solar radiation and differences between them, which are conditioned mainly by weather features and relief. Throughout our country, august takes the largest amount of solar radiation with an average value of 580W/m², representing 12% of annual amount and 32% of the summer season. Larger quantities of this radiation, meet in the Western Lowland Shkodra as 630W/m^2, Lezha 622W/m^2, Tirana 619W/m^2, Lushnje 626W/m^2, Fier 633W/m, Durres 638W/m^2. Relatively small values meet in the Central Highlands province and NE part of Northern province, such as Korca 604W/m^2, Bilisht 601W/m^2, Erseka 586W/m^2, Pogradec 585W/m^2, Permeti 591W/m^2, Gjirokastra 576W/m^2, Saranda 564W/m^2, Kukes 520W/m^2, B.Curri 550W/m^2, Peshkopia 586W/m^2. Relatively significant quantities of radiation, also are presented in August during the western part of Central Highlands province of Burrel Basin 622W/m^2, Elbasan area 611W/m^2, Krujë 585W/m^2 etc. Smaller quantities of solar radiation, meet in December with an average value of about 170W/m^2, representing 3.5% of annual amount. Greater amount of radiation during December takes Shkodra 204W/m^2, Durres 203W/m^2, Fier 195W/m^2, Lushnja 190W/m^2, Tirana 189W/m^2, Vlora 194W/m^2, Elbasan 195W/m^2 etc. Is worth noted that larger values belong to the months from October to March, which reach 34% between October and November. The frame time of these changes relates to the complete dominance of a cyclonic weather and the start of that uncyclonic, and its full impact is reflected to much smaller intermonth changes during the period from April to October with 1-15%, especially between the months of summer. In very small size of these changes characterize the average annual amount of radiation dispersed on a mobile plan in two axes, which is 8.5% greater than in a fixed plan, reaching average 134W/m^2. Biggest dispersed radiation changes are those between the period from October to March and the April-September, representing respectively 34% and 66% of the total amount of this radiation. With the changes expressed in dispersed radiation is presented its seasonal distribution, particularly between winter and summer, which directly conditioned by the two individual types of weather. Quite characteristic also shows the change of dispersed radiation values between spring, autumn and winter, which reach respectively 30%, 22% and 14% of the annual amount of this radiation, having a difference 8%, so twice the difference between winter and summer. Another phenomenon of this radiation is relatively small seasonal change of seasonal values between all points of observation, especially the winter, during which the difference between the smallest (Berat and Kukes 212W/m^2) and the largest (Vlora 252W/m^2) amounts to 40W/m^2, while the season of summer, autumn and spring respectively reach 58W/m^2, 63W/m^2 and 55W/m^2 between these points. Regarding the regional distribution of seasonal values of the radiation, we note that the largest amount has the Mountainous Region. In summer arrives Vlora 576W/m^2, Fier 567W/m^2, Lushnje 557W/m^2, in spring arrives South 511W/m^2, Fier 506W/m^2, Lushnje 500W/m^2, autumn in South 372W/m^2, Lushnje 365W/m^2, Fier 368W/m, Durres 360W/m^2. During the winter season arrives in Vlora 252W/m^2, Durres 246W/m^2, Lushnja 246W/m^2 etc. Approximate values appear also part of the SE Hill Central, such as Korca 240W/m^2, Bilisht 242W/m^2, Erseka 237W/m^2. But to the northern part of this region differes Burrel 234W/m^2, B.Curri 244W/m^2, Peshkopia 229W/m^2 etc. Monthly performance of the reflected radiation, looks generally similar to yearly seasonal changes observed, especially between two periods of the year. In the reflected radiation, the most value for all points of observation, is June, which reaches an average of 190W/m^2, representing 12% of the annual amount. Greater amount of this radiation has Mountainous (Vlora 200W/m^2, Lushnja 194W/m^2, Fier 197W/m^2, Durres 190W/m^2. Then comes the part of the hill Central JL (Korca 195W/m^2, Erseka 192W/m^2, Bilisht 195W/m^2) and the Northern part (B.Curri 195W/m^2, Kukes 193W/m^2, Burreli191W/m^2, Peshkopia 189W/m^2), etc. All data points of survey distinguish small changes this radiation. In contrast to the largest amount of reflected radiation, the minimum as regarded in month of December is with an average value of 60W/m^2, representing 3.7% of annual amount of this radiation and 32% of June, so the changes highlighted to June. Larger values are met in December, the Western Lowland, Vlora and Fier with 67W/m^2, Shkodra 64W/m^2, Tirana 63W/m^2, Lushnje 65W/m2, Durres 66W/m2, comparing those on the lower northern province of NE of Central part, Kruja 54W/m^2, Pogradec 56W/m^2, Kukes 53W/m^2. Regarding the differences between the amount of radiation in December with November and January, in contrast with those of May and July, we emphasize that they appear much larger, having a value of respectively 28% and 22%

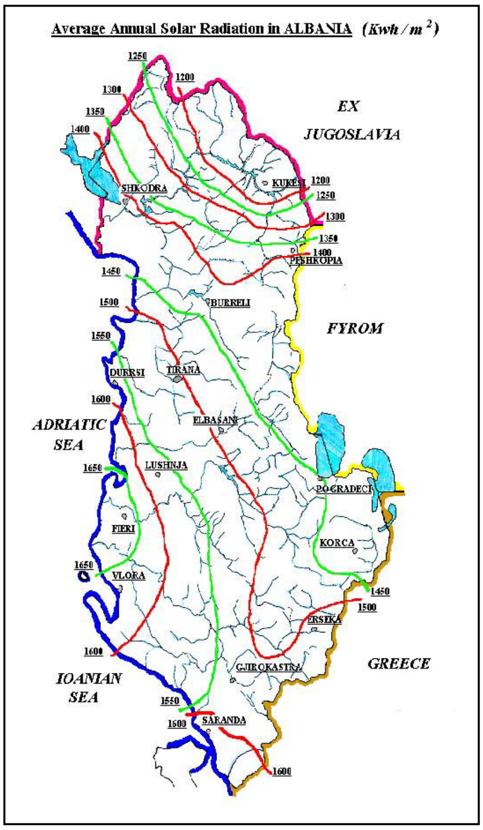


Fig. 1 Average annual solar radiation in Albania

smaller to them. Changes in these reflected proportions radiation, related to the fact that the action culminates in November cyclonic weather, so both the greatest number of days and reduction cloudy day. While in January begins the fall weather and the impact of this starts at the same time and duration of the day. In the context of photovoltaic energy, the largest amount of this energy constitutes average annual radiation in the blue sky for a mobile plan in two axes, which takes on average 28% - 31% more than he plans to rest in the same conditions of weather. Another phenomenon of this radiation is the large amount all year around, which is reflected in the approximate ratio between the above two periods of the year, so that from October to March the weather conditions and it cyclonic April to September with prevailing weather uncyclonic, which represent respectively 42% and 58% of the annual amount the general radiation. Besides the annual distribution, another indicator of the wealth of this radiation, are relatively small seasonal changes (1-5%) in comparison with the annual amount. Obviously, the zonal distribution of this radiation also observed significant changes which distinguish between the Western Lowlands, as Vlora 688W/m², Lushnja 676W/m², Fier 685W/m^2, Durres 675W/m², Shkodra 659W/m², Kuçova 651W/m², Tirana 656W/m². Immediately after it, comes the southern part of Central SE province Bilisht with 670W/m², Kora 659W/m², Erseka 649W/m², Pogradec 635W/m², northern Kukes 651W/m² , Burrel 663W/m², Peshkopia 651W/m². Maximum value of the annual change reaches between Vlora and Berat (581W/m² to 107W/m², representing 17% of the average monthly amount. Almost the same distribution of this radiation represents its annual amount, which stands Vlora with 8258W/m², Lushnja 8107W/m², Durres 8098W/m², while the difference between the smallest amount in Berat 6971W/m^2 with the largest amounts in Vlora 1287W/m², represent 17% of the annual average. In another part of distribution of this radiation, there is a direct link with two characteristic types of weather, such as the period from October to March and that April to September, where it gains 28% of radiation more. But in relation to the annual amount they represent respectively 43% and 57% of this amount reflecting primarily the direct impact of a weather conditions and uncyclonic conditiones. During the first period there's an average 3432W/m² of irradiation in clear sky on a mobile plan in two axes, and in the second period 4386W/m². Impact area is expressed in greater amount of radiation in the Western Lowland, as in Fier 4587W/m², Lushnje 4517W/m², Durres 4549W/m², South 4607W/m² etc. Then rankes southern part of the SE of hill Central as Erseka 4401W/m², Bilisht 4491W/m², Pogradec 4321W/m², Korça 4360W/m², etc., while in the NE of this region is distinguished Burrel with 4505W/m², Peshkopia 4447W/m², Kruja 4244W/m², Kukes 4650W/m², B.Curri 4183W/m^2 etc. Almost the same values also characterizes the zonal distribution of this radiation during the period from October to March, despite from the shortest amount of this radiation (43% of annual amount), then about 14% less than that from April to September, due to weather cyclonic dominance of this period. An approximate distribution of radiation represents seasonal average value summer 2235W/m², spring 2176W/m², fall 1805W/m² and winter 1539W/m², which distinguished the utmost two seasons, summer and winter so very different amount of this radiation. This phenomenon is highlighted by the fact that they constitute about 29% and 20% of the annual amount, whereas the ratio between them, winter represents 69% of the total summer radiation, so 31% less than this season. With less pronounced changes appear transitional seasons of spring and autumn, which constitute the averaged respectively 28% and 23% of the annual amount, whereas the ratio between them in this change is noted that the spring gets 5% more than autumn, whereas the latter 3% more than in winter. In these conditions there is a interseasoned reduction relatively small, especially between summer and spring with an average value of 1% more to the front, thanks to the influence of weather generally uncyclonic in these seasons. In general, seasonal changes of the radiation performance, as they have an annual regional character expressed in which separated Western Lowlands, where larger values of the radiation observed in Vlora 2351W/m², Lushnje 2301W/m², Fier 2337W/m², Durres 2319W/m², Kuçova 2244W/m², Lezha 2257W/m², Shkodra 2274W/m² etc. With relatively few changes appear SE and southern part of the Central Highlands province as Korçë 2283W/m², Heartland 2289W/m², Erseka 2124W/m², while in the northern part of NE of this region separated Kukes 2389W/m², Peshkopia 2266W/m², B.Curri 2129W/m², Burrel 2300W/m², Kruja 2163W/m² etc. Small changes between summer and spring worth on account of 2.3% in summer, characteristic of Shkodra, Permeti, Vlora, Tirana, Lushnjen, Fier, Durres and Saranda, so Western Lowland. But in all other points of observation, this difference amounts to 4-6%, influenced mainly by the conditions of relief, especially the extent and direction of slopes. Larger values of the difference between spring and autumn meet in Southern Highlands province, like in Permet 23%, 18% Gjirokastra Kruja 20%, 15% Berat, Saranda, 15% for the account of the former, then ranks the Central Highlands province of Kukes 19% B.Curri 17%, Peshkopia 16%, Burrel, 20% etc. But autumn takes a greater radiation then winter in the same zonal order, so by 17% in Gjirokastra, Saranda, Permet of up to 14-17% of survey points Peshkopi, Burrel, Kukes, Korce, Erseka, Bilisht etc. It is worth mentioning that the most significant changes to the amount of radiation between winter and summer meet almost all over the country and expressed the great values at all observation points, which amounts to 28-42% for the account of the summer, while average value amounts to 31%. Monthly performance of this radiation prevails December immediately with smaller quantity, which takes an average of 462W/m², representing 6% of annual amount and 10% of June with the largest amount of radiation, and in relation to energy in a fixed plan is 26% more than the same month. Naturally, smaller values of radiation during this month as noted above, is directly related to the peak of cyclonic action conditions, which entails both the highest number of days cloudy (average 13 days). The difference between the minimum value in the December survey points Kruja 382W/m^2, Kukes 390W/m^2 and Lezha 378W/m^2 with Vlora 517W/m^2, Lushnja 507W/m^2, Fier 521W/m^2 and Durres 505W/m^2 averaging 395W/m^2, the character code indicating the extent of these values is treated as above. With no significant changes, appear in June values, which is the month with the largest amount of radiation across the country, being directly related to uncyclonic weather conditions. In these circumstances our country, during June, takes an average of 768W/m^2, which represents 10% of annual amount and 40% more than the month of December. These values indicate substantial amounts of radiation during this month thanks to the large number of brighter days and relatively small values of average relative air humidity (64%), ranking immediately after the July and August. In general, the amount of radiation changes of June, as noted above, not great value between points of observation, however, noted that the character of their area code is visible. One of the characteristic features of June is the highest value of solar energy in comparison with July and August, which amounts respectively 8% and 2% indicating the crucial role of an uncyclonic weather condition values this month. With special features also appears more pronounced change in the amount of solar radiation between November to December and February to March, which reaches respectively 10% and 18% different to that with a fixed plan, where the values are small (report 12% and 15%). From the above data clearly that the largest amount of photovoltaic power in our country achieved in terms of an uncyclonic clear weather for a mobile plan in two axes, which takes on average each year 27% more than for radiation a plan to rest in the same weather conditions. In the context of photovoltaic property, with particular importance distribution also shows the average daily total radiation, which serves both for a more rational use of its economy. Daily performance of this radiation is almost the same annual and monthly, despite some quantitative differences in different months of the year. This means that the months with the largest amount of general radiation have days during April-September, while those with smaller amount from October to March, distinguishing two year periods in direct relation to two types of weather. During the period from October to March our country takes on average 78 W/m^2 per day, whereas in the April-September 184 W/m^2 per day, representing respectively 15% and 12% of the annual average amount of this energy. Greater amount of radiation with average daily value of 100 KWh/m^2 per day was October and March, 28 KWh/m^2 during the first period, which are partly under the influence of an uncyclonic weather condition. But in the second, obviously the larger quantity has June and July with average values respectively 214 and 216 KWh/m^2, which is conditioned by the presence of peak action, that the prevalence of uncyclonic weather in these months. Approximate value of these months occures between May and August, respectively, 194 KWh/m^2 and 180 KWh/m^2 per day, representing June and July with 28% of the annual amount average of this radiation. On the distribution of average daily radiation values are also observed notable character codes, where larger quantities of this radiation takes Mountainous Shkodra as 130 KWh/m^2, Lezha 133 KWh/m^2, Tirana 134 KWh/m^2, Durres 137 KWh/m^2, Lushnja 136 KWh/m^2, Kucova 137 KWh/m^2, Fier 143 KWh/m^2, Vlora 141 KWh/m^2 etc. Approximate values to Southern Mountain Region is also presented with Gjirokastra 134 KWh/m^2, Saranda 138 KWh/m^2, Xara 140 KWh/m^2, followed by the southern part of the SE of the Central Mountain Region, as Erseka 134 KWh/m^2, 132 Pogradec KWh/m^2, Korça 131 KWh/m^2, Sheqerasi 133 KWh/m^2 northern part of the province such as Burreli 130 KWh/m^2, Peshkopia 127 KWh/m^2, Kukes 121 KWh/m^2 etc. The largest average daily value changes of solar radiation meet between March and April with a quantity 51 KWh/m^2, then it comes between August and September with 42 KWh/m^2, so the transitional seasons where the transition between two types of weather. Particularly important for evaluating the average daily solar radiation in the plan of fixed and movable in two axes, also represents the amount of this radiation with the increase from hour to hour for months reduction of radiation characteristic of each season, so January, April, July and October, adding both his annual values. From the data of stations of observation, highlighted immediately change expressed between hours growth (10^{30} - 12^{oo}) and a decrease to (15^{oo} - 17^{oo}) of this radiation, which is reflected in the fact that growth has different values between the radiation general and distributed. Average daily amount of total radiation on a fixed plan in normal weather conditions for growing hours of this radiation is 48% greater than that of hours decreased, having a fair value 605 KWh/m^2 and directly reflect the crucial role of uncyclonic and cyclonic weather to it. At smaller values appears simultaneously dispersed radiation daily average, which for hours with his growth is 24% more than during the hours with a drop of this radiation. With characteristic features appear simultaneously daily average values of radiation to a fixed plane in the cloudless sky in the early hours of increases radiation, which is 53% more than those in classes with his fall, saying the real value respectively 930 KWh/m^2 and 437 KWh/m^2, then twice of it. It is worth mentioning that the average daily values of radiation during his hours is increased in both plans, with significant changes occured between them. In these hours of increased radiation is observed the total of daily average radiation in distributed and portable plan. They are respectively 9% and 8% more than a static plan, the real value, so 194 KWh/m^2 and 1018 KWh/m^2 in terms of a clear weather. Daily average values of this radiation are known for major changes in both plans and simultaneously within the same plan, so that fixed and mobile. So the average daily amount of daily radiation in really weather conditions amounts to 318 KWh/m^2, while during a clear weather 437 KWh/m^2, while the hours with these increase values respectively reach 605 KWh/m^2 and 930 KWh/m^2, while distributed 137 KWh/m^2. Also are observed changes in average daily amount of radiation in mobile plan and conditions of clear weather, which reach respectively 486 KWh/m^2 and 742 KWh/m^2, comprising 23% and 17% less than those with increasing radiation, whereas the amount of radiation dispersed in the same conditions amounts to 161 KWh/m^2. From the above treatment is clarified the fact that the most pronounced changes of the average daily values of solar radiation are those between the increase and decrease hours with the amount of this radiation, then comes the differences between a fixed plan and a mobile one to the account of the last. In these conditions the decisive role in the amount of photovoltaic system, plays the hours with maximum values of solar radiation, which are closely related and directly with uncyclonic weather conditions, stating quite well in most daily amount for July and April, at the peak of action of this weather. Obviously, the character code of the distribution of average daily quantity of solar radiation, as it's next month's annual season, highlighted by the fact that the Western Lowlands is characterized by larger values of this radiation as Fier 632 KWh/m^2, Lezha, Tirana and Shkodra 612 KWh/m^2, Durres 614 KWh/m^2, Lushnja and Vlora 608 KWh/m^2. At relatively small values appear SE of southern part of the Central Mountain Region as Bilisht, Korça and Erseka with 604 KWh/m^2, Pogradec 599 KWh/m^2.

In northern continuation ranks this with Kruja 613 KWh/m^2, Burrel 608 KWh/m^2, Peshkopia 594 KWh/m^2 and finally as the Southern Mountain Region Gjirokastra 574 KWh/m^2, Saranda 572 KWh/m^2, Permeti 585 KWh/m^2 etc. A particularly important trend presents average daily values during the transitional seasons of spring and fall for a fixed and mobile plan, which lie exactly on the borders of two weather types mentioned above. Data from the surveys, point that average daily radiation amounts are 13% -16% higher during the spring season than to fall for both plans. The most pronounced differences between these two seasons are observed in particular to deliver average daily radiation, which is 25% greater in spring to a fixed plan and 27% in a mobile plan to autumn, indicating the real values respectively 143 KWh/m^2 and 161 KWh/m^2, while in autumn 108 KWh/m^2 and 117 KWh/m^2. Naturally, larger values of average daily radiation are characteristic for the sobering days in a fixed plan and in particular that in mobile with two axes, which reach respectively 519 KWh/m^2 and 716 KWh/m^2, for the spring season, and 451 KWh/m^2 and 604 KWh/m^2 in autumn. In daily average radiation, a significance present also the values of the intensity of direct and spread radiation around true solar midday (TSM), which account for hours 9^{3o} -15^{3o}, culminating at 12^{3o}. Generally observed that the intensity values of these two components for 9^{3o} and 15^{3o} are the same. However it must be said that the changes more pronounced intensity average annual meet for radiation directly between the hours of 9^{3o} and 15^{3o} with 12^{3o}, which reaches respectively 70 W/m^2 at a minimum in December to 521 W/m^2 to a maximum of July, so seven times greater. But the intensity of dispersed radiation varies day to 72 W/m^2 in December to 366 W/m^2 to month in May, having a value five times greater between the aforementioned hours. On the distribution of intensity day directly observed the changes more pronounced between the hours of 9^{3o} and 15^{3o} to noon true solar (12^{oo} -12^{3o}) observed in the days of the months November, December and January with values respectively 27 W/m^2, 9 W/m^2 and 13 W/m^2, comprising an amount of radiation respectively 5, 8.7 and 9.7 times smaller than the TSM. But small changes partein days from March to October 1.5 -2.2 times smaller than to TSM, which correlate with the prevalence of an uncyclonic weather. In such conditions of weather, the largest amount of daily radiation intensity directly TSM, meet during the month of July with 522 W/m^2, then comes June with 503 W/m^2, while smaller values belong to the December and January, respectively 150 W/m^2 and 196 W/m^2, so 2.6 times smaller. Very small in size changes occur daily average intensity of dispersed radiation between 9^{3o} and 15^{3o} , which have a quantity 1.3-2.6 times smaller TSM. Obviously the smallest amount of this radiation was December and February respectively, values 72 W/m^2 and 83 W/m^2, so 2.6-2.3 times less than the TSM, and the largest amount for the days of April 238 W/m^2, and May and June with 252 W/m^2 or 1.4 times smaller. With significant changes occur the days TSM values of different months of the year, where larger quantities of this radiation meet during April to September totaling over 300 W/m^2.

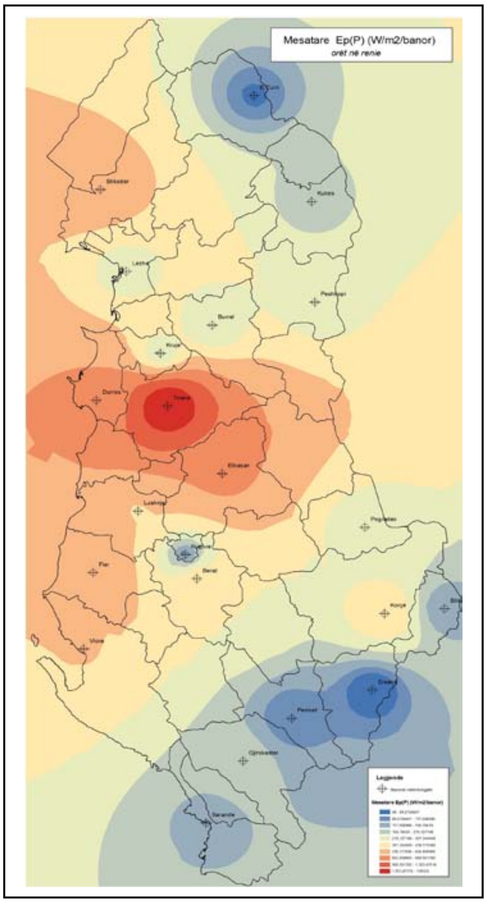


Fig. 2 of solar energy per capita, in decreasing hours of solar radiation (period 2001-2010) for 22 stationes.

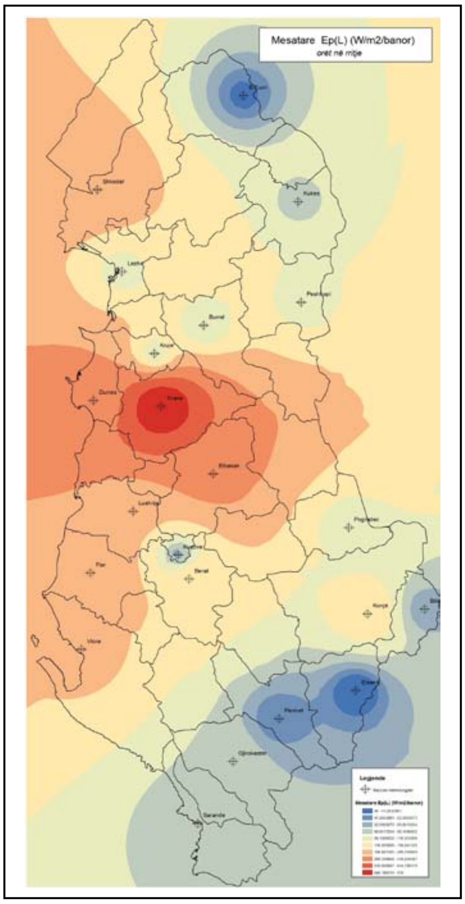


Fig. 3 Annual amount average of solar energy per capita, in increasing hours of solar radiation (period 2001-2010) for 22 stationes.

While in the months from October to March meet quantities less of this radiation, particularly from November to February, which have respectively the values 97 W/m², 72W/m² and 83 W/m^2 during the hours 9³º and 15³º, representing 2-2.6 times less than the TSM. From the above data is shown that the highest values of intensity of the components of solar wave radiation, actually achieved around solar noon (TSM). The highest amount of dispersed radiation intensity is achieved during the month of May with 372 W/m², that of direct radiation on horizontal surface meets in July 523 W/m², while smaller values for the two components during the year are in December, so 190 W/m² and 70 W/m², which means 2.7 and 5.3 times smaller than those of TSM. In the context of solar energy and assessing its potential, also appears important long-term distribution and monthly amounts of photosynthetic active radiation (PHAR), especially for Western Lowland (Fig.1). Photovoltaic active radiation increased significantly by decade of first to third in all months of the year, but larger values of this growth, meet in the decades of the months October to March. During the Western plains, between Shkodra and Vlora, are noted that smaller quantities belong to PHAR's decades of months from October to March predominantly cyclonic weather conditions, which vary from 14848.34 Wh/m^2 during the first decade of October to 13408.08 Wh/m² in the March. It seemes rather peculiar fact that during October and November, the PHAR's quantities increase from the first decades of value to third, respectively 6-10% and 7-17%, but the biggest changes, 10% and 17% meet between decade before and the second of these months. With less pronounced changes occur three decades in December (3-6%), and to those of January, February and March, rather, there is a noticeable increase of PHAR values respectively 10-15% and 4-7%, indicating to unstable weather during these months. Months from April to September generally distinguished for relatively small changes in the values of the long-term PHAR's, which moved up to 2-9%, especially during summer months, a complete dominance of an uncyclonic weather.

2. Conclusion

By data extracted from surveys conducted stations installed throughout the Albanian territory and time series for 30 years, which are dispersed throughout the country, is clearly expressed solar energy property all year, despite growing differences between the average daily values, the monthly and annual energy. In setting survey points is generally observed such a spatial distribution, which creates an opportunity for a more realistic assessment of solar radiation. A particularly important analysis represents the average annual radiation values, which are characterized by a spatially explicit distribution, being divided into two periods characteristic such as that from October to March and April to September. Regarding the monthly distribution of average radiation in cloudless sky in August, we must emphasize that the country receives the largest amount of this radiation 530W/m^2, representing 9.5% of annual amount. December, stands for the smallest quantity of solar radiation value 345W/m², representing 1.5 times less than that of August and 6% of annual amount. In seasonal distribution of this radiation in terms of a clear weather, are seen changes much smaller than those in terms of a real weather, even the most characteristic phenomenon is, that the spring season takes on average 2% more power than the summer.

Acknowledgments

The author expresses gratitude to Prof.Gjovalin Gruda, Polytechnic University, Faculty of History and Filology, Department of Geography, for his constant encouragement and motivation on this paper.

References

[1] Energy Efficiency and Renewable Energy – Consumer's Guide. U.S Department of Energy (December, 2003).

[2] Photosynthetic active radiation regime in the Western Lowland of Albania. Monthly measurements of solar radiation (in hours) in the meteorological stations in Albania, 2002.

[3] Measurements of meteorological stations in the district (Source: Institute of Energy, Water& Environment), 2008-2010.

[4] Instat database for 2001-2010 period.

The study of Interferogram denoising method Based on Empirical Mode Decomposition

Changjun Huang[1, 2], Jiming Guo[3], Xiaodong Yu[4] and Changzheng Yuan[5]

[1] School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China

[2] School of Municipal and Surveying Engineering, Hunan City University, Yiyang, 413000, China

[3,4,5] School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China

Abstract

This paper proposes a new filter based on empirical mode decomposition that is based on different characteristics of signal with noise in different IMFS for suppressing speckle in SAR interferogram is proposed. At first empirical mode decomposition is used to divide signal and processed high-frequency IMF signals separately by adaptive filter. The denoising effect of the proposed method, usual filter and multiscale EMD filter was investigated by experiment. When the part related to the speckle is subtracted from the original interferogram, the speckle noise is reduced. The result is compared with the four other methods of mean filter, median filter and the adaptive filter, which shows that EMD filter method is powerful to interferogram speckle noise reduction, as well as it can preserve fine details in the interferogram that are directly related to the ground topography and maintain phase values distribution.

Keywords: *Empirical Mode Decomposition, Interferogram, noise, filter*

1. Introduction

Synthetic aperture radar (SAR) is a powerful tool to get geophysical characters of the earth and imaging with high resolution. A key problem of the radar image is the presence of speckle noise which is formed by the coherence of radar echoes from different scatters in an element. In the data processing of SAR interferometry, the interferogram is formed by conjugate multiplying of two coregistered SAR complex images. Because of the speckle noise of SAR image, the phase image of the interferogram is also degraded and many residues will be produced in phase unwrapping which can induce a inaccurate evaluation of the true phase values. In order to obtain a more accurate phase model, as a consequence, a better topographic model, a filtering step must be performed before the solution of phase ambiguities in the interferogram.

Some domestic and foreign scholars put forward some interferogram denoising methods, such as Seymour proposed the phase multiple optic filter of the interferometric complex [1], Eichel P.H and Lanar I.R proposed the circular cycle mean filtering and median filtering method [2]-[3], Lee proposed the adaptive filter [4], Zhu Daiying proposed Chirp-Z transform denoising method [5], and Goldstein and Werner proposed the classical frequency domain adaptive filter algorithm [6]. In general, these methods can be classified into two categories, there are two popular approaches to phase noise filter which are space domain filter and frequency domain filter; generally, these algorithms have adaptive filter window or bandwidth based on the local statistic character of the noise [1]. But due to INSAR interference noise and signal distribution in the data have its own characteristics, simple smoothing processing cannot achieve the good results.

Based on the above-mentioned shortcomings, this paper proposed a kind of filter algorithm based on the empirical mode decomposition (EMD) filter of interferogram phase noise suppression[7], which first decompose the real and imaginary parts of interferogram with the empirical mode decomposition method, and then determine phase value contribution for each pixel within the phase value of the filtered pixel phase template center in the complex domain, according to the interferogram gradient, achieve strong filter in low SNR region and weak filter in high SNR region, so that the edges of interferogram are preserved. The experimental results show that, the algorithm not only has the strong ability to suppress the speckle noise, and better maintain the edges and details of the interferogram, but also effectively reduces the loss of information in the interferogram, and ensure the phase purity of the phase image.

2. EMD Algorithm

The EMD involves the adaptive decomposition of given signal, $x(t)$, into a series of oscillating components, IMFs, by means of a decomposition process called sifting algorithm. The name IMF is adapted because it represents the oscillation mode embedded in the data. With this definition, the IMF in each cycle, defined by the zero crossings of, involves only one mode of oscillation, no complex riding waves are allowed. The essence of the EMD is to identify the IMF by characteristic time scales, which can be defined locally by the time lapse between two extrema of an oscillatory mode or by the time lapse between two zero crossings of such mode [8].

The EMD picks out the highest frequency oscillation that remains in the signal. Thus, locally, each IMF contains lower frequency oscillations than the one extracted just before. Furthermore, the EMD does not use any pre-determined filter or Wavelet function. It is fully data driven method. Since the decomposition of the EMD is based on the local characteristics time scale of the data, it is applicable to nonlinear and non-stationary processes. The EMD decomposes into a sum of IMFs that [9]: (1) have the same numbers of zero crossings and extrema; and (2) are symmetric with respect to the local mean. The first condition is similar to the narrow-band requirement for a stationary Gaussian process. The second condition modifies a global requirement to a local one, and is necessary to ensure that the IF will not have unwanted fluctuations as induced by the symmetric waveforms [9]. The sifting process is defined by the following steps:

Step 1) Fix $\varepsilon, j \leftarrow 1(j^{th} IMF)$

Step 2) $r_{j-1}(t) \leftarrow x(t)(residual)$

Step 3) Extract the $j-th$ IMF:

(a) $h_{j,i-1}(t) \leftarrow r_{j-1}(t), i \leftarrow 1$ (i number of sifts);

(b) Extract local maxima/minima of h $h_{j,i-1}(t)$;

(c) Compute upper envelope and lower envelope functions $U_{j,i-1}(t)$ and $L_{j,i-1}(t)$ by interpolating respectively local maxima and minima of $h_{j,i-1}(t)$;

(d) Compute the envelopes mean:
$\mu_{j,i-1}(t) \leftarrow (U_{j,i-1}(t) + L_{j,i-1}(t))/2$;

(e) Update:
$h_{j,i}(t) \leftarrow h_{j,i-1}(t) - \mu_{j,i,-1}(t), i \leftarrow i+1$;

(f) Calculate stopping criterion:

$$SD(i) = \sum_{t=0}^{T} \frac{\left|h_{j,i-1}(t) - h_{j,i}(t)\right|^2}{(h_{j,i-1}(t))^2} \qquad (1)$$

(g) Decision: Repeat Step (b)-(f) until $SD(i) < \varepsilon$ and then put $IMF_j(t) \leftarrow h_{j,i}(t)(j^{th} IMF)$

Step 4) Update residual: $r_j(t) \leftarrow r_{j-1}(t) - IMF_j(t)$

Step5) Repeat Step 3 with $j \leftarrow j+1$ until the number of extrema in $r_j(t) \le 2$ where T is the time duration. The sifting is repeated several times (i) in order to get h to be a true IMF that fulfills the requirements R1 and R2. The result of the sifting procedure is that $x(t)$ will be decomposed into $IMF_j(t), j = 1, \cdots, N$ and residual $r_N(t)$:

$$x(t) = \sum_{j=1}^{N} IMF_j(t) + r_N(t) \qquad (2)$$

To guarantee that the IMF components retain enough physical sense of both amplitude and frequency modulations, we have to determine a criterion for the sifting process to stop. This is accomplished by limiting the size of the standard deviation SD computed from the two consecutive sifting results [10]. Usually, SD is set between 0.2 to 0.3. Note that the EMD does not use any pre-determined filter or Wavelet function. It is a fully data driven method.

3. Denoising Principle

According to the property of the decomposition procedures, the data are decomposed into n IMFs (fundamental components), each with distinct time scale. More specifically, the first component associated with the smallest time scale corresponds to the fastest time variation of data. As the decomposition process proceeds, the time scale is increasing, and hence, the mean frequency of the mode is decreasing. Based on this observation, we may devise a general purpose time-space filter as

$$x_{lh}(t) = \sum_{j=1}^{h} IMF_j(t) \qquad (3)$$

where $l, h\{1, \cdots, \}, l \le h$. For example, when $l = 1$ and $h < n$, it is a high-pass filtered signal; when $l > 1$ and $h = n$, it is a low-pass filtered signal; when $1 < l \le h < n$, it is a band-pass filtered signal. In this paper, Eq. (3) forms the basis functions for representing interferogram data as described below, where we use it as a low-pass filter.

The EMD algorithm extracts the oscillatory mode which exhibits the highest local information from the data ("detail" in the wavelet context), and leaves the remainder as a "residual" ("approximation" in wavelet analysis). According to the major merits of EMD, the process of deriving the basic functions is empirical and the basic functions are obtained dynamically from the signal itself [11]. Therefore, it is reasonable to consider that the residual presents the basic characteristics of the interferogram and the detail denotes the variation of the noise represented by the highest local information. This is the motivation we use the EMD as a low-pass filter and only the distinct interferogram characteristics are utilized as discriminating features for accurate interferogram recognition.

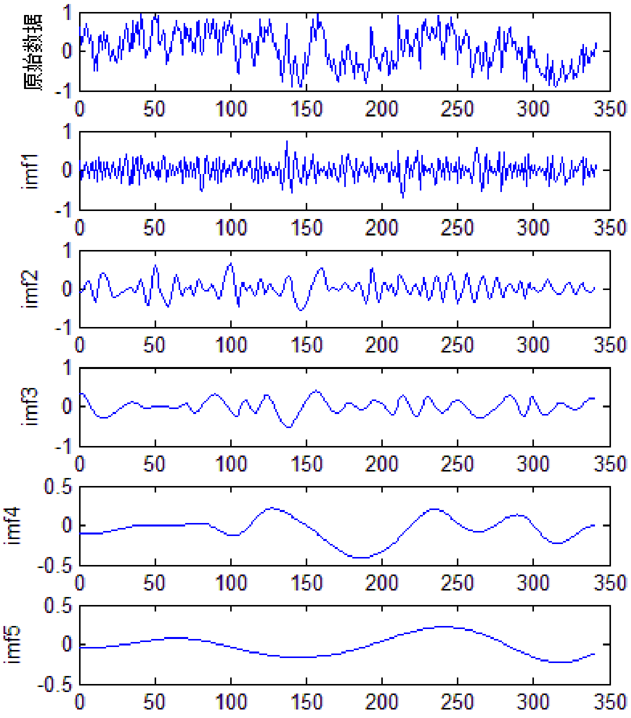
In this work different kinds of preprocessing are used: temporal filtering using Savitzky-Golay [10], Averaging, Median, and nonlinear transformation (hard and soft thresholding) [12].Accordingly, EMD can be extended to SAR Interferogram denoising. The different spatial scale information can be effectively separated by EMD which can process non-stationary, nonlinear information. Meanwhile the results of processing about spatial-frequency to singular signal can be controlled in a very small range, so that the abnormal vibration only impact the local, and will not spread to the whole region. Therefore, the methods of EMD can effectively separate scale images.

4. Experimental Results and Analysis

4.1. Experimental Data

The experimental data, the ERS-1/2, interval of 1 day and repeated track SLC data, whose size is 1800 x 2500.we obtain experimental interferograms after experimental data are removed the ground effect by the Swiss GAMMA software in this paper. After experimental interferogram data filtered by the empirical mode decomposition (EMD) method, we analyzed and compared the results with mean filter, median filter, and adaptive filter.

Taking the real component and imaginary component of original interferogram to compose the two data sets, we respectively decompose the real and imaginary parts of the original interferogram with the empirical mode decomposition (EMD) method, and choose different number of IMF to filter according to different needs and different form of noise. We can get filtered images after the real and imaginary parts filtered by EMD will be reconstructed, the filtered results shown in Fig.5. In order to analyze the EMD decomposition results, we select the 200th line of first 340 columns in the real and imaginary components of original interferogram that include both the region with more intensive interference fringes and the relatively sparse interference fringes, which have very strong representative to analysis for further. The EMD decomposition effect diagrams are shown in Fig.1 and Fig. 2.



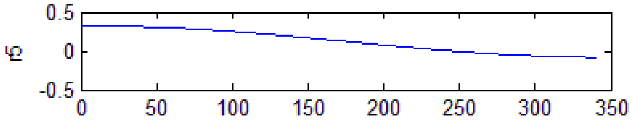


Fig.1 Obtained five IMF components and the residual (r5 on the bottom) from the real component of the original interferogram after applying the EMD method

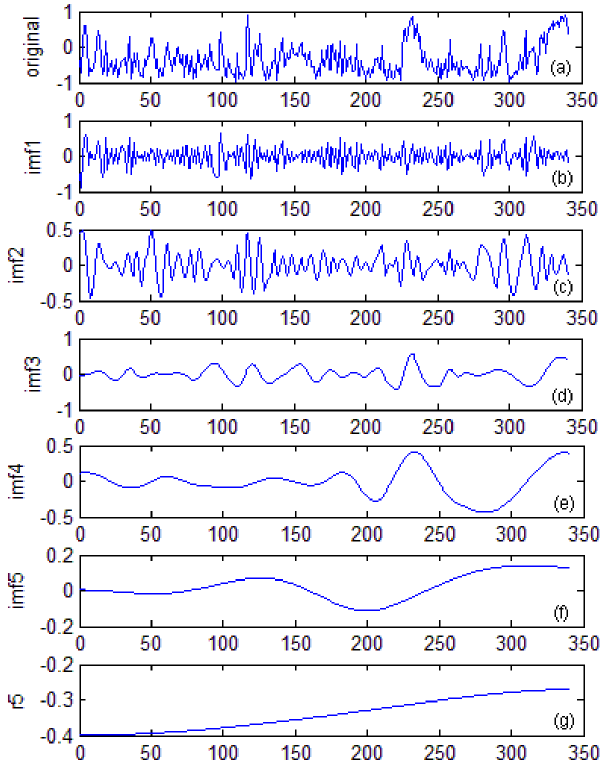
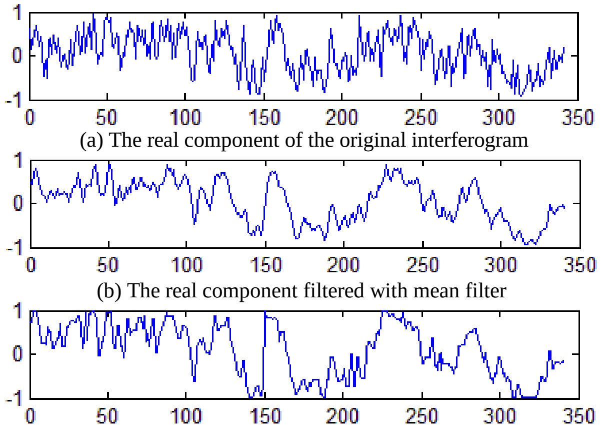


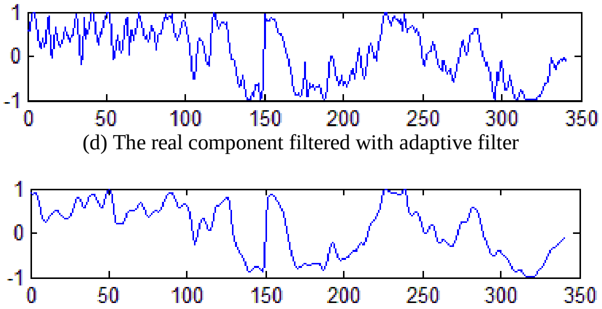
Fig.2 Obtained five IMF components and the residual (r5 on the bottom) from the imaginary component of the original interferogram after applying the EMD method



(a) The real component of the original interferogram

(b) The real component filtered with mean filter

(c) The real component filtered with median filter

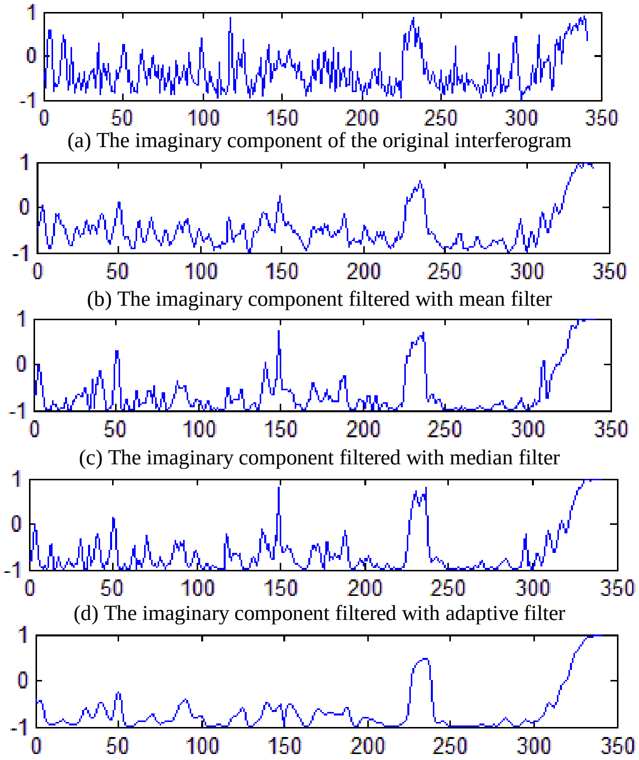


(d) The real component filtered with adaptive filter

(e) The real component filtered with EMD

Fig.3. Real component filtered with different filters compared with original real component. (a) is the real components of the original data, (b),(c), (b) and (e)are the real components results filtered by the four filters.

From Fig. 3 and Fig. 4, we can know that the image curves after filter denoising is smooth than the original real and imaginary parts information, which demonstrate the four filters remove a lot of noises. In (b) to (d) graphs, mean filter, median filter and adaptive filter method had some smoothing effect, but there still is difficult to remove some burrs, the effect of the three filter is similar; as can be seen in Fig.3 and Fig.4 (e), the empirical mode decomposition (EMD) method is obviously better than the former several filters methods whether in removing noises, or image smoothing degree, which remove the burrs, and achieve filtering smoothing effect.



(a) The imaginary component of the original interferogram

(b) The imaginary component filtered with mean filter

(c) The imaginary component filtered with median filter

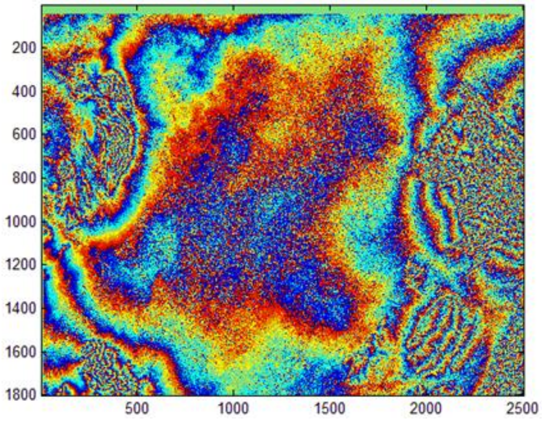
(d) The imaginary component filtered with adaptive filter

(e) The imaginary component filtered with EMD

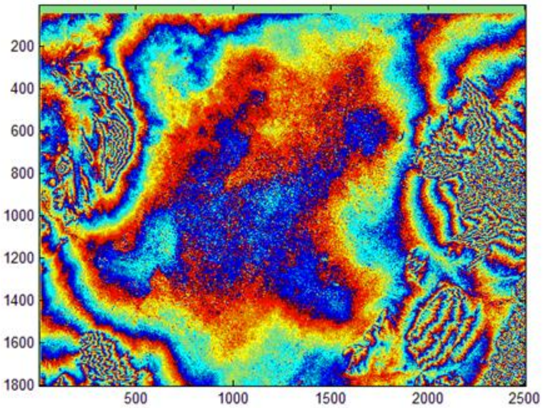
Fig.4. Imaginary component filtered with different filters compared with original imaginary component. (a) is the imaginary components of the original data, (b),(c), (b) and (e) are the imaginary components results filtered by the four filters.

4.2. Experiments Compare and Analysis

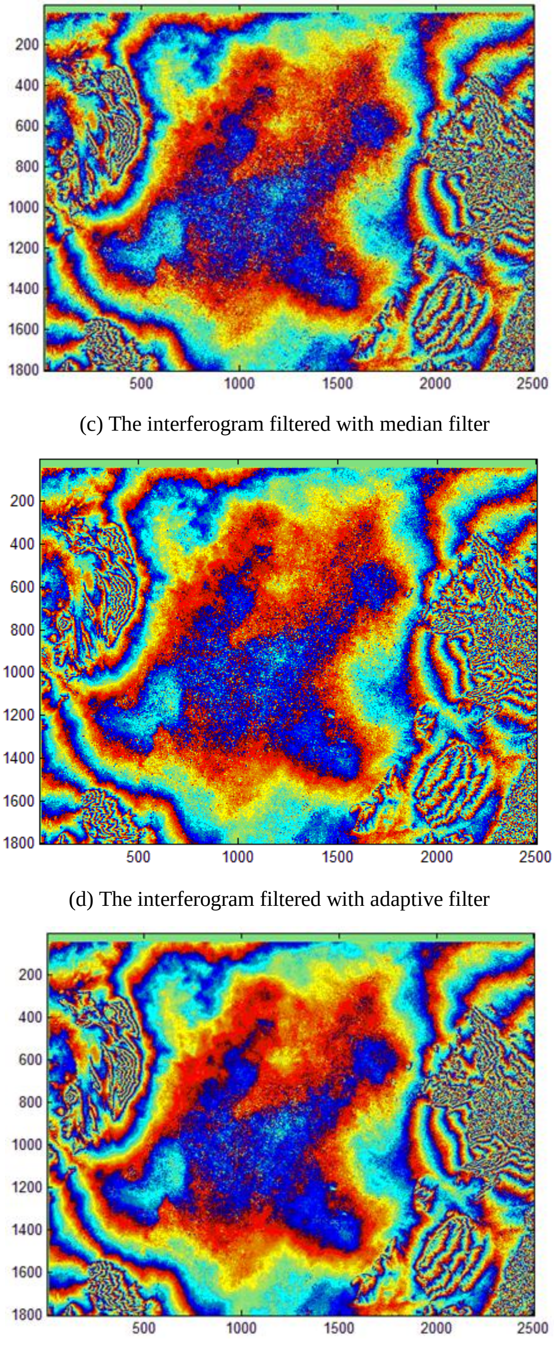
This paper chooses interferogram filtering quantitative evaluation indexes of RMS index, phase standard deviation (PSD) [13], Sum of Phase Difference (SPD) index [14] and residual index [15] to evaluate the above-mentioned four filter methods [16]. Fig.5 is interferogram filtered with different filters compared with original interferogram. In the interferogram filtered by the four filters, we select the phase diagrams of the 200th rows of 340 columns to further comparative and research the results of the four filters; the cross sections over the filtered interferogram are shown in Fig.6.



(a) The original interferogram



(b) The interferogram filtered with mean filter



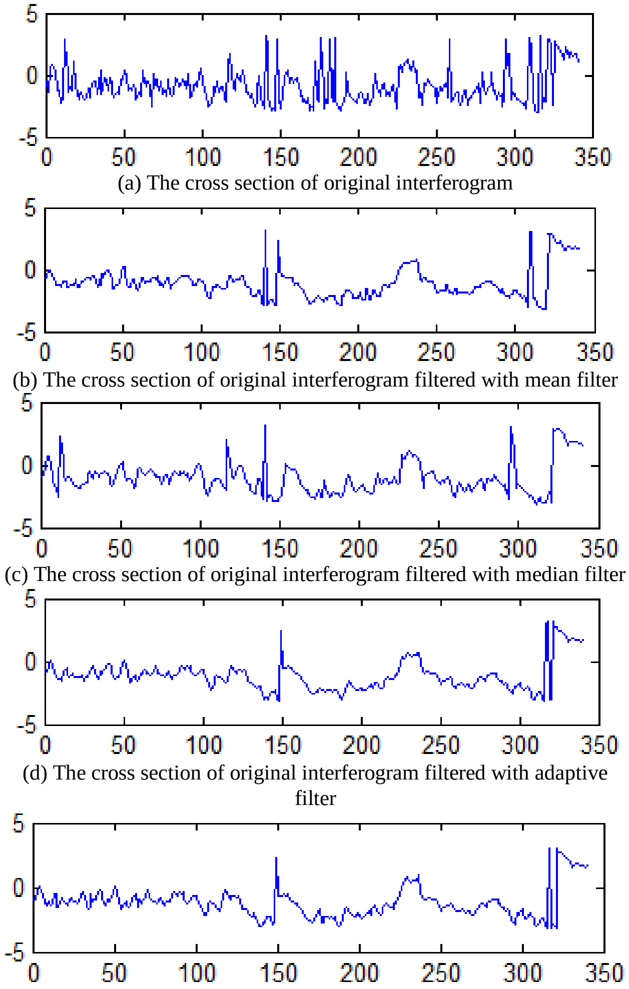
(c) The interferogram filtered with median filter

(d) The interferogram filtered with adaptive filter

(e) The interferogram filtered with EMD

Fig.5 Interferogram filtered with different filters compared with original interferogram

From Fig.5, we can know that the speckle noises of denoising interferograms are reducing in (b) to (d), but there are still some spots existing; from visual effect, the denoising interferograms of the mean filter and median filter have obvious speckle noise that is not eliminated; in (e), the denoising interferograms of EMD is very smoothing, no obvious speckles, where stripes are clear, and feature, structure characteristic and small target have been well maintained. From above-mentioned, the empirical mode decomposition (EMD) filter method is obviously better than the preceding three filters, whether in removing the noise, or image smoothing degree [17].



(a) The cross section of original interferogram

(b) The cross section of original interferogram filtered with mean filter

(c) The cross section of original interferogram filtered with median filter

(d) The cross section of original interferogram filtered with adaptive filter

(e) The cross section of original interferogram filtered with EMD

Fig.6. Cross section over the filtered interferogram

From shown in Fig.6, Compared with the other three filters, the interferogram fringes filtered by empirical mode decomposition (EMD) filter method have better continuity[18], whose noise suppression effect is very obvious, which are more consistent with the cross sections of the original interferogram.

Table 1 is statistics of various filter evaluation criterions. As can be seen, the RMS, PSD and SPD of interference phase diagram filtered by the mean filter, median filter and adaptive filter is reduced, which illustrate the 3 kinds of filtering algorithm play a smoothing effect to interferometric phase images, but the empirical mode decomposition filter method is superior to the 3 algorithms in keeping of the edge and phase details. So the denoising ability of the empirical mode decomposition filtering method is better than the three kinds filter methods.

Table.1. Statistics of various filter evaluation criterions

Denoising method	RMS	PSD	SPD	Residual points
original interferogram	1.8765	1.7608	4.7741E+005	201240
mean filter	1.0672	0.8888	3.5366E+005	5872
median filter	0.9275	0.7856	3.4568E+005	4217
adaptive filter	0.8536	0.6193	3.2767E+005	1028
EMD	0.3417	0.3569	2.3139E+005	685

5. Conclusions

A large number of noises in interferogram seriously affect the efficiency and accuracy of phase unwrapping algorithm. Therefore, in the processing of InSAR interferogram, we must effectively remove interference noise, and improve the operation efficiency and required accuracy. According to the characteristics of EMD, this paper introduces the empirical mode decomposition (EMD) method to SAR interferogram filtering. The experimental results show, the empirical mode decomposition is powerful to suppress speckle noise and phase noise while preserving edges than the classical filtering method, whether from the visual interpretation, or quantitative evaluation index. Our next research is to develop two-dimension filter based on EMD method.

Acknowledgments

The authors wish to thank the helpful comments and suggestions from my teachers and colleagues in Wuhan University and Hunan City University. This work is supported by the central college basic scientific research business expenses special fund (No.2012214020209) and the scientific research fund of Hunan provincial education department (No.12C0566).

References

[1].Seymour, M.S. Gumming, I.G. Maximum likelihood estimation for SAR Interferometry in Processing . IGARSS, 1994, 12, 2272-2278.

[2]. Eichel P.H, Ghiglia D.C. Spotlight SAR Interferometry for Terrain Elevation Mapping and Interferometric Change Detection. Sand: Sandia National Labs technology, 1993, 12, 2529-2546.

[3].Lanari R, Fornaro G. Generation of Digital Elevation Models by Using SIR_C/X_SAR Multifrequency Two-pass Interferometry: The Etna Case Study. IEEE Trans on GRS, 1996, 34, 1096-1115.

[4].GOLDSTEINRM, WERNERCL. Radar Interferogram Filtering for Geophysical Applications. Geophysical Research Letters, 1998, 25, 4035-4038

[5].Lee,J.S., Papathannassion, K.P. A. New Technique for Noise Filtering of SAR Interferometric Phase Images. IFFF. Trans. on GRS, 1998, 34, 1455-1459.

[6].Zhu Daingying, Scheiber, R, Zhu Zhaoda. Impacts of an efficient topography adaptive filter on coherence estimation and phase unwrapping. EUSAR, 2000,23, 318-320.

[7]. Yue Huanyin, Guo Huadong, Han Chunming, et al. A SAR Interferogram Filter Based on the Empirical Mode Decomposition Method. IEEE AGCS, 2011, 13, 2061-2063.

[8]. Boudraa A O,Cexus J C. Denoising via empirical mode decomposition. IEEE International Symposium on Control, Communication and Signal Processing (ISCCSP06), Morocco,2006.

[9]. N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shin,Q. Zheng, N.C. Yen, C.C. Tung and H.H. Liu, ”The Em-pirical Mode Decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis,”Proc. Royal Soc. London A , vol. 454, pp. 903-995, 1998.

[10].D.L. Donoho and I.M. Johnstone, ”Ideal spatial adaptation via wavelet shrinkage,” Biometrica, vol. 81, pp.425-455, 1994.

[11]. M.M. Goodwin and M. Vetterli, ”Matching pursuit and atomic signal models based on recursive filter banks,” IEEE Trans. Sig. Process., vol. 47, no. 7, pp. 1890-1901,1999.

[12].A. Savitzky and M.J.E. Golay, “Smoothing and differentiation of data by simplified least squares procedures”. Analytical Chemistry, vol. 36, pp. 1627-1639, 1964.

[13].Tan Shanwen, Qin Shuren, Tang Baoping. Hilbert-Huang transforms filter and its application. Journal of Chongqing University, 2004.27, 109-120.

[14].BO YC, et al. A Wavelet-Based Filter for SAR Speckle Reduction and the Comparative Evaluation on Its Performance.Journal of Remote Sensing, 2003,7, 393-399.

[15].LEEJ S, JURKEVICHI. Speckle Filtering of Synthetic Aperture Radar Images: a Review. Remote Sensing Reviews, 1994, 12, 13-340.

[16].OLIVER CJ, QUEGAN Understanding Synthetic Aperture Radar Images. London: Artech House Inc., UK, 1998.

[17]. LEEJ S. Digital Image Enhancement and Noise Filtering by Use of Local Statistics. IEEE Trans. Pattern Analysis and Machine Intelligence, 1980, 2, 165-168.

[18]. BO YC, et al. A Wavelet-Based Filter for SAR Speckle Reduction and the Comparative Evaluation on Its Performance. Journal of Remote Sensing, 2003,7, 393-399.

Changjun Huang obtained the B.S. and M.S. degrees in Geodesy and Survey Engineering from East China Institute of Technology, China, in 2003 and 2006, respectively. Since September 2011, he is a PhD student in School of Geodesy and Geomatics, Wuhan University, China. His current research interests include InSAR interference measurement and application of InSAR in high precision deformation monitoring.

Jiming Guo received the Ph.D. degree in Geodesy and Survey Engineering from Wuhan University, China, in 1985.
From June 2002 to June 2003, he was a Visiting Researcher in GPS Group, Geodesy and Geomatics Engineering Department, University of New Brunswick, Canada. He is currently a professor, Ph.D supervisor in Wuhan University, China; he is concentrated on the research and education in engineering survey and GPS application.

Xiaodong Yu received the B.S degree in Geodesy and Survey Engineering from China University of Mining and technology, China, in 2011. Since September 2011, he is a master in School of Geodesy and Geomatics, Wuhan University, China. Currently, he researches the theory of InSAR and application of InSAR in high precision deformation monitoring.

Changzheng Yuan obtained the B.S degree in School of Geodesy and Geomatics, Wuhan University, China. Since September 2007 to June 2011.Currently, he is studying for a master's degree in the same school. The direction of his research is InSAR interference measurement and its application in deformation monitoring.

Achieving Load Balance by Separating IP Address Spaces

Sanqi Zhou[1], Jia Chen[1], Huachun Zhou[1] and Hongke Zhang[1]

[1] National Engineering Laboratory for Next Generation Internet Interconnection Infrastructure
Beijing JiaoTong University
Beijing, 100044, China

Abstract

In this paper, we propose a load balance approach by separating the host and router IP addresses into two spaces. In addition, in our approach, we propose a scheduling algorithm, named Edge Stream Balance (ESB), which is used by the proposed multipath routing scheme based on the address space separation. Each router can schedule each stream that is initiated by the connected host onto the proper path to the destination host by ESB dynamically. The multiple paths between any pair of hosts can be obtained by the connected routers by using the address separating mechanism. The merit of our approach is that: it balances the network traffic dynamically while being free of traffic demand assumption and offline flow optimization. The path of each stream is selected by each router individually other than using central system based on the address separation. The time complexity of ESB is much lower than the linear programming (LP) and integer linear programming (ILP) which are used in flow optimization. Simulation results show that on average of all simulated scenarios, compared to the existing single path routing which is based on the address separating, the unused link ratio (ULR) reduces by 82%. And in the relative sense, the traffic across the network is balanced 31%.

Keywords: *Load Balance, Address Space Separation, Multipath.*

1. Introduction

In the existing Internet, load balance is a critical issue which is not still solved elegantly [1]. This is caused by two reasons: one is that the hosts and routers which are growing rapidly are distributed by power-law in the topology [2, 3], the other is the routing scheme is executed independently in each router based on the Shortest Path First (SPF) policy which may overload certain paths while underloading some others [4].

Previous works have explored load balance in some different ways. Fortz et. al. [5] proposed a solution of optimizing OSPF weights. Sridharan et. al. [4] and Wang et. al. [6] exploited more effective ways to split traffic over the shortest paths. Xu et. al. [7] optimized the link loads by using only *E* link weights against *O(NE)* parameters in [4] and [6], where *N* is the number of routers. These schemes are all depending on assuming having knowledge of the network traffic demands. Furthermore in [4] and [6], a central controller is used to compute and configure the flow splitting ratio instead of routers. Thus, it sacrifices the main benefit of running a distributed protocol. Both Antic et. al. [8] and Tsunoda et. al. [9] proposed the shortest path routing (SPR) solutions respectively, in which each source router selects different intermediate routers for forwarding each packet at a time. In the both solutions, the traffic demand bound of each router should be known by all routers. In [8], the high time complexity of the realtime linear programming (LP) may affect practicing in a large topology. Keller et. al. [10] optimized the traffic infiltrated into other autonomous systems (AS) by migrating the intradomain edge links. However, the realtime traffic (e.g., the Audio stream) cannot be kept in this approach. Both Saucez et. al. [11] and Paul et. al. [12] proposed traffic engineering solutions in the Identifier(ID)/Locator separating context. However, in these solutions, the path selection is totally based on a central controller in each AS. Meanwhile, the topology and traffic information must be gathered in time or periodically by the controller.

Therefore, although the research on load balance has made a great progress [4-12], to our best knowledge, the following question is still to be addressed: *Is it possible to achieve load balance in a wholly distributed system without traffic demand assumption and high time complexity?* To answer, we present a novel approach.

As [11] and [12], our approach employs the principle of separating IP address space. That is because a key merit of separating IP space is to make possible to associate multiple router addresses to a unique host address. Thus, it enables each router choose different routers (i.e., the paths) to route the packets between the same host pair for balancing the traffic. In this paper, we address the question presented above by the following two steps.

First, we propose a multipath routing scheme based on a generalized address space separating mechanism. By this scheme, each router can encapsulate a set of appropriate router addresses into the packets of the same stream when each packet initially enters the network from the router, and can also decapsulate the addresses from the packet when it leaves the network. These addresses are obtained from the response messages of address space separating mechanism (described in Section II). Second, we propose a scheduling algorithm, which is called Edge Stream Balance (ESB). In each router, it is used to select the appropriate router addresses to encapsulate into packets.

This approach brings three main merits. 1) The network traffic is balanced without a central controller. Because each router gets all host-to-routers addresses mapping information by the address space separating mechanism, and some parts of the realtime traffic information by dumping the packets it received, each router can select the appropriate router addresses while without centralized service. 2) The time complexity of ESB is much lower than the LP and integer linear programming (ILP) such that the traffic can be balanced timely. 3) ESB runs without any transcendental knowledge of traffic demand.

We also simulate our approach in a large number of scenarios and analyze the average performance in each topology model.

The rest of the paper is organized as follows. Section II describes the generalized IP address space separating mechanism. Section III describes the multipath routing scheme and the ESB algorithm. In section IV, we perform simulations and analyze the results. Section IV concludes the paper.

2. Generalized IP Address Space Separation

Several address space separating solutions, which are generally in the principle of ID/Locator separation, were proposed in recent years [13-15]. In this paper, we consider a generalized IP address space separating mechanism (without specific packet format definition and etc.) named *One-hop Distributed Hash Table (One-hop DHT) based address separation* [16], which can support the proposed multipath routing scheme and ESB algorithm to achieve load balance. Fig. 1 [17] shows how it works.

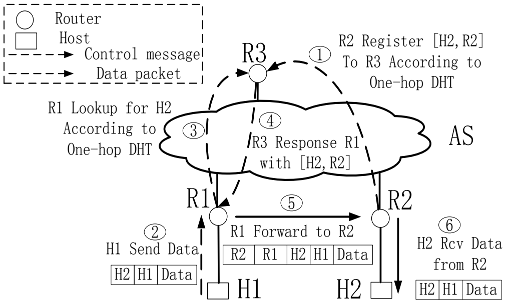


Fig. 1 One-hop DHT based address separation.

In Fig. 1, it shows the 6 primary steps of forwarding a packet from one host to another (i.e., H1 to H2) based on the address separation. Each router creates a hash ring with the specific addresses of all routers (e.g., the maximum address of each router obtained from Link State Database (LSDB) of OSPF [18]) by using an identical hash function (e.g., Secure Hash Algorithm 1, SHA-1 [19]). When a host accesses the network, the mapping entry of the host IP, its connected router's maximum IP (here we call it Router ID) and the router's other IPes are stored in a certain router according to the hash function (i.e., step 1). Thus, when a router forwards a packet received from its connected host (i.e., step 2), one of the destination router IPes can be obtained by using the hash function (i.e., step 3 and 4). Then, the source/destination router IPes are encapsulated into the packet header to forward to the destination router (i.e., step 5) which decapsulates the router IPes and forwards the packet to the destination host (i.e., step 6). Meanwhile, the mapping entry is cached by the source router for accessing the same host afterwards.

3. Load Balance Solution

In this section, we first describe the multipath routing scheme in Subsection 3.1. Then, ESB algorithm is described in Subsection 3.2.

3.1 Multipath Routing Scheme

Based on the separating mechanism mentioned above, each router can register its IPes and the neighbors' IPes which can be got by a certain protocol, e.g., the Hello protocol in OSPF, to another router which can be found in the hash ring created by the hash function. When the packets, sent from the same source host and destined to the same destination host, are received by the source router, they may be encapsulated with different IPes of the source router and its neighbors, destination router and its neighbors (which can be got from the router in which the IPes are registered). Then, the packets are forwarded along the paths indicated by these IPes. Each time the router, which is the intermediate destination indicated by one of the IPes, receives the packet, it changes the current destination IP into a new one which is contained in the packet, and then forwards it to the next intermediate destination. The process continues until the packet is received by the destination host.

Fig. 2 shows the multipath routing scheme based on the introduced separating mechanism. The two packets sent from Hs to Hd are routed on different paths that are indicated by solid and dash arrows respectively. The original protocol type in the IP header is temporarily saved between the data field and the IP header (i.e., the "**y**") by router A when the packet is forwarded from Hs to router A, and is restored by router G before being forwarded to Hd. In the forwarding process, some other specific values are filled in the protocol field sequentially as shown in the underlined font in Fig. 2. These values can take the reserved protocol types defined in the IP standard [20], and are used by the multipath routing procedure in the packet forwarding process. Procedure 1 shows the pseudocode of the multipath routing scheme.

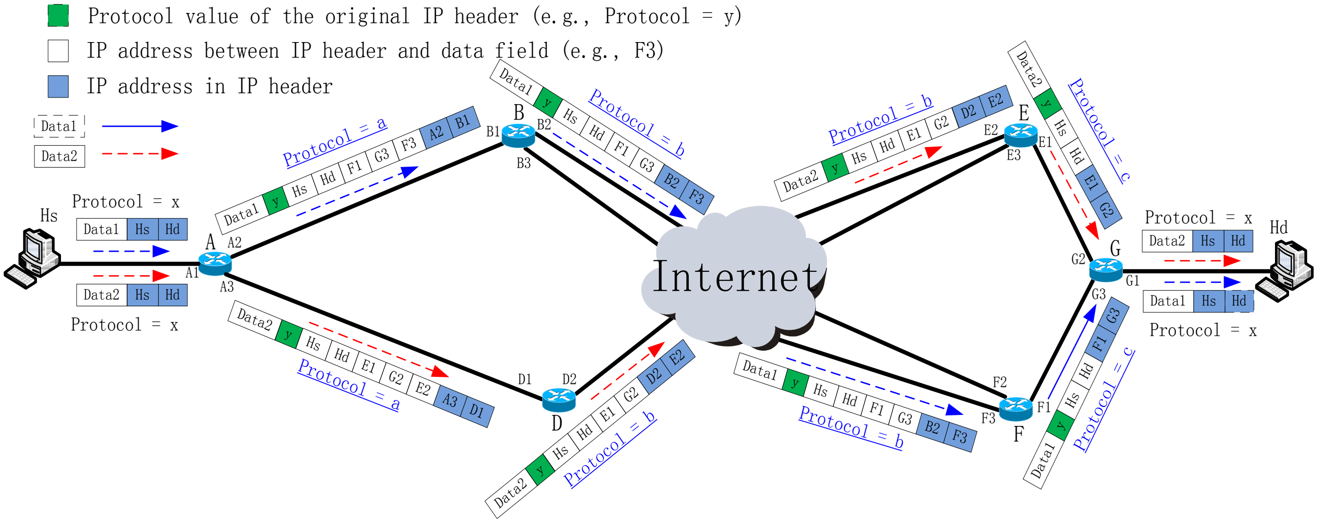

🟩 Protocol value of the original IP header (e.g., Protocol = y)
□ IP address between IP header and data field (e.g., F3)
🟦 IP address in IP header
Data1
Data2
Hs
Hd
Protocol = x
Internet
A A1 A2 A3
B B1 B2 B3
D D1 D2
E E1 E2 E3
F F1 F2 F3
G G1 G2 G3
Protocol = a
Protocol = b
Protocol = c
Data1 y Hs Hd F1 G3 F3 A2 B1
Data1 y Hs Hd F1 G3 B2 F3
Data2 y Hs Hd E1 G2 E2 A3 D1
Data2 y Hs Hd E1 G2 D2 E2
Data2 y Hs Hd E1 G2
Data1 y Hs Hd F1 G3

Fig. 2 Multipath routing scheme.

Procedure 1 Pseudocode of multipath routing scheme

//In this procedure, ***pkt*** is the received packet, ***pkt.hdr*** is the standard IP header, ***pkt.hdr.ptl*** is the protocol field in the IP header and the local mapping interfaces (***LMIF***) is the set of host accessible router's interface IPes, which are connected with other routers, and the router's neighbor IPes. We use ***DstLmifEnt*** to point to the <dest IP, ***LMIF***> entry which is the set of IPes of the destination host, routers and router neighbors. It is cached in the current router according to the introduced separating mechanism.

When a packet received by a router and the IP header has been checked:

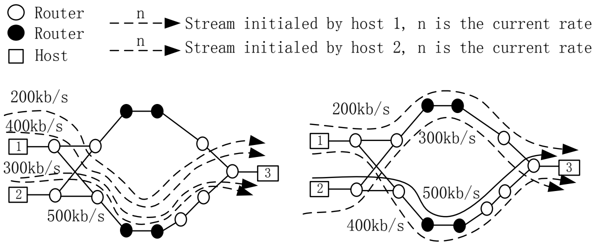
```
1:  if (pkt.hdr.destIP != the current interface IP) then
2:    if (the interface is connected with hosts) then
3:      if(LookupCacheEntry(pkt.hdr.destIP,DstLmifEnt)==true) then
4:        ESB(pkt, LMIF, DstLmifEnt); //select IPes from LMIF and
          DstLmifEnt to be encapsulated into pkt.
5:      else //the <dest IP, LMIF> entry has not been cached
6:        Send a lookup packet based on the address separating
          mechanism and drop pkt;
7:        return;
8:      end if
9:    end if
10: else //pkt destines to the interface
11:   switch (pkt.hdr.ptl)
12:   {  case a: decapsulate the first IP behind the header (i.e., F3 in
          the data1 packet in Fig. 2) into pkt.hdr.destIP, and then lookup
          the routing entries to get the output interface IP which is to be
          taken as pkt.hdr.srcIP;
13:            pkt.hdr.ptl = b; break;
14:      case b: decapsulate the first IP behind the header into
          pkt.hdr.destIP;
15:              decapsulate the second IP behind the header into
          pkt.hdr.srcIP;
16:            pkt.hdr.ptl = c; break;
17:      case c: restore the original source and destination IPes, and
          decapsulate the saved protocol field (i.e., "Protocol=x" in Fig. 2)
          into pkt.hdr.ptl; break; }
18:   Recalculate the other fields (i.e., checksum and etc.) in the IP
          header.
19: end if
20: Pass the packet to the existing forwarding procedure which is
          already modified to consider the protocol types a, b and c the
          same as regular IP packet;
```

In statement "3", "*LookupCacheEntry*(***pkt.hdr.destIP***, ***DstLmifEnt***)" looks up all cached <dest IP, ***LMIF***> entries to find out the one that matches the first parameter, and then to point it with the second parameter.

Number of Paths. Here, we assume that there are averagely *k* neighbors per router and two bidirectional links between two routers at most. Thus, there are at most *2k* source router neighbor interfaces and *2k(k-1)* destination router neighbor interfaces (another two interfaces per destination router neighbor are connected with the destination router) can be selected to indicate different paths. Therefore, on average, there are $O(k^3)$ magnitude paths between each pair of routers.

3.2 Edge Stream Balance (ESB)

In Procedure 1, ESB is invoked to choose the appropriate IPes which are to be encapsulated in the packet to indicate the routers to be forwarded to (i.e., the statement "4"). The principle of ESB is: when the link utilization (LU) on the links between routers, one of which is connected with hosts is balanced (Edge Stream), the LU across the network is tended to be balanced (in the means of stream, that is, each time a new stream is initiated by host). The principle can hold for this reason: In the real network, the edge link with higher bandwidth is always connected with the core link with higher bandwidth. This is to ensure each link can be used as transiting the traffic effectively as possible. Fig. 3 shows the single path routing scheme and the multipath routing scheme with ESB.


Router
Router
Host
n
Stream initialed by host 1, n is the current rate
n
Stream initialed by host 2, n is the current rate
200kb/s
400kb/s
1
300kb/s
2
500kb/s
3
200kb/s
1
300kb/s
2
500kb/s
400kb/s
3

(a) Single Path Routing (b) ESB

Fig. 3 The schemes of the single path routing and the multipath routing with ESB. Each arrow indicates the path that a stream is forwarded along. The different colors of the nodes are used to illustrate the result of ESB compared to that of single path routing.

Assuming all links are with the same bandwidth, in Fig. 3(a), each stream initiated by either host 1 or host 2 is routed on the shortest path such that the lower link between the black nodes undertakes all 4 streams while the upper one does not carry any of them. However, in Fig. 3(b), the streams initiated by each host are separated on two paths by selecting the router interfaces which currently have the minimum LU. Therefore, the links between the black nodes are more approximately balanced. Notice that, ESB doesn't know the rate of a new initiated stream when the first packet received by the router. ESB only chooses the interfaces with minimum current LU for the stream, and then all the following packets of this stream are totally routed on the indicated path even if the LU of other interfaces may be lower. This is because the traffic can be balanced approximately in the means of stream level while keeping the time complexity of packet forwarding as lower as possible (i.e., ESB only takes more time when it is called for choosing a path for a stream for the first time). The algorithm includes two main phases. One is to balance the LU generated by the output traffic on the current router interfaces which connected with the neighbor routers, the other is to balance the LU of the input traffic on the destination router's neighbor interfaces which are not connected with the destination router. The first phase is achieved by selecting the interfaces of the minimum total current output LU among the local router interfaces and its neighbor interfaces. The second phase is achieved by selecting the interfaces of the minimum input LU in the cached <dest IP, ***LMIF***> entry that is corresponding with the destination IP of the packet. That is, the first phase records and uses the LU in the global scope while the second phase does that on each cached entry. Thus to the router that invokes ESB, the output LU on each of its link and the input LU on each link, which is not connected with the router, of all its neighbors, are balanced respectively. Algorithm 1 shows the pseudocode of ESB.

Algorithm 1 Pseudocode of ESB

```
ESB(pkt, LMIF, DstLmifEnt)
len = pkt.hdr.length;
$t_{PktDeque}$ = ($t_{now}$ = GetCurrentTime()) - $\boldsymbol{T_{LUinterval}}$; //$\boldsymbol{T_{LUinterval}}$ is used
    to calculate the traffic rate, and hence the LU.
PktDeque(LMIF.ALLif.iface,DstLmifEnt.ALLnif.iface,$t_{PktDeque}$);
    //pop up all (len, $t_{rcv}$) before $t_{PktDeque}$ from each element of iface or
    niface arrays of the input parameters and subtract the
    corresponding len from the trf field in each element. The trf only
    contains the traffic from connected hosts.
if(GetCachePath(pkt, &IP1, &IP3, &IP2, &hd, &hs, &proto, a,
    $t_{now}$) == false) then //If GetCachePath() == false, a new stream is
    initiated. Otherwise, pkt belongs to an existing stream of which
    the path has been cached in the current router. Then, the
    combination of the source/destination IPes and TCP or UDP ports
    is used as the index to get the cached router IPes. Meanwhile, "a"
    is set into pkt.hdr.ptl and the len is accumulated on the
    correspondingly cached interfaces (i.e., the trf fields) such that the
    paths can be chosen reasonably when the next stream is initiated.
  la=minLUindex(LMIF.ALLif.iface); //obtain the minimum output
    LU interface, the bandwidth is employed.
  LMIF.ALLif.iface[la].trf += len;
  LMIF.ALLif.iface[la].PktEnque(len, $t_{now}$); //push in queue
    //accumulate the output traffic in $\boldsymbol{T_{interval}}$ on the interface.
  lb = 1; //init lb
  if(LMIF.ALLif.iface[la].NeighborIPnum>1) then
    //multiple interfaces of neighbors are connected to the router
    interface
    lb=minLUindex(LMIF.ALLif.iface[la].niface);
      //LMIF.ALLif.iface[la].niface is the interface array of the
      neighbor routers of la interface.
  end if
  LMIF.ALLif.iface[la].niface[lb].trf += len;
  LMIF.ALLif.iface[la].niface[lb].PktEnque(len, $t_{now}$);
  da=minLUindex(DstLmifEnt.ALLnif.iface);//obtain the minimum
    input LU interface of the destination router neighbors.
  DstLmifEnt.ALLnif.iface[da].trf += len;
  DstLmifEnt.ALLnif.iface[da].PktEnque(len, $t_{now}$);
  db = 1; //init db
  if(DstLmifEnt.ALLnif.iface[da].router.ifaceNum>1)then
    //DstLmifEnt.ALLnif.iface[da].router.ifaceNum is the number of
    the interfaces which are connected with the destination router
    and belongs to the router that has DstLmifEnt.ALLnif.iface[da].
    DstNeighbor = DstLmifEnt.ALLnif.iface[da].router;
    db = minLUindex(DstNeighbor.iface); //DstNeighbor.iface is
      the interface array of the selected destination router neighbor.
      The interfaces are connected with the destination router.
  end if
  DstNeighbor.iface[db].trf += len;
  DstNeighbor.iface[db].PktEnque(len, $t_{now}$);
  //Encapsulate pkt and modify pkt.hdr.ptl
  proto = pkt.hdr.ptl;
  pkt.hdr.ptl = a; //set it to "a" for neighbor receiving
  hs = pkt.hdr.srcIP;
  hd = pkt.hdr.destIP;
  pkt.hdr.srcIP = LMIF.ALLif.iface[la].IP;
  pkt.hdr.destIP = LMIF.ALLif.iface[la].niface[lb].IP;
  IP1 = DstLmifEnt.ALLnif.iface[da].IP; //destination neighbor
  IP2 = DstNeighbor.iface[db].IP; //destination neighbor
  IP3 = DstNeighbor.iface[db].peerIP; //destination router
  SetCachePath(pkt, IP1, IP3, IP2, la, lb, da, db, $t_{now}$); //Cache the
    router IPes. The corresponding interface indexes are also cached
    to record the following traffic of the same stream on the
```

```
        interfaces.
34:  end if
35:  Encapsulate (pkt, IP1, IP3, IP2, hd, hs, proto);
36:  return;

Delete the cached paths when the timer of the corresponding
streams are timeout:
     //GetCachePath() resets timer of the stream only if the stream has
     existed. SetCachePath() resets the stream timer when it is called,
     that is, when the stream is initiated.
1:   DelCachePath(pkt);
```

In the statement "32", the *peerIP* indicates the destination router interface that connected with *IP2*. Because two interfaces of a router can't be connected in the real network, the *peerIP* is unique for each *IP2*. The function "*DelCachePath*" is used to delete the path which has been indicated for a stream when the path is timeout for this stream. This is because, when a stream is released or cancelled by a host, or be even "silent" for some reason, the router which is connected with the host can't be aware. Thus, we should delete the cached stream path in the local router. When the router receives any new packet of this stream from its connected host, it takes the packet as a start of an initiating stream and assigns a new path for it.

We can see that, both the multipath routing scheme and ESB algorithm only process each packet locally, while without communicating to any other router or central system beyond the address separating mechanism and routing protocol. Furthermore, ESB only chooses path by collecting and analyzing the realtime traffic that is sent from or destined to the local hosts, rather than using global traffic information or static traffic demands.

Time Complexity. In Algorithm 1, the time complexity of ESB is mainly made up of three phases. The first is "*minLUindex*" which performs the selection of the interface with minimum input or output LU. In the worst case, it sequentially traverses the interface array with random stored elements. Thus, the worst time complexity will be proportional to the array length. Assuming there are *k* neighbors per router and *n* interfaces between two routers at most, the interface number (connected with router) of a router won't be more than *k*n*. The total time complexity of the four invoking of "*minLUindex*" is

$$O\left(TotalminLUindex\right)=O\left(kn\right)_{la}+O\left(k\right)_{lb}+O\left(k(k-1)n\right)_{da}+O\left(n\right)_{db}, \quad (1)$$

where each term with subscript is the corresponding time complexity of the invoking in Algorithm 1. The largest array is *DstLifEnt.ALLnif.iface* with at most *k* routers. The length is no more than *k*(k-1)*n* (the destination router is a neighbor of each destination router neighbor). Hence, *O(TotalminLUindex)* is at most $O(k^2n)$. Generally, *n* is always far less than *k* (In general, two links between two routers at most). Thus, *O(TotalminLUindex)* won't be higher than $O(k^3)$. The second phase is "*PktDeque*" which maintains the traffic records in the most recent $\boldsymbol{T_{interval}}$ for all interfaces in the ***LMIFes***. Similarly with "*minLUindex*", "*PktDeque*" also has to sequentially traverse the four interface arrays. However, each interface in the arrays has a First In First Out (FIFO) queue which contains the traffic elements of "(len, t_{rcv})" sorted by time. "*PktDeque*" finds the first element that succeeds $t_{PktDeque}$ in each queue and clear all elements which prior to the element. Thus, the realtime traffic rate and LU can be obtained. Since the FIFO is sequentially sorted and can be saved in array, we can use binary search to find the successor of $t_{PktDeque}$ in each queue. The complexity is log_2N_{pkt}, where N_{pkt} is the number of traffic elements per queue. Therefore, the time complexity of "*PktDeque*" is

$$O\left(PktDeque\right)=O\left(TotalminLUindex\right)\bullet O\left(log_2N_{pkt}\right)\cdot \quad (2)$$

The third phase is made up of "*SetCachePath*" and "*GetCachePath*". Generally, the packet number of a stream is much higher than one, and "*GetCachePath*" is called each time a packet is received while "*SetCachePath*" is only called when the first packet of a stream is received. Thus, we can use "*SetCachePath*" to create a completely ordered list, and do binary search in "*GetCachePath*". Therefore, the worst time complexity of "*GetCachePath*" is

$$O\left(GetCachePath\right)=log_2\left(\left(2^{l_{IP}}\bullet 2^{l_{TCP}}\right)/2+\left(2^{l_{IP}}\bullet 2^{l_{UDP}}\right)/2\right), \quad (3)$$

where l_{IP} is the length of IP address and l_{TCP}/l_{UDP} is the length of TCP/UDP port. Such we have

$$O\left(GetCachePath\right)=l_{IP}+max\left(l_{TCP},l_{UDP}\right)\cdot \quad (4)$$

Hence, the total time complexity of ESB is

$$O\left(ESB\right)=O\left(TotalminLUindex\right)+O\left(PktDeque\right)+O\left(GetCachePath\right), \quad (5)$$

Thus, the worst *O(ESB)* is $O(k^3log_2N_{pkt}+l_{IP}+l_{TCP})$. In the existing Internet, l_{IP} is at most 128 [21], and l_{TCP} is equal to l_{UDP} which is 16 [22][23]. To our knowledge, the maximum bandwidth of a router interface today is less than 160Gbps [24]. Assuming $\boldsymbol{T_{interval}}$ is 1 second (long enough to calculate LU), due to the IP packet size is at least 20Bytes, N_{pkt} is less than 2^{30}. Thus in the worst condition, *O(ESB)* is still very small (less than $30k^3$). Compared to LP and ILP which is usually a non-polynomial (NP) or high order polynomial (P) problem [8, 25], ESB is a low order P problem thus can work well in realtime.

4. Evaluation

In this section, we perform the simulations in numerous scenarios to evaluate the performance of our approach in different network conditions. Then, we measure and analyze the unused link ratio (ULR) and the traffic distribution.

4.1 Simulation Environment

In this paper, we simulate 432 scenarios (144 scenarios for each algorithm) in NS2 [28]. Each scenario is made up of a certain case of each dimension shown in Table 1. The ***Host Number*** has two values which are used to simulate the lower and higher total network traffic. The ***Traffic Source Type*** can either be Pareto of which the burst interval is distributed in power law (e.g., some Web traffic), or the Constant Bit Rate (CBR) of which the burst interval is constant (e.g., live video). We also attach 7 traffic sources per host on average according to ***Traffic Dist***. Each source rate is 1Mb/s and the bandwidth of each link is 7Mb/s. The ***Host Dist*** refers to the distribution of the hosts along the routers, and the ***Traffic Dist*** refers to that of the traffic sources along the hosts. All distribution types of ***Host Dist*** and ***Traffic Dist*** are used to simulate as many cases as possible in the real network.

Table 1: The Cases Of Dimensions Of The Scenarios

Dimensions	**Case**			
Synthetic Topology	randloose	randtight	powloose	powtight
Host Number	100		500	
Traffic Source type	Constant Bit Rate (CBR)		Pareto	
Host Dist	Uniform	Exponential		Pareto
Traffic Dist	Uniform	Exponential		Pareto
Scheduling Algorithm	NONE	Round Robin (RR)		ESB

Table 2: Synthetic Topology Information

Name	**Topology**	**Router #**	**Link #**	**Prob.**	**One-Nbr.**
randloose	Pure-random	100	152	0.03	Random
randtight	Pure-random	100	491	0.1	Random
powloose	Power Law	100	151	0.03	0.25
powtight	Power Law	100	497	0.1	0.25

Table 2 shows the detailed topology information in Table 1. We set 100 routers in each synthetic topology and the number of links is also listed. The **One-Nbr.** refers to the fraction of routers which only have one neighbor. The first and the second topologies are random graphs generated by GT-ITM [27] while the others are power law graphs generated by Inet-3.0 [28] with our modification on its supported ***Prob.***. The ***randloose*** and ***randtight*** are generated by the pure-random algorithm with constant connection probability (i.e., the ***Prob.***) between each pair of routers. Hence, there are more paths between the routers in ***randtight*** than in ***randloose*** according to Table 2. The ***Prob.*** of ***powloose*** and ***powtight*** are equal to that of the corresponding random graphs such that the influence of the degree distribution can be observed. Meanwhile, there are 25% routers in ***powloose*** and ***powtight*** with only one neighbor. Notice that, our approach works as single path routing while the stream is established between the hosts that connected with these routers.

Why only Synthetic Topologies and Scenarios are included? This is because we are to find out which specific topology model (other than the approximate ones, e.g., the real topology) is more appropriate for performing our approach on average of all specific traffic models. The results can be used as a reference to estimate the performance of our approach when using into practice, e.g., a real topology with predicted traffic demands.

4.2 Unused Link Ratio (ULR)

Fig. 4 shows the ULR of all scenarios. The ULR is defined as the ratio of the number of links without data traffic to the total link number in the topology. Due to round robin (RR) or ESB may select different source and destination router neighbors for the identical host pair, compared to single path routing (i.e., NONE in Fig. 4), more links can be used to transmit the traffic of the same host pair on the incompletely overlapped paths. Hence, the ULR of NONE is always larger than RR and ESB. Due to the average link number per router in ***randtight***/***powtight*** is higher than that in ***randloose***/***powloose***, the ULR in ***randtight***/***powtight*** is higher under NONE because higher ratio of links are not on the shortest path between any communication pair, and is lower under RR or ESB because the path number between each pair of routers is three order of neighbor (thus the link) number of each router on average such that higher ratio of links can be covered by the paths in ***randtight***/***powtight***. Compared ***powloose***/***powtight*** to ***randloose***/***randtight***, the ULR in ***powloose***/***powtight*** is always higher under NONE because most shortest paths between routers pass through the links that connected with the hub nodes in ***powloose/powtight*** such that the links (major part of all links) which are not connected with the hubs are much less probable to be used than all links in ***randloose***/***randtight***. The ULR in ***powloose***/***powtight*** is also higher under RR or ESB because most routers have lesser neighbors than those in ***randloose***/***randtight*** such that the path number between each pair of routers is much lower (due to the three order relationship mentioned above) than ***randloose***/***randtight*** on average, thus more links are not used. Fig. 4 also indicates that the ULR is higher when the host number is lower in the same condition. In each grid divided by the solid and dash lines, the front part of each curve is always lower than the back part. This is because the ***Host Dist*** of the front part is uniform. Although the ***Traffic Dist*** is Pareto in some cases of the front part while it is uniform in the back part, the traffic source number is 7 times of the host in each scenario (by our setting) such that all communication paths in the front part in each grid of each curve cover more links than the back ones. In Fig. 4, we can also easily find that the ***randtight*** is most appropriate for the multipath routing scheme (the ULR reduces most sharply when using the algorithms) while both ***powloose*** and ***randloose*** have worse effects. That means, the multipath routing schem takes better effect when the number of links among routers is higher and the links distribute more uniformly on the routers. Meanwhile, the scenarios in ***randloose*** have worse effect is because the ULR of NONE is much lower than other topologies. Since RR has used practically all neighbors of the source and destination router of a stream, the ULR performance of ESB is a little lower than RR. Compared to NONE, the average reduction of ULR of all scenarios is 82.4% when using RR, and is 82.0% when using ESB.

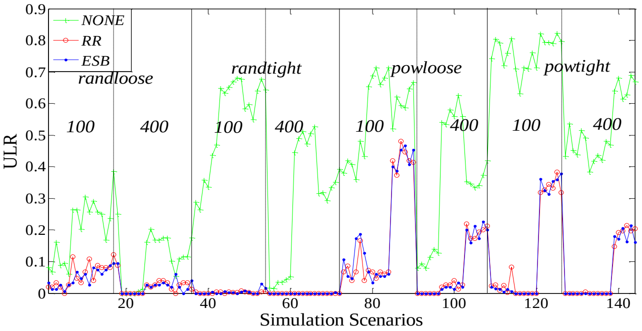


Fig. 4 ULR. The vertical solid lines separate different topologies while the dash ones separate different host number in each topology.

4.3 Traffic Distribution

Due to the total network traffic in different scenarios may not be equal (e.g., different ***Host Number***), we use the normalized link utilization (NLU) which is the ratio of LU to maximum link utilization (MLU), to present the traffic distribution of all scenarios in a same scale [17]. Figs. 5-8 show respectively the traffic ratio (TR) along NLU which of a scenario can be calculated as:

$$NLU_{link_i\,(j)}=\left(\frac{Trf_{link_i\,(j)}}{B_{link_i}\bullet T_{sim(j)}}\right)\Bigg/\left(\frac{maxLUTrf_j}{B_{link_{maxLUTrf}(j)}\bullet T_{sim(j)}}\right), \quad (6)$$

where $Trf_{link_i(j)}$ and $maxLUTrf_j$ are respectively the traffic on $link_i$ and the link with MLU during the simulation time $T_{sim(j)}$ in scenario j, B_{link_i} and $B_{link_{maxLUTrf}(j)}$ are respectively the bandwidth of $link_i$ and the link with $maxLUTrf_j$. Due to the bandwidth of each link in each topology and $T_{sim(j)}$ in each scenario are both set the same in our simulation, we have

$$NLU_{link_i\,(j)}=Trf_{link_i\,(j)}\big/maxLUTrf_j \ . \quad (7)$$

TR is defined as the ratio of the traffic on links of a certain NLU range to the total network traffic:

$$TR_{NLU_k}(j)=\sum_{i=1}^{link\#_{NLU_k}(j)}Trf_{link_{(i,k)}(j)}\Bigg/\sum_{i=1}^{n}Trf_{link_i\,(j)}\ , \quad (8)$$

where $Trf_{link_{(i,k)}(j)}$ is the traffic of the i^{th} link in $link\#_{NLU_k}(j)$ and n is the number of links in the topology of scenario j. Combining (6) and (8), we have

$$TR_{NLU_k}(j)=\sum_{i=1}^{link\#_{NLU_k}(j)}NLU_{link_{(i,k)}(j)}\Bigg/\sum_{i=1}^{n}NLU_{link_i\,(j)}\ , \quad (9)$$

where $NLU_{link_{(i,k)}(j)}$ is the NLU of the i^{th} link in $link\#_{NLU_k}(j)$. That is, TR can be calculated by using NLU.

In Figs. 5-8, since each point is the mean value of TR of all scenarios under the corresponding topology and algorithm, and the standard deviation (STDEV) is small, the average traffic distribution of the scenarios can be generally represented by the corresponding curve.

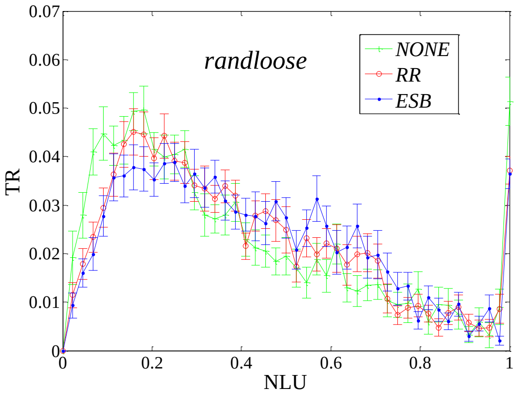


Fig. 5 Traffic Ratio.

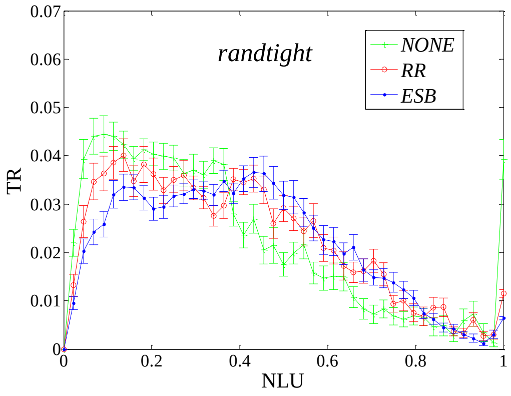


Fig. 6 Traffic Ratio.

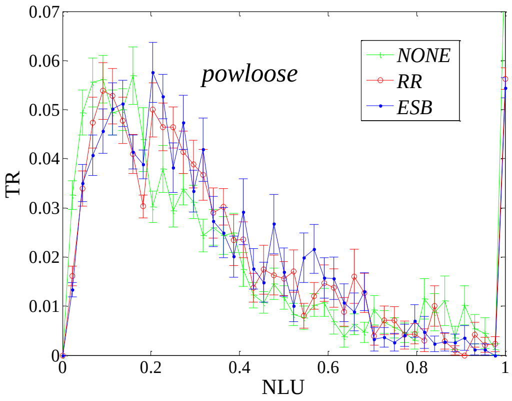


Fig. 7 Traffic Ratio.

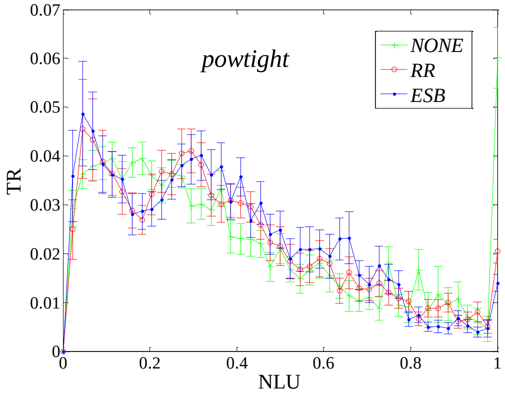


Fig. 8 Traffic Ratio.

In Figs. 5-8, compared to NONE, the TR of RR or ESB is higher when NLU nears to the middle and is lower when it nears to the two ends. This is because RR or ESB splits the traffic of the links with higher LU onto the links many of which are with lower LU, such that the LU and NLU of these links “move” towards each other. Hence, the increasing of TR in the middle part of NLU arises from balancing the LU of most links. Along with LU balancing, the increasing part of TR will be higher and move to right on NLU. That is, the LU of more links will be closer to MLU which will be smaller in each scenario accordingly. Ideally, if the TR is 1 when NLU is 1, MLU is the minimum.

In Fig. 8, we notice that the TR in the lower part of NONE is lower than RR and ESB. That is because the ULR of NONE is too high in ***powtight***. More links can only get much lower traffic in ***powtight*** than in other topologies when using RR or ESB. Therefore, RR or ESB can only make those links, which are actually “idle” in NONE, with low LU. And hence, RR or ESB make TR higher in the lower part of NLU than NONE. We use the sum of TR in the NLU range of [0, 0.1] and [0.9, 1] to present the effect of LU balance shown in Table 3. Since the LU of most links in each scenario are all still much lower than the MLU, the traffic is more balanced when the sum is lower.

Table 3: Sum of Traffic Ratio

Topology	TR (NLU in [0, 0.1] and [0.9, 1])		
	NONE	***RR***	***ESB***
randloose	24.4%	17.7%	16.1%
randtight	25.3%	17.6%	13.4%
powloose	36.5%	27.9%	25.7%
powtight	27.1%	23.4%	23.1%

In Table 3, ESB is better in each topology. This is because ESB always uses the path which has the minimum LU on the two ends currently for each stream. Compared to NONE of the 4 topologies, in the relative sense, the average reduction of the sum of TR is 23.7% when using RR, and is 31.3% when using ESB. The ***randtight*** has the best effect (reduces 30.5% in RR and 47.1% in ESB) because more paths can be used to carry the split traffic for most streams. The ***powtight*** has the worst effect (reduces 13.5% in RR and 14.6% in ESB) because the TR of RR or ESB is higher than NONE when NLU in [0, 0.1], which means more “idle” links just start to carry traffic when RR or ESB are used in ***powtight*** than in other topologies.

5. Conclusions

In this paper, we have proposed a load balance approach by separating the host and router IP addresses into two spaces. In our approach, we have proposed a scheduling algorithm, named ESB, which is used by the proposed multipath routing scheme based on the address space separation. Each router can schedule each stream that is initiated by the connected host onto the proper path to the destination host by ESB dynamically. The multiple paths between any pair of hosts can be obtained by the connected routers by using the address separating mechanism. The merit of our approach is that: it balances the network traffic dynamically while being free of traffic demand assumption and offline flow optimization. The path of each stream is selected by each router individually other than using central system based on the address separation. The time complexity of ESB is much lower than the LP and ILP which are used in flow optimization. Simulation results have shown that on average of all simulated scenarios, compared to the existing single path routing which is based on the address separating, our approach evidently reduces the ULR and in relative terms, balances the traffic across the network.

Acknowledgments

This work is supported in part by NSF of China under Grant No. 61232017 and 61102049, in part by National High Technology Research and Development

Program("863"Program) of China under Grant No. 2011AA010701 and 2011AA01A101, in part by Fundamental Research Funds for the Central Universities under Grant No. 2011JBM009 and in part by Innovation Funds for Excellent PhDs of Beijing JiaoTong University under Grant No. 2011YJS207.

References

[1] J. He and J. Rexford, "Toward Internet-Wide Multipath Routing", IEEE Network, 2008, Vol. 22, pp. 16 - 21.

[2] W. Willinger, V. Paxson, and M. S. Taqqu. “Self-similarity and heavy tails: Structural modeling of network traffic”, A Practical Guide to Heavy Tails: Statistical Techniques and Applications, 1998.

[3] M. Faloutsos, P. Faloutsos, and C. Faloutsos, “On power-law relationships of the internet topology”, in Proc. ACM SIGCOMM, 1999.

[4] A. Sridharan, R. Guerin, and C. Diot, “Achieving near-optimal traffic engineering solutions for current OSPF/IS-IS networks”, IEEE/ACM Trans. on Networking, Apr. 2005.

[5] B. Fortz and M. Thorup, “Internet Traffic Engineering by Optimizing OSPF Weights”, Proc. IEEE INFOCOM 2000.

[6] Z. Wang, Y. Wang, and L. Zhang, “Internet traffic engineering without full mesh overlaying”, Proc. IEEE INFOCOM 2001.

[7] D. Xu et. al., “Link-state routing with hop-by-hop forwarding can achieve optimal traffic engineering”, IEEE Trans. on Networking, Apr. 2011.

[8] M. Antic et. al., “Two Phase Load Balanced Routing using OSPF”, IEEE Jour. of Selected Area in Comm., Jan. 2010.

[9] S. Tsunoda, et. al., “Load-Balanced Shortest-Path-Based Routing Without Traffic Splitting in Hose Model”, Proc. IEEE ICC 2011.

[10]E. Keller, et. al., "Rehoming edge links for better traffic engineering,", ACM SIGCOMM CCR, Mar. 2012.

[11]D. Saucez, et. al., “Interdomain Traffic Engineering in a Locator/Identifier Separation Context”, Proc. INM, Oct. 2008.

[12]S. Paul, et. al., "An Identifier/Locator Split Architecture for Exploring Path Diversity through Site Multi-homing - A Hybrid Host-Network Cooperative Approach" Proc., IEEE ICC 2010.

[13]D. Farinacci, et. al., “Locator/ID separation protocol (LISP)”, IETF, Internet Draft draft-ietf-lisp-16, Nov. 2011.

[14]R. Moskowitz, et. al., “Host identity protocol”, RFC5201, Apr. 2008.

[15]E. Nordmark and M. Bagnulo, “Shim6: Level 3 multihoming shim protocol for IPv6”, IETF, RFC 5533, Jun. 2009.

[16]C. Kim et. al., “Floodless in SEATTLE: A Scalable Ethernet Architecture for Large Enterprises”, Proc. ACM SIGCOMM 2008.

[17]Sanqi Zhou, Jia Chen, Hongbin Luo and Hongke Zhang, Proceedings of 2012 World Congress on Information and Communication Technologies (WICT2012), Nov. 2012.

[18]J. Moy, “OSPF Version 2”, IETF, RFC 2328, Apr. 1998.

[19]D. Eastlake and P. Jones, “US Secure Hash Algorithm 1 (SHA1)”, IETF RFC3174, Sep. 2001.

[20] “Internet Protocol”, IETF, RFC 791, Sep. 1981.

[21] “Internet Protocol, Version 6”, IETF, RFC 2460, Dec. 1998.

[22] “Transmission Control Protocol”, IETF, RFC793, Sep. 1981.

[23] “User Datagram Protocol”, IETF, RFC 768, Aug. 1980.

[24]Cisco, U.S. [Online] http://www.cisco.com/en/US/prod/collateral/routers/ps5763/CRS-FP-140_DS.pdf.

[25]A. Elwalid et. al., “MATE: MPLS adaptive traffic engineering”, Proc. IEEE INFOCOM, 2001.

[26] “NS2”, USC/ISI, Xerox PARC, LBNL and UCB, U.S. [Online].

[27] “GT-ITM”, CC, Georgia Institute of Technology, U.S. [Online].

[28] “Inet-3.0”, CCES, University of Michigan, U.S. [Online].

Sanqi Zhou received the B.S. degree in electrical engineering and automation from North China Power Electric University, Beijing, China, in 2007. He received the M.S. degree in Traffic Information Engineering and Control from Beijing JiaoTong University, China, in 2009. He is pursuing the Ph.D. degree at national engineering laboratory for next generation Internet interconnection devices, Beijing Jiaotong University. His research interests include network traffic engineering, energy efficiency and the next generation Internet technology.

Jia Chen received her B.S. degree in Communication Engineering in 2005 from Beijing University of Posts and Telecommunications in China. She received Master of Research (M.R.) degree in Telecommunications in 2006, and the Ph.D degree in Electrical and Electronic Engineering 2010 from Department of Electrical and Electronic Engineering, University College London, UK. She worked in British Telecommunication (UK) for an industry fellowship position from Jan. 2009 to Apr. 2009. She joined Beijing Jiaotong University (Beijing, China) as a lecturer since July 2010. Her current research interests include architecture and protocol design and analysis for the future Internet.

Huachun Zhou received his B.S. degree from People’s PoliceOfficer University of China in 1986, and the M.S. and Ph.D. degree from Beijing Jiaotong University of China in 1989 and 2008, respectively. He is currently a professor with the Institute of Electronic Information Engineering, Beijing Jiaotong University of China. His main research interests are in the area of mobility management, mobile and secure computing, routing protocols, network management technologies and database applications.

Hongke Zhang received his M.S. and Ph.D. degrees in Electricaland Communication Systems from the University of Electronic Science and Technology of China in 1988 and 1992, respectively. From Sep. 1992 to June 1994, he was a post-doc research associate at Beijing Jiaotong University. In July 1994, he jointed Beijing Jiaotong University, where he is a professor. He has published more than 100 research papers in the areas of communications, computer networks and information theory. He is the director of the National Engineering Laboratory for Next Generation Internet Interconnection Devices.

Modeling a repository of modules for ports Terminals Operating System (TOS)

Ahmed Faouzi[1], Charif Mabrouki[2], Alami Semma[3]

[1] Department of Mathematics and Computer Science, Faculty of Science and Techniques,
Hassan 1 University
Settat, Morocco

[2] Department of Mathematics and Computer Science, Faculty of Science and Techniques,
Hassan 1 University
Settat, Morocco

[3] Department of Mathematics and Computer Science, Faculty of Science and Techniques,
Hassan 1 University
Settat, Morocco

Abstract

The purpose of this paper is the modeling of a repository for modules and interfaces that must include all integrated information system management of a port terminal.

Modules will provide a basic framework necessary for automatic management of internal operations and activities of all Port Terminals worldwide.

Interfaces will provide a basic framework for the management of external operations subject to standardized exchange with partner's Information Systems of the port community.
These modules and interfaces will be used for the implementation of all integrated systems for the management of a Terminal Port (TOS).

Keywords: Ports Terminals, Integrated System, Traffic, container, Vessel, TOS, Process

1. Introduction

In all countries the Port is located as a core around which we find a community in general called "Community Port": Port Operators, Port Authority, Customs, Bank, Officer Maritime Carrier, Clients ...
Overall, we can divide the activity of a port into two categories: Container traffic activity (goods carried in containers) and various traffic activity (goods made in other types of packaging: cereals, oil, wood ...).

Information System plays a very important role in the activity of a Terminal Port, firstly to ensure business operational in Port (Goods and Treatments Ships) and secondly to ensure the interface with members of the port community.
In Market of Port's Information Systems there is a lack of a common repository of all modules that must implement integrated information system for the management of container traffic, Also there is a lack of integrated system for the management of Various traffic.
This article addresses the modeling of a common repository of modules and interfaces that must include all information system for managing container traffic.

2. Problematic

✓ Lack of a common repository of modules that must include all integrated information system for Container Port Traffic,

✓ Difficulty to have a repository for electronic exchange interface to community's Information Systems

3. Terminal Port Activities

3.1 General Services

Below the mains activities in a Port Terminal :

✓ Services to goods handling aboard ship and shore handling : Container, Various good,

✓ Services to Ships: Pilotage, Towing, providing water / energy ...

✓ Storage, scoring, weighing, packing and unpacking of containers and trailers

✓ Other Services: hauling, stacking of goods, loading and unloading trucks;

3.2 Operating Procedures (Container Traffic)

3.2.1 Import process:
1. Arrival of the ship at the dock
2. Unloading import containers
3. Storage terminal
4. Delivery of containers

For this process we will be considered the following cycles:

- Cycle truck
- Cycle ship
- Movements in full land

3.2.2 Export process:
1. Arrival of the truck delivering the export container
2. Storage terminal
3. Loading the container on the vessel.

For this process we will be consider the following cycles:

- Cycle truck
- Cycle ship
- Movements in full land

3.2.3 Trucks Cycles (Import, Export)

Before the physical input of the truck terminal, paperwork must be completed, and the order "service request" delivery container inquired about the system, customs clearance must be paid.

Orders and authorizations can happen in paper format or in EDI messages.

Before the physical input of the truck, port terminal in general practice administrative control, and instruction document is generated for the truck, then follow the physical check.

Physical Control definition:

Physical control is defined as the visual inspection by a pointer that verifies container numbers, license plate information CSC (Container Safety Certificate), stickers IMO, presence and seal numbers, or any other inspection that may be required by the operator or by the contract between the operator and shipping line or other customer.

If the physical check is passed, the truck is directed to the interchange area designated by the business Information System.

In the interchange, the container will be unloaded or loaded on the truck. Of course, a trailer can load multiple containers, and therefore in the interchange area we can handle multiple containers.

Movements combined I / O must also be possible.

After treatment of the truck, it proceeds to the exit door.

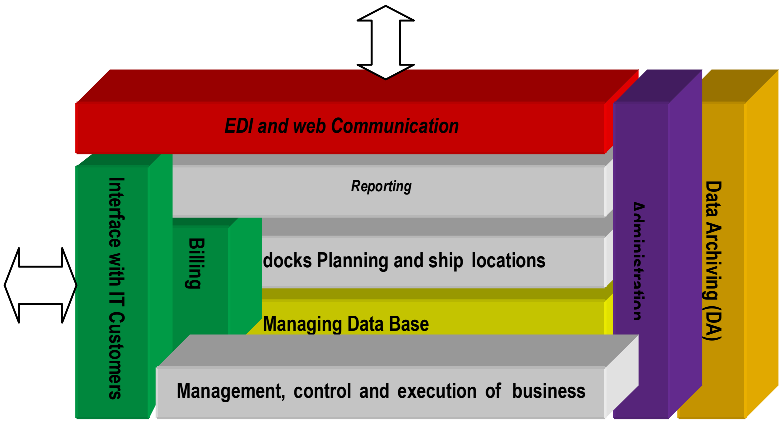
3.2.4 Ship Cycle: Import

1. Message BAPLIE: the terminal receives EDI BAPLIE plan. This BAPLIE will be associated with the ship's visit. Based on a list of containers BAPLIE discharging that may be generated.
2. Arrival of the ship after ship registration in the Information System and planning the sequence containers and cranes, unloading the ship can begin.
3. After execution of movements, the container reaches its planned position. The movements are recorded using a laptop or RTD or VMT.
4. After handling all containers the ship's visit will be closed. Handling means : all the operations by ships (import and export).

3.2.5 Ship Cycle: Export

1. Message MOVINS: the terminal receives EDI MOVINS plan. This plan MOVINS be associated with the ship's visit. Based on the presence and MOVINS containers in the terminal a list of containers to be loaded can be generated. MOVINS message contains instructions indicating the owner of the bays in the ship for some types of containers.
2. Arrival of the ship after the registration of the ship in Information System and planning the sequence containers and cranes, vessel loading can begin.
3. After execution of movements, containers reach their planned position in the ship. Movements are recorded by means of a portable computer,
4. After handling all the export and import containers, the ship's visit will be closed.

4 Modeling a repository of the Basic Modules required for an Integrated Information System of Port Terminal (TOS : Terminal Operating System)


EDI and web Communication
Reporting
docks Planning and ship locations
Managing Data Base
Management, control and execution of business
Interface with IT Customers
Billing
Administration
Data Archiving (DA)

Modeling TOS System modules

4.1 Module 1 : docks Planning and ship locations:
The TOS (Terminal Operating System) must integrate these modules:

- ✓ Planning Docks (Location vessel)
- ✓ Allocation of Human resources and equipments resources (the ship)
- ✓ Planning Ships
- ✓ Planning full land
- ✓ Planning and Control gear

4.1.1 Planning docks and Resource allocation
The planning docks module and ship locations (BP, Berth Planning) allow to plan (at short and at long term) the arrival and work of the ships "human and equipments resources'. It must give a correct view in real time about the current ships arrival and operative portal vessels.

- ➢ BP module consists of an array of planning. One axis represent ships arrival and second axis represents time. The BP should allow seeing the current time, the history and future situations.
- ➢ The physical data of each ships arrival must be registered in the BP system: water depth, maximum vessel tides;
- ➢ BP module simulates the effect of attracting additional customers, and whether the terminal has capacity to dock, where this ability is? and when this capacity is available ?

General objective: the planning should be in order of priority:

1. The gantry will work to maximum rate during loading or unloading. This means delivery / Evacuation containers synchronized with the rhythm of the gantry.
2. Distances traveled by vehicles should be as short as possible
3. Port Machines should not circulate without container: no unproductive trips
4. The capacity utilization of machines, including RTG machines should be balanced. This means that the number of movements assigned to RTGs will be more or less equal by RTG.
5. The number of unproductive movements should be as small as possible: the TOS must avoid "Shifting".

The basis for effective operations is the availability of accurate information before and during the operational cycles.

Computer systems must have interfaces in order to receive information from business partners and authorities. We mean by the authorities port authority, harbor and customs.

All movements of containers on the terminal are programmed and planned on the basis of information on container handling: orders for loading, unloading, transfer visit, delivery or receipt portal ....

All planning must be based on the "next container move", to avoid non-productive movements in the terminal. For example, planning a container transshipment received is based on trip data with which the container leaves the terminal.

The TOS should plan the container so that the inverse image of the planning will be obtained on the ship terminal. This prevents repositioning unproductive.

The TOS should allow the use of various planning strategies on full land, depending on the composition of stacked containers and means of transport

The general principle to plan the locations of containers is as follows:

1. The operator defines filters to group containers in collections. The criteria to filter the containers can be: container size, weight, service, port of loading, port of destination...
2. The operator sets for containers belonging to a collection where they will be stored, so the operator defines storage areas. The definition must be able to perform each type of transaction (import, export, transshipment), by service ship or a ship visit.
3. Defined area or by the whole terminal, the operator defined how to stack containers by area: containers heaviest over lighter containers, containers for the same destination port on the same destination port, containers for the same container ship on the same ship ...

The operator must therefore assign a storage area for each collection. In the case where a container cannot be "filtered" in a collection, the operator must manually plan the container

Regarding the strategy to assign zones collections, editor must implement at least the following strategies (based on logical storage areas in the TOS or virtual, not on physical breakdown on the terminal):

Strategy 1: Group all containers belonging to the same collection in an area.
This option makes sense and is a good strategy for Import Operations. This strategy puts a lot of pressure on the RTG machine, it must handle a large number of containers, which will limit the performance of the gantry as the technical performance of an RTG is less than the technical performance of a gantry, the RTG will not keep pace with the gantry. In addition this strategy has a consequence: in case of failure of the RTG machine all operations may stop.
Strategy 2: Distribute containers belonging to the same collection to areas distributed: Scattered Planning (SC)
In the case of SC, the system distributes the containers through several zones. Consequently, the amount of work or number of movements will be divided by number of RTG machine. This is logical, because the technical performance of an RTG is lower than the performance of a gantry. The TOS must ensure that the distances between gantry and RTGs is not too high because it will meet the second principle of the Strategy 1.
Strategy 1 is considered the best for a terminal in which traffic Import premium
Strategy 2 is considered the best for a transshipment terminal.

Planning and storage IMO containers (International Maritime Organization) on container terminals is based on the rules of segregation IMO and if applicable, on the national or local rules.
IMO (International Maritime Organization) rules classify containers, based on their content in classes. For each class, rules of segregation were defined by the IMO.
These rules essentially segregation specify if a minimum distance must be maintained between the containers belonging to certain categories of IMO, if the containers can be stored on the terminal, etc.. These containers will be routed to the dedicated slots for storage containers IMO.
Containers "non-waterproof" are another special case. For these containers dedicated physical installation is required.

Operational procedures associated with planning strategies allow the terminal benefit from the expertise of TOS.

4.1.2 Planning Ships (PS)
The TOS must allow planning ships as follow:
- Registration of ship characteristics: physical dimensions, capacity, crane positions, cockpit, stability data and requirements in order to optimize the loading of the ship, heights and maximum weight per container stack.
- Integration of EDI messages (BAPLIE and MOVINS) editing and corrections if necessary EDI messages before applying them on the ship's structure to processing
- Simulation of different work scenarios. The operator will receive information on the effectiveness of the calculated planning systems. The operator can adjust the settings accordingly by planning, taking into account the parameters and stability requirements. Final planning will be stored on CD or USB stick and will be transmitted to the master for verification and approval
- Assignment of an optimal number of gantries to the ship in order to achieve the contractual terms of the handling operation to finish before the ETD (Estimate Time Departure) of the ship.

PS module will consider the positions of the containers terminal and minimize repositioning of the terminal and reduce the distance by port equipment.
PS module will consider in planning the availability of gantries and port equipment and notify operators of resources to achieve handling in a time.
The module should include the ability to plan and manually change the schedule for automatic containers selected by the operator.
The module will consider the IMO rules, adapted by operators during the planning of ships.
PS module allows the monitoring in real time all operations vessel.

4.1.3 Planning full land (PL):

The TOS should allow the distribution of terminal storage areas following names adapted to the requirements of the operator: buildings, roads, intersections, arrival ship, areas Reefers portals, Interchange area, and any other structure on the terminal.

The operator must have functions to modify the plan of the terminal in real time and without the intervention of the software. PL module allows you to define areas not available due to a scheduled maintenance.

PL module should allow the definition of rules for the planning of containers on the terminal based on different criteria such as:

- Fixed Fields import, export
- Hotspots next service "vessel call" etc…
- Way storage container
- Strategies for group containers following example weight, port of destination, consignee
- Storage Strategies to avoid port congestion gear
- Storage Strategies to reduce distances by the craft port.

PL module should have functions for:
- Show real-time capacity on the terminal
- The future capacity of the terminal
- Adapt the rules for selecting storage areas, as well as how to stack containers in real time and to notify the operator before unloading if, based on simulation, the rules are not sufficient to storing the containers.
- The recording positions of containers using GPS or similar

PL module provides functionality "drag and drop" to manually change the positions of containers.
PL module provides functionality to reorganize the terminal operations to prepare the ship loading or unloading and trucks operations.

PL module communicates in real time with other TOS modules, and contains modules for specific Reporting planning positions of containers. The module will consider the IMO rules adapted by operators.
5.1.4 Planning and control gear (PG):
Module planning and control gear allows the detailed specification of all types of port equipment such as RTG machines, the reach stacker, trucks, trailers.. with all their technical characteristics and handling capacity (maximum stacking height, maximum speed vacuum speed with full container, lifting capacity, GPS sensors and weighing..).

The definition of gear also contains data on the patterns movement and links between machines in the circuit control PG module connections between individual movements and synchronizes the steps most effectively "Example To move a container, in the first RTG machine must load the container on a chassis, The chassis must be transported by a tractor to the gantry, The gantry will unload the container from the chassis"

The PG module transmits instructions to terminals RTD Radio Frequency (Radio Data Terminal) installed on the gantry. The driver will confirm the execution of an instruction and receive the next instruction.

The PG module contains the history of all the movements of each device port, the duration of each movement as well as dependencies.

PG module evaluates in real time the position of movements and optimizes the individual instructions to be made and the use of all equipments.
In case the machine wills no instructions, the system informs the operator immediately so that they can assign equipments to work when needed. PG module allows to assign equipments to specific operations.

The workflows are automatically assigned according to deadlines or manually by operators. The PG system receives other sequences from the planning vessel module.

The PG module allows generating internal reports productivity and provides an interface with CMMS (Computer Maintenance Management Systems).

4.2 Module: Managing Data Base (Kernels)
Data bases (DB) should be centrally managed and controlled, and basic data will be used by all modules. Data such as a directory of ports, ISO containers codes …
The TOS must include categories of data:
1. Commercials partners: ships owners (operators of ships, NVOCC Non-Vessel Operating Container Carriers), shipping lines, agents, trucking companies; any other party may receive invoices for port terminal services.
2. Transport Data: ships, fields, trucks, Ship structure, such as physical characteristics, weight…
3. Maritime services and port rotation
4. Ports directory following conventions
5. Container directory types according to ISO rules
6. Terminal Operators
7. Work shifts, hours of operation standard and overtime
8. Definition equipment and technical
9. Staff
10. System users: roles, access authority (editing, creation, consultation) and functional safety
11. Overall system configuration
12. Terminal Definition

4.3 Module: Data Archiving (DA)
This DA module should include Archiving Strategy:
The purpose of archiving is to maximize the system's disk space of the TOS and also optimize TOS efficiency and response time.
The archival frequency times depend on legislation in the country.
The DA module must contain 3 archiving strategies:
1. Operational archiving (data date: 2 to 3 years): Data are available on the operating system and are available through the user TOS interface.

2. Mid-term archiving (data Date: 3 to 10 years): Data are available on a medium that is not part of the operating system, but can be consulted by the user TOS interface.
3. Long-term archiving (date of data > 10 years): Archiving is done in an outside medium searchable by SQL or TXT tools.

5.4 Module: Management, control and execution of business processes
This module must include:

✓ Administration of orders
✓ Monitoring and execution

The execution part is an integral part of operational modules as described in the previous pages. This module manages a comprehensive operations and procedures of container terminals Mainly it provides all the user interfaces of recording, and allows the operating procedures. All cycles, links with the authority, customers and partners are scheduled at this level.

4.4.1 Administration orders (AD)
Administration orders <or Services request > allows to create, modify and cancel orders if necessary.
Orders are in effect requests services from partners. Service requests are typically the list of services to clients:
1. Receiving a full container delivered by truck
2. Delivery of an empty container to a truck
3. Full container delivery to a truck
4. Receiving an empty container
5. Request to load a container on a ship
6. Request of unloading a container from a ship
7. Request to receive a ship terminal, arrival time, departure time,...
8. Request for verification of lead
9. Demand for warehousing of lead line shipping
10. Application for the condition of a container
11. Application control and warehousing stickers IMO
12. Request CFS operations
13. Request fumigation, cleaning etc..
14. Demand for transportation equipment repositioning or land full
15. Change request travel / service vessel
16. Inspection request (initiated by the customs)
17. Asked to provide specific services to vessels
18. Preparation request refrigerated containers (Pre-tripping)
19. For reorganization cargo ship (Restows)
20. The ability to register and manage bottlenecks imposed by customs or other authority, interface with Customs

In principle, any service provided by the port terminal must be registered in the system TOS. Registration application is the basis for the planning, implementation and ongoing operations.

4.4.2 Monitoring and Execution: (ME)
All automation process and management for these orders must be implemented in the TOS, in this part will be integrated orchestration, editing interchange document, compliance monitoring, management rules…

The ME module is integrated and interfaced with planning module, EDI communication, WEB Client Access, Reporting, and billing.

The module must allow the execution of any instruction generated by the TOS in real time and plan subsequent movements based on the confirmation of execution of the previous movement.

Examples:
1 / for the AD order on the Reefers This ME module will provide interfaces to allow RTD terminal to receive instructions for connecting and disconnecting refrigerated. The interfaces must also input temperatures of refrigerated; it will return data to the client.
2 / For the vessel treatment, ME module will integrate automatically EDI messages like : Bayplan MOVINS, PRESTOW. It will provide user interfaces for stops input, board and land score, orchestration modules with planning, communicate orders for tractors and RTG, validate positions containers, refunds performance handling, editing documents and reports, return EDI messages like : COARRI, CODECO (if direct outputs), …

4.5 Module: Communication and web access

✓ EDI (Electronic Data Interchange) Communications
✓ Interface WEB Extranet

The TOS must integrates EDI communication modules (Electronic Data Interchange) and WEB.
The EDI module should allow the exchange of structured messages between the port terminal and its partners and can handle all standard EDIFACT messages used in practice in the field of handling and processing of container terminal (Import / Export / Transshipment).

The EDI module should define the partners with which the port terminal communicates. The EDI module allows you to modify EDI messages according to need and standards in accordance with the partners of the port terminal.

The EDI module must have a web interface allowing partners to not only view data on their operations, but the interface must also allow partners to record data and orders (service requests).

WEB features contain all devices to ensure a good systems protection: applications, databases, users profiles, access permissions to some or several TOS modules and data

4.6 Module: Reporting (RE)

4.6.1 Reporting of each module

Each TOS module must have its own Reporting functions. The TOS should have standard reports and must also allow authorized users to create their own reports.

The TOS should allow the printing of reports and the saving reports in standard formats (Microsoft Office ®) and sending reports by e-mail.

Reports should be able to compile information from various TOS modules.

As Port Terminal must have the ability to generate their own reports using a reporting tool or Business Intelligence (BI) of his choice, the TOS must provide complete documentation of the TOS database.

for a real-time Reporting, the TOS must have modules showing in real time the status of operations. This module allows users to define their own key performance indicators (KPI).

This module must generate views of the terminal indicating the performance by area and by operation type. The user can click a view details for domain, area of operation, equipment movement.

RE module must be available through an internet connection (HTTPS) for only authorized users. The terminal can manage users without the intervention of the TOS editor.

Port Authority (under the agreement) requested statistics like :

✓ Volumes handled,

✓ Dangerous Materials,

✓ Work accidents,

✓ Productivity and performance according to the guidelines and definitions established by Port Authority.

4.7 Module: Billing (BI)

The system must allow TOS to charge all customers services according to the contracts and prices recorded in the TOS.

✓ Billing by customer contract

✓ Billing Services recorded in the TOS

✓ Billing Services not registered in the TOS

✓ Sending Electronic Invoice

✓ Generate Report and Statistics

Billing module allows recording all customer data and all identifiers according to the country concerned.

Contracts must be recorded in the module BI and contain at least:

- Types of services to be billed,
- Validity of the contract, discounts, ...
- Benefits
- Rates
- Periods
- Discounts

The BI module allows managing a standard contract (default) which will be used if there is no contract for a client or a group of clients.

BI module allows billing in advance and receive advances. The BI will consider the advances at the time of the creation of the final invoice.

BI modules allow billing occasional services which are not recorded in the operational modules of the TOS. BI modules can send invoices in an electronic format or by EDI.

Functions statistics and reports are available in the BI modules.

The BI module allows analyzes based on cost and revenue referring to customers contracts, and allows evaluating the profitability of currents and futures customers.

4.8 Module: Interfaces with Information System of Customer

In general several TOS clients have already their own billing Information System and they will not use the integrated BI TOS module so TOS should allow a billing interface with this Information System,

Exchanges between the Billing Information System of Port Authority and the Information System of terminal (TOS) is mainly related to billing requirements and integration management control, the TOS will provide to the Billing Information System all data needed to billing :

✓ Handling

✓ Storage

✓ Special Operations

✓ Vessels and containers services

The system shall provide to the Billing Information System an access to customer data, specific contracts, to allow customization of billing.

The system will provide in real-time a file data interface, and will provide to the billing information system special orders to unlocks/locks "customers, container"

The system will manage the invoice sequence number between the Billing Information System and the TOS module BI.

4.9 Module: Administration

This module must include:

- ✓ Create Users,
- ✓ Create Groups of users
- ✓ Create Roles with Modules and sub modules
- ✓ Access Control modules
- ✓ Control access to the database (see Fields)
- ✓ Logging Operations
- ✓ Module Performance and High Availability
- ✓ Load Sharing Mode
- ✓ Backup Mode
- ✓ Error Handling
- ✓ Administration and Supervision
- ✓ On line documentation

5 Conclusion

This work was performed on findings observed in the field of the Port Terminal and following loan business operations at the Port.

We met several difficulties along our realization, e.g.: access to business data, process development activity, the release of critical activities, research common modules between the ports of the world...

The common repository described in this article, which set the modules that must incorporate all Port Information System is the result of our work, firstly this work gives a hand to any editor -wishing to invest in port’s Information Systems development- by having a repository of basic modules to implement, and secondly this work gives to customers ‘Port Operators’ buyers of integrated software solution a repository for help choosing of a market solution.

Modeling modules for an information system of various traffic is a perspective that can be the subject of another research.

References

[1] ERP et conduite des changements, Jean-Louis Tomas, Yossi Gal Collection: Dunod, 2011 - 6ème édition
[2] Norme EDIFACT, Nations Unies
[3] L‘Efficience commerciale en B to B, Christophe Bénaroya , Henri Lagrasse , Editions EMS
[4] www.portstrategy.com
[5] A new approach to supply chain management based on pooling ITIL and APICS Principles and Practices. Abdelaali Himi, mustapha el masbahi, Samir Bahsani, Alami Semma in IJCSI
[6] The concept of the value chain - http://chohmann.free.
[7] IT Service Management According To The ITIL Framework Applied To The enterprise value chain. Abdelaâli HIMI, Samir BAHSANI, Alami SEMMAA in IJCSI
[8] Gervais M., 1995, "Business Strategies", Economica
[9] www.imo.org International Maritime Organization
[10] Marsa Maroc 2012, Manuel of procedures.

Research on the Multi-view Point 3-D Clouds Splicing Algorithm based on Local Registration

Daoming FENG

College of mathematics and computer science, Xinyu University
Jiangxi 338000, china

Abstract

The paper proposed a new 3-D measurement point cloud splicing algorithm. The algorithm utilizes registration ideal in model identification technology to realize unbound and accurate splicing of 3-D data. First, sample the overlapping areas in the two 3-D point clouds which need to be spliced. Carry out pre-processing over the sampled point cloud with principal analysis method based on the statistic theory. Through extracting the feature vector that could best indicate the point cloud information, it realizes the dimension reduction for data. Then, apply improved iterate corresponding point algorithm to the sampled point cloud data which has realized pre-registration to achieve accurate registration. In the process, the set of progressive decrease of iterate condition by different levels reduced the iterate times. The utilization of new comprehensive distance measurement function effectively increases the accuracy and robustness of overall iterated convergence. Finally, apply the transformation parameter based on local sampled point cloud calculation to the entire point cloud splicing and achieve the accurate registration of multiple sampled point cloud. In the end, the actual test proved that the algorithm boasts high splicing accuracy with high overall convergence robustness, few convergence iterate times and strong anti-noise capacity.

Keywords: *3-D scanning, registration, free view point, iterate closest point method.*

1. Introduction

With the development of non-contact optical 3D scanning technology and the application expansion in the field of medical care, entertainment, manufacturing, testing and reverse engineering, panoramic holistic measurement becomes more and more important [1]. However, optical scanning system based on the characteristics of WYSIWYG, limits the use when it could not obtain the entire information of an object surface caused by obscuration or blind spots. So, in order to obtain full information of the measured object, all-round ad multi-angle scan measurements for local measurement data splicing has been the focus of the study, which is also a difficulty.

For some special applications, such as the human oral teeth three-dimensional information direct collection, three-dimensional scanning of the human body, tiny objects measurement and multi-angle measurements which do not support marked point pasting or shaft fixation, it is very difficult to splice and using traditional methods could not be very satisfactory. In this paper, based on analyzing the special nature of problems, it proposed a free camera-oriented unconstrained "two-step" point cloud splicing algorithm. The method does not require any auxiliary signs pasted on the surface. It stitches multiple point cloud data for free angle scanning and uses N side method for pre-splicing. The obtained conversion parameters provides subsequent fine splicing with a good initial value, and then it uses improved ICP algorithm to ensure global splicing convergence accuracy. Finally, the algorithm is validated by actual tests that the algorithm boasts reliability and splicing accuracy for free angle scanning point cloud and situations without distinct feature information

2. Relevant Algorithm Study and Analysis

Current splicing techniques can be broadly divided into mechanical shaft splicing paste the identified splicing and splicing algorithm based on closet point convergence algorithm. These three methods have their own advantages and application conditions.

Iterative closest point (ICP) splicing algorithm is proposed by Besl and McKay [7]. It is a geometric structure-based splicing and registration algorithm. The algorithm first identifies the correspondence between the point clouds that need splicing. It then uses repeated iterate method to reduce the distance between two groups of point clouds and gradually reduces errors. It then finds a group of geometric transformation matrix with minimum square variance soas to make the two point cloud to be spliced after geometric conversion [8]. The method does not require the additional features and it could achieve two objects splicing. However, ICP algorithm still has its own shortcomings: (A) in order to avoid falling into the local minimum mis-registration, the algorithm needs to have a good initial estimate for conversion parameters (B) when the two curve surfaces(which need registration) data are large, there would be large calculation for searching the closest point, leading to time-consuming consequence; (C) on noise-free data, it could not guarantee to obtain the correct results, to converge to the global (or even local) minimum, so it has poor robustness.

3. Algorithm Descriptions

This paper presents a highly efficient, high-quality splicing algorithm. The algorithm generally requires two steps. First, carry out N side pre-splicing for sampling point cloud and largely overlap the two point clouds to achieve the rough splicing. Then use ICPP(iterative closest plane and point algorithm) for pre-splicing model to ensure global convergence, reduce iterate times, realize accurate splicing and make the two point clouds fully overlap. Finally, convert the two splicing into parameter and use it to the entire data to realize the highly accurate seamless splicing.

3.1 Improved principal component analysis to extract feature vector

The idea of principal component analysis originates from the Karhunen-Loeve transform. The purpose is to find the optimal set of unit orthogonal vector base (that is the principal element) by the linear transformation. Use their linear combination to reconstruct the original sample, and to minimize difference between rebuild samples and the original sample. The principal component analysis is based on the training sample set second-order statistics. In fact, it refers to the de-correlation on second-order statistics.

PCA (principal components analysis) is based on the Euclidean distance metric unsupervised learning methods. As the most commonly used data dimensionality reduction method, PCA is widely used in pattern recognition [1, 2], face recognition, which is an effective statistic technique i image understanding field. It is also used for low-dimensional data to indicate features of high-dimensional data. It is used to distinguish the characteristics of data and highlight similarities and differences of the data. Therefore, for high-dimensional data, which is not easily distinguishable from the intuitive way, PCA is a powerful tool to analyze these data. With PCA calculation, you can find out about the spindle distribution of the point cloud model distribution.

The whole calculation process of PCA is presented as follows: If R is the correlation matrix of a n-dimensional input vector X, i.e., $R = E[xx^T]$, the eigenvalues $\lambda_i \geq \lambda_{i+1}$ are in descending order. The corresponding feature vectors are $\omega_1, \omega_2, ..., \omega_n$ respectively. We wish to find an orthogonal matrix W so that W is the diagonal matrix of the transformed matrix of X:

$$W = [\omega_1, \omega_2, ..., \omega_n]^T, \quad \omega_i \omega_j^T = \{1\ i=j; 0\ i|=j\}$$

Then $y_j = \omega_j^T x$, j=1,2,…,n, in which y $_j$ is the projection of vector X in the principle direction represented by ω_1, namely, the principal component. By selecting m feature vectors corresponding with larger eigenvalues, we transform the low-dimensional vector y<Rm into a high-dimensional x<Rn, m<n.

Step 1: get the data.
Take samples of two three-dimensional point sets and make one of them the targeted point set $Pt1(x_i,y_i,z_i)$, another the reference point set $Pt2(x_j,y_j,z_j)$, the number of point in the two point sets are N_i and N_j respectively.

Step 2: calculate the barycenter of the targeted point set $Pt1$ and the reference point set $Pt2$.

$$W_{pt1} = \frac{1}{N_i}\sum_{i=1}^{N_i} Pt1_i$$

$$W_{pt2} = \frac{1}{N_j}\sum_{j=1}^{N_j} Pt1_j$$

Step 3: calculate the covariance matrix C_{pt1} and C_{pt2} of the targeted point set $Pt1$ and the reference point set $Pt2$.

$$C_{Pt1} = \frac{1}{N_i}\sum_{i=1}^{N_i}(Pt1_i - W_{pt1})(Pt1_i - W_{pt1})^T$$

$$C_{Pt2} = \frac{1}{N_j}\sum_{j=1}^{N_j}(Pt1_j - W_{pt2})(Pt1_j - W_{pt2})^T$$

Step 4: calculate the eigenvalues and feature vectors of the covariance matrix of the two point sets.

$$Evu \cdot C = Evr \cdot C$$

Sort the eigenvalues in descending order $Evu_i \geq Evu_{i+1}$. As this system deals with three dimensional point set, we choose feature vectors *Pt1_Evr*$_m$, *Pt2_Evr*$_m$(m=1,2,3) that corresponds with the top 3 non-zero eigenvalues as pivots. Then we can use low-dimensional subspace of principal component to describe the former space.

Step 5: create two transformed matrix TR1 and TR2 in three-dimensional vector direction

$$TR_1 = \begin{bmatrix} Pt1\_Evr_1.x, Pt1\_Evr_1.y, Pt1\_Evr_1.z \\ Pt1\_Evr_2.x, Pt1\_Evr_2.y, Pt1\_Evr_2.z \\ Pt1\_Evr_3.x, Pt1\_Evr_3.y, Pt1\_Evr_3.z \end{bmatrix}$$

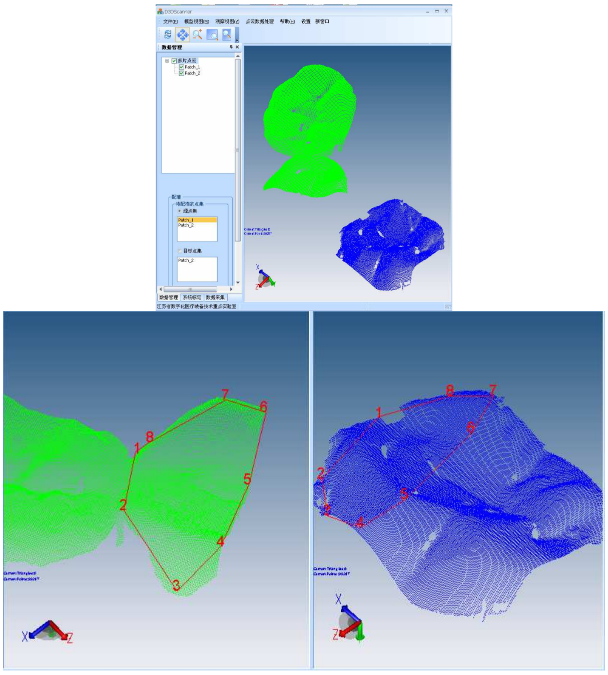
$$TR_2 = \begin{bmatrix} Pt2\_Evr_1.x, Pt2\_Evr_1.y, Pt2\_Evr_1.z \\ Pt2\_Evr_2.x, Pt2\_Evr_2.y, Pt2\_Evr_2.z \\ Pt2\_Evr_3.x, Pt2\_Evr_3.y, Pt2\_Evr_3.z \end{bmatrix}$$

Step 6: describe targeted point set Pt1 and reference point set Pt2 in space of principal component by Pt1' and Pt2' _temp:

$$Pt1' = TR_1 \cdot Pt1, \quad Pt2'\_temp = TR_2 \cdot Pt2$$

Calculate the translation matrix T1=Wpt2- Wpt1 between the two point sets and we can get the needed point set by coinciding the two bary-centers in the pivotal coordinate system.

Based on this principle, this system interactively selects two parts which point clouds overlap as targeted point set and reference point set. As picture 1 demonstrates, the two pictures are parts of the point clouds. Image a.b is the sample data stored in targeted point set Pt1(xi,yi,zi) and reference point set Pt2(xj,yj,zj)



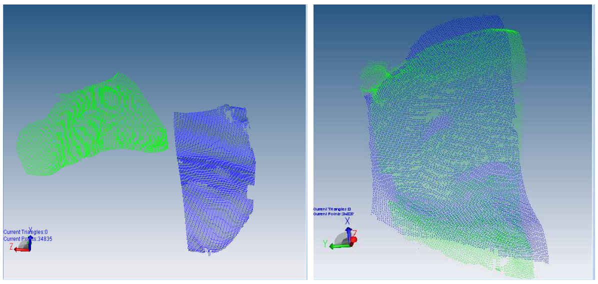


Figure 1 image of samples in data system of those needing registration

Given the statistics of the proportion and calculative proportion of principal components in total variance, we can see that the dominant accumulated contribution are made by the top 3 pivotal components, providing enough pre-splicing conditions for follow-up Iterative closest point splicing.

After primary location registration, we compare the two point clouds generally. Due to the complexity of three-dimensional geometric model and the accuracy of registration, fine-tuning is required after general registration.

3.2 Improved ICP Algorithm Used for Accurate Splicing

Traditional ICP (iterative closest point) algorithm [7] is highly restrictive, with the limitations of the initial positions. It is not applicable to two point cloud with large distance. In addition, it is slow in convergence, with poor efficiency. Therefore, many scholars have proposed improved ICP algorithm. Document [10] proposed an ICL (iterative closest line) algorithm. The registration is performed through a direct connection and finding of the corresponding segment of the two points clouds. But it could not guarantee to the correspondence relationship between the line segments; Document [11] uses the normal plane of the point in the first surface as the distance metric function, due to the contradictions among different corresponding control points, the convergence rate is relatively slow; document [12] uses the cut plane of the point to approximate the point cloud. The target distance metric function is simplified to the least squares distance from the point to the tangent plan. In this case, through the small angle approximation rotation, it could achieve the conversion. When the initial positions of two points set are very close, with some relatively low noise, the pint-to-plane error distance measure can achieve the second convergence. However, if the distance of the initial positions of the two point clouds are distant, or point cloud noise is large, the algorithm will lead to the error convergence of the ICP algorithm; document [13] proposed to regard the distance from the point to a local triangle food point of another curve surface as the distance metric function; document [14] proposed the mixed local curve rate and weighting distance among points, the two distance measuring methods. It obtains a second-order approximated square distance function. The distance is from point pi to the curve surface Q the distance. It is able to obtain quadratic convergence when two point clouds are close. It could also realize relative fine linear convergence for point clouds with distant initial positions, but it could not guarantee the global optimum.

According to the study, ICP algorithm relies heavily on selection of corresponding point and setting of target distance measurement function to effectively converge and achieve global optimum minimum of objective function. Therefore, we propose the point-to-point distance metrics and point-to-plane and point overall distance metrics. By setting iterative termination condition through different levels, it reduces the iterative termination times. It uses kd tree-based search algorithm [15] to repeatedly sample corresponding point set to reduce the search time of the corresponding point, ensuring the correctness of the corresponding point selection and ultimately achieving correct global minimum convergence acquisition.

(1) The Initial Level Convergence

As for *P*$_{1,}$ N the local point cloud data of N deformation sampling, Q1 gets the proper amount of point set based on geometric curvature s and regards *P*$_2$,*Q2* as set pair of corresponding points. First, it sets a relatively large termination threshold 0.1mm in order to reduce the number of iterations and further narrow the distance between the point clouds.

Suppose pi, Qi (i = 1, ..., N) are points in point clouds P2 and Q2. Theoretically, it is believed there exists such rigid body transformation between (p_i,q_i) when two surfaces splicing is completed.

$$\|Tp_i - q_j\| = 0$$

However, the equation ca not be established because the actual measurement data obtained can be influenced by calibration error, system accuracy, and noise data. Therefore, we need to first set an infinitesimal value accepted by the precision so that the equation can hold water. Set the limit value e and you can easily obtain the transformation matrix space by minimizing the following value:

$$e = \sum_{i=1}^{N} \|Tp_i - q_i)\|^2$$

However, the seeking of corresponding point pairs is difficult and time consuming, especially for non-regular curved surface. The paper first fixes the curved surface P_2. Deploy the k-d tree search algorithm [15] to find the nearest point of the minimum distance in P2 to Q2:

$$q_i = q \mid \min_{q\in Q} \|Tp_i - q\|$$

Assume that TR0 is the obtained position after the initial N-gon method. With the increase of iteration times, you should substitute the obtained data from last iteration T (k-1) into the iterative formula so as to seek q_j^k :

$$e^k = \sum_{i=1}^{N} \|T^k p_i - q_j^k)\|^2 ,$$

In the above formula, $q_j^k = q \mid \min_{q\in Q} \|T^{k-1} p_i - q\|$. From the above analysis, we can see that the point set obtained through minimization is a digitalized surface model. If we approximately replace the curved surface with a point, then the problem will be greatly simplified. During the initial iteration cycle, if distance between the point clouds is unknown or relatively big, we can deploy the point-to-plane distance metric function [12] in that it can renders better linear convergence in the case of far distance to point cloud position. As shown in the following Diagram 4, we can achieve conversion through the small-angle approximation rotation. Assume that the tangent plane at point q_i is represented by s_i , then the distance metric function ca be expressed as follows:

$$e = \sum_{i=1}^{N} \|Tp_i - q_j^{'})\|^2 \text{ Wherein, } q_j^{'} = q \mid \min_{q\in S_J} \|Tp_i - q\|$$

The distance between the point and the surface can be expressed by a linear function:

$$e^k = \sum_{i=1}^{N} d_s^2 (T^k p_i, S_j^k) ,$$

$$S_j^k = \{s \mid n_{qj}^k \cdot (q_j^{'k} - s) = 0\} ,\quad q_j^{'k} = (T^{k-1} l_i) \cap Q$$

In the above formula, d_s is the distance between the point and the plane; $n_{qj}{}^k$ stands for the normal vector of surface Q_2 at the point $q^{'k}{}_j$; $l = a \mid (p_i - a) \times n_{pi} = 0$ refers to the linear tangent vector at the point p_i, n_{pi} represents the normal vector of the surface *P* at the point p_i; $(Tl_i) \cap Q_2$ is the insertion point of the line *l on the* surface of Q2.

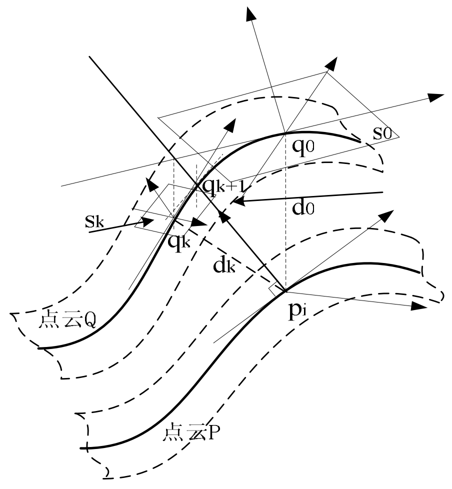


Figure 2 Analog Figure of the Convergence Approximation of Metric Function

This algorithm gives no brooding to special corresponding points, thus eliminating the problem of slow convergence. Besides, the system has set a relatively relax condition for convergence termination: 0.1 mm, which greatly reduces the number of iterations. After the step-by-step iterative approximation, we can obtain the final conversion matrix parameters. In order to verify the efficiency of the algorithm, we compare the algorithm of this system and the metric function algorithm based on point-to-plane distance as well as the traditional ICP algorithm.

(2) Second-hierarchy Convergence

Establish corresponding relation between the corresponding points seeking and the corresponding transformation parameters. If the correspondence relationship parameters $\{c_1,\ldots,c_k\}$ are known, we can easily find the nearest corresponding point of each point with these parameters. On the contrary, if the cloud data of the two point clouds are in the best conversion splicing state, the conversion relation of the two point clouds is to find the splicing conversion parameters. Therefore, based on the established splicing conversion relation $TR_{1,}$ this system re-samples corresponding point sets P3 and Q3. Then it searches for the corresponding nearest points of these sampling points among the local parts around the reference sampling set. After these steps, we can reduce the search time to less than the O(LogNx)of the global search by k-d tree. The detailed implementation is as follows:

Points p_i'=(x_1,y_1,z_1) and q_j'=(x_2,y_2,z_2), $(i=1,\ldots N_q)$ are respectively two corresponding point set pair based on the approximate conversion re-sampling. The Euclidean distance between the two points between the is $d(p_i',q_i')$, $d(p_{i,}q_i)=\|p_i-q_i\|=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$,. The initial iterative is set as T_0=[1,0,0,0,0,0,0]^T; the number of iterations, k = 0; the spliced vector obtained in the iterative splicing process is defined as a sequence: $t_1,t_2,t_3,\ldots t_k$. The specific implementation steps can be expressed as:

1) Obtain the target point cloud P_3(p_i<P,(i=1…Np))and the reference point cloud Q_3(qi<Q,(i=1,…Nq))
2) Based on the initial transformation parameters TR_1, resample N_p points of P_3 in Q3 and mark the corresponding point set Q4;
3) Initialize the data: X0=P3, T_0=[1,0,0,0,0,0,0]^T,k=0;
4) Around the local area of corresponding point set Q_4, use k-d tree to search for the nearest point *Y of each point of* P_3.. The k-th iteration is defined as: Y_k=T(X_k,Q_4);
5) Calculate the spliced vector: (t_k,d_k)=T(X_0,Y_k);
6) Apply the k-th splicing vector to the initial point set X$_0$: X_{k+1}=t_k(X$_0$);
7) Simplify the point-to-point quadratic distance function to the multiple unconstrained minimization based on the linear search. The angle between iteration vectors in the splicing space determines the direction of convergence the. In the formula, $\Delta q_k = q_k - q_{k-1}$
8) Determine whether the error converges. If the dk-dk +1 < δ , converge. Otherwise, come to step 4. The termination of this iteration renders us the splicing conversion parameters TR2. The termination condition δ is set as 0.02mm.

The efficiency of the traditional ICP algorithm based on the point-to-point distance function is low, and, in the worst case, the cost time can reach $O(N_pN_q)$. In addition, it itself can not guarantee the global minimum convergence. This paper obtains the optimum initial position through the preliminary steps. After several re-sampling, the number of corresponding point is significantly reduced. k-d tree is adopted here to search in the neighboring filed, which successfully solves the slowness of the traditional algorithm and brings the advantages, for example precision, of this proposed algorithm into full play. In this way, we can reduce the difficulty of solving complex nonlinear problems.

The proposed splicing algorithm consists of two steps: initial pre-splicing and fine splicing. The purpose of the pre-splicing is to minimize positional relationship between two point cloud data so that two data can have the same accuracy alignment splicing conditions; the purpose of fine splicing is to minimize the distance and error of two point cloud sampling data to be spliced. The pre-splicing guarantees that the two depending approximate conversion parameters are known. Through the ICPP procedures to search the nearest point so as to obtain the approximate conversion parameter matrix space, which provides the premise for the point-to -surface ICP search. Thus, we can reduce the difficulty in solving complex nonlinear problem and facilitate the search of global minimum distance d (P, Q). The proposed algorithm not only has good splicing precision but also has good convergence efficiency. Compared with the original point-to-point-distance-based ICP algorithm and point-to-surface-distance-based ICP algorithm, the improved multi-sampling point algorithm deploys the k-d tree global search to reduce the number of iterations in local area, improve the convergence efficiency, and guarantee the splicing accuracy. The following Table 1 shows the comparison of data of various algorithms:

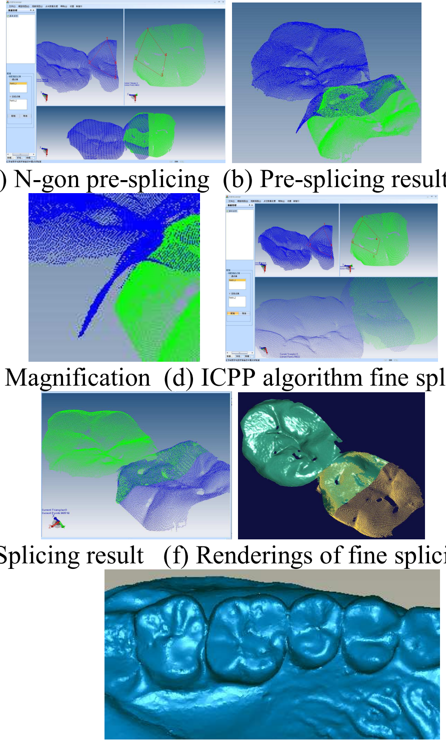
Table 1 Comparison of Convergence Efficiency and Accuracy between Various Algorithms

	Amount of Point Cloud Data	Point-to-point-distance-based ICP Algorithm	Point-to-surface-distance-based ICP Algorithm	The Proposed ICPP Algorithm
Number of Iteration	99022	23	18	13
	111385	35	32	16
Splicing Accuracy (mm)	99022	0.018	0.026	0.02
	111385	0.019	0.028	0.019

This paper takes the photos of tooth and head, whose point cloud data is respectively 99,022 and 111,385 points, for example. Three algorithms are applied and compared in terms of the number of iterations and splicing accuracy, which are specifically shown in Table 1. We can learn that the proposed algorithm is better than the other two algorithms in terms of convergence speed and the overall splicing precision. It is notable that the point-to-point algorithm can result in local convergence and splicing errors in the case of head photo.

4. Algorithm Example Verification and Splicing Assessment

In order to realize the unconstrained multi-cloud 3-D splicing based on free-view angle scanning, this paper develops a 3-D point cloud collection and splicing system according to VC++.NET20003. We developed the measurement system based on the structured light viewing angle measuring system D3Dscanner. The tooth in Prosthodontics is taken as the measurement object for multi-view point cloud scanning. The proposed algorithm is applied to splice multi-chip point cloud and display the process. Figure 4 is the software interface using this system for the pre-splicing; Figure 3a shows the polygon sampling pre-splicing of two point clouds to be spliced. Figure 3b shows the pre-splicing results. From the magnified partial view of the splicing gap 4c, we can see that precise splicing is needed. On the basis of the pre-splicing, the photo after the ICPP fine splicing in this paper is displayed in Figure 6d. Figure 3e shows the high-precise splicing. From the renderings of Figure 3f, we can know that the splicing precision can meet the system requirements. Figure 3g demonstrates the splicing of multiple point clouds by using the proposed algorithm.



(a) N-gon pre-splicing (b) Pre-splicing result

(c) Magnification (d) ICPP algorithm fine splicing

(e) Splicing result (f) Renderings of fine splicing

(g) Rendering of multiple teeth point cloud

Figure 3 System splicing process of software interface

In terms of splicing results, we can assess the splicing result from the point cloud direct display. However, we can not judge the splicing precision. Therefore, we use the similarity assessment function S mentioned by the reference [16] to quantify precision. The mathematical function of the degree of similarity S is expressed as:

$$S = (e^{-\frac{SSD_{D,M}}{(L_D+L_M)^{1/2}}})\times 100\%,$$

Where, $SSD = |x_b - T(x_a)|^2$,XP = (XP, YP, ZP), Xq = (xq, YQ, ZQ) are respectively the corresponding coordinates of point sets P and Q; $SSD_{D,M}$ is the quadric sum of the smallest difference between two point cloud models; L_D is quadratic sum of the distance between vertex value and the original point of the point cloud P model; L_M is the square of the distance between the vertex value and origin point of cloud Q model. The following table shows the similarity assessment of the proposed algorithm on randomly chosen multiple cloud points of corresponding point.

Table 2 Similarity of Point Cloud Splicing

Corresponding two point clouds	Similarity
The 1st pair	94.523%
The 2nd pair	95.146%
The 3rd pair	93.259%
The 4th pair	94.842%
The 5th pair	95.024%

5. Conclusion

This paper introduces a point cloud splicing method oriented for scanning point clouds with special needs. It solves such problems with global measurement in terms of no obvious characteristics, no use of shaft measuring and paste mark point. As splicing is a problem of local matching, this paper, according to ideas about matching, realizes local splicing. Firstly, we broadly sample the overlapping parts of the two point clouds. Then, it employs the proposed method of N-gon to minimize the distance between each other and proposes necessary initial position for fine splicing. Through the setting of hierarchy conditions for iteration termination and the improved iterative closest point fusion algorithm, we can reduce the number of iterations and ensure the obtaining global minimum convergence. Examples prove that the proposed method has better splicing accuracy and convergence efficiency with simple operation. Finally, the three-dimensional splicing technology and our self-developed three-dimensional measurement system are jointly deployed to realize accurate information acquisition from complex objects. In a word, the proposed method boasts pretty high practical value.

References

[1] Michael Zimba, Sun Xingming, "Digital Image Splicing Detection Based on Local Complexity of Local Maximum Partial Gradient of DWT Coefficients.", JDCTA, Vol. 6, No. 5, pp. 1 ~ 9, 2012

[2] Bo Wang, Xiangwei Kong, Lanying Wu, "Different-quality Re-demosaicing in Digital Image Forensics", JCIT, Vol. 7, No. 17, pp. 492 ~ 505, 2012

[3] Li Yao, Dong-Xiao Li, , Ming Zhang, "A Temporally Streamlined Optimization Method for Stereo Video Correspondence", IJACT, Vol. 4, No. 2, pp. 238 ~ 246, 2012

[4] HE Huaiqing , , YANG Lei , XU Qing , "Multidimensional Uncertainty Visualization with Parallel Coordinate and Star Glyph", JDCTA, Vol. 5, No. 6, pp. 412 ~ 420, 2011

[5] Kwangmu Shin, Sunghwan Chun, Kidong Chung, "An Efficient Mode Selection Exploiting Property of Region in Multi-view Video Coding", IJIPM, Vol. 2, No. 3, pp. 44 ~ 51, 2011

[6] Zhi Liu, Hongjun Wang, Hui Xu, Shangling Song, "3D Tongue Reconstruction Based on Multi-view Images and Finite Element", AISS, Vol. 3, No. 11, pp. 60 ~ 66, 2011

[7] Xin Wang, Guofang Lv, Huibin Wang, "Multi-View Tracking of Occluded Targets by Scenic Feature Modeling", AISS, Vol. 4, No. 22, pp. 312 ~ 319, 2012

[1] [8] Guifang Duan, Shuci Wu, Yen-Wei Chen, "Robust Facial Feature Point Extraction via Viewpoint-Specific Active Appearance Model", JNIT, Vol. 3, No. 3, pp. 10 ~ 18, 2012

Computerized Simulation of Automotive Air-Conditioning System: A Parametric Study

Haslinda Mohamed Kamar[1], Nazri Kamsah[2] and Mohd Yusoff Senawi[3]

**[1] Department of Thermo-Fluid, Universiti Teknologi Malaysia
81310 UTM Skudai, Johor, Malaysia**

**[2] Department of Thermo-Fluid, Universiti Teknologi Malaysia
81310 UTM Skudai, Johor, Malaysia**

**[3] Department of Thermo-Fluid, Universiti Teknologi Malaysia
81310 UTM Skudai, Johor, Malaysia**

Abstract

This paper presents results of a parametric study performed on an automotive air-conditioning (AAC) system of a passenger car. The goals are to assess the effects of varying the volumetric flow rate of supply air, number of occupants, vehicle speed, and the fractional ventilation air intake (XOA), on the dry-bulb temperature and specific humidity of the air inside the passenger's cabin, and on the evaporator coil cooling load of the AAC system. Results of the parametric study show that increasing the supply air flow rate reduces the dry-bulb temperature of the cabin air, increases both the specific humidity of the air and the evaporator coil load. Increasing the number of occupants in the passenger cabin causes the cabin air temperature, specific humidity and the evaporator coil load to increase. Increasing the vehicle speed causes the specific humidity of the cabin air and the evaporator coil cooling load to increase but the dry-bulb temperature of the air is not significantly affected. Increasing the fractional fresh air intake (XOA) also increases the cabin air specific humidity and the evaporator coil cooling load.

Keywords: *Automotive air conditioning (AAC), passenger car cabin, parametric study*

1. Introduction

Automotive air-conditioning (AAC) is a necessity for thermal comfort in the cabin of a passenger vehicle especially for people who are living in countries with hot and humid climate. However, the extra weight added to the vehicle and the operation of the AAC system cause the fuel consumption of the vehicle to increase. The additional fuel consumption in turn results in higher emission of greenhouse gases that pollute the environments. Therefore augmentation of the AAC system efficiency and evaluation of its thermal performance has become important. The AAC system is often operated under varying conditions thus substantial efforts are required to evaluate its performance. These conditions include the temperature of the air entering the evaporator and condenser, the evaporator air volumetric flow rate and ventilation mode, the compressor speed, the condenser face air velocity, passenger cabin's material, the number of occupants, and the weather conditions which affect the internal and external sensible thermal loads.

The largest auxiliary load on a passenger car's engine is input power required by the air-conditioning system's compressor. During a peak load the compressor would consume up to 5 to 6 kW of power from the vehicle's engine power output. This is equivalent to a vehicle being driven at a speed of 56 km/h. The additional fuel consumption when the air-conditioning system is in-used is quite substantial. One study indicated that the air-conditioner usage reduces fuel economy by about 20%. It also increases the emissions of nitrogen oxides by about 80% and carbon dioxide (CO) by about 70%, although the actual numbers depend on the actual driving conditions.

A semi-empirical computer simulation program (CARSIM) for simulating thermal and energy performance of an automotive air-conditioning (AAC) system of a passenger car has been developed [1]. The empirical correlations for evaporator sensible and latent heat transfer were embedded in the loads calculation program to enable the determination of evaporator inlet and outlet air conditions and the passenger cabin air conditions. The computer program has been validated by comparing its predicted outputs with the data obtained from an actual road test on a Proton Wira Aeroback passenger car. The results predicted by the CARSIM computer program were found to have very good agreement with the actual road test data, with errors ranging between 2 to 4%.

This paper presents results of a parametric study performed on the automotive air-conditioning (AAC) system of the Proton Wira Aeroback passenger car using the CARSIM computer simulation program. The goal is to investigate the effects of: (1) supply air volumetric flow rate, (2) number of occupants, (3) vehicle speed, and (4) the fractional outside air intake (XOA), on the cabin's air dry-bulb temperature, specific humidity and the total evaporator coil (cooling) load of the AAC system. A base case simulation was performed on the system to find out the trend of variation of the cabin air dry-bulb temperature, specific humidity and the total evaporator coil (cooling) load as the time is varied from 10 am to 2 pm.

2. Base Case Conditions

A computer simulation for the base case conditions was performed on the automotive air-conditioning (AAC) system by using the CARSIM computer simulation program developed earlier and reported in [1]. The input data used for the computer simulation are given in Table 1.

Table 1 Parameters for the base case condition.

Parameters	Value
Vehicle speed	~ 90 km/h
Occupant	~ one (1)
Colour of the car body	~ light yellow (α = 0.5)
Glass thickness	~ 3 mm (shading coefficient = 1)
A/C blower speed	~ maximum
Travelling duration	~ 4 hours (10 am to 2 pm)
Fraction of outside air (XOA)	~ 0.16 (or 16% of outside air)

The computer simulation was performed as if the vehicle is driven at a constant speed of 90 km/h with only one occupant. The travelling duration is four hours i.e. from 10 am (or 600 minutes past midnight) to 2 pm (or 840 minutes past midnight). The car air-conditioning system is turned-on during this period in which the blower fan speed is set to a maximum. The fractional outside air (XOA) is set at a fixed value of 16%. The air in the passenger's cabin gains sensible heat from the surrounding air and latent heat from the passenger. As a result, the cabin air dry-bulb temperature, specific humidity and the evaporator coil cooling load will vary with time during the simulated journey. The simulation for the base case conditions was carried out to find out the trend in which these parameters vary with the time. Figure 1 shows a schematic diagram of the simplified air-conditioning system of the passenger car considered in this study.

Figure 2 shows the variation of dry-bulb temperature of the cabin air with time in minutes, measured after midnight. The time t = 600 minutes corresponds to 10 am while t = 840 minutes corresponds to 2 pm. It can be seen that during the first five minutes after the air-conditioning system is being turned-on the cabin air temperature drops slightly by about 0.05°C. This is due to a sudden cooling of the air when cooled air is suddenly blown into the cabin. Thereafter, the cabin air temperature rises steadily with time, from 18.4°C until it reaches about 18.8°C at about 1 pm. The cabin air temperature is affected by the temperature of the ambient air, t_0. The ambient air temperature rises steadily with time as the intensity of solar radiation increases. This causes the cabin air temperature to rise in a fashion seen in the figure. However, as the solar radiation intensity decreases after 1 pm, the cabin air temperature also decreases steadily, from 18.8°C to about 18.6°C at 2 pm. Although the dry-bulb temperature of the cabin air appears to vary with time, the range of its variation is quite small, i.e. less than 0.5°C, and hence can be considered insignificant.

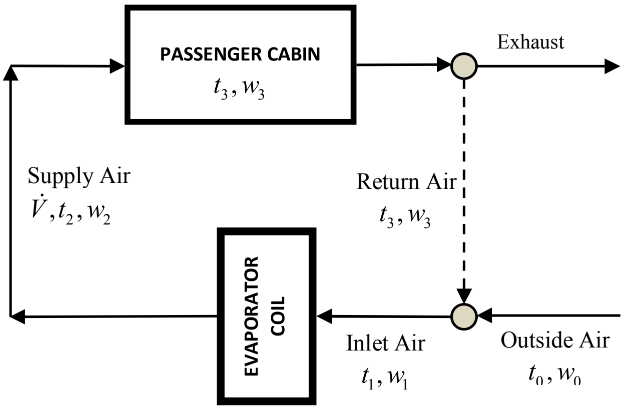


Figure 1 Schematic diagram of an air-conditioning system of a passenger car [1].

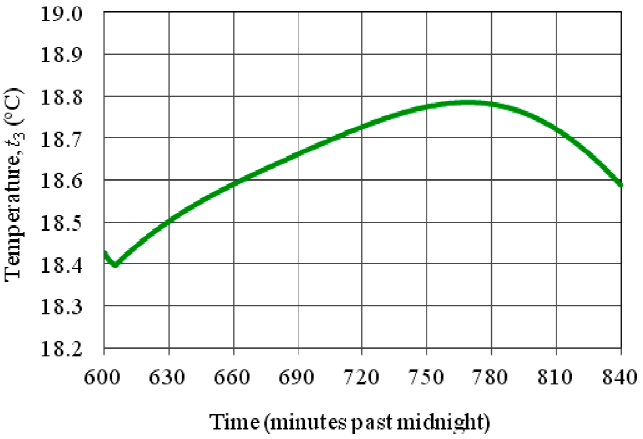


Figure 2 Variation of dry-bulb temperature of the cabin air for the base case condition.

The variation of specific humidity of the cabin air with time is shown in Figure 3. The air specific humidity is related to the saturation pressure of water vapor that exists in the air. The saturation pressure increases dramatically with the dry-bulb temperature of the air. Hence, when the temperature of the cabin air rises, the specific humidity of the air also increases in a nearly similar fashion. It can be noted that the specific humidity of the air increases steadily from about 6.26 g/kg at 10 am (600 minutes after midnight) to a highest value of about 6.45 g/kg at 1 pm (780 minutes after midnight). Thereafter, the specific humidity of the cabin air decreases with time, as the dry-bulb temperature falls. At 2 pm (840 minutes after midnight) the specific humidity of the cabin air drops to about 6.40 g/kg.

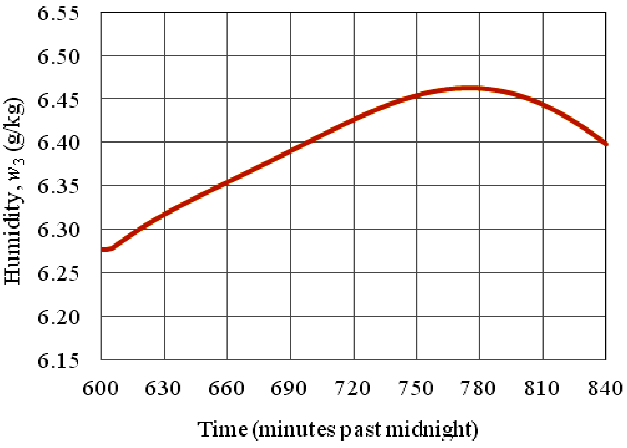


Figure 3 Variation of specific humidity of the cabin air for base case condition.

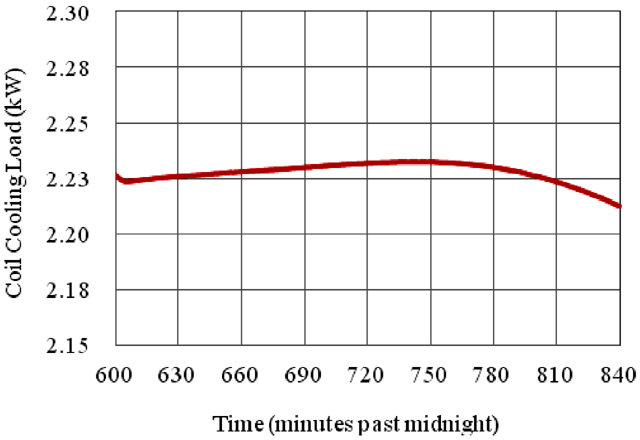


Figure 4 Variation of evaporator coil cooling load with time for base case condition.

Figure 4 shows variations of the evaporator coil cooling load with time. The cooling load is the sum of sensible heat load, $Q_{C,S}$ and latent heat load, $Q_{C,L}$. The sensible heat is influenced by the temperature t_1 of the air coming into the cooling coil. The latent heat is influenced by the specific humidity w_1 of the incoming air. The incoming air temperature t_1 is in turn affected by the temperature of the incoming outside air, t_0 and the temperature of the return air, t_3. Both the t_0 and t_3 increases with time causing the temperature t_1 to continuously increases from 10 am to 1 pm. This in turn causes a steady increase in the sensible heat load of the coil during the same period. As both the t_0 and t_3 decreases after 1 pm, the sensible heat load also decreases. The specific humidity of the air coming into the cooling coil, w_2 is influenced by the humidity of the incoming outside air, w_0 and the humidity of the return air, w_3. Both the w_0 and w_3 increase steadily from 10 am to 1 pm causing the latent heat load to rise steadily during the same period. When both the w_0 and w_3 decrease after 1 pm, the latent heat load of the evaporator coil also decreases. Although the specific humidity varies with time, the range of its variation is quite small.

3. Parametric Study

A parametric study was performed by computer simulation to investigate the effects of varying the supply-air volume flow rate, number of occupants, vehicle speed and the fractional outside air (XOA) on the cabin air dry-bulb temperature, specific humidity and the evaporator coil cooling load. The supply-air volume flow rate was varied from 70 L/s to 100 L/s with a 10 L/s interval. The number of occupants of the vehicle was varied from one to four persons. The vehicle speed was varied from 60 km/h to 105 km/h with an interval of 15 km/h. Finally, the fractional outside air (XOA) which is the ratio between the ventilation-air volume flow rate and the cooled-air volume flow rate, was varied from 0.2 to 0.3, with a 0.05 interval. When the cooled-air volume flow rate was varied during the simulation, other parameters that are listed in Table 1 were held constant at the prescribed values. Similar procedure was followed when the other parameters were varied.

4. Results and Discussion

4.1 Effects of Supply-Air Volume Flow Rate

Figure 5 shows the effect of cooled-air volume flow rate on the dry-bulb temperature of the cabin air. The simulation results show that for a given air flow rate, the dry-bulb temperature varies with time in a very similar manner as that for the base case. At a given cooled-air volume flow rate, the cabin air temperature increases steadily from 10 am (600 minutes past midnight) and reaches the highest value at about 1 pm (780 minutes after midnight). After reaching a maximum value, the cabin air temperature decreases with time. However, increasing the cooled-air volume flow rate decreases the dry-bulb temperature of the cabin air. On average, the cabin air temperature decreases by about 0.5°C (or 2.5%) for every 10 L/s increment of the cooled-air volume flow rate. The highest temperature of the cabin air temperature is 20.2°C, at 1 pm when the cooled-air volume flow rate is 70 L/s. The temperature falls to about 18.7°C when the cooled-air volume flow rate is increased to 100 L/s, which can be considered as a significant temperature change.

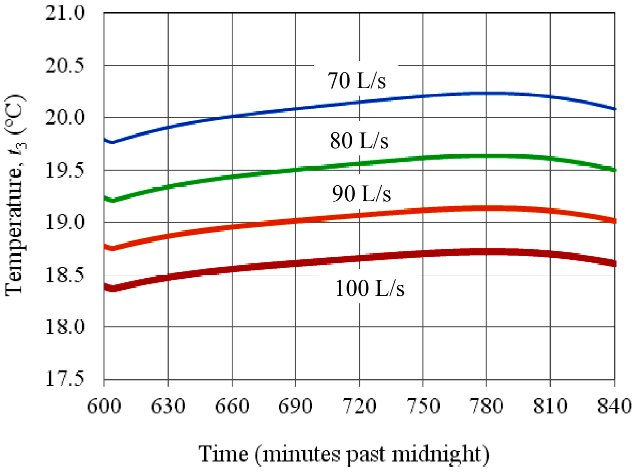


Figure 5 Variation of evaporator coil cooling load with time for base case condition.

Figure 6 shows the effect of cooled-air volume flow rate on the specific humidity of the air inside the passenger cabin. Again, the simulation results show that the humidity of the cabin air varies with time in more or less similar manner, regardless of the cooled-air volume flow rate. However, the specific humidity of the cabin air increases when the cooled-air volume flow rate is increased. Increasing the cooled-air volume flow rate means the velocity of the air flowing through the evaporator coil is increased. This reduces the ability of the evaporator coil to remove moisture from the air that is passing through it. Consequently, the specific humidity of the cabin air will rise. On average, results of the simulation results show that the specific humidity of the cabin air rises by about 5.4% (or 0.34 g/kg) for each 10 L/s increase in the cooled-air volume flow rate.

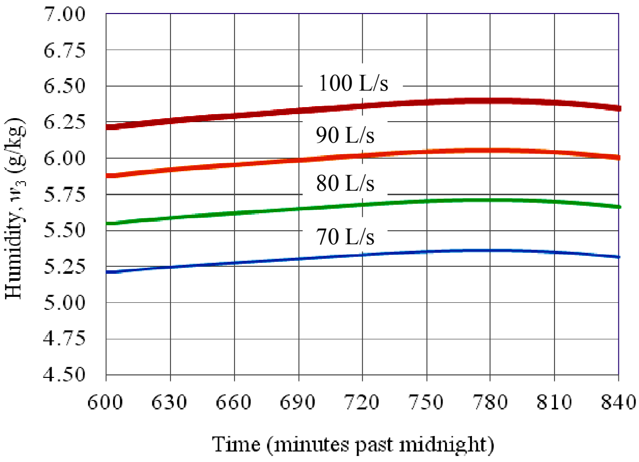


Figure 6 Effect of cooled-air volume flow rate on specific humidity of the cabin air.

The effect of cooled-air volume flow rate on the evaporator coil total cooling load is shown in Figure 7. It can be seen that, for a given air volume flow rate, the variation of the coil cooling load with time is very much similar to that for the base case. However, increasing the volumetric flow rate of the cooled-air increases the coil cooling load, at any given time. On average, the simulation results show that the cooling load increases by about 6 to 8% for each 10 L/s increase in the cooled-air volume flow rate.

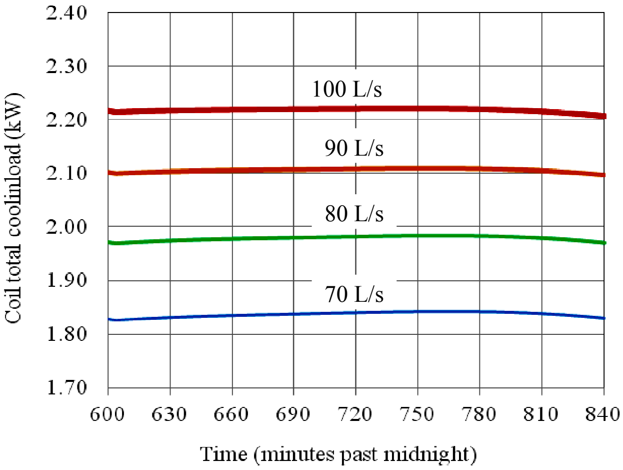


Figure 7 Effect of cooled-air volume flow rate on the evaporator coil cooling load.

4.2 Effects of Number of Occupant

The computer simulation results on the effect of the number of occupants in the passenger cabin on the dry-bulb temperature of the cabin air is shown in Figure 8.

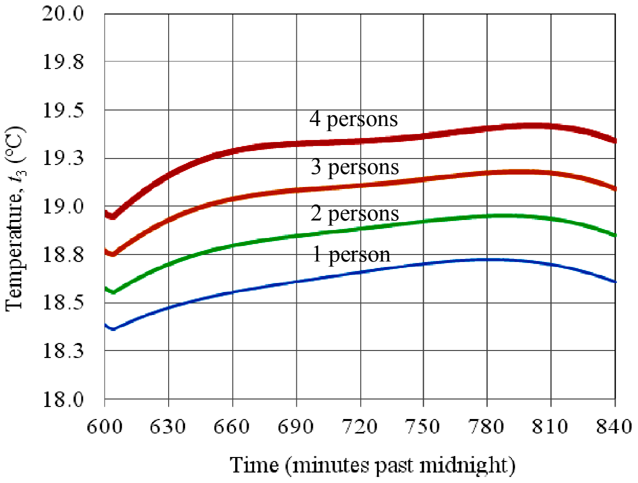


Figure 8 Effect of number of occupants on the dry-bulb temperature of the cabin air.

Note that the curve for one person represents that for the "base case" condition (see Figure 2). The simulation results indicate that the variation of cabin air temperature with time is not much affected by the number of occupant in the cabin. However, at a any given time, when the number of occupants is increased the cabin air dry-bulb temperature increases. This is because human body continuously transfers energy in the form of sensible heat which will cause the cabin air temperature to rise. As more passengers occupy the cabin space, more sensible heat is transferred to the cabin air causing greater temperature rise of the air. On average, the simulation results show that the cabin air temperature increases by about 1.2 % or 0.23°C for each additional person occupying the passenger cabin. This can be considered as a significant increment.

The effect of the number of occupants on the specific humidity of the cabin air is shown in Figure 9. Note that the curve for one person represents the simulation result for a base case condition (see Figure 3). As seen in the figure, the variation of specific humidity of the cabin air with time is generally not affected by the number of occupants present in the passenger cabin. However, the simulation results show that at any given time, increasing the number of occupants will significantly increases the specific humidity of the cabin air. This is because human releases energy into the cabin air in the form of latent heat, through perspiration process and breathing. Hence when more people occupy the cabin space, the more latent heat and moisture are added to the cabin air resulting in significant increase in the specific humidity of the air. On average, the humidity of the cabin air rises by about 5.4% or 0.38 g/kg for each additional occupant in the passenger cabin.

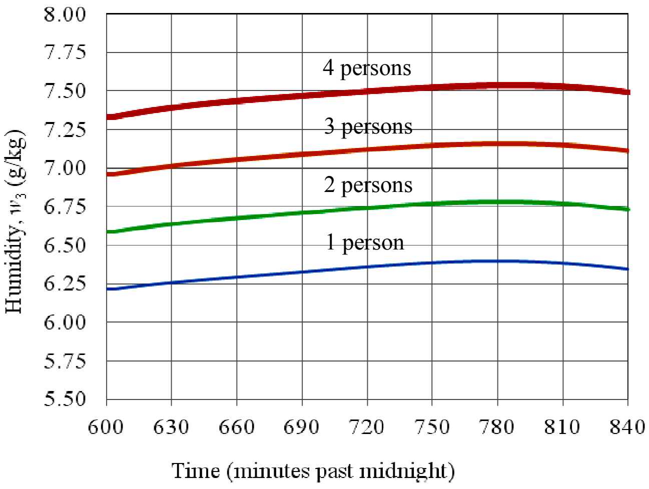


Figure 9 Effect of number of occupants on cabin air specific humidity.

Figure 10 shows the effect of the number of occupants on the evaporator coil cooling load. The curve for one person represents the simulation result for the base case condition (see Figure 4). It is seen that the variation of the coil cooling load with time appears to be unaffected by the number of occupants in the cabin. However, increasing the number of occupant will increase the amount of sensible and latent heat in the cabin air. This represents the additional thermal load that needs to be absorbed from the cabin air by the air-conditioning system. The mass flow rate of the refrigerant through the evaporator must be increase to achieve this. This is accomplished through the increase in the compressor speed. The figure shows that, on average, for each additional occupant in the cabin, the evaporator coil cooling load rises by 2.5% or about 0.06 kW.

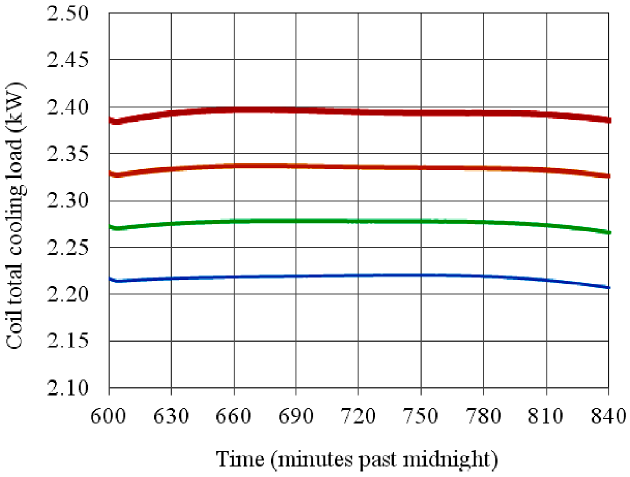


Figure 10 Effect of number of occupants on the evaporator coil cooling load.

4.3 Effects of Vehicle Speed

Computer simulation was also carried out to determine the effects of varying the vehicle speed on the dry-bulb temperature and specific humidity of the cabin air, and the evaporator coil cooling load. The vehicle speed was varied from 60 km/h to 105 km/h with 15 km/h interval.

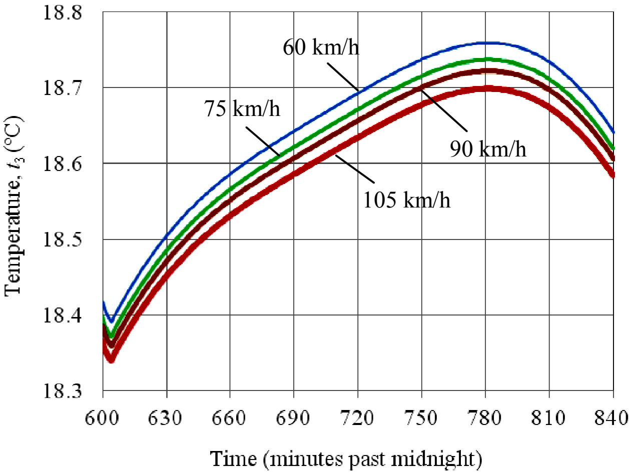


Figure 11 Effect of vehicle speed on the dry-bulb temperature of the cabin air.

Increasing the vehicle speed will cause the compressor of the air-conditioning system to operate at higher speed. This in turn leads to a higher mass flow rate of the refrigerant. As a result, more heat is absorbed by the refrigerant from the air passing through the evaporator coil [5]. The dry-bulb temperature of the cabin air will be reduced since the supply air enters the cabin at lower temperature. This is situation is shown in Figure 11. However, as seen from the figure, the reduction in the cabin air dry-bulb temperature is very small. On average, the simulation results show that the cabin air temperature drops only by about 0.1% or 0.02°C for every 15 km/h increment of the vehicle speed. This result is however consistent with that reported in the literature [6].

The effect of varying vehicle speed on the specific humidity of the cabin air is shown in Figure 12. In general, the simulation results show that the vehicle speed does not have significant effects on the trend of variation of specific humidity of the cabin air with time. However, at any given time, the specific humidity of the cabin air decreases quite significantly as the vehicle speed is increased. This results suggests that as the vehicle moves at higher speed, more moisture is absorbed from the air that is passing through the evaporator coil. As a result, the air that is supplied into the cabin air is dryer. This causes the reduction in the specific humidity of the cabin air. On average, the specific humidity drops by about 5.4% or 0.33 g/kg for every 15 km/h increment of the vehicle speed.

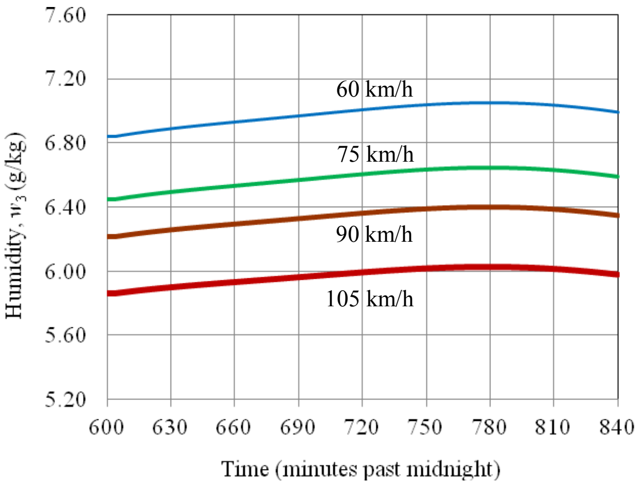


Figure 12 Effect of vehicle speed on the specific humidity of the cabin air.

The results of computer simulation shows that the vehicle speed has insignificant effects on the variation of evaporator coil cooling load with time. This is shown in Figure 13. However, at any given time, the coil cooling load increases as the speed of the vehicle is increased. On average, for every increment of 15 km/h, the coil load increases by 3.9% or 0.09 kW. However the increase in the cooling load is somewhat smaller when the vehicle speed is increased from 75 km/h to 90 km/h.

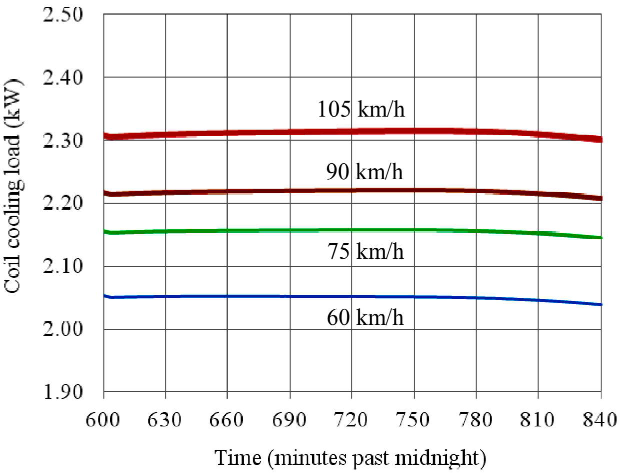


Figure 13 Effect of vehicle speed on the evaporator coil cooling load.

4.4 Effects of Fractional Ventilation Air Intake

Fractional ventilation air intake (XOA) is the ratio between the ventilation-air flow rate and the air-conditioning system's supply-air flow rate. It's value ranges from 0 for air recirculation mode (no ventilation-air) and 1 for a ventilation-air full usage (no air recirculation).

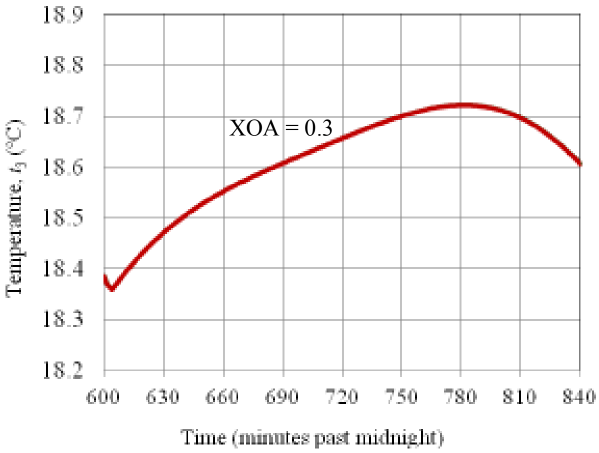


Figure 14 Effect of XOA on the dry-bulb temperature of the cabin air.

In this study the XOA was varied from 0.2 to 0.3 with a 0.05 increment the effects of doing this on the dry-bulb temperature and specific humidity of the cabin air, and the cooling load of the evaporator coil were investigated. The effects of varying the XOA on the dry-bulb temperature of the cabin air is shown in Figure 14. It is seen from the figure that the variation of the cabin air temperature with time is not significantly affected by the value of the XOA. Also, increasing the XOA from 0.2 to 0.3 does not seem to have appreciable impact on the cabin air dry-bulb temperature. This could be because the variation of the XOA considered in this study is too small.

The effect of varying the XOA on the specific humidity of the cabin air is shown in Figure 15. The simulation results show that the variation of specific humidity with time is generally not much affected by the value of the XOA. However, at any given time, increasing the value of XOA causes the specific humidity of the cabin air to increase. When the XOA is increased the specific humidity of the air at the inlet of the evaporator coil is also increased because the outside air has higher moisture content compared with the returned air from the cabin. When the moisture content of the air at the inlet of the evaporator coil is higher, the ability of the coil to remove the moisture is somewhat reduced. As a result, the air leaving the evaporator coil, which is supplied into the cabin, will have slightly higher moisture content. On average, the specific humidity of the cabin air is increased by about 5.2% or 0.37 g/kg for every 0.05 increment of the XOA.

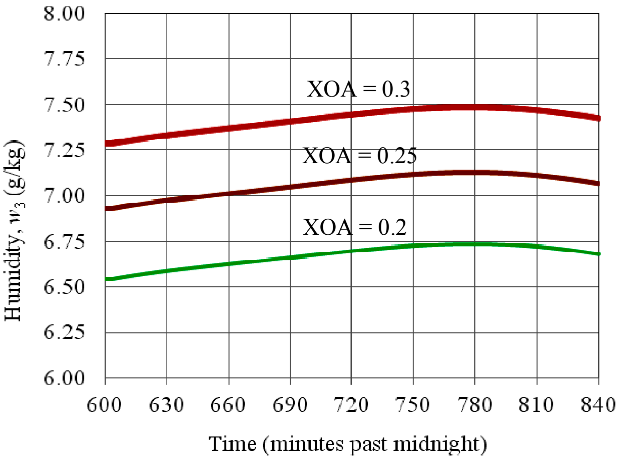


Figure 15 Effect of varying the XOA on the specific humidity of the cabin air.

The simulation result on the effects of varying the XOA on the evaporator coil cooling load is shown in Figure 16. It is seen from the figure that the value of XOA does not have any significant effect on the variation of the cooling load with time. However, at any given time, increasing the XOA causes the evaporator cooling load to increase quite significantly. The increase in the coil load is a direct consequence of the increase in temperature and specific humidity of the air inlet to the evaporator coil, as more outside air is introduced to the system. The outside air has higher temperature and humidity than the cabin. On average, for every 0.05 increment of the XOA, the coil cooling total load rises by about 6.8% (or 180 Watt), which can be considered as significant increment.

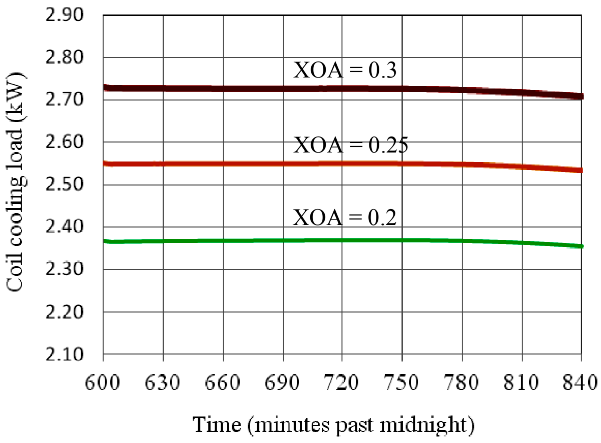


Figure 16 Effect of varying the XOA on the evaporator coil cooling load.

5. Conclusion

A semi-empirical computer simulation program (CARSIM) has been developed for simulating thermal and energy performance of an automotive air-conditioning (AAC) system. The computer program was used to perform a parametric study to investigate the effects of varying the volume flow rate of supply air, the number of occupants in the passenger cabin, the vehicle speed and the fractional outside air intake (XOA) on the dry-bulb temperature and specific humidity of the cabin air, and on the evaporator coil cooling load of a 1.6 L Proton Wira passenger car. Results of the parametric study show that for each 10 L/s increment of the supply air flow rate the cabin air temperature is reduced by about 2.5%, the specific humidity increases by 5.4%, and the evaporator coil cooling load increases by about 6%. For each additional occupant in the passenger compartment, the cabin air temperature, specific humidity, and evaporator coil cooling load are increased by 1.2%, 5.4%, and 2.5%, respectively. For every 15 km/h increase of the vehicle speed, the specific humidity of the cabin air is increased by 5.4% and the evaporator coil cooling load by 3.9%. The temperature of the cabin air appears to be not affected by the vehicle speed. The fractional fresh air intake (XOA) has no significant influence on the cabin air temperature. However, for every 0.05 increase of XOA, the interior air specific humidity increases by 5.2% and the evaporator coil cooling load also increased, by about 6.8%.

Acknowledgement

The authors would like to acknowledge the financial supports from the Universiti Teknologi Malaysia and research funding given by Ministry of Science, Technology and Innovation, (MOSTI) Malaysia and the Research Management Centre (RMC), Universiti Teknologi Malaysia, through an IRPA grant no. 74062.

References

[1] Haslinda Mohamed Kamar, Mohd Yusoff Senawi and Nazri Kamsah, Computerized Simulation of Automotive Air-Conditioning System: Development of Mathematical Model and Its Validation, The International Journal of Computer Science Issues, Vol. 9, Issue 2, No 2, March 2012.

[2] Moller, S. K. and Wooldridge, M. J., User's Guide for the Computer Program BUNYIP: Building Energy Investigation Package (Version 2.0), Highett, Victoria, Australia. (1985).

[3] Kohler, J., Kuhn, B. and Beer, H., Numerical Calculation of The Distribution of Temperature And Heat Flux In Buses Under The Influence of The Vehicle Air-Conditioning System. ASHRAE Transactions. 96 (Part 1) (1990) 432 – 446.

[4] H. Zhang, L. Dai, G. Xu, Y. Li, W. Chen, W-Q., Tao Studies of Air-flow and Temperature Fields Inside a Passenger Compartment for Improving Thermal Comfort and Saving Energy: Part I Test/Numerical Model and Validation. Applied Thermal Engineering. 29 (2009) 2022–2027.

[5] Somchai Wongwises, Amnouy Kamboon and Banchob Orachon, Experimental Investigation of Hydrocarbon Mixtures to Replace HFC-134a in an Automotive Air Conditioning System. Energy Conversion and Management. 47 (2006) 1644–1659.

[6] H. Zhang, L. Dai, G. Xu, Y. Li, W. Chen, W-Q., Tao. Studies of Air-flow and Temperature Fields Inside a Passenger Compartment for Improving Thermal Comfort and Saving Energy: Part II Test/Numerical Model and Validation. Applied Thermal Engineering. 29 (2009) 2028–2036.

[7] Mezrhab, A. and Bouzidi, M., Computation of Thermal Comfort Inside a Passenger Car Compartment. Applied Thermal Engineering. 26 (2006) 1697–1704.

Haslinda Mohamed Kamar, is a member of ASHRAE. She received her Bachelor's Degree in Mechanical Engineering from University of Glasgow, Scotland in 1993, Master and Ph.D in Mechanical Engineering from University Teknologi Malaysia in 1997 and 2009, respectively. She is now a senior lecturer in the Faculty of Mechanical Engineering, Universiti Teknologi Malaysia. Her areas of interest are automotive air-conditioning system, thermal comfort & energy efficiency in hot climates, indoor air quality (IAQ), natural ventilation as passive cooling strategy in buildings and computational fluid dynamics (CFD) modeling and simulation.

Nazri Kamsah is a member of ASHRAE. He received his Bachelor's Degree in Mechanical Engineering from University of Sunderland, United Kingdom in 1983, Masters of Engineering (Mechanical) from Universiti Teknologi Malaysia in 1988, and Ph.D in Mechanical Engineering from University of New Hampshire, Durham, USA in 2001. He is currently a senior lecturer in the Faculty of Mechanical Engineering, Universiti Teknologi Malaysia. His areas of interest include computational solid mechanics, finite element modeling and simulation, thermal management in microelectronics, thermal comfort and energy efficiency in buildings, natural ventilation as passive cooling strategy for buildings, indoor air quality (IAQ) and computational fluid dynamics (CFD) modeling and simulation.

Mohd Yusoff Senawi, graduated with Bachelor of Mechanical Engineering from The University of New South Wales in 1987. Received Post Graduate Diploma in Computer Science, Master and Ph.D in Mechanical Engineering from Universiti Teknologi Malaysia (UTM) in 1989, 1993 and 2000, respectively. Currently a Senior Lecturer at UTM. Research interests include cooling loads calculation and energy analysis of air conditioning systems.

A New Approach for Quality Management in Pervasive Computing Environments

ALTI Adel[1], ROOSE Phillipe[2]

**[1] Computer Science Department, Science Faculty
Ferhat Abbas University, Sétif B.P. 19000, Algeria**

**[2] LIUPPA / IUT Bayonne
2 Allée du Parc Montaury 64600 Anglet – France**

Abstract

This paper provides an extension of MDA called Context-aware Quality Model Driven Architecture (CQ-MDA) which can be used for quality control in pervasive computing environments. The proposed CQ-MDA approach based on ContextualArchRQMM (Contextual ARCHitecture Quality Requirement MetaModel), being an extension to the MDA, allows for considering quality and resources-awareness while conducting the design process. The contributions of this paper are a meta-model for architecture quality control of context-aware applications and a model driven approach to separate architecture concerns from context and quality concerns and to configure reconfigurable software architectures of distributed systems. To demonstrate the utility of our approach, we use a videoconference system.

Keywords: *MDA, Context, Quality Model, Dynamic reconfiguration, ADL.*

1. Introduction

Model Driven Approach (MDA) [5] has been proposed by the OMG (Object management Group). The basic models of MDA are entities able to unify and support the development of computer systems by providing interoperability and portability. MDA approach does not address how to consider non-functional demands, i.e. how to represent and transform them.

An application for heterogeneous mobile embedded and limited (low bandwidth, power consumption, etc.) device has to firstly prevent interaction and mobility limitation. The heterogeneity of components regarding embedded sensors, CPU power, communication mechanisms (GPRS, WIFI, Bluetooth, ZigBee, etc.), speed of transmission as well as the media variety (sound, video, text and image) requires taking into account adaptation to an abstract level in order to avoid the ad hoc solutions which are not reusable and/or generalized. This is due to the following points:

- The separation of concerns met in software architecture is the separation of communications supported by first class connector from the business logic supported by components. However, communication is not the unique non-functional concern found in software design. Data adaptation, context-awareness, resource-awareness and QoS are other non-functional concerns which cut across component's business logic. Introducing in software architecture will make design of complex software an easier task and will yield clear and lucid specification.
- Few ADLs are able to define new connectors' types that ensure the non-functional concerns of the components (*security, communication, conversion, etc.*).
- Few ADLs support the elaboration of quality model explicitly and facilitate the system architecture quality control with the continuous evolution of its context.

In this paper, we present an extended Model Driven Architecture which includes support for software architecture quality control and resources requirements changes, in the framework of CQ-MDA (Context-aware Quality Model Driven Architecture). Some other works concentrate only on quality system architecture or context-aware system architecture [8, 9]. Our approach focuses on separation of two concerns: the architecture and the implementation contexts. This enables us to support them with the elaboration of quality model explicitly and to facilitate the system architecture quality control with the continuous evolution of its context. To cope with a serious gap in styles quality control, we have previously introduced the *ArchRQMM* (ARCHitecture Requirement Quality MetaModel) [3]. One of the strengths of *ArchRQMM* relies in its ability to separate architecture concerns from requirement and quality concerns and to automatically perform formal architecture quality analysis at architecture stage using OCL [12]. However, our metamodel does not support the definition of a context-awareness and a resource-awareness metamodel.

We begin this paper by introducing ArchRQMM metamodel. Section 3 proposes the main element of CQ-MDA approach, i.e. ContextualArchRQMM metamodel which it is an ArchRQMM extension used as support for context model description and quality model definition. Section 4 describes the CQ-MDA itself. Section 5 shows an example of applying CQ-MDA for VideoConference system development [15]. Section 6 summarizes related works. Section 7 concludes this article and presents some future works.

2. An Overview of ArchRQMM (Architecture Requirement Quality Metamodel)

ArchRQMM metamodel enables architectural styles quality evaluation and selection at the architecture design step and ensures formal verification of the properties' quality of architectures on modelling styles. The metamodel was described in details in [3, 4]. It was developed according to ISO/IEC 9126 standard [7]. *ArchRQMM* is based on a set of meta-classes for the common concepts of architectures descriptions languages (ADLs) and a set of quality characteristics based on a standard ISO quality model [10] which can be investigated and evaluated in the architecture level (maintenability, reusability, efficiency, etc.) . Fig. 1 presents a MOF metamodel of the *ArchRQMM*. One of the strengths of *ArchRQMM* relies in its ability to separate architecture concerns from requirement and quality concerns and to automatically perform formal architecture quality analysis at architecture stage using OCL [12]. The focus of rigorous architecture quality analysis is to prevent the non-required affections before the early phases of system development. The use of *ArchRQMM* metamodel offers number of advantages compared to other related works using UML profiling mechanisms like MARTE [18] including: 1) – architectures, requirements and quality models are explicitly represented, 2) – a formal support to prove the quality properties of architectural styles at the architecture level using OCL[12], 3)- support for model non-functional aspects of software architecture through architecture properties and measurable standards [7,4] , and 4) – automatic evaluation and selection of styles that best meet architects' needs using *QualiStyle* tool [4].

3. ContextualArchRQMM Metamodel

3.1 Objectives and Motivations

The main idea of this proposal is to take into consideration the non-functional concerns (*adaptation service*, *communication protocol*, *security*, *QoS*, *etc.*) of the components by connectors at the software architecture level. In our approach, the two types of preoccupations are ensured respectively by the components and the connectors. Thus, the connectors ensure the communication and the connection of components that realize the functional part (*business logic components*). Their execution within adequate configurations also requires taking into account of the non-functional aspects.

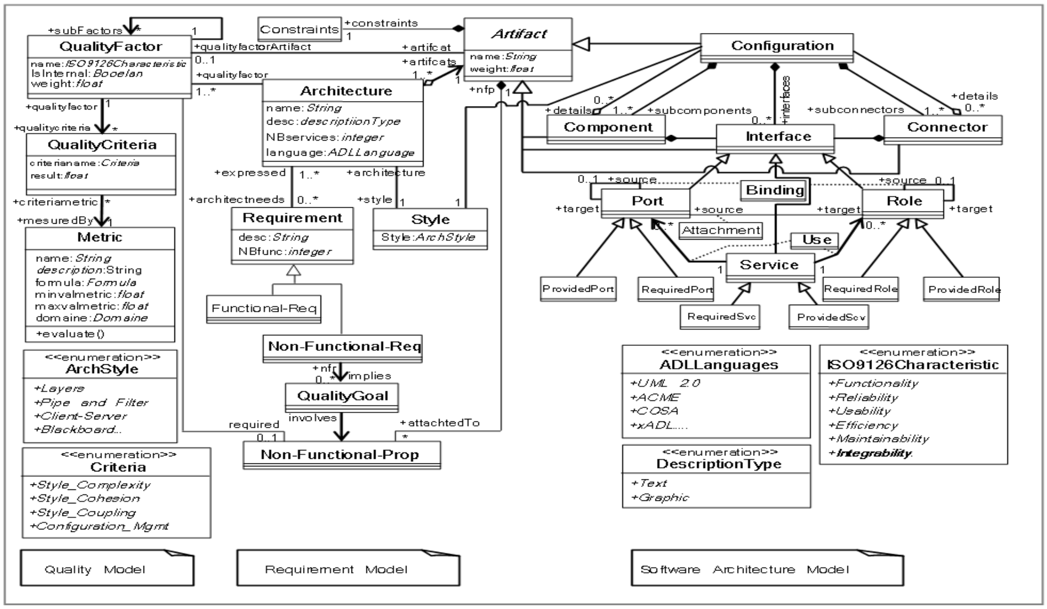


Fig. 1 A MOF Metamodel of ArchRQMM.

3.2 Context-awareness Metamodel

We extend our software architecture metamodel, with a context metamodel (Fig. 2). The goal is to represent context information of system architecture at model level. Context is any information that can be collected from artefact needs, resources capacities and user preferences [20]. *ContextualArchRQMM* uses these informations to perform a software architecture quality evaluation and selection in software development process. We have identified two types of context, i.e., required context (user preferences, artifacts needs) and provided context that encompasses the properties of the execution environment of an application. Context elements are realized through *Context* class, are expressed as QoS properties of the contextual artifacts (*Non-Functional-Prop* class).

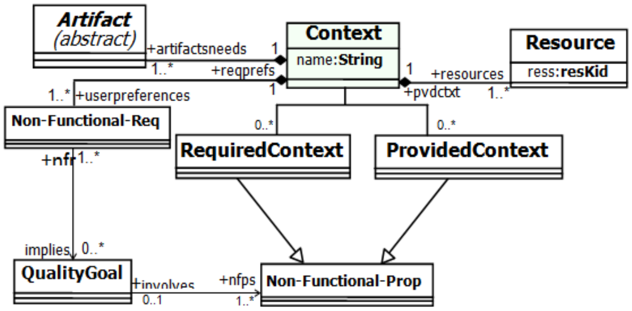


Fig. 2 The context metamodel of ContextualArchRQMM

3.3 Resource-awareness Metamodel

Fig. 3 depicts a resource-awareness metamodel. The hardware components are mobile devices (Class *Device*) like PDAs, PC Portables or smart phone, are constrained in their resources (memory size, CPU power, bandwith, battery, etc) and act as execution environment for architectural artefact (Class *Artifact*). Network connections (Class *Node*) connect hardware components having a limited bandwith. A resource-awareness about current usage of processing power, network bandwith, etc. is a prerequisite to guarantee a minimum quality of service.

3.4 Contextual Architectural Artifacts

For an efficient and clear specification of connection points, we have introduced more precise port according to their global roles in a component: the *DataPort*, the *ContextPort*, information available at run-time when the service is active. The *ServiceControlPort* is a standard dedicated port for controlling a service. It allows the service to be (re)started, updated, relocated, stopped and uninstalled.

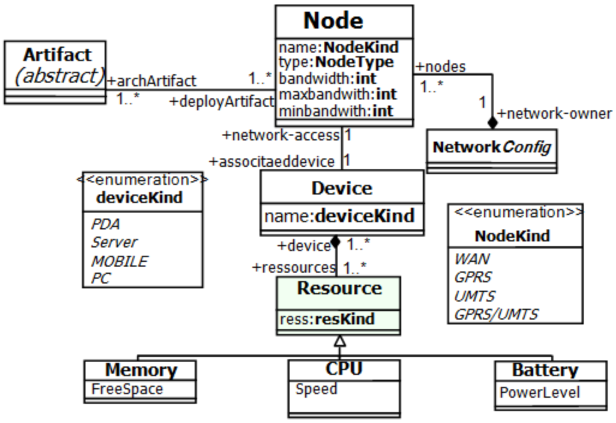


Fig. 3 The resource metamodel of ContextualArchRQMM

The *QoSNotificationPort* is responsible for sending QoS information to execution platform in order to decide if a service reconfiguration is needed. As software architecture descriptions rely on a *connector* to express interactions between components, an equivalent abstraction must be used to express a contextual and a heterogeneous interaction (i.e. various interactions paradigms). We extend an architectural connector with a contextual concern in a heterogeneous interaction (Fig. 4). Three auto-adaptative mechanisms are distinguished: communication (i.e. clarify the connection between various components regarding the communications paradigms), service adaptation (i.e. adding, suppression and substitution of adaptation services), and QoS adaptation (selecting parameters of service to provide adequate quality to component needs at runtime). The business logic component is adapted explicitly and automatically by a *contextual connector*. This means that context ports of *business logic* components instances, related to the context managed by a contextual connector, are all connected to that contextual connector. The *data role* may be connected to the *data port* of a component (provided or required) and the *contextual role* may be connected to the *contextual port* of a component. The distinction between a data and context roles (and also between a data and context ports) addresses the constraint typically imposed by many ADLs about the clear separation between functional and non-functional aspects. This ensures a quality of the components assembly by inserting a contextual connectors, as well as management of adaptation service quality.

3.5 Metamodel for Dynamic Reconfiguration

Dynamic reconfiguration is defined by transitions between configuration families (Fig.5.). Our metamodel proposes to define *configuration family* to capture a non-predefined number of configurations having close adaptation services.

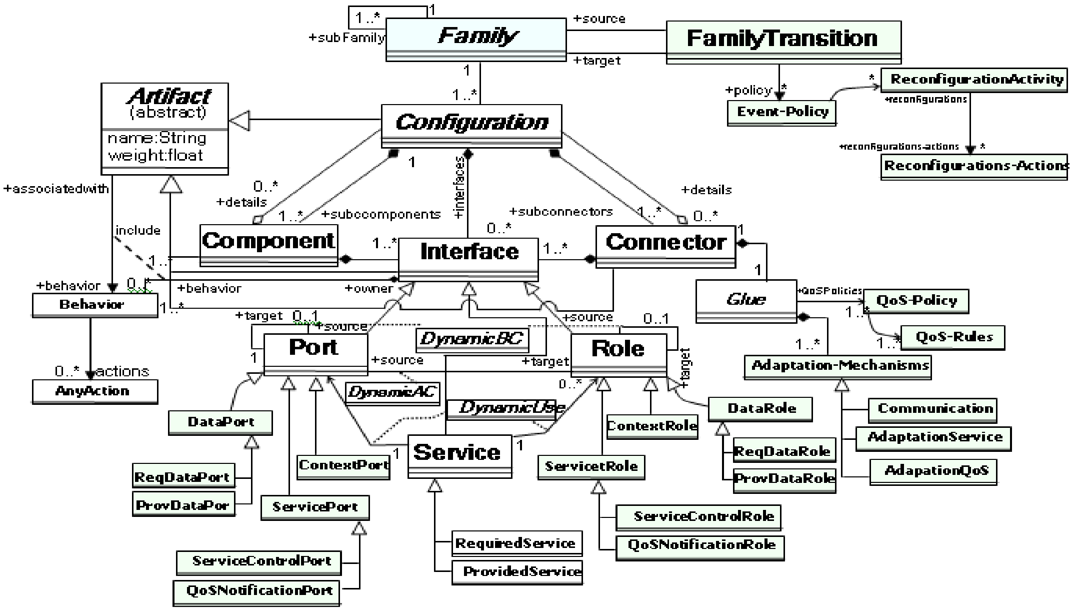


Fig. 4 Contextual architectural artifacts in ContextualArchRQMM

For each family, a specific set of adaptation services defined. For example at *image family* which includes connectors offering services of the same nature (i.e. image adaptation services) but only differs by their adaptability to the context.

A transition allows switching the system from the source configuration family to another new target configuration family. A transition can be triggered by different events, like changes in the environment, changes in the applications to be executed, or changes in the system operational conditions (e.g., a battery operated system detects a change in the battery status, or a component that becomes faulty). We can have a transition into the same configuration family; it is a transition between two configurations of the same family. For each transition, a reconfiguration activity presenting a set of reconfiguration actions is associated. It represents a set of actions switching from the current configuration to the target one. In our approach we have a non-predefined number of configurations, but we have statically predefined families. To answer to an adaptation task, on a mobile device system at the run-time, one needs to satisfy a new need related to a new execution context. The ideal solution is to install, update or remove an adaptation service at the connector's configuration. This contribution of reconfiguration is similar to other work described in a paper [11] but our work concentrates on connector reconfiguration and insisted on the separation of the two concerns: software architecture model and context model.
Four possible adaptations in *ContextualArchRQMM* are: parametric adaptations (i.e. an update parameter value command is sent along with the name and the new value of the parameter to the command queue of the connector), services adaptations (i.e. call to another available service provider by *composing* and/or *decomposing* of services using the *DynamicUse* concept), sub-family (re) assembly: (i.e. attach/detach several subfamilies into a family), move and re-routing**:** (i.e. we use the routing service to lookup another relay to deploy the desired service).

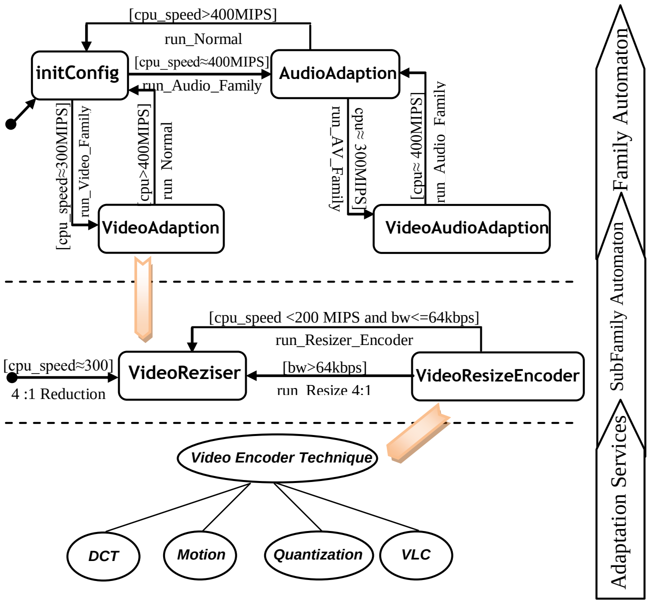


Fig. 5 Autmoaton hirerachy in the adaptation connector

4. Context-aware Quality – Model Driven Architecture (CQ-MDA)

The general structure of Context-aware Quality – Model Driven Architecture (CQ-MDA) is presented in Fig. 6. We consider the full software development cycle within MDA, i.e. from formulation of needs up to the code generation. The proposed structure consists in five levels representing CIM, PIM, Contextual Platform Independent Model (CPIM), Contextual Platform Specific Model (CPSM), and code. Each level is decomposed into three parts: the left part represents architectural artifacts and context concepts; the right part represents quality model and measurements done for these artifacts while the center part represents requirements.

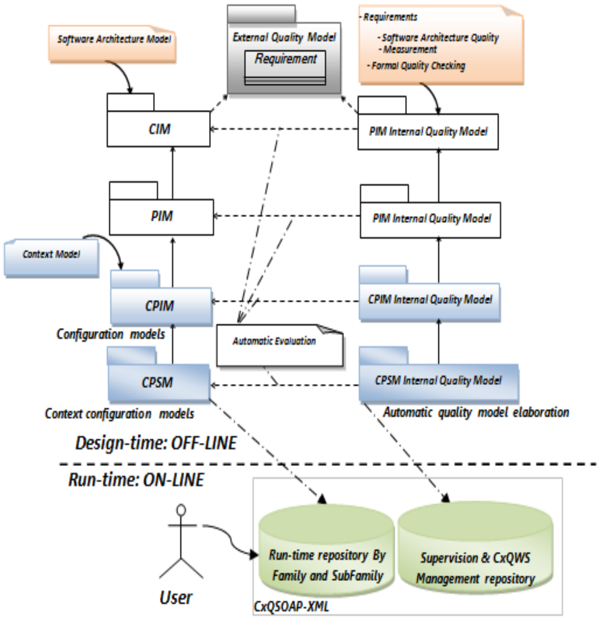

Software Architecture Model
External Quality Model
Requirement
- Requirements
- Software Architecture Quality
- Measurement
- Formal Quality Checking
CIM
PIM Internal Quality Model
PIM
PIM Internal Quality Model
Context Model
CPIM
CPIM Internal Quality Model
Configuration models
Automatic Evaluation
CPSM
CPSM Internal Quality Model
Context configuration models
Automatic quality model elaboration
Design-time: OFF-LINE
Run-time: ON-LINE
User
Run-time repository By Family and SubFamily
Supervision & CxQWS Management repository
CxQSOAP-XML

Fig. 6 Context-aware Quality Driven Model Architecture

4.1 Architecture Quality Control at the Design-Time

Architecture quality should be controlled at each steps of the design. *External requirement*s of the system are transformed into *internal ones* for the architecture and its components. *Internal requirements* are needed for assessing designed architecture models. So, particular internal models, being instances of *ContextualArchRQMM* metamodel, are used to assess particular models of CQ-MDA. The software architecture quality model is produced by measurement done for each architectural artefact for a given factor in the context of associated requirement, for a given criteria with associated metric. Two ways of using our meta-model are possible:

- The first one assumes that the software architecture quality metamodel is used for evaluating an architecture model. The architecture model is tested and validated with the semantic constraints defined by the metamodel. If the verified architecture model gets bad marks then the design process can be stopped or it can go back to the previous stage either to change requirements or to elaborate a different (better) architectural model.
- The second one, using software architecture quality metamodel considers the case when the metamodel is used for selecting the best architectural model from different choices. In this case the values of a metric are used to classify the models. A metric formula gives a note for the architecture model. The values of the metric function are used to classify the models and to choose the suitable one and we select a first model if we have the same value. After that, the selected architectural model is evaluated by the OCL constraints to remove any quality semantic violation.

4.2 Architecture Adaptation at the Run-Time

We can say that two configurations provide a close service if and only if their marks of the architecture quality criteria (i.e. context-independent) and contextual architectural quality criteria (which are related to run-time context) are close. Because context-independent quality criterion variation is more perceptible by users, platform will begin its research with the evaluation of the configurations having the same mark of context-independent quality criterion as the current configuration. In response to events notifying about changes in the environment (less bandwith, less available memory…), or in the running application (overflow/underflow of the buffer, increased transmission time…), the *Adaptation Manager* will be notified by set of probes which constitute the monitoring framework, update configurations and annotate the events to these configurations. Our platform use configurations families and subfamilies described in XML format from a preliminary analysis of the application (i.e. at the design step) in terms of QoS and update it in real time.

For an efficient and better implementation of self-management process (Fig. 7), we have used *"poisson"* simulation and formal methods (OCL) to assess the degradation of quality attributes due to movement of devices and employ runtime adaptation to mitigate such problems.

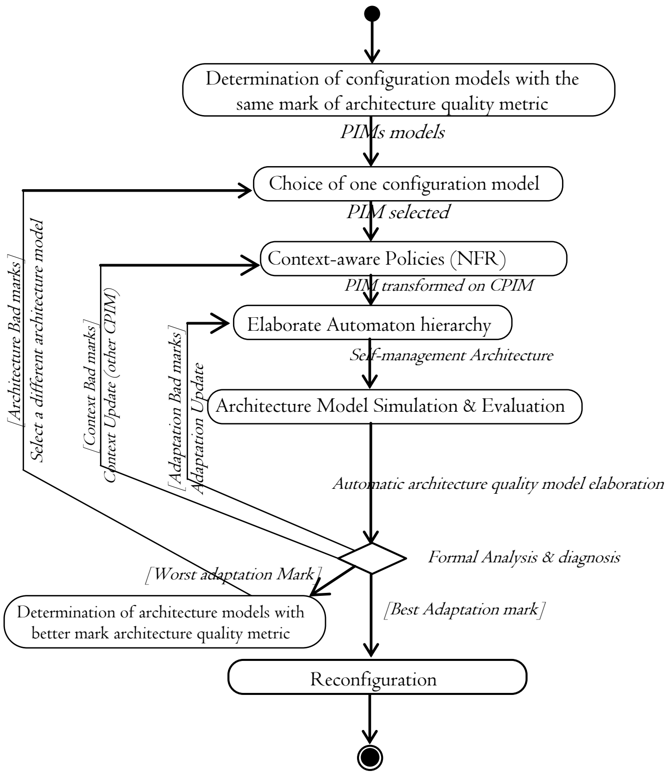

Determination of configuration models with the same mark of architecture quality metric
PIMs models
Choice of one configuration model
PIM selected
Context-aware Policies (NFR)
PIM transformed on CPIM
Elaborate Automaton hierarchy
Self-management Architecture
Architecture Model Simulation & Evaluation
Automatic architecture quality model elaboration
Formal Analysis & diagnosis
[Architecture Bad marks]
Select a different architecture model
[Context Bad marks]
Context Update (other CPIM)
[Adaptation Bad marks]
Adaptation Update
[Worst adaptation Mark]
[Best Adaptation mark]
Determination of architecture models with better mark architecture quality metric
Reconfiguration

Fig. 7 Process reconfiguration model of CQ-MDA

Our process started with the evaluation of the configurations having the same mark of structural architecture criterion *(coupling, cohesion, structural complexity…,)* as the current configuration. That will only modify the mark of the adaptative criterion (*response time, adaptation effort…*). As soon as a reconfiguration event is received, the *Quality-Manager* search for a better configuration model to using successively by analyzing finite sets of configurations *having the same mark of structural architecture quality metric and differ only by their adaptability cost to the context.* Firstly, the platform will be able to restrict the scope of the search into the range of configurations, which differ from the current configuration only by the adaptation service (or component) at the origin of the reconfiguration event. But when this approach does not give any solution, we face the issue of the deployment of a sub-family or a family. The *Adaptation Manager* receives the new selected configuration model and starts-up the reconfiguration.

5. Case Study: Video Conference System

A case study given below is intended to show applicability of CQ-MDA both for evaluation, for selection and for reconfiguration of the best architectural model from some alternatives.

A case study deals with *VideoConference* System [15]. *VideoConference* has the following optional services:

- Audio Encoder: (de-)compressing the audio stream.
- Video Encoder: (de-)compressing the video stream.
- Audio Filter: components for changing the frame size.
- Video Filter: reducing the video frame rate.

The following user preferences are considered:

- Recording, reviewing user' video and creating respective reports.
- Video should be delivered in quality and in period no longer than one minute from their request.

According to *ContextualArchRQMM*, all these requirements should be associated with a respective architecture quality model with selected quality factors. In our example, for illustration, only non-functional requirements are taken into account. It is proposed to use the efficiency factor with time-behavior sub-factor [4]. On the CIM level some internal requirements may be specified additionally to external ones. We propose "an easy maintenance of software architecture model: internal requirement" as we consider it to be important factor from architect point of view. This additional requirement can be expressed more precisely as "low complexity, high cohesion and low coupling these requirements are the main facts to take into account for achieving easy maintainability architecture (subfactors of the maintainability factor [4])." The time behaviour sub-factor for software architecture model artefact cannot be evaluated at CIM level (as the software architecture is not defined yet) and should be forwarded to the next level i.e. PIM level. Therefore the CQ-MDA approach will be shown in details using the transformation of the PIM model with respective internal quality model into CPIM model with its internal quality model and the CPIM model with respective internal quality model into CPSM model with its internal quality model.

5.1 PIM Level – Quality Control at the Design-Time

PIM model is the starting point for the considered transformation. Several architectural models can be used to design a given system. For the *VideoConference* system, the model is designed with *PipesAndFilters* style as shown in Figure 8. At PIM level we have also formally defined set of architectural artifacts that are traced from CIM model.

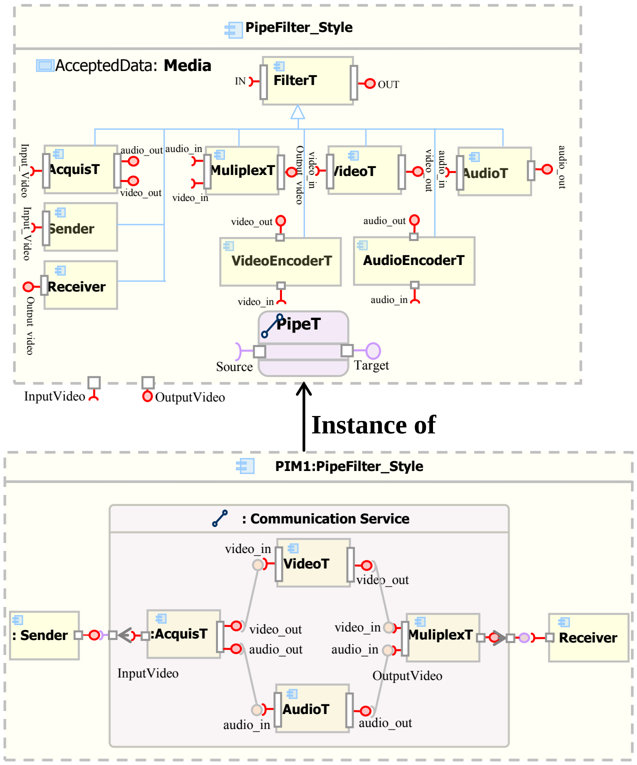


Fig. 8 PIM software architecture model

Internal quality model on this level is traced from the upper quality level model. So, we have to consider the factors from CIM level, i.e. efficiency factor with time-behaviour sub-factor and maintainability factor with modularity, analyzability sub-factors. The first factor is efficiency with sub-factor Time-behavior cannot be evaluated at this level as we have not found accepted metrics for evaluation of the PIM model. This factor must be still forwarded for evaluation to the next modeling level. The second factor is maintainability with modularity and analyzability sub-factors [4]. The first sub-factor, modularity, depends on the configuration, component and connector modularity. If the system has been divided correctly to suitable modular, the software system can be analyzed more easily. At the architecture level, this factor can be measured with criteria, named *coupling* and *cohesion*. In [4] these two metrics are proposed for measuring architecture modularity. We used these metrics in our model. We have evaluated each kind of models with similar measurements of the whole architecture of the basic metrics (i.e. coupling, cohesion and complexity).

The evaluation results are given in Tab. 1 using a prototype implemented written in Java called *QualiStyle* [4]. The architecture model should be tested and validated with the semantic constraints defined by the meta-model. If the verified architecture model gets bad marks then the design process can be stopped or it returned to the previous stage (i.e. CIM) either to change requirements or to elaborate a different (better) architectural model. High cohesion, low coupling and low complexity are the main facts to take into account for making a design understandable, maintainable, and of higher quality. All these basic metrics are in *[0, 1]*. The higher cohesion's value (resp. lower complexity's value) is the better for architecture quality. As for the architecture model from Table 1 the values of coupling is equal 0.482 and a threshold of coupling is equal 0.66, the value of cohesion is equal 0.341 and a threshold of coupling is equal 0.5 and the value of complexity is equal 0.362 and a threshold of complexity is equal 1, the architectural model provides an acceptable maintainability (a high level of cohesion, a low level of coupling, a low level of complexity). This architectural model is accepted for further transformation. This result is practically significant as well related to maintainability effort, e.g. low level of coupling, dependencies among all architectural artifacts are loss, high number of reused artifacts (i.e. number of Pipe connector instances, m = 4).

Table 1: PIM evaluation results.

PIM	*Coupling*	*Cohesion*	*Complexity*
Pipe-Filter	0.482	0.341	0.362

5.2 CPIM Level–Quality Control at the Design-Time

PIM software architecture model may be transformed, manually or automatically, into different CPIM models. The software architecture model from Fig. 8 is transformed into five CPIMs models (Fig. 9) and the total resource requirements are given in Table 2. Fig. 10 depicts our automaton for the video adaptation family.

At this level analyzability, time-behavior sub-factors taken from upper level are evaluated (it is worth to mention – different metrics can be used for this purpose). The evaluation results should be helpful in choosing the best CPIM model for further transformation.

Table 2: resources requirements

Component	*User preferences*	*CPU speed*	*Bandwith*
RateAudioT	-	≈ 100 MIPS	4:1 Reduction
ResizeVideoT	-	≈ 400 MIPS	2:1 Reduction
AudioEncoderT	High Quality Medium Quality Low Quality	≈ 300 MIPS	64 kbps 32 kbps 8 kbps
VideoEncoderT	High Quality Medium Quality Low Quality	≈200 MIPS	10:1 Reduction 20:1 Reduction 30:1 Reduction

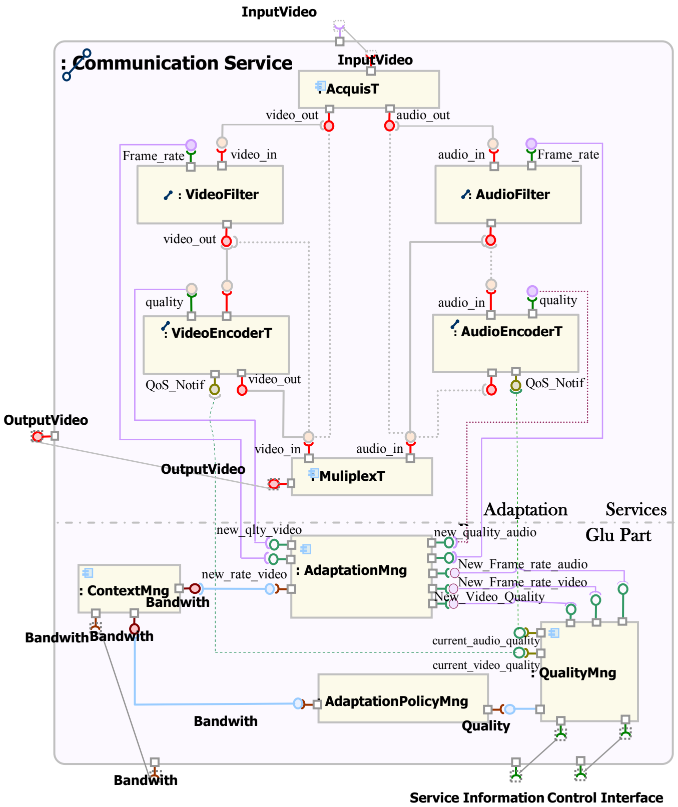


Fig. 9 Alternatives versions of CPIM models

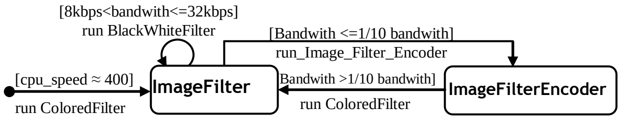


Fig.10 Video adaptation automaton

For time-behavior, three metrics proposed in [7], one of them is selected and adapted in our case. The estimated Time Behavior Metric (*TBM*) for a set *A* of artifacts of a given configuration performed with a given time in a certain context calculated as the weighted sum of TB_a metric counted for every artefact instance *"a"*:

$$TBM^{Benefit}_{Memory_{size},CPU_{speed},Network_{bw}}(config)=\sum_{a\in A} w_a * TB_a \qquad (1)$$

Apart from the evaluation of time behavior sub-factor we evaluate the analyzability sub-factor to select the best CPIM model. In [16] two metrics were proposed for the dynamic adaptivity at the architectural level, but only one, *MaAC* (Minimum architectural Adaptive Cost) was used and validated for analysability assessment in our example. According to the choice made of the sub-factors of quality and their measurement, we define the *Quality* function which measures the quality of a given configuration:

$$Quality(config)=\frac{TBM^{Benefit}_{Memory_{size},CPU_{speed},Network_{bw}}(config)}{MaAC^{Cost}_{Memory_{size},CPU_{speed},Network_{bw}}(config)} \qquad (2)$$

Table 3 shows the evaluation results, meaning that CPIM5 turns out to be the best. Differences can be seen in the adaptation cost of this CPIM and other CPIMs, which is due to the low adaptation effort compared to other CPIMs. This result is practically significant as well related to adaptation effort e.g. number of artifacts which should be added to make a system adaptive are very loss as consequence of self- management for environment evolution (i.e. *CPU usage*, *bandwith*) guided by the adaptation policies.

Table 3: CPIMs evaluation results

Adaptable and optional services	*TBM (ms)*	*MaAC (artifact nb)*
Video Resize, High Quality Video Encoder/Decoder	200 ~ 400	0 ~ 16
Video Resize, Medium Quality Video Encoder/Decoder	200 ~ 330	0 ~ 16
Video Resize, Low Quality Video Encoder/Decoder	350 ~ 500	0 ~ 8
Video Resize, Audio Encoder/Decoder	470 ~ 800	0 ~ 8
All Adaptable Services	420 ~ 930	0

5.3 Architecture Adaptation at the Run-Time

Participants to the video conference are interested for service quality in the face of device heterogeneity. We distinguish two Participants' families: speaker and auditor. The service quality requirement can be satisfied by using our context quality management strategy. The goal for a given mobile device is to achieve qualities and allocate resources to result in the best configuration such that the system quality is maximized subject to device resource constraints, user preferences constraints. The platform is capable of adding/removing/updating/moving services at the execution time. The important task of our platform is to perform the dynamic changes at the run-time and, more precisely, with minimum length of time and decision making. It is necessary to have a mechanism for media flow measurement which will detect when the application must be reconfigured for reasons of lower available bandwith. In addition, it is necessary to know when the bandwidth is sufficient to switch to another configuration. So, we propose to use our context quality management. We can see the different adaptations in the following scenarios:

Scenario # 1. The application is first of all deployed in a favorable context, where neither the stations nor the network are saturated. Initially, the context is sufficient to provide both video and audio. If we receive a video stream packaged with *RealVideo* in a 120 x120 window at 10 frames / second with phone audio quality, rate of 56Kbits is sufficient.

Scenario # 2. The supervisor has noticed a problem of bandwith, and thinks that the bandwith will not hold until the end of the video. To detect a decrease in media throughput, the *Adaptation Manager* receives two events of the buffering connector corresponding to *overflow*/*underflow* of the buffer size (i.e. 20% - 80% of the buffer size) from the supervisor. When an event of underflow is received, it indicates a problem of the video transmission (loss of information transmission, increased transmission time). Since, an overflow event implies that the current bandwith is not sufficient. To alleviate too many changes (i.e. minimum reconfiguration cost) in the current configuration, the application can switch to the ideal configuration if the video stream of data can be supported for long enough time (*depending on the size of the buffer*). The ideal management on bandwith degradation is to follow a minor change by the replacement of a service connector (*Video Encoder/Decoder with High quality)* by another connector service (*Video Encoder/Decoder with Lower quality*).

Scenario # 3. In another scenario, due to movement of devices, the network throughput connecting the devices is very loss, making it difficult for communication service to interact with auditor. The platform looks for a new configuration to use, starting by looking for a new relay allowing the moving a video resize connector to a suitable device.

6. Related Works and Discussion

The first related area of research are ADLs that have been proposed for representing dynamic architectures including: ACME [14], π-ADL [6], C3 [2] and AADL [1]. However, except for ACME, most ADLs do not support the concept of evaluation function. In addition, most of them are not contextual defined. MARTE [17] does not treat the problem of heterogeneity by a meta-model which verifies the adequacy of service regarding its context and research of the adaptation strategy [19]. Π-ADL [6] is a formal architecture description language based on the π-calculus. It does not support contextual connectors and not integrate quality metrics. Recently, Garlan and al. [14] extended ACME ADL in order to support evaluation function in evolution styles and their multiple decision forms. However, this work does not consider exploiting contextual connectors in heterogeneous environment where entities of different nature collaborate: software and hardware components. The second related area of research are some works involving quality in MDA approach, like QADA (Quality-driven Architecture Design and Quality Analysis) [8] – a methodology targeted at the development of service architectures. Other works involving Context in MDA approach, e.g. Context-aware Model Driven Architecture Model Transformation [13] – a methodology targeted at the development of context-aware applications and other networked systems. These works concentrate only on quality system architecture or context-aware system architecture, while CQ-MDA insisted on the separation of the two concerns: software architecture model and context model.

7. Conclusion and Future Works

This paper proposed *ContextualArchRQMM* metamodel centred on the concept of contextual connector, which take advantage of traditional architectural connectors and provides a lightweight support for the definition of some composition facilities such as heterogeneous interfaces at the connector level. The paper proposed also CQ-MDA approach based on ContextualArchRQMM, being an extension to the MDA, allows for considering quality and resources-awareness while conducting the design process. The main idea of presented extension consists of three abstractions levels: PIM, CPIM and CPSM. At the PIM level, a model is decomposed on two interrelated models: software architecture artifacts, which reflect functional requirements and quality model. At the CPIM level a simultaneous transformation of these two models with contextual information details are elaborated and then refined to a specific platform at the CPSM level. Such a procedure ensures that the transformation decisions should be based on the quality assessment of the created models. At design-time, our approach is used to assess the quality attributes of the system's architectures. At run-time, the framework copes with the challenges posed by the highly dynamic nature of mobile systems through continuous monitoring and calculation of the most suitable architecture. If a better architecture is found, the framework adapts at run-time the software, potentially via connector adaptation and mobility. We presented an illustrative example to show the applicability of the proposed CQ-MDA approach. The results of the experiments (based on the example of *VideoConference* with four CPIMs) are encouraging. The experiment shows that our approach outperforms two abstractions level in terms of some quality metrics such as adaptation ratio and time response. In the future, we will consider moving our approach to a real execution platform to validate its feasibility.

Acknowledgments

We would like to thank our colleagues and students for testing our simulation tool and specially *Hamza Reffad* and *Bekakchi Youcef*. The authors appreciate the in-depth comments given by the anonymous reviewers to improve this work.

References

[1] B. Berthomieu1, J.P. Bodeveix, C. Chaudet, F. Vernadat, “Formal Verification of AADL Specifications in the Topcased Environment,” 14th Ada-Europe International Conference, 2009, pp. 207 – 221.

[2] A. Amirat and M. Oussalah, “First-Class Connectors to Support Systematic Construction of Hierarchical Software Architecture,” Journal of Object Technology, Vol. 8. N°.7, 2009, pp. 107-130.

[3] A. Alti, A. Boukerram and A. Smeda, “Architectural Styles Quality Evaluation and Selection,” 9th International Conference NOTERE’09, Montréal (Canada), 2009.

[4] A. Alti, A. Smeda, “Architectural Styles Quality Evaluation and Selection,” Proceeding of 4th International Conference on Software and Technologies (ICSOFT’2009), Barcelona (Spain), 2009, pp. 74 - 82.

[5] J. Miller, J. Mujerki, editors. “MDA Guide, Version 1.0. OMG Technical Report,”, http://www.omg.org/docs/ptc/03-05-01.pdf, 2003.

[6] F. Oquendo, “π-ADL: an architecture description language based en the higher order typed π-calculus for specifying dynamic and mobile software architecture,” ACM Soft. Eng., vol. 29, n°. 4, 2004, pp. 1 - 13.

[7] ISO/IEC 9126-3. In Software Engineering – Product quality – Part 3: Internal metrics, ISO-IEC, 2003.

[8] QADA, http://virtual.vtt.fi/qada , 2007.

[9] P. Tarvainen, “Adaptability Evaluation at Software Architecture Level, ” The Open Soft. Eng. J. vol. 2, Bentham Sc. Pub. Ltd., 2008, pp. 1-30.

[10] F. Losavio, L. Chirinos, N. Lévy, and A. Ramdane Cherif, “Quality characteristics for software architecture,” JOT, 2(2), 2003, pp. 133-150.

[11] F. Kritchen, B. Hamid, B. Zalila and B. Coulette, “Designing Dynamic Reconfiguration for Distributed Real Time Embedded Systems,” 10th International Conference NOTERE’2010, Tozeur (Tunisia) ,2010, pp. 249-254.

[12] OMG. UML OCL 2.0 Specification: Revised Final Adopted Specification. http://www.omg.org/docs/ptc/05-06-06.pdf, June 2005.

[13] S. Vale, S. Hammoudi, Context-aware Model Driven Development by Parameterized Transformation. MDISIS’2008, pp. 167–180.

[14] D. Garlan, J.M. Barnes, B. Schmerl, O. Celiku., “Evolution Styles: Foundations and Tool Support for Software Architecture Evolution,” WICSA’09, 2009, pp. 16-25.

[15] S. Laplace, M. Dalmau, P. Roose, Prise en compte de la qualité de service dans la conception et l’exploitation d’applications réparties, In the Workshop GEDSIP@Inforsid 2009, Toulouse, 26 mai 2009.

[16] C. Raibulet, L. Masciadri, “Evaluation of Dynamic Adaptivity through Metrics: an Achievable Target?” WICSA’09, 2009, pp. 65-71.

[17] S. Gérard, D. Petriu and J. Medina. “MARTE: A New Standard for Modeling and Analysis of Real-Time and Embedded Systems”, 19th Euromicro Conf. on Real-Time Systems (ECRTS 07), Pisa, Italy, 2007.

[18] OMG. A UML Profile for MARTE: Modeling and Analysis of Real-Time Embedded systems, June 2008, http://www.omg.org/docs/ptc /09-06-08.pdf , 2008.

[19] C. Marcel, R Michel, Christian M. “Autonomic Adaptation based on Service-Context Adequacy Determination”. In ENTCS, p. 35-50, 2007.

[20] M. Dalmau, P. Roose, S. Laplace. “Context Aware Adaptable Applications - A global approach”, Special Issue on Pervasive Computing Systems and Technologies - International Journal of Computer Science - IJCSI Vol. 1, Issue 1, 2009 - ISSN 1694-0784

Adel Alti obtained the Master degree from the University of Setif (UFAS), Algeria, in 1998. He is holding a Ph.D. degree in software engineering from UFAS university of Sétif, Algeria, 2011. Right now he is an associate professor at University of Sétif. He is a member of the research group LRSD. His area of interests includes automated software engineering, mapping multimedia concepts into UML, semantic integration of architectural description into MDA platforms, context-aware quality software architectures and automated service management, Context and QoS. During his work he has published number of publications concerning these subjects.

Roose Philippe is an associate professor at the LIUPPA/UPPA – FRANCE. He obtained his PhD degree in computer science from university of Bayonne, France, 2001. He head of the MOJITO and AEXIUM projects. His research interests are software architecture and platforms, pervasifs and ubiquitous computing, mobility, software components services, context and QoS, multi-parts profiles. He is the co-author of three books on software component technologies.

IJCSI CALL FOR PAPERS SEPTEMBER 2013 ISSUE

Volume 10, Issue 5

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.
See authors guide for manuscript preparation and submission guidelines.

Accepted papers will be published online and indexed by Google Scholar, Cornell's University Library, DBLP, ScientificCommons, CiteSeerX, Bielefeld Academic Search Engine (BASE), SCIRUS, EBSCO, ProQuest and more.

Deadline: 10th September 2013
Online Publication: 30th September 2013

- Evolutionary computation
- Industrial systems
- Evolutionary computation
- Autonomic and autonomous systems
- Bio-technologies
- Knowledge data systems
- Mobile and distance education
- Intelligent techniques, logics, and systems
- Knowledge processing
- Information technologies
- Internet and web technologies
- Digital information processing
- Cognitive science and knowledge agent-based systems
- Mobility and multimedia systems
- Systems performance
- Networking and telecommunications
- Software development and deployment
- Knowledge virtualization
- Systems and networks on the chip
- Context-aware systems
- Networking technologies
- Security in network, systems, and applications
- Knowledge for global defense
- Information Systems [IS]
- IPv6 Today - Technology and deployment
- Modeling
- Optimization
- Complexity
- Natural Language Processing
- Speech Synthesis
- Data Mining

For more topics, please see http://www.ijcsi.org/call-for-papers.php

All submitted papers will be judged based on their quality by the technical committee and reviewers. Papers that describe on-going research and experimentation are encouraged.
All paper submissions will be handled electronically and detailed instructions on submission procedure are available on IJCSI website (www.IJCSI.org).

For more information, please visit the journal website (www.IJCSI.org)

www.IJCSI.org

IJCSI

The International Journal of Computer Science Issues (IJCSI) is a well-established and notable venue for publishing high quality research papers as recognized by various universities and international professional bodies. IJCSI is a refereed open access international journal for publishing scientific papers in all areas of computer science research. The purpose of establishing IJCSI is to provide assistance in the development of science, fast operative publication and storage of materials and results of scientific researches and representation of the scientific conception of the society.

It also provides a venue for researchers, students and professionals to submit ongoing research and developments in these areas. Authors are encouraged to contribute to the journal by submitting articles that illustrate new research results, projects, surveying works and industrial experiences that describe significant advances in field of computer science.

Indexing of IJCSI

1. Google Scholar
2. Bielefeld Academic Search Engine (BASE)
3. CiteSeerX
4. SCIRUS
5. Docstoc
6. Scribd
7. Cornell's University Library
8. SciRate
9. ScientificCommons
10. DBLP
11. EBSCO
12. ProQuest

© IJCSI PUBLICATION
www.IJCSI.org