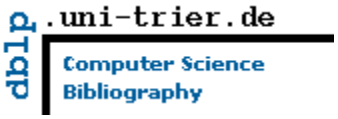
IJCSI

International Journal of Computer Science Issues

Volume 8, Issue 5, No 1, September 2011
ISSN (Online): 1694-0814

www.IJCSI.org

IJCSI proceedings are currently indexed by:



IJCSI Publicity Board 2011

IJCSI Editorial Board 2011

EDITORIAL

In this fifth edition of 2011, 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 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 8, Issue 5, No 1, September 2011 (IJCSI Vol. 8, Issue 5, No 1). The acceptance rate for this issue is 31.7%.

IJCSI Editorial Board
September 2011 Issue
ISSN (Online): 1694-0814

www.IJCSI.org

IJCSI Reviewers Committee 2011

• Mr. Markus Schatten, University of Zagreb, Faculty of Organization and Informatics, Croatia
• Mr. Vassilis Papataxiarhis, Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece
• Dr Modestos Stavrakis, University of the Aegean, Greece
• Dr Fadi KHALIL, LAAS -- CNRS Laboratory, France
• Dr Dimitar Trajanov, Faculty of Electrical Engineering and Information technologies, ss. Cyril and Methodius Univesity - Skopje, Macedonia
• Dr Jinping Yuan, College of Information System and Management,National Univ. of Defense Tech., China
• Dr Alexis Lazanas, Ministry of Education, Greece
• Dr Stavroula Mougiakakou, University of Bern, ARTORG Center for Biomedical Engineering Research, Switzerland
• Dr Cyril de Runz, CReSTIC-SIC, IUT de Reims, University of Reims, France
• Mr. Pramodkumar P. Gupta, Dept of Bioinformatics, Dr D Y Patil University, India
• Dr Alireza Fereidunian, School of ECE, University of Tehran, Iran
• Mr. Fred Viezens, Otto-Von-Guericke-University Magdeburg, Germany
• Dr. Richard G. Bush, Lawrence Technological University, United States
• Dr. Ola Osunkoya, Information Security Architect, USA
• Mr. Kotsokostas N.Antonios, TEI Piraeus, Hellas
• Prof Steven Totosy de Zepetnek, U of Halle-Wittenberg & Purdue U & National Sun Yat-sen U, Germany, USA, Taiwan
• Mr. M Arif Siddiqui, Najran University, Saudi Arabia
• Ms. Ilknur Icke, The Graduate Center, City University of New York, USA
• Prof Miroslav Baca, Faculty of Organization and Informatics, University of Zagreb, Croatia
• Dr. Elvia Ruiz Beltrán, Instituto Tecnológico de Aguascalientes, Mexico
• Mr. Moustafa Banbouk, Engineer du Telecom, UAE
• Mr. Kevin P. Monaghan, Wayne State University, Detroit, Michigan, USA
• Ms. Moira Stephens, University of Sydney, Australia
• Ms. Maryam Feily, National Advanced IPv6 Centre of Excellence (NAV6) , Universiti Sains Malaysia (USM), Malaysia
• Dr. Constantine YIALOURIS, Informatics Laboratory Agricultural University of Athens, Greece
• Mrs. Angeles Abella, U. de Montreal, Canada
• Dr. Patrizio Arrigo, CNR ISMAC, italy
• Mr. Anirban Mukhopadhyay, B.P.Poddar Institute of Management & Technology, India
• Mr. Dinesh Kumar, DAV Institute of Engineering & Technology, India
• Mr. Jorge L. Hernandez-Ardieta, INDRA SISTEMAS / University Carlos III of Madrid, Spain
• Mr. AliReza Shahrestani, University of Malaya (UM), National Advanced IPv6 Centre of Excellence (NAv6), Malaysia
• Mr. Blagoj Ristevski, Faculty of Administration and Information Systems Management - Bitola, Republic of Macedonia
• Mr. Mauricio Egidio Cantão, Department of Computer Science / University of São Paulo, Brazil
• Mr. Jules Ruis, Fractal Consultancy, The netherlands
• Mr. Mohammad Iftekhar Husain, University at Buffalo, USA
• Dr. Deepak Laxmi Narasimha, Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

• Dr. Paola Di Maio, DMEM University of Strathclyde, UK
• Dr. Bhanu Pratap Singh, Institute of Instrumentation Engineering, Kurukshetra University Kurukshetra, India
• Mr. Sana Ullah, Inha University, South Korea
• Mr. Cornelis Pieter Pieters, Condast, The Netherlands
• Dr. Amogh Kavimandan, The MathWorks Inc., USA
• Dr. Zhinan Zhou, Samsung Telecommunications America, USA
• Mr. Alberto de Santos Sierra, Universidad Politécnica de Madrid, Spain
• Dr. Md. Atiqur Rahman Ahad, Department of Applied Physics, Electronics & Communication Engineering (APECE), University of Dhaka, Bangladesh
• Dr. Charalampos Bratsas, Lab of Medical Informatics, Medical Faculty, Aristotle University, Thessaloniki, Greece
• Ms. Alexia Dini Kounoudes, Cyprus University of Technology, Cyprus
• Dr. Jorge A. Ruiz-Vanoye, Universidad Juárez Autónoma de Tabasco, Mexico
• Dr. Alejandro Fuentes Penna, Universidad Popular Autónoma del Estado de Puebla, México
• Dr. Ocotlán Díaz-Parra, Universidad Juárez Autónoma de Tabasco, México
• Mrs. Nantia Iakovidou, Aristotle University of Thessaloniki, Greece
• Mr. Vinay Chopra, DAV Institute of Engineering & Technology, Jalandhar
• Ms. Carmen Lastres, Universidad Politécnica de Madrid - Centre for Smart Environments, Spain
• Dr. Sanja Lazarova-Molnar, United Arab Emirates University, UAE
• Mr. Srikrishna Nudurumati, Imaging & Printing Group R&D Hub, Hewlett-Packard, India
• Dr. Olivier Nocent, CReSTIC/SIC, University of Reims, France
• Mr. Burak Cizmeci, Isik University, Turkey
• Dr. Carlos Jaime Barrios Hernandez, LIG (Laboratory Of Informatics of Grenoble), France
• Mr. Md. Rabiul Islam, Rajshahi university of Engineering & Technology (RUET), Bangladesh
• Dr. LAKHOUA Mohamed Najeh, ISSAT - Laboratory of Analysis and Control of Systems, Tunisia
• Dr. Alessandro Lavacchi, Department of Chemistry - University of Firenze, Italy
• Mr. Mungwe, University of Oldenburg, Germany
• Mr. Somnath Tagore, Dr D Y Patil University, India
• Ms. Xueqin Wang, ATCS, USA
• Dr. Borislav D Dimitrov, Department of General Practice, Royal College of Surgeons in Ireland, Dublin, Ireland
• Dr. Fondjo Fotou Franklin, Langston University, USA
• Dr. Vishal Goyal, Department of Computer Science, Punjabi University, Patiala, India
• Mr. Thomas J. Clancy, ACM, United States
• Dr. Ahmed Nabih Zaki Rashed, Dr. in Electronic Engineering, Faculty of Electronic Engineering, menouf 32951, Electronics and Electrical Communication Engineering Department, Menoufia university, EGYPT, EGYPT
• Dr. Rushed Kanawati, LIPN, France
• Mr. Koteshwar Rao, K G Reddy College Of ENGG.&TECH,CHILKUR, RR DIST.,AP, India
• Mr. M. Nagesh Kumar, Department of Electronics and Communication, J.S.S. research foundation, Mysore University, Mysore-6, India
• Dr. Ibrahim Noha, Grenoble Informatics Laboratory, France
• Mr. Muhammad Yasir Qadri, University of Essex, UK
• Mr. Annadurai .P, KMCPGS, Lawspet, Pondicherry, India, (Aff. Pondicherry Univeristy, India
• Mr. E Munivel , CEDTI (Govt. of India), India
• Dr. Chitra Ganesh Desai, University of Pune, India
• Mr. Syed, Analytical Services & Materials, Inc., USA

• Mrs. Payal N. Raj, Veer South Gujarat University, India
• Mrs. Priti Maheshwary, Maulana Azad National Institute of Technology, Bhopal, India
• Mr. Mahesh Goyani, S.P. University, India, India
• Mr. Vinay Verma, Defence Avionics Research Establishment, DRDO, India
• Dr. George A. Papakostas, Democritus University of Thrace, Greece
• Mr. Abhijit Sanjiv Kulkarni, DARE, DRDO, India
• Mr. Kavi Kumar Khedo, University of Mauritius, Mauritius
• Dr. B. Sivaselvan, Indian Institute of Information Technology, Design & Manufacturing, Kancheepuram, IIT Madras Campus, India
• Dr. Partha Pratim Bhattacharya, Greater Kolkata College of Engineering and Management, West Bengal University of Technology, India
• Mr. Manish Maheshwari, Makhanlal C University of Journalism & Communication, India
• Dr. Siddhartha Kumar Khaitan, Iowa State University, USA
• Dr. Mandhapati Raju, General Motors Inc, USA
• Dr. M.Iqbal Saripan, Universiti Putra Malaysia, Malaysia
• Mr. Ahmad Shukri Mohd Noor, University Malaysia Terengganu, Malaysia
• Mr. Selvakuberan K, TATA Consultancy Services, India
• Dr. Smita Rajpal, Institute of Technology and Management, Gurgaon, India
• Mr. Rakesh Kachroo, Tata Consultancy Services, India
• Mr. Raman Kumar, National Institute of Technology, Jalandhar, Punjab., India
• Mr. Nitesh Sureja, S.P.University, India
• Dr. M. Emre Celebi, Louisiana State University, Shreveport, USA
• Dr. Aung Kyaw Oo, Defence Services Academy, Myanmar
• Mr. Sanjay P. Patel, Sankalchand Patel College of Engineering, Visnagar, Gujarat, India
• Dr. Pascal Fallavollita, Queens University, Canada
• Mr. Jitendra Agrawal, Rajiv Gandhi Technological University, Bhopal, MP, India
• Mr. Ismael Rafael Ponce Medellín, Cenidet (Centro Nacional de Investigación y Desarrollo Tecnológico), Mexico
• Mr. Supheakmungkol SARIN, Waseda University, Japan
• Mr. Shoukat Ullah, Govt. Post Graduate College Bannu, Pakistan
• Dr. Vivian Augustine, Telecom Zimbabwe, Zimbabwe
• Mrs. Mutalli Vatila, Offshore Business Philipines, Philipines
• Mr. Pankaj Kumar, SAMA, India
• Dr. Himanshu Aggarwal, Punjabi University,Patiala, India
• Dr. Vauvert Guillaume, Europages, France
• Prof Yee Ming Chen, Department of Industrial Engineering and Management, Yuan Ze University, Taiwan
• Dr. Constantino Malagón, Nebrija University, Spain
• Prof Kanwalvir Singh Dhindsa, B.B.S.B.Engg.College, Fatehgarh Sahib (Punjab), India
• Mr. Angkoon Phinyomark, Prince of Singkla University, Thailand
• Ms. Nital H. Mistry, Veer Narmad South Gujarat University, Surat, India
• Dr. M.R.Sumalatha, Anna University, India
• Mr. Somesh Kumar Dewangan, Disha Institute of Management and Technology, India
• Mr. Raman Maini, Punjabi University, Patiala(Punjab)-147002, India
• Dr. Abdelkader Outtagarts, Alcatel-Lucent Bell-Labs, France
• Prof Dr. Abdul Wahid, AKG Engg. College, Ghaziabad, India
• Mr. Prabu Mohandas, Anna University/Adhiyamaan College of Engineering, india
• Dr. Manish Kumar Jindal, Panjab University Regional Centre, Muktsar, India

• Mr. Neetesh Gupta, Technocrats Inst. of Technology, Bhopal, India
• Ms. A. Lavanya, Manipal University, Karnataka, India
• Ms. D. Pravallika, Manipal University, Karnataka, India
• Prof. Vuda Sreenivasarao, St. Mary's college of Engg & Tech, India
• Prof. Ashutosh Kumar Dubey, Assistant Professor, India
• Mr. Ranjit Singh, Apeejay Institute of Management, Jalandhar, India
• Mr. Prasad S.Halgaonkar, MIT, Pune University, India
• Mr. Anand Sharma, MITS, Lakshmangarh, Sikar (Rajasthan), India
• Mr. Amit Kumar, Jaypee University of Engineering and Technology, India
• Prof. Vasavi Bande, Computer Science and Engneering, Hyderabad Institute of Technology and Management, India
• Dr. Jagdish Lal Raheja, Central Electronics Engineering Research Institute, India
• Mr G. Appasami, Dept. of CSE, Dr. Pauls Engineering College, Anna University - Chennai, India
• Mr Vimal Mishra, U.P. Technical Education, Allahabad, India
• Dr. Arti Arya, PES School of Engineering, Bangalore (under VTU, Belgaum, Karnataka), India
• Mr. Pawan Jindal, J.U.E.T. Guna, M.P., India
• Prof. Santhosh.P.Mathew, Saintgits College of Engineering, Kottayam, India
• Dr. P. K. Suri, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, India
• Dr. Syed Akhter Hossain, Daffodil International University, Bangladesh
• Mr. Nasim Qaisar, Federal Urdu Univetrsity of Arts , Science and Technology, Pakistan
• Mr. Mohit Jain, Maharaja Surajmal Institute of Technology (Affiliated to Guru Gobind Singh Indraprastha University, New Delhi), India
• Dr. Shaveta Rani, GZS College of Engineering & Technology, India
• Dr. Paramjeet Singh, GZS College of Engineering & Technology, India
• Prof. T Venkat Narayana Rao, Department of CSE, Hyderabad Institute of Technology and Management , India
• Mr. Vikas Gupta, CDLM Government Engineering College, Panniwala Mota, India
• Dr Juan José Martínez Castillo, University of Yacambu, Venezuela
• Mr Kunwar S. Vaisla, Department of Computer Science & Engineering, BCT Kumaon Engineering College, India
• Prof. Manpreet Singh, M. M. Engg. College, M. M. University, Haryana, India
• Mr. Syed Imran, University College Cork, Ireland
• Dr. Namfon Assawamekin, University of the Thai Chamber of Commerce, Thailand
• Dr. Shahaboddin Shamshirband, Islamic Azad University, Iran
• Dr. Mohamed Ali Mahjoub, University of Monastir, Tunisia
• Mr. Adis Medic, Infosys ltd, Bosnia and Herzegovina
• Mr Swarup Roy, Department of Information Technology, North Eastern Hill University, Umshing, Shillong 793022, Meghalaya, India
• Mr. Suresh Kallam, East China University of Technology, Nanchang, China
• Dr. Mohammed Ali Hussain, Sai Madhavi Institute of Science & Technology, Rajahmundry, India
• Mr. Vikas Gupta, Adesh Instutute of Engineering & Technology, India
• Dr. Anuraag Awasthi, JV Womens University, Jaipur, India
• Dr. Mathura Prasad Thapliyal, Department of Computer Science, HNB Garhwal University (Centr al University), Srinagar (Garhwal), India
• Mr. Md. Rajibul Islam, Ibnu Sina Institute, University Technology Malaysia, Malaysia
• Mr. Adnan Qureshi, University of Jinan, Shandong, P.R.China, P.R.China
• Dr. Jatinderkumar R. Saini, Narmada College of Computer Application, India

TABLE OF CONTENTS

An efficient algorithm for the nearest neighbourhood search for point clouds

Luca Di Angelo[1] and Luigi Giaccari[2]

[1] Department of Industrial Engineering, University of L'Aquila
L'Aquila, 67100, Italy

[2] ANSYS Germany Gmbh
Otterfing, 83624, Germany

Abstract

This paper presents a high-performance method for the k-nearest neighbourhood search. Starting from a point cloud, first the method carries out the space division by the typical cubic grid partition of the bounding box; then a new data structure is constructed. Based on these two previous steps, an efficient implementation of the k-nearest neighbourhood is proposed. The performance of the method here presented is compared with that of the *kd-tree* and *bd-tree* algorithms taken from the ANN library [1] as regards the computing time for some benchmarking point clouds and artificially generated test cases. The results are analysed and critically discussed.

Keywords: *k-nearest neighbour, point cloud, space partition.*

1. Introduction

For the last few years the use of points as the representational primitives of geometric models has spread out in computer graphics and geometric modelling applications ([2], [3] and [4]). This is also due to the recent introduction on the market of 3D scanning systems offering high resolutions with a measuring accuracy as high as 10 μm, which make it possible to capture the smallest surface features. However, these devices generate very large data sets. Some point clouds, such as those obtained by means of 3D scanning, cannot be directly used in the previously defined applications; they need to be processed in order to reconstruct high-level information starting from the only Cartesian coordinates. Typically, point clouds are processed to remove any residual noise, change the sampling rate, estimate the points' normal and/or proceed to their tessellation. All these operations require the computation of the k-nearest neighbourhoods (*knn*) for each point in the cloud. As pointed out by Sankaranarayanan et al in [4], the correct computation of neighbourhoods is important both for algorithms that estimate properties that are common in the neighbourhood and for algorithms that analyse variations in these properties. It is evident that these neighbourhoods must be obtained at the lowest computational cost as possible, so that even clouds with several millions points can be easily managed. When analysing the related literature, it seems evident that the methods, used with the typical computing powers, show such a performance that they constitute the major bottleneck in the implementation of the above-mentioned applications.

In order to make a useful contribution to this field, this paper proposes a simple algorithm for the *knn* search. It is based on a new data structure applied to the typical space division approach, which makes possible a more efficient search for the nearest neighbourhoods. This method is tested for the *knn* search in some benchmarking point clouds and artificially generated test cases. The results derived from these experiments are critically discussed hereinafter.

2. Related works

The more recent exhaustive overview of the *knn* search methods are presented in [4]. In what follows we will be considering some of the most important papers which are related to the method here being proposed, leaving out, for example, algorithms which work with multiple processors CPU and GPU and for the approximation of the k–nearest neighbourhoods.

The simplest method to construct the k–nearest neighbours of datasets is based on the simple brute-force algorithm [5]: first the Euclidean distances between each point and all the other ones are calculated; then, the k-nearest neighbours are found as those k points with the shortest distances. This algorithm is computationally inefficient since, for each data point, its time complexity is $O(n_p^2)$, where n_p is the number of points. In order to reduce this computational cost, many methods are proposed in literature; the most important ones can be divided into three main categories:

- Voronoi-point based;
- space division strategy based;
- pivot – based.

2.1 Voronoi point based approaches

The methods belonging to the first category are mainly used in two-dimensional datasets and are based on the consideration that the Voronoi diagram decomposes the plane into cells, each of which contains a point. For a point **p** contained in the cell *C*, the points located in the cells sharing edges with *C* are the nearest to **p**.

The first algorithm that uses this approach in a three-dimensional dataset is presented by Dey et al. in [6] which proposed a method that is based on the dual of the Voronoi graph: it determines the k–nearest neighbours in a three-dimensional dataset by constructing a Dirichlet triangulation. As pointed out by Li et al in [7], it takes $O(n_p^2 \log_2 n_p)$ time, making it impractical for use in reverse engineering where, more and more often, the clouds have over one million points.

In order to improve the efficiency of the search, recently Goodsell in [8] has proposed a new method for two–dimensional datasets, which is based on the Voronoi points; the results reported show that the timing of the algorithm is quadratic ($O(n_p^2)$).

2.2 Space division strategy based approaches

Typically, with the space division strategy based methods determining whether a point is a member of the k-nearest neighbours permits to work with a small subset of the data; this way, computational costs are strongly reduced. Some of the most widely used algorithms for the *knn* search belonging to this category are the *kd-tree* and *bd-tree*. The first one is based on a k – dimensional binary search tree ([9], [10]). By the *kd-tree* the space is hierarchically partitioned into hyper – rectangular regions (*buckets*) by using hyper – planes perpendicular to the coordinate axes to form a tree. Once this structure is constructed, the search for the nearest neighbour is done by descending the tree to find the bucket containing the query point. The search for the *knn* is limited to the points within that bucket or those contained in the near buckets. Optimally, the *kd-tree* requires $O(n_p log_2 n_p)$ operations for its construction and an $O(log_2 n_p)$ operation for the search ([10] and [12]). The *box – decomposition tree* (*bd-tree*) ([13]) is a variant of the *kd-tree* that was introduced to provide greater robustness for highly clustered datasets. Above all, the *bd-tree* differs from the *kd-tree* in the fact that, in addition to the *splitting* operation, there is another decomposition operation called *shrinking*. According to the *shrinking* rule, it is possible to further divide a box containing more points than the bucket size.

Piegl and Tiller in [14] proposed a much simple algorithm for computing all the k-nearest neighbours in 2-D. Firstly, the dataset is partitioned by a rectangular grid and the points are binned in appropriate cells. If several points fall under the same bin, they are stored in a linked list. The search is extended to the rings (in ascending order) around the cell containing the query point. The search stops when the k-th shortest distance is smaller than the distance between the query point and the closest wall of the outer cell ring. The empirical tests show that the algorithm is sub-linear for small k (around 1-5% of the data); it is linear for medium k (up to about 10-20% of data) and quadric for large k (over 20% of data). Furthermore, the algorithm seems to not be practically affected by the topology of the point cloud and by the grid size.

Li and Cripps in [7] proposed a method for which the bounding box containing the points is first partitioned by a cubic grid, whose grid size (ρ_2) is estimated by the following empirical formula:

$$\rho_2 = \alpha \left(\sqrt[3]{\frac{(x_{max} - x_{min})(y_{max} - y_{min})(z_{max} - z_{min})}{n_p}} \right) \qquad (1)$$

where:

- α is a user – defined scalar factor;
- n_p is the number of points.

The points are stored by using the following cube structure:

$$cube[i][j][k] \text{ with } i=1,\ldots,n_x;\ j=1,\ldots,n_y;\ k=1,\ldots,n_z; \qquad (2)$$

where n_x, n_y and n_z are the number of divisions along **x**, **y** and **z** directions, respectively. For each point **p** of the cloud, the search for the k-nearest points is carried out among those (*candidate points*) which are inside the inner and intersecting cubes of a sphere with centre at **p** and a radius $r = \rho_2 \cdot \min(n_x, n_y, n_z) \cdot \sqrt[3]{\frac{k}{n_p}}$. If the number of *candidate points* is less than *k*, the search is carried out inside a sphere of larger radius. The experimental results show that the timings are not significantly affected by the structure of the point clouds and that they are approximately linear for $k<0.05n_p$.

An efficient and simple method is proposed by Franklin in [15]. Concerning the data structure, it essentially consists of a ragged array, containing the points belonging to each cubical cell, and of a dope vector pointing to the first point of the cell. The search of closest points of a query point **q** is carried out inside a rectangular blocks of cells around that containing **q**, by using a sorted cells list.

In order to improve the efficiency of the k-nearest neighbourhood search, Gejun et al in [16] put forth a new strategy for space division. After a preliminary division where the side-length grid is chosen by the user, a secondary division is done on the basis of an empirical formula. The experiments' results show, if we focus only on the *knn* search speed, that after the second division the search range has been reduced and the searching efficiency has improved.

Several techniques are used in order to transform the d-dimensional data points in 1-d values. Some solutions of this type are based on the *pyramid technique* ([17], [18], [19]). This technique, proposed by Berchtold et al. in [20], consists in the partitioning the d-dimensional space $[0,1]^d$ (called *unit hypercube*) into 2d pyramids with the tops at (0,5; 0,5; : : : ; 0,5) and bases on each of the 2d faces of the *unit hypercube*. At each point a hash value, that is the sum between the identification number of the pyramid to which the point belongs and the distance of the point from the pyramid vertex, is assigned. All the points are stored, according to hash values, in a B^+-tree for optimal querying. The reported results in [18] show that the proposed method has a speed-up factor over the *kd-tree* between 1.6 and 2.9.
The previously presented approaches compute the neighbourhood of each point of a cloud, one point at a time. Sankaranarayanan et al. in [4] presented a more sophisticated approach that reuses point neighbourhoods already calculated to determine neighbourhoods of adjacent points. Moreover, in order to manage a large amount of points, the authors use a disk-based data structure. The results reported show that the method's performance is promising, above all, in terms of capability to elaborate clouds with 50 millions points and not in terms of computational times.

2.3 Pivot – based approaches

Generally, the pivot based methods select some *pivots* from the database and classify all the other elements according to their distance from the *pivots*. The distances $d(\mathbf{s}_j, \mathbf{p}_i)$ between elements ($\mathbf{s}_j$) and pivots ($\mathbf{p}_i$) and between the query $\mathbf{q_k}$ and the pivots ($d(\mathbf{s}_j, \mathbf{q}_k)$) are used to filter out elements. Typical algorithms belonging to this group are the *AESA* ([16]), the *LAESA* ([22] and [23]) and its variants ([24] and [25]) and the *Fixed Queries Array* ([26]). These algorithms are based on the common idea that if for some pivots $\mathbf{p}_i$ $\left|d(\mathbf{s}_j, \mathbf{p}_i) - d\left(\mathbf{s}_j, \mathbf{q}_k\right)\right| > r$ then, by the triangular inequality, $d(\mathbf{q}_k, \mathbf{p}_i) > r$ without explicitly evaluating $d(\mathbf{q}_k, \mathbf{p}_i)$. All the points which do not verify the first previous inequality must be directly compared against the query point. By increasing the number of pivots, distance evaluations increase but so does the number of elements being filtered out. As pointed out by Chavez et al. in [27], the optimum value of the pivots cannot be normally reached because it is too high in terms of space requirements. Recently some improvements to these algorithms have been proposed by Chavez et al. in [27] and Fredriksson in [28].

3. Algorithm description

In this section we describe our algorithm in detail. Generally speaking, it consists of two main steps:

- the data structure construction;
- the nearest-neighbourhood search.

3.1 The data structure construction

Efficient k-nearest neighbourhood search requires an efficient data structure which prevents from searching the entire dataset for each candidate point. For this purpose, when using *SDS* (acronym of *Search Data Structure*), the points are first indexed as {1, ..., n_p}, where n_p is the number of points. Then, similarly to the methods based on the space division strategy, the axes-aligned bounding box of the points is partitioned by a cubic grid. The box edge size (*step*) is then evaluated by the density parameter ρ of the points ($\rho = \frac{n_p}{n_{box}}$, where n_p is the number of points and n_{box} is the number of boxes), by using the formula (1) proposed by Li and Cripps in [7]. Every point is assigned to its corresponding box by the following hashing function which performs a double to integer conversion:

$$box_{id} = \frac{x}{d} + \frac{y}{d} \cdot n_x + \frac{z}{d} \cdot n_x \cdot n_y \qquad (3)$$

where:

- *x*, *y* and *z* are the point coordinates;
- n_x, n_y are the number of divisions along **x** and **y** directions, respectively.

The box_{id} is the attribute to each data point that performs the assignment of a point to a box. All the points in a box are implicitly sorted by the index (from *1* to n_p).
The new data structure consists of two different arrays (*First* and *Next*). *First* has dimension $n_{box}x1$, where n_{box} is the number of boxes partitioning the point cloud. In the *i-th* row of *First*, of all the points contained in the *i-th* box, the point having the lowest index is recorded. The flag -1 is used for an empty box. *Next* has dimension $n_p x1$, where n_p is the number of points. In the *j-th* row of *Next*, the point contained in the *i-th* box and having an index value immediately higher than *j* is recorded. The flag -1 is used for the last point in the box.
By using this structure, it is possible to access all the points in one box with the following simple and fast operation:

idPoint=Next[IdPoint] (4)

Figure 1 shows an example of decomposition of a cubic box (figure 1a) and also reports the corresponding data structure (figure 1b) for six 3D points.
In the case of very large scanned point clouds, this data structure produces a great number of empty boxes which occupy a lot of memory space and may cause the search in many useless boxes. In order to overcome this problem, a modified data structure is also proposed (henceforth *SDS_m*). In particular, the *first_m* is an associative container in which the *key value* is the number of non –

empty box and the *mapped value* is the corresponding first point included within (figure 1c). In this case, the access to the first point of the box is obtained by a binary search:

idPoint=BynarySearch(first, boxid) (5)

This access is intrinsically slower than that in Eq. (4); *SDS_m* is convenient, as opposed to the previous version, in cases of a very small percentage of non–empty boxes.

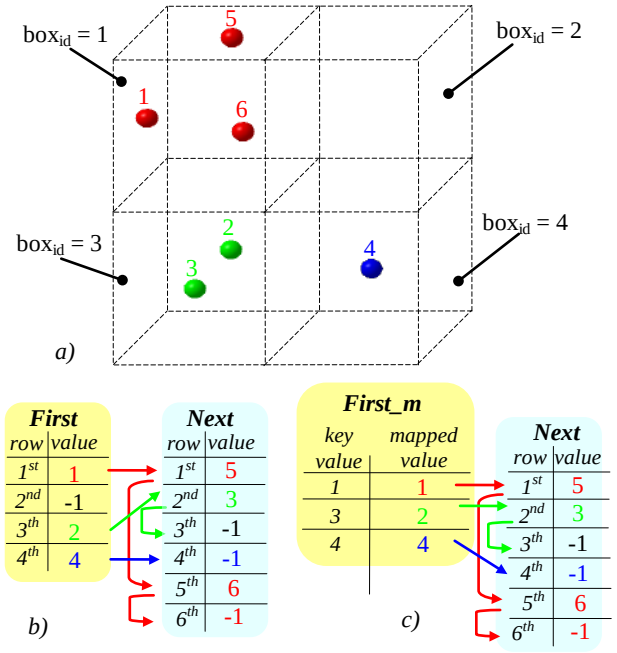


Figure 1. An example of decomposition of a cubic box (a) and the corresponding data structures (b) and (c) for six 3D points

3.2 The nearest-neighbourhood search

In order to better explain the algorithm for the nearest neighbourhood search, figure 2 proposes the pseudo – code for the *SDS* data structure and the 2D case. The reported considerations are easily extendible to the 3D space and the *SDS_m*.

The first box to be analysed is the one including the query point (**q**); for each point contained in that box the distance from **q** is evaluated. Then, the distances between **q** and the walls of the box are calculated (figure 3b). These distances are concerned with which boxes need to be further analysed (figure 3c). Contrary to the algorithm proposed by Piegl and Tiller in [14] and by Franklin in [15], this strategy prevents that all the cells belonging to the i-th ring of the cell containing the query point are explored. The search is performed until the k-th shortest distance is smaller than that existing between the query point and the closest wall of the outer box being considered. In a similar way to the typical methods presented in literature, *k* distances are stored in a priority queue with standard insertion, a structure suited for small values of *k*. The algorithm has also some controls to prevent any search attempts outside the axes-aligned bounding box of the points.

3.3 The input parameter of the method

The proposed method requires that the box size parameter should be set. The box size parameter affects the performance of the query operations of the SDS. Generally speaking, the ideal box size identifies boxes with just one point (no "overloaded boxes"). The number of boxes affects memory usage, and the number of points inside a box affects the number of distances which need to be computed during the nearest point search time. In practical cases, the optimal box size minimises empty and overloaded boxes. In what follows specific experimentations are carried out so as to investigate, in different typical applications, the values that best satisfy these conflicting constraints.

4. Performance of SDS

As pointed out by Hoppe et al. in [29], it is difficult to analyse analytically the time complexity of the search for the k-nearest neighbourhood since it strongly depends on the input data. For this reason, we have empirically analysed the performance of the *SDS* in the data structure construction and in the nearest neighbour search for some benchmarking scanned point clouds and artificially generated test cases Furthermore, in order to qualify how the proposed method compares with the state-of-art, its performances are compared with that of the *kd-tree* and the *bd-tree* algorithms taken from the ANN library [1]. Any other methods have not been analysed because their implementations were not available. Obviously, all the methods which have been analysed perform the same identification of the nearest points but they show different time complexity. All the tests have been run on a WorkStation with 3.0 GHz Intel Xeon processor and 16.0 Gb RAM.

4.1 Scanned point clouds

The test cases being considered are the typical benchmarks used in the related literature to evaluate the *knn* search methods (table 1). They consist of 16 point clouds acquired with different sampling rate from objects having different number of points and geometries ([30], [31] and [32]). Some of these point clouds are very large datasets (*Amphora*, *Neptune*, *Asian Dragon* and *Thai Statue*).

Let **q**(x_q, y) be the query point;
Let *k* be the number of the nearest points to **q** being needed;
Let *d*(**q**,**p**[i]) be the distance between **q** and i-th point of the dataset;
Let d_{min} *[1:k]* be the sorted list in ascending order containing the k-smallest distances between **q** and the data points;
Let *indices[1:k]* be pointers to the k-nearest points from **q**;
Let *dx_left* be the distance between **q** and the vertical left wall of the box containing **q**;
Let *dx_right* be the distance between **q** and the vertical right wall of the box containing **q**;
Let *dy_up* be the distance between **q** and the horizontal upper wall of the box containing **q**;
Let *dy_down* be the distance between **q** and the horizontal bottom wall of the box containing **q**;
Let *r* be the ring level;
Let *rx_max*, *rx_min*, *ry_max*, *ry_min* be the offset reached by the ring;
Let n_x and n_y be the number of divisions along **x** and **y** directions;
Let *step* be the cell dimensions;
Let *BoxSearch* be the function to calculate the distances between the points contained in a box and **q** and to upgrade the d_{min}*[1:k]*:

```
void BoxSearch(ix ,iy)
{
    id= ix+ iy*ny; /* get box id from coordinates */
    idPoint=First[id]; /* first point of the box */
    while (idPoint>-1) /* loop all the points in the box */
    {
        dist = d(q,p[idPoint]);
        if (dist < d_min[k]) /* compute squared distance between query points and reference point idp */
        {  insertion of dist in d_min; /* update the array of the smallest distances between q and the data points */
           insertion of idPoint in indices; /* update nearest neighbour pointer*/}
        idPoint = Next[idPoint];} /* get the next reference point in the box */
}
```

```
initialise r=1, d_min[1:k] = big number;
initialise rx_max, rx_min, ry_max, ry_min = 0;
evaluation of dx_left, dx_right, dy_up, dy_down;
evaluation of ix = (x_q /d) and iy = (y_q /d)
BoxSearch(ix, iy);
goon=true;
while (goon=true)
{  goon=false;

   if ((d_min[k] > dx_left) AND (ix-r)>=0)
    {  // dx_left upgrade
       goon = true;
       rx_min=-r;
       for j=ry_min to ry_max
       {   BoxSearch(ix-r ,iy+j);}
       dx_left = dx_left + step;
    }

   if ((d_min[k] > dx_right) AND (ix+r)< n_x)
    {  // dx_right upgrade
       goon = true;
       rx_max=r;
       for j=ry_min to ry_max
       {   BoxSearch(ix+r ,iy+j);}
       dx_right = dx_right + step;
    }

   if ((d_min[k] > dy_up) AND (iy+r)< n_y)
    {  // dx_up upgrade
       goon = true;
       ry_max=r;
       for i=rx_min to rx_max
       {   BoxSearch(ix+i ,iy+r);}
       dy_up = dy_up + step;
    }

   if ((d_min[k] > dy_down) AND (iy+r)< n_y)
    {  dy_ up upgrade
       goon = true;
       ry_min=-r;
       for i=rx_min to rx_max
       {   BoxSearch(ix+i ,iy-r);}
       dy_down = dy_down + step;
    }

   r=r+1; */ go to the next ring */
} /* end while*/
```

Figure 2. The nearest-neighbourhood search algorithm

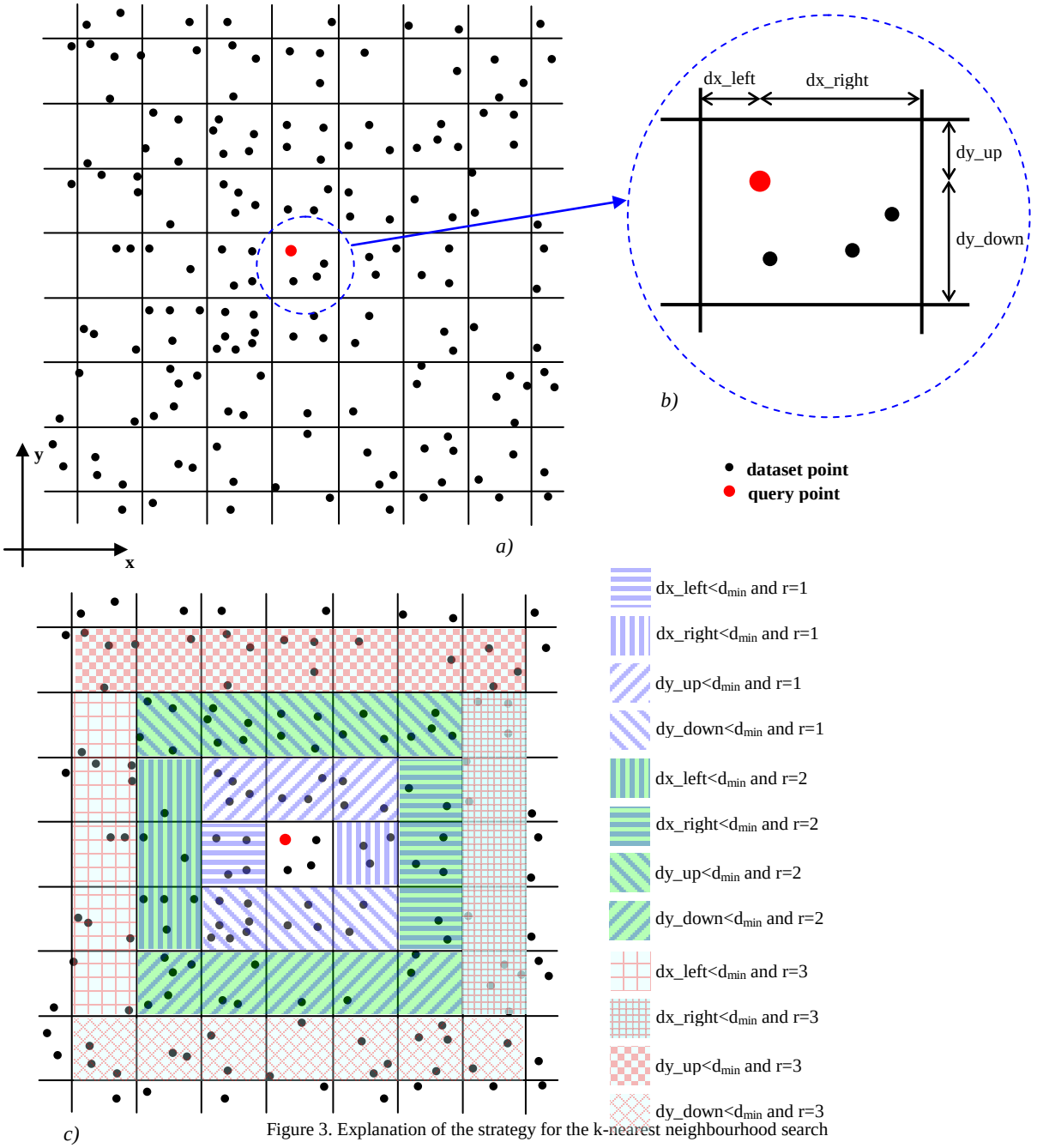


Figure 3. Explanation of the strategy for the k-nearest neighbourhood search

A first experiment is carried out in order to verify the influence on *SDS* of the density parameter ρ by varying the number of nearest points (k) needed. As it was to be expected, as the value of ρ decreases, the computational time for the data structure construction and the amount of memory usage increase. But, on the other hand, when analysing the total computational time it is possible to verify that, for each k, the best results are obtained when the density value decreases as the number of points in the cloud increases. The obtained results show that ρ can be successfully approximated as a function of n_p, according to the following expression:

$$\rho = \max\left(-0.0618 \cdot \ln\left(n_p\right) + 0.969; 0.05\right) \qquad (6)$$

The *coefficient of determination* of the logarithmic regression approximating the 16 scanned point clouds is $R^2 = 0.8618$. No explicit dependence has been noticed on the k value. The choice to keep down the value of ρ to 0.05

comes from the necessity to contain memory usage in the case of very large datasets. Eq. (6) has proved to be effective for ρ estimation also in the case of point clouds whose results are not here reported since they have not been used to build the previous regression. Thus, in what follows, the experiment is carried out by using the value of ρ resulting from (6).

Table 1: Scanned point clouds used to evaluate the *knn* search methods

Name	*n. of points*
Rocker-arm	10,044
Stanford Bunny	35,947
Horse	48,485
Armadillo	172,975
Pulley	293,672
Dragon	435,545
Bimba	502,694
Happy Buddha	543,652
Rolling Stage	596,903
Chinese Dragon	655,980
Turbine Blade	882,954
Nicolò da Uzzano	946,760
Amphora	1,317,152
Neptune	2,003,933
Asian Dragon	3,609,601
Thai Statue	4,999,997

The table 2 reports the ratio between the number of empty boxes and the number of boxes ($n_{empty-boxes}/n_{boxes}$) and also the average computational time per query point for the scanned point clouds of table 1. The value of the ratio can be considered to increase with the number of points, reaching a maximum value of 0.9940. Except for the three largest datasets (namely, *Neptune*, *Asian Dragon* and *Thai Statue*), the value of the average computational time for each k remains almost constant for the different clouds. The last result verifies that the *SDS* is not practically affected by the topology of the point cloud. Furthermore, by analysing separately the two contributions of the computational time, it is found that the *SDS* structure provides a time complexity $O(n_p)$ for the data structure construction and $O(n_p \bullet log_2(n_p))$ for the nearest neighbour search, versus the time complexity $O(n_p \bullet log_2(n_p))$ for both phases provided by the two other methods. The table 3 reports the mean values of the speed-ups obtained by comparing the total computational time (structure construction and nearest neighbour search) of the *SDS* with those of the *kd-tree* and the *bd-tree*, for the different *k* values. It is evident that the *SDS* is always computationally convenient, but the gain value decreases as the value of *k* increases. This is due to the fact that when the *k* value is increased the three methods tend toward the brute search.

If we use expression (6) in the case of very large datasets, the best value of ρ decreases to values which require a great amount of memory. Therefore, a further experimentation is carried out in order to verify the effectiveness of the *SDS*_m in these cases. The table 4 reports, by way of an example, the performance comparison between the *SDS* and the *SDS_m* in the cases of *Neptune*, *Asian Dragon* and *Thai Statue* with varying ρ and for $k = 8$. With equal ρ, an average decrease of 19% of speed-ups produces an average decrease of 86% in memory usage by the data structure. Very similar conclusions are reached when analysing the results for the other values of *k* considered.

4.2 Uniform point clouds

The test cases analysed in this section consist of different clouds whose n_p points are randomly generated with a uniform probability density and are contained in a cube. All the time values reported further down have been obtained as the mean time values of 30 cases analysed.

On analysing the influence of the density parameter ρ when varying the number of nearest points (*k*) needed, obtained results show that ρ can be approximated as a function of k/n_p, according to the following expressions:

$$\begin{cases} \rho = 0.1021 \cdot \log\left(\frac{k}{n_p}\right) + 1.99 & \text{for } \frac{k}{n_p} \le 0.04 \\ \rho = 0.1609 \cdot \log\left(\frac{k}{n_p}\right) + 1.51 & \text{for } \frac{k}{n_p} > 0.04 \end{cases} \quad (7)$$

The *coefficient of determination* of the two logarithmic regressions are $R^2 = 0.5312$ and $R^2 = 0.6842$ respectively. In what follows, the experiment is performed by using the value of ρ resulting from Eq. (7). After analysing separately the two contributions of the computational time, it is found that, similarly to what happens with scanned point clouds, the *SDS* provides a time complexity $O(n_p)$ for the data structure construction and $O(n_p \bullet log_2(n_p))$ for the nearest neighbour search, versus the time complexity $O(n_p \bullet log_2(n_p))$ for both phases provided by the two other methods. The table 5 reports the speed-ups mean value obtained by comparing the total computational time (structure construction and nearest neighbour search) of the *SDS* with those of the *kd-tree* and the *bd-tree*, for some *k* values. As it has been previously highlighted, the structure of *SDS* and the expression of ρ, as proposed in Eq. (7), make the method being proposed particularly efficient in the *knn* search for small values of *k*.

With a view to verifying the efficiency of the data structure and the *knn* search method which is here presented, a further experimentation is carried out so as to compare the time performance of *SDS*, *kd-tree* and *bd-tree* with that of the brute search in the case of a small number of points. The table 6 shows the mean value of the speed up for different *k*. It is evident that the use of SDS is convenient also for uniform point clouds with few points.

4.3 Very pathological cases

There are cases in which the *SDS*, like the rest of methods based on a cubic grid partition of the bounding box, cannot be used because it turns out to be strongly inefficient as opposed to the *kd-tree* and the *bd-tree*. A typical example is a scanned point cloud with one point that is very distant from the others. By using in this case the value of ρ obtained according to expression (6), both $n_{empty-boxes}/n_{boxes}$ and the maximum number of points contained in a box strongly increase. In the worst case, all the points, except for the outlier, are within a single box: *SDS* degenerates into the brute-search algorithm with a computational time complexity of $O(n_p^3)$, as opposed to the $O(n_p \bullet log_2(n_p))$ of the worst case with the *kd-tree* and *bd-tree*.

The final experiment is carried out in order to find a parameter that efficiently measures (without affecting the computational time) the dataset uniformity and its limit value for which *SDS* is not convenient as opposed to the other methods. The test cases here considered are artificially generated starting from the 16 scanned point clouds reported in table 1 by adding, to each cloud, a point at the minimum distance from the others for which the *SDS* is inconvenient, for each *k*, as opposed to the other two methods. The results show that the ratio $n_{empty-boxes}/n_{boxes}$ is a satisfactory parameter that does not affect the computational time; furthermore, the obtained value made it possible to define the limit values according to the following expressions:

$$if \left\{ \left[(n_p > 2*10^5)\ AND \left(\frac{n_{empty-boxes}}{n_{boxes}} \geq 0.996 \right) \right] \text{ OR } \left[(n_p \leq 2*10^5)\ AND \left(\frac{n_{empty-boxes}}{n_{boxes}} \geq 0,0021 \ln(n_p) + 0.97 \right) \right] \right\}$$

$$\Rightarrow \text{ switch to } kd\text{-}tree \text{ or } bd\text{-}tree \qquad (8)$$

Since in this paper we have demonstrated that the construction of its data structure is very efficient, the *SDS* can be used to measure the uniformity of the dataset; if n_p and $n_{empty-boxes}/n_{boxes}$ satisfy any of the two conditions in (8), then the *knn* search needs to be done by the *kd-tree* or the *bd-tree*.

4. Conclusion

This paper has presented a high performance method for the *k*-nearest neighbourhood search. Said method is based on a data structure founded on the typical cubic grid partition of the bounding box. Like the principal methods presented in literature, also *SDS* requires that a parameter ($\rho = n_p/n_{box}$) should be empirically set; expressions which furnish its best value are proposed in typical applications. The performance of the *SDS* is verified by a comprehensive experiment which analyses both typical scanned and uniform point clouds. The results obtained show that the *SDS* structure provides, for both types of clouds, a time complexity $O(n_p)$ for the data structure construction and $O(n_p \bullet log_2(n_p))$ for the nearest neighbour search. Furthermore, for the scanned point clouds, the results verify that the *SDS* is not practically affected by the topology of point cloud.

When comparing the performance of the *SDS* with that of the *kd-tree* and the *bd-tree* algorithms taken from the ANN library [1], it is evident that the *SDS* is always computationally convenient; the gain value decreases as the value of *k* increases. This is due to the fact that when the *k* value is increased, the three methods tend toward the brute search.

Finally, the use of the *SDS* is also convenient, as opposed to the brute search algorithm, for uniform point clouds with few points.

References

[1] http://www.cs.umd.edu/~mount/ANN/

[2] M. Andersson, J. Giesen, M. Pauly, B. Speckmann, “Bounds on the k-neighbourhood for locally uniformly sampled surfaces”. In Proceedings of the Euro – graphics symposium on point-based graphics, June 2-4, 2004, Zurich, Switzerland, pp. 167–171.

[3] M. Pauly, R. Keiser, L. P. Kobbelt, M. Gross, “Shape modelling with point-sampled geometry”, ACM Transactions on Graphics, Vol. 22, No. 3, 2003, pp. 641–50.

[4] J. Sankaranarayanan, H. Samet, A. Varshney, “A fast all nearest neighbour algorithm for applications involving large point-clouds”, Comp. & Graphics, Vol. 31, No. 2, 2007, pp. 157–174.

[5] M. Sarkar and T. Leong, “Application of k-nearest neighbours algorithm on breast cancer diagnosis problem”, In Proceedings of the 2000 AMIA Annual Symposium, November 4–8, 2000, Los Angeles, California, USA.

[6] T. Dey, C. Bajaj, and K. Sugihara, “On good triangulation in three dimensions”, In Proceedings of the ACM Symposium on Solid modelling and application, 1991 (ACM Press, New York), pp. 431–441.

[7] X. Li and R. J. Cripps, “Algorithm for finding all k-nearest neighbours in three dimensional scattered data points and its application in reverse engineering”, Proceedings of the Institution of Mech. Engineers, Part B: Journal of Engineering Manufacture, Vol. 221, No. 9, 2007, 1467-1472.

[8] G. Goodsell, “On finding p-th nearest neighbours of scattered points in two dimensions for small p”, Computer Aided Geometric Design, Vol. 17, No. 4, 2000, pp. 387–392.

[9] J.L. Bentley, “Multidimensional Binary Search Trees Used for Associative Searching”, Communications of the ACM, Vol. 18, No. 9, 1975, pp. 509-517.

Table 2. Percentage of the number of empty boxes and average computational time per query point for the 16 scanned point clouds of Table 1

Name	$\frac{n_{empty-boxes}}{n_{boxes}}$	Average computational time per query point [μs] k=1	k=2	k=4	k=8	k=16	k=32	k=64	k=128	k=256
Rocker-arm	0.8636	0.45	0.79	1.04	1.19	2.12	4.10	8.34	18.66	41.83
Stanford Bunny	0.9304	0.46	0.51	0.69	1.21	2.08	3.97	8.02	17.81	42.31
Horse	0.9551	0.52	0.54	0.76	1.32	2.30	4.20	8.30	18.75	40.70
Armadillo	0.9690	0.39	0.54	0.71	1.19	2.25	3.98	8.03	17.81	40.27
Pulley	0.9445	0.35	0.58	0.81	1.41	2.37	4.57	9.38	21.37	48.83
Dragon	0.9681	0.37	0.51	0.76	1.24	2.23	4.01	8.14	19.19	41.65
Bimba	0.9747	0.38	0.60	0.81	1.37	2.41	4.33	8.82	19.21	42.49
Happy Buddha	0.9687	0.37	0.53	0.78	1.28	2.27	4.24	8.68	19.72	45.83
Rolling Stage	0.9781	0.37	0.49	0.74	1.18	2.09	3.81	7.83	18.23	39.35
Chinese Dragon	0.9702	0.37	0.64	0.85	1.45	2.54	4.62	9.43	21.07	46.91
Turbine Blade	0.9597	0.33	0.46	0.72	1.15	1.98	4.04	8.53	19.84	45.48
Nicolò da Uzzano	0.9871	0.43	0.72	0.89	1.44	2.66	5.02	10.52	22.01	48.09
Amphora	0.9802	0.36	0.49	0.75	1.21	2.09	3.88	7.99	19.05	41.81
Neptune	0.9926	0.54	0.84	1.09	1.71	2.91	5.57	11.69	26.12	59.07
Asian Dragon	0.9940	0.45	0.74	0.93	1.57	2.86	5.40	11.81	24.83	114.04
Thai Statue	0.9922	0.45	0.73	1.05	1.76	3.22	6.47	13.87	38.81	173.23

Table 3. Speed-up mean value obtained for the scanned point clouds used to evaluate the *knn* search methods

		k=1	k=2	k=4	k=8	k=16	k=32	k=64	k=128	k=256
SDS	$time_{kd-tree}/time_{SDS}$	6.84	5.98	4.87	3.72	2.88	2.38	2.03	1.88	1.94
	$time_{bd-tree}/time_{SDS}$	11.28	9.69	7.89	5.90	4.55	3.72	3.12	2.71	2.66

Table 4. Performance comparison between the *SDS* and the *SDS_m* in the case of very large datasets.

Name	ρ	Memory usage by the data structure [Mb]		Speed ups			
		SDS	*SDS_m*	$time_{kd\text{-}tree}/time_{SDS}$	$time_{bd\text{-}tree}/time_{SDS}$	$time_{kd\text{-}tree}/time_{SDS\_m}$	$time_{bd\text{-}tree}/time_{SDS\_m}$
Neptune	0.05	160,53	8,64	4,05	6,56	3,26	5,29
	0.1	84,09	8,30	3,73	6,05	2,98	4,83
	0.15	58,61	8,15	3,50	5,68	2,78	4,52
	0.2	45,87	8,07	3,30	5,36	2,62	4,24
	0.25	38,22	8,01	3,13	5,07	2,45	3,97
Asian Dragon	0.05	289,16	15,43	4,26	6,85	3,53	5,68
	0.1	151,46	14,84	4,06	6,52	3,32	5,33
	0.15	105,57	14,60	3,90	6,27	3,16	5,07
	0.2	82,62	14,46	3,70	5,94	2,98	4,79
	0.25	68,85	14,37	3,52	5,66	2,85	4,59
Thai Statue	0.05	400,54	22,04	5,21	8,36	4,31	6,91
	0.1	209,81	21,02	4,91	7,87	4,01	6,43
	0.15	146,23	20,59	4,62	7,41	3,72	5,97
	0.2	114,44	20,34	4,34	6,96	3,57	5,72
	0.25	95,37	20,17	4,13	6,62	3,42	5,48

Table 5. Speed-up mean value obtained for the uniform point clouds used to evaluate the *knn* search methods

k	Speed-ups	n_p = 100	1,000	10,000	100,000	250,000	500,000	1,000,000
1	$time_{kd-tree}/time_{SDS}$	11.4	7.3	7.4	8.9	9.3	9.3	12.2
	$time_{bd-tree}/time_{SDS}$	14.8	12.6	13.9	14.8	14.8	14.7	19.1
16	$time_{kd-tree}/time_{SDS}$	3.6	2.4	2.3	5.5	3.3	2.8	2.8
	$time_{bd-tree}/time_{SDS}$	5.1	3.5	3.4	7.8	4.5	3.7	3.6

25 6	$time_{kd-tree}/time_{SDS}$	--	1.7	1.9	3.3	2.6	2.2	2.1
	$time_{bd-tree}/time_{SDS}$	--	2.1	2.2	4.4	3.2	2.7	2.5

Table 6. Speed-up mean value obtained for the scanned point clouds used to evaluate the *knn* search methods

	Speed ups		
$\mathbf{n_p}$	$time_{brute-search}/time_{SDS}$	$time_{brute-search}/time_{kd-tree}$	$time_{brute-search}/time_{bd-tree}$
10	1.16	0.12	0.11
25	1.19	0.14	0.12
50	1.29	0.26	0.21
100	1.70	0.39	0.27

[10] J.L. Bentley, "Multidimensional Binary Search Trees in Database Applications", IEEE Transactions on Software Engineering, Vol. 5, No. 4, 1979, pp. 333-340.
[11] J. L. Bentley, B.W. Weide, "Optimal Expected-Time Algorithms for Closest Point Problems", ACM Transactions on. Mathematical Software, Vol. 6, No. 4, 1980, pp. 563-580.
[12] J. L. Bentley, "Multidimensional Divide-and-Conquer", Communications of the ACM, Vol. 23, No. 4, 1980, pp. 214-229.
[13] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, A. Wu, "An optimal algorithm for approximate nearest neighbour searching", Journal of the ACM, Vol. 45, No. 6, 1998, pp. 891–923.
[14] L. A. Piegl and W. Tiller, "Algorithm for finding all k nearest neighbours", Computer Aided Design, Vol. 34, No. 2, 2002, pp. 167-172.
[15] W. R. Franklin, "NearPt3: Nearest Point Query on 184M Points in E3 with a Uniform Grid", In Proceedings of the 17th Canadian Conference on Computational Geometry (CCCG), 2005, pp. 239-242.
[16] Z. Gejun, M. Changsheng, X. Feng, "The K-nearest neighbour fast searching algorithm of scattered data", In proceedings of the Intern. Conf. on Future Information Technology and Management Engineering (FITME), October 9-10, 2010, Changzhou, China, pp. 125 – 128.
[17] D. H .Lee and H. J .Kim, "An efficient technique for nearest-neighbour query processing on the SPY-TEC", IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No 6, 2003, pp. 1472-1486.
[18] B. G. Nickerson and Q. Shi, "K-nearest neighbour search using the pyramid technique", In Proceedings of the 18th Canadian Conference on Computational Geometry (CCCG), 2006, pp. 155-158.
[19] R. Zhang, P. Kalnis, B. C. Ooi, and K. L. Tan, "Generalized multidimensional data mapping and query processing", ACM Transactions on Database Systems, Vol. 30, No 3, 2005, pp. 661-697.
[20] S. Berchtold, C. Böhm, and H. P. Kriegel, "The pyramid-technique: towards breaking the curse of dimensionality", In Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, 1998, pp. 142-153.
[21] E. Vidal, "An algorithm for finding nearest neighbours in (approximately) constant average time", Pattern Recognition Letters, Vol. 4, No. 3, 1986, pp. 145–157.
[22] L. Mico', J. Oncina, E. Vidal, "A new version of the nearest-neighbour approximating and eliminating search (AESA) with linear pre-processing-time and memory requirements", Pattern Recognition Letters, Vol. 15, No. 1, 1994, pp. 9–17.
[23] L. Mico', J. Oncina, R. C. Carrasco, "A fast branch and bound nearest neighbour classifier in metric spaces", Pattern Recognition Letters, Vol. 17, No. 7, 1996, pp. 731–739.
[24] S. Nene and S. Nayar, "A simple algorithm for nearest neighbour search in high dimensions", IEEE Trans. on Pattern Analysis and Machine Intel., Vol. 19, No. 9, 1997, 989–1003.
[25] E. Chavez, J. Marroquìn, R. Baeza-Yates, "Spaghettis: an array based algorithm for similarity queries in metric spaces", In Proceedings of String Processing and Information Retrieval Symposium, September 21 – 24, 1999, Cancun, Mexico.
[26] E. Chavez, J. L. Marroquìn, G. Navarro, "Fixed queries array: A fast and economical data structure for proximity searching", Multimedia Tools and Applications, Vol. 14, No. 2, 2001, pp. 113– 135.
[27] E. Chávez, G. Navarro, "A compact space decomposition for effective metric indexing", Pattern Recognition Letters, Vol. 26, No. 9, 2005, pp. 1363–1376.
[28] K. Fredriksson, "Engineering efficient metric indexes", Pattern Recognition Letters, Vol. 28, No.1, 2007, pp. 75-84.
[29] H. Hoppe, T. Derose, T. Duchamp, J. McDonald, W. Stuetzle, "Surface reconstruction from unorganized point clouds", In ACM SIGGRAPH, 1992, pp. 71-78.
[30] http://www.graphics.stanford.edu/data/3Dscanrep/
[31] http://shapes.aimatshape.net/
[32] http://www.lodbook.com/models/

Dr Luca Di Angelo obtained his degree in Mechanical Engineering in 1999 at the Faculty of Engineering of L'Aquila and his PhD in Mechanical Engineering in 2002 at the University of 'Tor Vergata' in Rome. Since the 2005, he has been a researcher at Faculty of Engineering of L'Aquila, Italy. His research interests include: computational geometry, geometric modelling of functional geometric shape, shape errors modelling and simulation and features based CAD technology. Dr. Luca Di Angelo is co-author over fifty papers in international journals and international conferences.

Luigi Giaccari received B.S degree and M.S degree in Mechanical Engineering at the Faculty of Engineering of L'Aquila, in 2007 and 2009. Since May 2011, he has been a software developer at ANSYS Germany Gmbh. His research interests include computational geometry and mesh generation. Giaccari is

co-author of two papers in international journals and international conferences.

A Collective Intelligence Based Approach to Business-to-Business E-Marketplaces

Youssef Iguider and Hiroyoshi Morita

**The University of Electro-Communications. Graduate School of Information Systems.
Tokyo 182-8585, Japan**

Abstract

This paper describes a novel approach for addressing the issue of business-matching and recommending in business-to-business e-marketplaces. The presented matching system makes use of Collective Intelligence (CI) means for identifying and recommending business opportunities that are best correlated with the user's need. The objective from this approach is to shift the focus from search-oriented matching, toward assistance-providing business matching systems for the next-generation e-marketplaces. At first, the content of business proposals which are submitted to the e-marketplace, along with their companies' profiles are parsed and indexed. The indexed data may be collected from electronically submitted entries, as well as from scanned and handwritten documents. The search engine starts by expanding the queried keywords, to enable an intuitive-like search. The look-up results are then filtered based on compatibility scoring mechanisms, based on CI techniques. The personalized business-matching results and recommendations are later served to the user via a novel visual interactive graphical interface. An experimental system-prototype applying the proposed and described approach is developed and now being experimentally tested, to fully demonstrate the capabilities of the proposed system on real-world data.

Keywords: *Collective Intelligence, e-marketplace, small business, handwriting recognition.*

1. Introduction

1.1 Collective Intelligence (CI)

This research effort is inspired by the vision of Dr. Douglass Engelbart [1] about the use of technology to improve our collective intelligence for the betterment of humanity, by integrating social-cultural strategies with new technology to create a new way to portray information [1]. Dr. Engelbart is often considered the main founder of the field of Collective Intelligence (CI) [2][3]. CI is defined as the capacity of human collectives to engage in intellectual cooperation, in order to build new conclusions from independent contributors [4]. This study is an attempt to apply CI approach to e-marketplaces for small businesses.

1.2 Why Small Business?

The sector of small business has a big impact on nations' economies. Because it usually accounts for over half of all industrial activities; and it is the major source of employment in most countries. In the United States, for instance, small businesses provide 55% of all jobs, and contribute with 54% of all USA sales [5]. In Japan also, the small business sector is the economy's engine [6]. Small and mid-size enterprises (SME) represent 99.7% of the 4,973 thousands companies registered in Japan. These SMEs employ 70.2% of Japan's workforce, and their contribution share to the shipment in the manufacturing market is estimated at 51.1% [6].

1.3 Challenges Faced by Small Business

Although the sector of small business has such an essential role to play in nations' economies, small businesses are too often severely treated by the market's difficult realities. It is a fact that only about 50% of small businesses remain in business after their first 3 years [7]. Small businesses are exposed to bigger threats than larger companies, because they do not have the back-up of extra finance and resources that larger companies possess. Difficulties to commercialize their products; mismatched trading, and the lack of funding partners are often listed among the top challenges that are commonly faced by small businesses, which often lead to their failure [5][7][8].

On the other hand, On a global level, because of the global digital divide and the troubling gaps created by unequal socio-economic levels, small business, especially in rural areas of developing countries, often have less or no access to regional and international trading opportunities, which are offered by Internet and its means of information flow, such as e-marketplaces, electronic commerce, and online social networks, etc. [9][10]. It is feared that the rapid developments in Information and Communication Technology (ICT) which opens up further new global business opportunities in the form of e-commerce, may widen this digital divide and lead small businesses in the

rural areas of developing countries to lag even further behind and lose in the race [11][12].

1.4 Existing e-Marketplaces

Business consulting and business matching for small and large businesses is mostly provided offline. A service that usually comes with more costly expenses than what small business can usually afford. Moreover, the recommended business matches are often geographically limited to local or regional partners. Few online e-marketplaces provide international business matching services. Most of these platforms simply list the same static information from their databases similarly to all users. Without truly taking into account the specific needs and background of the user [13]. In the present paper, we attempt to introduce and discuss a new business-matching and recommending system, which would enable e-marketplaces to provide each user with personalized results that are customized to the specific needs of her/his business. The recommended business matches are served via a novel visual interactive graphical interface. The following section discusses the evolution of information systems, and anticipates the coming trend of information systems, and therefore the motivation behind this research. The section III introduces and describes the proposed system. The section 4 provides an overview about the initial experimental studies. The section V outlines our conclusions and discusses further strategies for addressing the challenges faced by small businesses on a global level.

2. Targeted Information System

Information systems (IS) are combinations of information technology and people's activities using that technology to support operations, management, and decision-making.

2.1 Hardware-Oriented Information Systems

Early IS solutions were operating based on primitive systems that used machine codes and data, to have the central processing unit of computing devises execute specific instructions. These primitive systems often executed one program at a time, and operated mainly as specific hardware-dependent systems. We represent this era of information systems as the “Hardware Management” stage, in Figure 1.

2.2 Software-Oriented Information Systems

Later, in the 60s through the 80s, after hardware capabilities evolved to allow similar software to run on more than one platform, advanced operating systems were born. Which enabled multi-tasking information systems to operate a large amount of software applications. Which help the users perform common tasks and activities in the real world. We represent this era of information systems as the “Software Management” stage, in Figure 1.

2.3 Knowledge-Oriented Information Systems

The expansion of the World Wide Web and the explosion of Internet interactions, led to the constantly increasing production of a huge amount of online data, which is doubling approximately every six months [14]. Therefore, we believe that there is an emerging need for new data management systems, able to take advantage of these large amounts of data, by uncovering new, implicit and potentially useful knowledge from them. And also creating new knowledge out of their interlinked characteristics. This knowledge operating information systems would play a vital role in the information industry. We represent this era of information systems as the “Knowledge Management” stage in Figure 1.

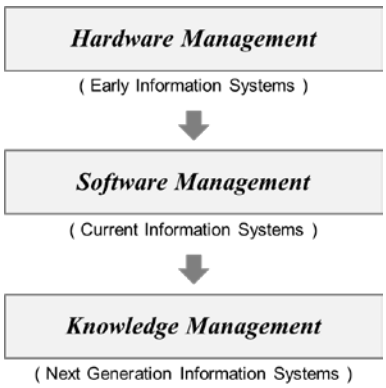


Fig. 1 Evolution of information systems

2.3 CI Based Knowledge-Oriented IS

This work attempts to propose a knowledge managing IS model, based on collective intelligence, which we later apply to develop a new business-matching system for next generation e-marketplaces.
“CI is a multidisciplinary philosophical framework, which integrates social-cultural strategies with new technology to create a way to portray information, with the goal to include, view, and aggregate as much information as possible in order to enable people to act strategically to solve complex problems” [1][15][16].
The basic idea behind our proposed knowledge-managing system is illustrated below in Figure 2. And the process is described below:

Collective-Intelligence Collection: Acquiring data and information from independent users, then provisioning that

data in a way which ensures a later optimal processing.

Intelligence Processing: Converting the collected data into a form suitable for producing intelligence. By conducting various detailed analysis, comparisons and information correlation among the collected data.

Personalized Services: Reducing information overload, by focusing on the consumer's specific need, to interpret the processed information into a finished intelligence product that may help the user draw analytical conclusions.

New Knowledge Creation: Aggregating the collected data and processed information, to systematically and dynamically create new knowledge that may convert lacking information into expanded intelligence.

Customized Expertise Servicing: Conveying expanded intelligence in a usable form, to support user's decision-making with personalized and relevant insights.

Smart & Intuitive User Interface: Creating new ways of structuring facts, and new ways of interacting with the system, is key to extending people's capability to create, manipulate and share knowledge [1].

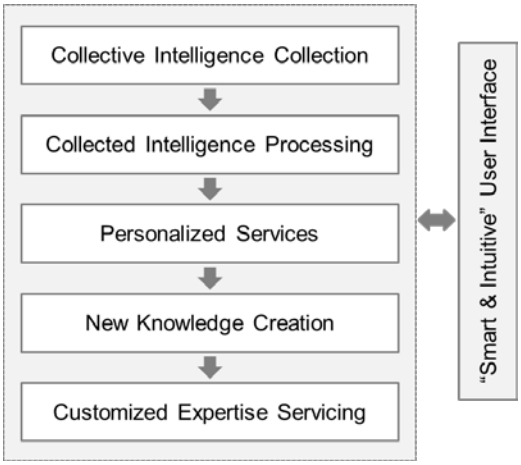


Fig. 2 CI based knowledge management

Actually, we recently discovered that our proposed approach to knowledge-managing systems has an analogical similarity with the five-step approach, called Intelligence Cycle (Figure 3), which is used by the Central Intelligence Agency (CIA) in the United States and many other Intelligence communities. Their approach is apparently well proven for producing and reporting highly accurate Intelligence services [17].

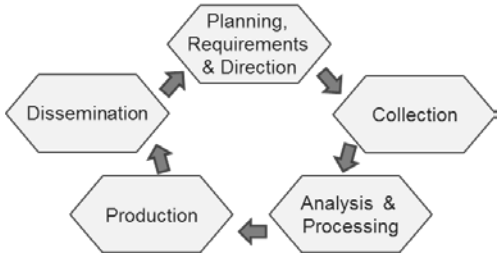


Fig. 3 CIA's Intelligence Cycle.

3. A CI Based System For E-Marketplaces

This section attempts to apply the approach discussed in the previous section, toward designing a CI based new business-matching and recommending system model for e-marketplaces.

Unlike most existing e-marketplaces, where the system usually simply matches companies and lists to the user straightforward search results from available static databases [13], the proposed system model (Figure 4) uses a CI approach toward the process of business-matching and business opportunity recommending. As shown in Figure 4, at first, various data are collected from the collectively submitted users' business (selling or buying) proposals, as well as from their company's profile and background. With the goal to allow traders to have access to the untapped business opportunities which are not available electronically, but on paper. And also with the goal to help bridging the digital divide with traders in many developing countries, especially in rural areas. The system may accept also the information that is captured from handwritten document and transformed to digital data. Moreover the recent advances in on-line data capturing technologies and its widespread deployment in devices like PDAs and notebook PCs is creating large amounts of handwritten data that need to be archived and retrieved efficiently, especially that recognition algorithms and engines are already available for all major language scripts [18][19]. The collected data are then processed and carefully indexed. Then, based on the user's query, and also based on his/her recorded business background and company's profile, the system conducts various analysis and correlation operations, on the user's data vs. the data of the potential candidate partners, and their business proposals and needs. With the goal to reduce the search result overload, and instead convey to the user personalized business recommendations specific to his/her needs, interest and background. The goal is also to enable the user uncover new relevant business opportunities that might not be easily reachable though straightforward search of a static database.

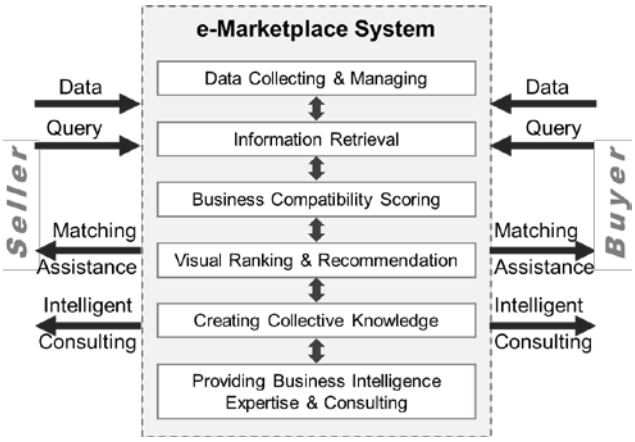


Fig. 4 CI based system for e-marketplaces

Table 1 illustrates an example, where a part of the user's information is scored and matched against the data of other identified business-partner candidates. With the goal to evaluate the user's compatibility with the identified business-partner candidates, and qualify the potential of their proposed business opportunities, the user's attributes are mapped against the attributes of the identified partner-candidates, via several correlation means, including the matching via Euclidean Distance Scoring (1).

Table 1: Compatibility scoring

Attribute #	Attributes Description	User	Candidate Partners				
			P-1	P-2	P-3	…	P-n
A_1	Company Country	$S_0(A_1)$	$S_1(A_1)$	$S_2(A_1)$	$S_3(A_1)$	…	$S_n(A_1)$
A_2	Business Category	$S_0(A_2)$	$S_1(A_2)$	$S_2(A_2)$	$S_3(A_2)$	…	$S_n(A_2)$
A_3	Company Age	$S_0(A_3)$	$S_1(A_3)$	$S_2(A_3)$	$S_3(A_3)$	…	$S_n(A_3)$
A_4	Capital	$S_0(A_4)$	$S_1(A_4)$	$S_2(A_4)$	$S_3(A_4)$	…	$S_n(A_4)$
A_5	Business Volume	$S_0(A_5)$	$S_1(A_5)$	$S_2(A_5)$	$S_3(A_5)$	…	$S_n(A_5)$
A_6	Employees	$S_0(A_6)$	$S_1(A_6)$	$S_2(A_6)$	$S_3(A_6)$	…	$S_n(A_6)$
A_7	Offer type (Proposal)	$S_0(A_7)$	$S_1(A_7)$	$S_2(A_7)$	$S_3(A_7)$	…	$S_n(A_7)$
A_8	Product/Service category	$S_0(A_8)$	$S_1(A_8)$	$S_2(A_8)$	$S_3(A_8)$	…	$S_n(A_8)$
A_9	Minimum order ($)	$S_0(A_9)$	$S_1(A_9)$	$S_2(A_9)$	$S_3(A_9)$	…	$S_n(A_9)$
A_{10}	Targeted region/s	$S_0(A_{10})$	$S_1(A_{10})$	$S_2(A_{10})$	$S_3(A_{10})$	…	$S_n(A_{10})$
A_{11}	Non acceptable countries	$S_0(A_{11})$	$S_1(A_{11})$	$S_2(A_{11})$	$S_0(A_{11})$	…	$S_n(A_{11})$

$$C_{(i,j)} = \sqrt{\sum_{k=1}^{11} \left(S_i(A_k) - S_j(A_k)\right)^2} \qquad (1)$$

Where

$C_{(i,j)}$ - Compatibility between two partner-candidates i and j

$S_i(A_k)$ - Score of a partner-candidate i with regard to the attribute A_k

To comply with the recommendation of Dr. Engelbart (CI's founder) about the importance of creating new ways and symbols for structuring facts, to extend the user's capability of manipulating the created knowledge [1], the matching results and personalized recommendations are later conveyed to the user via a new visual graphic interface, as illustrated by Figure 5 and Table 2.

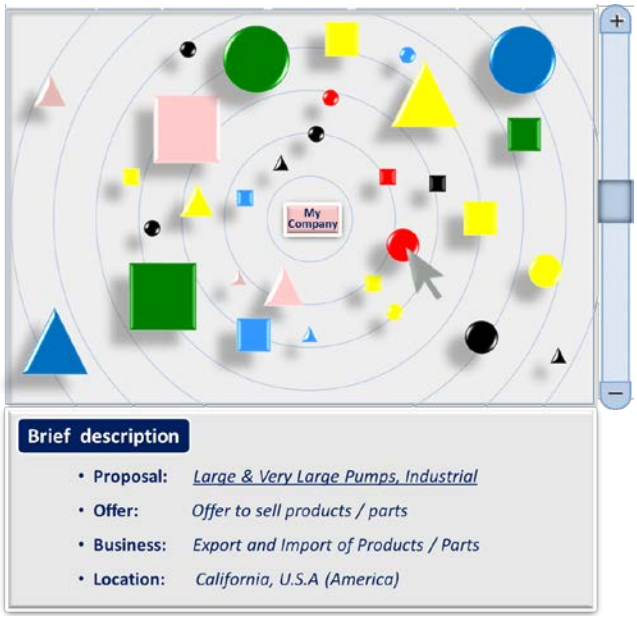


Fig. 5 Visual representations of the business-matching results

Colors, shapes and sizes are used to symbolize characteristics of the candidate partners, to help the user visually explore and interact with the recommended business opportunities via the graphical navigational interface.

Table 2: Graphic symbols interpretation

Shape		"Offer Type" Attribute
Circle	●	SELLING
Triangle	▲	BUYING
Square	■	OTHERS
Color		**"Country Area" Attribute**
Blue	●	EUROPE
Yellow	●	ASIA
Black	●	AFRICA
Green	●	OCEANIA
Red	●	AMERICA
Pink	●	JAPAN
Size		**"Company Size" Attribute**
Small	●	SMALL
Medium	●	MEDIUM
Large	●	LARGE

Automated Business Consulting System: At a later stage, our research aims at creating a CI business-matching system that would enable e-marketplaces to provide automated consulting. The system aggregates the collected data to create new relevant knowledge. By systematically applying best practices used by business experts. A personalized expertize would be generated to support users' decision-making and enhance their market insights and business intelligence. Figure 6 illustrates an example where the targeted automated system would systematically uncover strategic partnering opportunities for the user.

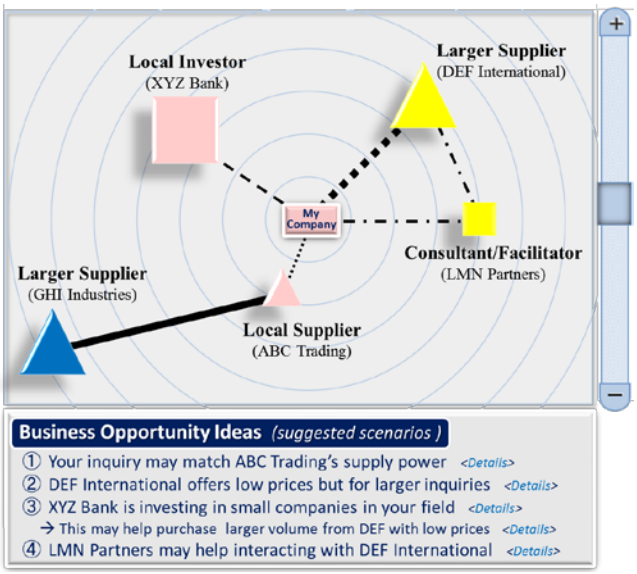


Fig. 6 CI Based Automated Consulting

4. Experimental Studies

To experimentally study the proposed CI based business-matching and recommending system, we are currently developing a prototype that can process advanced business matching and generate recommendations, based on real-word business opportunities.

4.1 JETRO's online business matching database

To study our prototype using real-word data, experimental simulations were conducted based on data collected from JETRO (Japan External Trade Organization)'s online business matching database. JETRO is a Japanese governmental organization, which promotes mutual trade between Japan and the rest of the world. JETRO is running a free online business matching service "Trade Tie-up Promotion Program" <www.jetro.go.jp/ttppoas/> [20]. Which allows business people (especially small and medium companies) to browse through over 20,000 business proposals in various fields.

4.2 Experimental prototype

The components of the experimental prototype are shown in Figure 7. At first an experimental simulation was developed using Python programing language [21], for parsing JETRO's proposals. BeautifulSoup classes were used to screen-scrap and parse the content of JETRO's entries. The result of this parsing led to compiling information from each entry (proposal) in JETRO database.

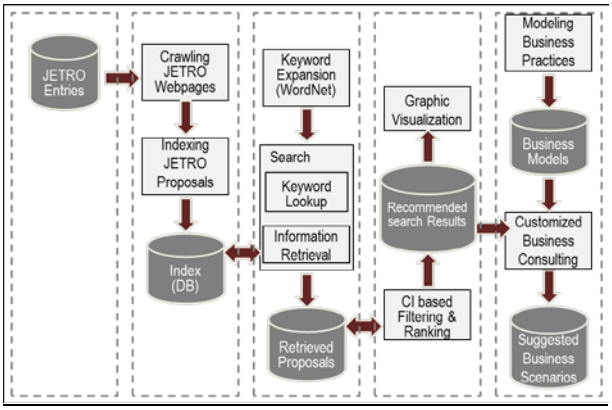


Fig. 7 Experimental business-matching and recommending prototype

4.3 Indexing TTPP business proposals

To process TTPP business proposals in depth, a set of entries (Figure 8) was indexed into a SQLite database. To proceed, the contents were parsed by randomly looping through TTPP URLs to crawl their entry contents. The experimental crawler was designed to ignore a set of words which carry no important meaning. The system was also designed to identify and ignore URL of the proposals that were either expired or deleted by the users. As results, the crawled and parsed TTPP business proposals were automatically indexed into a SQLite database, which was automatically created and saved for further use.

#	Proposal Description	Format
1	Proposal title	Text
2	Proposal URL	URL
3	Proposal No	Number
4	Business Type	Text
5	Offer Type	Text
6	Proposal Category	Text
7	Country / Area	Text

#	Node Attributes	Format
8	Ranking Distance	Number
9	"Offer Type" attribute	Text
10	"Company Size" attribute	Text
11	"Country Area" attribute	Text

Fig. 8 An experimental set of business proposal entries

Figure 9 show samples of indexed data, where the database was populated with, (Figure 9.a) a list of the automatically generated and processed URLs, (Figure 9.b) parsed and processed words, and (Figure 9.c) information about how each word is associated to its corresponding proposal, along with the position of that word within the proposal page.

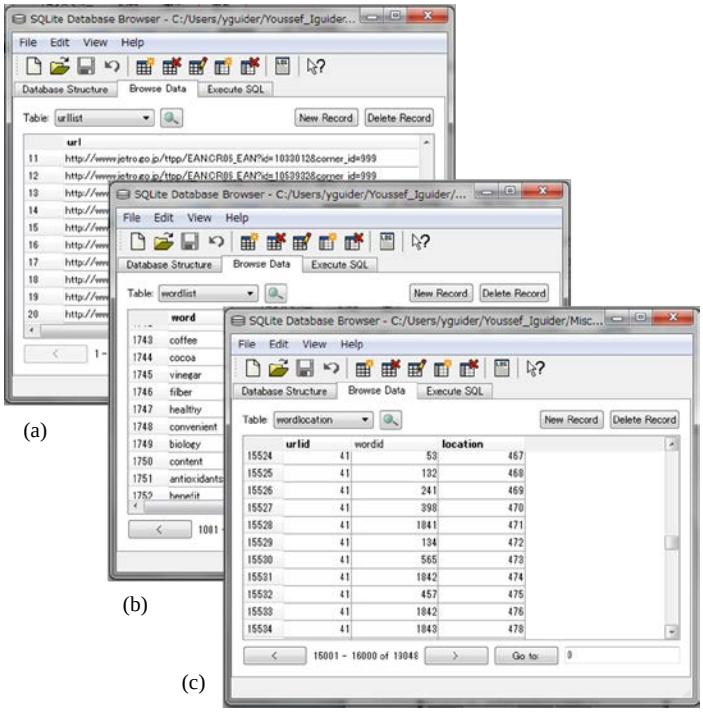


Fig. 9 Sample data indexed into SQLite database: (a) Indexed URLs; (b) Indexed words; (c) Linking words to proposals.

4.4 Keyword expansion based on WordNet

To enable an intuitive-like search, the system uses WordNet to allow identifying additional words that may be relevant to the user's query. This may help uncovering hidden business opportunities that could be of interest to the user. Figures 10,11 illustrate an example where the system is trying to expand a submitted keyword - "carpet".

The expansion of the word "carpet" (Figures 10,11) led to the following:

Similar words: carpet, rug, carpeting
Related Words: Brussels_carpet, Kurdistan, Wilton, Wilton_carpet, broadloom, drugget, flying_carpet, hearthrug, nammad, numdah, numdah_rug, prayer_mat, prayer_rug, red_carpet, runner, scatter_rug, shag_rug, stair-carpet, throw_rug
Related Parts: Edging

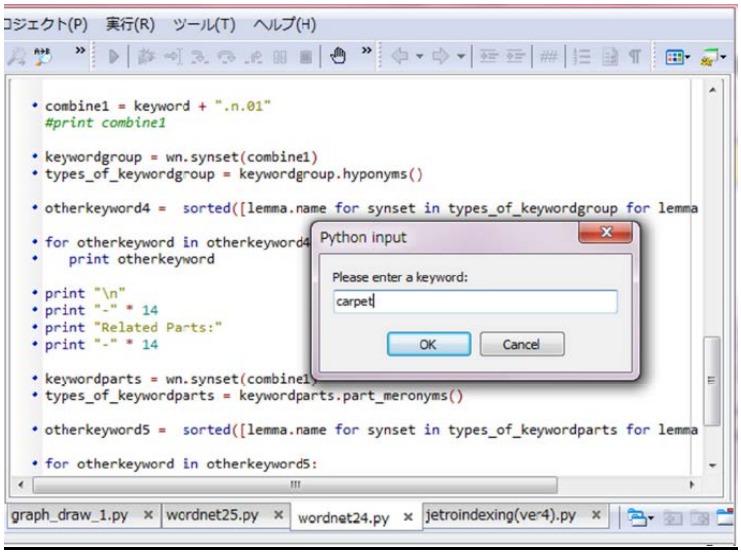


Fig. 10 Submitting a query keyword

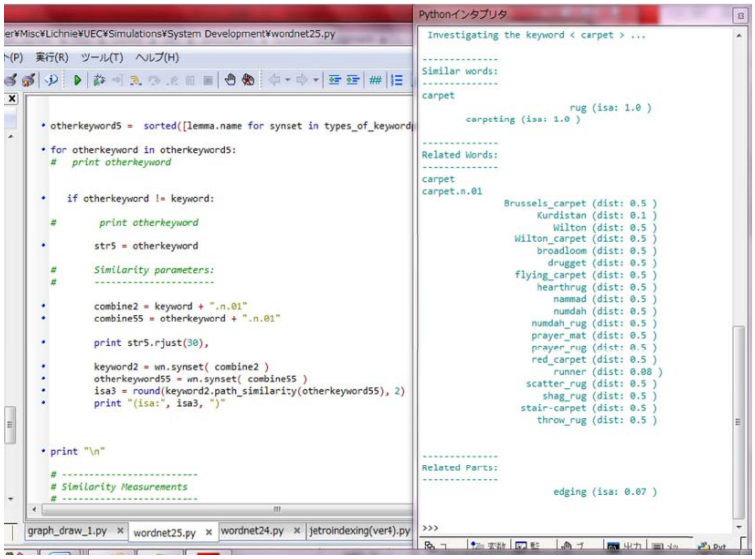
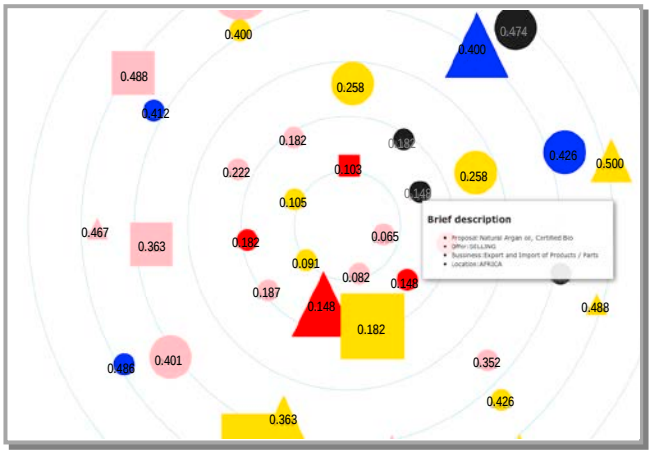
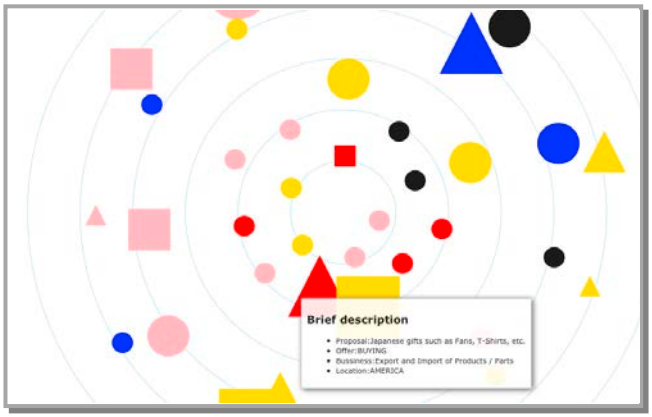


Fig. 11 Expanding the keyword "carpet"

Further, the prototype is simulated to look up the submitted keyword along with the expanded keywords in the indexed SQLite database. The data and content of the identified business proposals are then mapped to the attribute related to the user, by using the CI approach discussed above.
The initial experimental results look quiet promising. Figure 12 presents an experimental visual representation of the business-matching results obtained from processing the experimental set of business proposal entries mentioned above. The results are automatically visualized via an experimental interactive graphical user interface, which was developed using JavaScript InfoVis Toolkit.



(a)



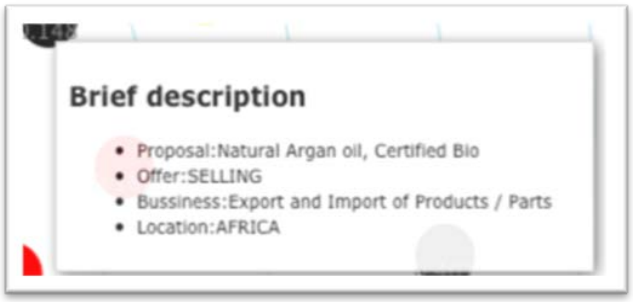
(b)

Fig. 12 Experimental visual representation of business-matching results

The graph is automatically generated based on attributes related to the recommended business proposals. Each proposal is displayed in the form of an interactive node which has a specific shape, color and size. The shape, color and size of nodes are automatically assigned according to the description in Table 2. The nodes are automatically placed away from the center, based on their Distance Scoring according to (1). This enables an intuitive visual navigation of the search results. The nodes are designed to have interactive capabilities with the goal to allow visually exploring and interacting with the recommended business opportunities. For instance, by mouseovering a node, a brief description of its business proposal is displayed, as shown in Figure 13.

Clicking a node, leads to automatically placing in the center of the graph. Meanwhile information about the clicked node is transmitted (as feedback data) back to the system for further processing and compatibility scoring. Several other interactive features are currently being developed to be added to the capabilities of this interactive graphical user interface.

With the goal to have the interactive graphical user interface allow accessing the details of business proposals at any step, our experimental prototype is designed to provide a direct access back to the original business proposals as published on JETRO's TTPP website.



(a) Mouseovering a "selling" node.

(b) Mouseovering a "buying" node.

Fig. 13 Mouseovering an interactive node.

We currently are finalizing the development of the prototype and its visualization system, with the goal to allow us convey the business-matching results and recommendations, via a visual graphic interactive interface, similar to what was described in Figure 5 and Table 2.

5. Conclusion and Discussion

We expect the presented and discussed business-matching and recommending system for e-marketplaces to take the business matching process to a new level. The system makes use of a Collective Intelligence (CI) approach to identify and recommend the best matching and correlated business opportunities for the user. The results are served via a novel visual interactive graphical interface. A system prototype applying the proposed and described approach is being developed and experimentally tested, to fully demonstrate the capabilities of the proposed system on real-world data. Although the prototype is at an early stage, the initial experiments show promising results. We believe therefore that the proposed and simulated approach can addresses many of the marketing challenges, discussed above, that are typically faced by most small businesses.

Handwritten business documents for e-marketplaces: The recent advances in on-line data capturing technologies and its widespread deployment in devices like PDAs and notebook PCs, is creating large amounts of handwritten data that need to be archived and retrieved efficiently. It is important that next generation marketplaces could make use of OCR technologies and handwriting recognition solutions for converting handwritten business documents and other scanned documents (catalogues etc.) into indexable and retrievable data. The increase of such relevant information would enable better correlation, and therefore improve the matching and recommending of business partnerships. Moreover, since handwritten signature remains the most widely accepted biometric means for identity verification in business transaction agreements [22], it makes sense to apply the proven technologies and algorithms which enable author verification and identification via handwriting analysis. To this purpose we are exploring the feasibility to apply our previous works on handwriting recognition [19], and analysis of human handwriting [23][24], with the goal to enable our proposed business-matching and recommending system to make use of the key business information whether it is available electronically or from handwritten documents.

The treats of Digital Divide on small business: Beside the marketing issues discussed earlier, the global digital divide is severely treating many small businesses, especially in the rural areas of developing countries. The Okinawa Charter on Global Information Society (GIS) by the G8, and also the Japan Ministry of International Trade and Industry (MITI) in their Proposal warned that the rapid developments in Information and Communication Technology (ICT) have opened up new global business opportunities in the form of e-commerce, however it is feared that these developments may widen the digital divide, and underdeveloped countries may lag further behind [11][12]. Therefore, to help solving this complicated issue, from our angle, by bridging the digital divide in the e-Marketplaces, we consider allowing the business information available on handwritten documents to be converted into electronic data. Especially that OCR technologies as well as handwriting recognition algorithms and engines are now available for the major language scripts [18][19]. This would not only assist the traders in less developed areas by narrowing down the treat of the digital divide on their small business. But it would also assist the users of e-Marketplaces in industrialized countries to open up and access new markets full of still untapped business opportunities.

References

[1] D. Engelbart , “Boosting Collective IQ,” The Doug Engelbart Institute (www.dougengelbart.org/about/collective-iq.html)

[2] P. Lévy, “From Social Computing to Reflexive Collective intelligence,” IEML Research Program, CRC, FRSC, University of Ottawa. Aug.09.

[3] Stanford University’s archive “Doug Engelbart 1968 Demo,” (http://sloan.stanford.edu/mousesite/1968Demo.html)

[4] T. Segaran, “Programming Collective Intelligence,” O’REILLY ISBN:978-0-596-52932-1, 2007.

[5] The United States Department of The Treasury. Various reports about the Economic Recovery (www.treas.gov)

[6] Japan’s Ministry of Economy, Trade and Industry, “Census of Manufactures,” 2008. (www.meti.go.jp/english/statistics/)

[7] E.J. O’Donoghue, N. Key, and M.J. Roberts, “Does risk matter for farm businesses? The effect of crop insurance on production and diversification,” Report of Economic Research Service, USDA.

[8] B.S. Cromie, “Entrepreneurial Networks,” International Small Business Journal 9, 57-74, 1991.

[9] T.P. Rama Rao, “E-Commerce and Digital Divide: Impact on Consumers, ” Presented at the Regional Meeting for the Asia-Pacific: New Dimensions of Consumer Protection in the Era of Globalisation, Goa, India, September 10-11, 2001.

[10] “Digital Opportunities and the Missing Link for Developing Countries,” Oxford University Press 2001: Oxford Review of Economic Policy, Volume17, Issue2, pp. 282-295, June 1, 2001 .

[11] “Okinawa Charter on Global Information Society”, http://www.g8.utoronto.ca/summit/2000okinawa/gis.htm

[12] MITI’s Proposal for WTO E-Commerce Initiative, “Towards eQuality: Global E-Commerce Presents Digital Opportunity to Close the Divide Between Developed and Developing Countries (2nd Draft)”, http://wwwmeti.go.jp

[13] Y. Iguider and H. Morita, “A Collective Intelligence Based Business-Matching and Recommending System for Next Generation E-Marketplaces,” Proceedings of 2011 IEEE Symp. on Computers & Informatics, pp.489-494, Mar. 2011.

[14] Federal Communications Commission (FCC 98-225), “MEMORANDUM OPINION AND ORDER,” Sep. 14, 1998.

[15] MIT Center for Collective Intelligence (http://cci.mit.edu/)

[16] Wikipedia - The Free Encyclopedia, on “Collective Intelligence” (http://en.wikipedia.org/wiki/Collective_Intelligence)

[17] “INTELLIGENCE COLLECTION ACTIVITIES AND DISCIPLINES,” Operations Security INTELLIGENCE THREAT HANDBOOK, OPSEC, April 1996

[18] S. Jaeger, C.-L. Liu and M. Nakagawa, “The state of the art in Japanese online handwriting recognition compared to techniques in western handwriting recognition,” Internat-l Journal on Document Analysis and Recognition - IJDAR , vol.6, no.2, pp.75-88, 2003.

[19] Y. Iguider and M. Yasuhara, “An Active Recognition of Handwritten Isolated Arabic Characters,” Transaction of the Society of Instrument and Control Engineers. Vol.32, No.8, pp1267/1276, 1996.

[20] Trade Tie-up Promotion Program. The Japan External Trade Organization (JETRO) (http://www.jetro.go.jp/ttppe/)

[21] M. Lutz. “Programming Python,” 3rd Edit. O'Reilly Media, Aug.06.

[22] M.M. Fahmy, “Online handwritten signature verification system based on DWT features extraction and neural network classification,” Ain Shams Engineering Journal, Vol.1, Issue 1, Sep. 2010, pp.59-70.

[23] Y. Iguider and M. Yasuhara, “Extracting Control Pulses of Handwriting movement,” Transaction of the Society of Instrument and Control Engineers. Vol.31, No.8, pp1175/1184, 1995.

[24] Y. Iguider and M. Yasuhara, “A Control-Theoretic Description of Handwriting Process,” Preprint of the 24th Symposium of the Institute of Systems Control and Information Engineers, pp45-48, Nov. 1992.

A PKI-based Track and Trace Network for Cross-boundary Container Security

S.L. Ting[1], W.H. Ip[1], W.H.K. Lam[2] and E.W.T. Ngai[3]

**[1] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University
Hung Hum, Hong Kong, China**

**[2] Department of Civil and Structural Engineering, The Hong Kong Polytechnic University
Hung Hum, Hong Kong, China**

**[3] Department of Management and Marketing, The Hong Kong Polytechnic University
Hung Hum, Hong Kong, China**

Abstract

Challenges in container management, such as theft of products, cross-border smuggling, long checking time in processes of customs clearance and unable to track locations of products in real time, are currently faced by supply chain parties. In recent year, Electronic Seal (eSeal) technology is introduced to protect containers in a securely way. However, concerns related to data security increase when eSeal is applied in the track and trace infrastructure. To overcome this issue, this paper presents an idea to integrate the Public Key Infrastructure (PKI) in an eSeal device to enhance the cross-boundary container security.

Keywords: *Cross-boundary Container, Cryptography, Electronic Seal (eSeal), Public Key Infrastructure (PKI), Radio Frequency Identification (RFID), Track and Trace.*

1. Introduction

Logistics enterprises face tough competitions in an increasingly globalized market with new high and increased security requirements that response the efficiency of container movements in recent years. With millions of containers shipping around the world, containers are manually sealed so as to prevent cross-border smuggling, theft of products, and terrorist threats. With current practices, manual metal bolts, weld lock truck seals, cable seals and barcode seals are commonly used to prevent products to be stolen or broken when transporting from the point of origin to the destination point [1]. However, Zhang and Zhang [2] argued that these existing capabilities featuring in container security are vulnerable to guarantee the shipment integrity since there are numerous ways to defeat the manual seal, such as cutting holes in the side or top of a container and then repairing it. Thus, there is a pressing need to increase the security requirements of cross-boundary containers.

Attempted to address this issue, Radio Frequency Identification (RFID) technology is proposed to promote the container security [3, 4, 5]. RFID is an emerging technology that uses wireless radio to identify objects from a distance without requiring line of sight or physical contact. It can track the movements of objects through a network of radio enabled scanning devices over a distance of several meters [6, 7]. As stated by the academic study of Ngai et al. [8], RFID have been widely adopted in various industries like manufacturing, healthcare, logistics, and transportation. When it applies to the container security, an RFID-enabled seal, called an Electronic Seal (eSeal), is developed to allow importers, shipping companies, port officials and customs inspectors to determine, without a physical inspection, whether the seal has been tampered with and the security of the container compromised [1]. As stated by Kwok et al. [9], eSeal is a new way for improving issues of container management in global. For example, eSeal with RFID technology was rolled out at 20 lanes at Kaohsiung Harbor in Taiwan in 2009. Passive ultra-high frequency RFID tags embedded in eSeals, complied with ISO 17712, were used for identification of trucks in a port. Accuracy for read rate is higher than 95% with a truck travelling at a speed of 60 kilometers per hour and read range is more than 7 meters. The result showed that use of RFID-based seal in a port is able to reduce 6000 man-hours for escorts annually [10]. This implies RFID technology embedded in the eSeal is an effective approach to protect containers in a secure way.

RFID and eSeal technology has the potential to increase the container security; but this raises concerns regarding data security issues when implementing in a track and trace infrastructure (i.e. the container en route from the point of origin to the destination point) [11]. This is because all the product movement can be recorded by RFID technology among the entire supply chain in which such information (e.g. design of the distribution channel) may reveal private information regarding a particular business, like the business secrets about future strategic process

restructurings [12]. However, the academic literature addressing the security of the stored information in eSeal is scare. In particular, cryptographic measures in securing RFID data exchange in a track and trace infrastructure are almost absent. Yet, most studies were based on the electronic document which satisfies a pedigree requirement (i.e. ePedgree) [13, 14, 15]. Such approach can provide full supply chain visibility with detailed track and trace information in electronic format; but it cannot guarantee whether the container is moving on the right pedigree or not.

As a result of increased security requirements on RFID data protection in the track and trace infrastructure, cryptographic algorithms can be an important bridge between the efficiency of and security of data exchange process. Public Key Infrastructure (PKI) technology has such potential, in which it enables users to securely and privately exchange data through the use of a public and a private cryptographic key pair [16]. As discussed by Liping and Lei [17], PKI is a new security technology to establish trust relationship between parties in a secure network communications and online transactions. In addition to the applications domains like e-commerce and financial transaction environment, PKI is now applied in various areas. For example, Wilson [18] presented the use of public key certificates to secret the Unique Health Identifiers within an anonymous digital certificate. With the expected growth of m-commerce, Dankers et al. [19] tailored the PKI to enable end-to-end security for mobile systems. Considering the ensure the confidentiality and secrecy of the patient information, Brandner et al. [20] determined the user-oriented and legal requirements for a PKI for electronic signatures for medical documents.

With these successful applications mentioned in the literature, this paper presents an attempt to apply the asymmetric cryptography method (i.e. PKI technology) to strengthen the security in the data exchange of eSeal within the entire supply chain network (i.e. from their point of origin, while en route, through customs, and to their final destination). Apart from elaborating the use of the PKI in eSeal, this paper also introduces a PKI-based Track and Trace Network (PKI-TTN) to enhance the effectiveness of the cross-boundary container security.

The rest of this paper is organized as follows: Section 2 reviews the current track and trace infrastructure, eSeal technology and its security considerations in the data protection. The cryptographic method of the PKI-based Track and Trace Network (PKI-TTN) is presented in Section 3. Section 4 evaluates the performance of PKI-TTN and the conclusion is drawn in Section 5.

2. Track and Trace Infrastructure and eSeal Technology: Current Situation and Security Challenges

Track and trace infrastructure reflects the concept in tracking the current location of objects in the supply chain and tracing their past locations. In today's business world, the track and trace infrastructure of cross-boundary container has a tendency to get complex as the inspection tasks are done by multiple parties (e.g. shippers, forwarders, and consignees) [12]. Generally, there are three tasks to perform in each container terminals, namely container identification, seal checking, and damage checking. As reported by the study of Tsilingiris et al. [21], all these checking activities are inevitably repeated as the involved parties do not share the information to each other. As a result, checking time in processes of customs clearance is increased significantly.

Furthermore, another major problem encountered in the existing track and trace infrastructure is the theft of products and cross-border smuggling. Although manual seals are used to govern the control and management of containers, the influx of smuggling is constantly revoking [1]. As discussed by Bastia [22], the insufficient information transmission among supply chain parties makes them difficult in assuring the product lifecycle safety, especially the product authentication problems.

In order to fulfill the need for electronic infrastructure and information sharing platform, EPCglobal network was established for supply chain information sharing, especially for the flow of products in the supply chain [23]. Each product tagged with RFID label is assigned a universal unique ID called an Electronic Product Code (EPC) for the unique identification and tracking by different parties in the worldwide supply chain. It provides and promotes a standardized product identification mechanism and technology. To further satisfy the need for cost-effective and reliable hardware for cargo and container authentication and management, eSeal devices have been introduced during these last few years. Nowadays, there are a number of successful cases [2, 4, 9, 10], in many parts of the world showing that using RFID-based eSeal and information sharing platform can improve the efficiency and reduce cost of daily land logistic operations. For example, Kazakhstan and Lithuania customs authorities have been using RFID-based electronic seals on trucks passing through their countries since 2006. The system ensures that there is no unauthorized opening of cargo; and the location and status of the trucks transporting goods within their borders can be traced. It successfully helps to prevent smuggling and theft of cargo [24]. An increasing number of studies have been

conducted to evaluate the potential and benefits of the use of eSeal in container management.

So far, one problem arises when implementing eSeal technology in track and trace infrastructure is the data protection issues [12, 25]. Since much information related to the companies is contained in the RFID tag, therefore a secure solution in the data exchange process is essential to develop within the infrastructure. In particular, two security elements, namely confidentiality and integrity, are taken into consideration in the data protection issue. Confidentiality is concerned with the authentication for the access to data (i.e. only authorized persons or organizations can share the information); whereas the integrity is concerned with the completeness of the shared information (i.e. whether the information is being altered by others). Compared with the confidentiality, integrity is much more difficult to achieve as it requires to assess the trustworthiness of the source of the information [16]. Consequently, an increasing number of cryptographic methods have been developed for accomplishing these two security elements. Among these approaches, the PKI has been widely used because of its strong encrypted tunnel during the data exchange process as well as its well defined standard [26]. The distinguishing feature of a PKI is the use of a digital certificate issued by a trusted party, named Certification Authority (CA). The issued certificate is digitally signed using a CA's secret key to the messages specifying a user and the corresponding public key (Figure 1). Once the certification service is discontinued by a CA's key compromise, it will result in considerable degradation of the overall level of reliability and finally the fatal impact to the overall PKI [27].

Fig. 1 Digital certification

Regarding piracy and security are two issues to hurdle the RFID implementation [28], an increasing number of research activities [29, 30, 31, 32, 33] are focusing on employing PKI in the RFID systems, like the studies. Vaudenay [34] even pointed out that the use of public-key cryptography outperforms other security protocols. However, all these above mentioned research works are concerned with the identification protocols between the tag and the reader; in which the network (i.e. product flow) in the track and trace infrastructure is scare. Therefore, the present study focuses on the discussion of how PKI approach can be implemented as the authentication mechanism in the track and trace infrastructure.

3. PKI-based Track and Trace Network (PKI-TTN)

In order to fulfill the security requirements of eSeal and customs clearance applications, a PKI-based Track and Trace Network (PKI-TTN) including an eSeal device [9], eSeal middleware and public key infrastructure is developed. This network is developed base upon eSeal device that is designed and integrated with different logistics enabling technologies of RFID, Bluetooth, Wireless Network, Sensor and Global System for Mobile Communications (GSM). It is developed using standardization and modularization approaches, so that the eSeal device can be assembled by putting together different function modules to perform different features as well as adapt to different application requirements.

3.1 eSeal Device

An eSeal device is an integrated automatic identification device. It is developed in a modular-based structure (Figure 2). There are totally four modules, namely: Identification and Control Module, Communication Module, Sensor Module and Seal Module.

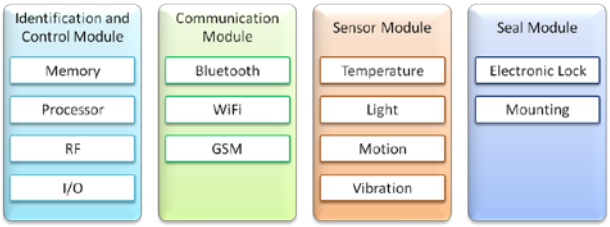


Fig. 2 Modular-based structure of an eSeal device [9]

Identification and Control Module: In Identification and Control Module, there is a processor and a memory for automatic identification purpose. This module also contains RF communication ability itself and serves basic identification purposes. It includes several input and output channels for providing additional features such as sound, light and Liquid Crystal Display (LCD). For additional functions, specific modules are designed (e.g. Communication, Sensor and Seal Modules). These can be plugged into the Identification and Control Module to extend the functionality.

Communication Module: The Communication Module is designed to extend the communication ability of the eSeal

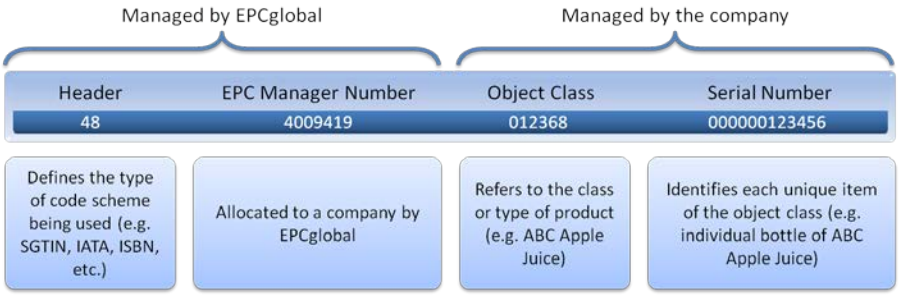


Fig. 3 EPC structure

device. By providing with different communication protocols, such as Bluetooth, GSM, and WiFi, eSeal device can communicate with different systems and devices.

Sensor Module: Senor Module is the plug and play modules of temperature sensors, light sensors, vibration sensors and motion sensors. It enables the eSeal device to monitor the environmental factors and record unexpected changes in the environment.

Seal Module: The Seal Module is a mechanical part and casting consisting of an Electronic Lock and Mounting. The Electronic Lock is designed as a changeable ID electronic lock. When the lock is opened, the ID of the lock will be changed. By tracking the ID change events in the log, the user can see whether the lock was opened authentically. In addition, the confidentiality and integrity of the information of this module is encrypted in a public key cryptography (see Section 3.3).

Standard Adoption of eSeal Device: In order to record the RFID information in a standardized and unique structure, RFID tag that features in EPCglobal specification is adopted in this study. It offers extended data capacity and employs a numbering scheme called EPC, which serves as a global standard to provide a unique ID for any item in the world [35]. With the license-plate type of identifier in EPC capability, real time tracking information on a product within its supply chain can be achieved [36]. Figure 3 depicts the structure of an EPC tag. Other than the four specific elements highlighted in the EPC structure, there is a reserve memory for advancing the security in data protection [37]. Thus, a PKI-based cryptographic approach is used to encrypt the tag information and the encrypted information is stored in such reserve space to protect the tag from malicious access.

3.2 eSeal Middleware

eSeal Middleware is developed to control and manage the functions and data exchange of eSeal device. It is a type of software which acts as a bridge between hardware devices and software systems, by providing connection, communication, configuration and management functions. With the aid of this, the data connected by the sensor module of eSeal device will be filtered, manipulated and transferred to the back end of the system. In general, eSeal Middleware consists of three components, namely adapters, filters, and loggers.

Adapters: Adapters are the components to connect various eSeal devices. In other words, they can control not only eSeal device mentioned in Section 3.1 but also the market available RFID devices.

Filters: Filters are the components used to filter and provide preliminary manipulation of the collected data from the hardware.

Loggers: Loggers are the output interface channels to connect to different backend systems. All the processed data will be sent to the desired destination via loggers. In order to support fast integration with applications already existing in the enterprises and organizations, the loggers provide standard data interchange protocols like Extensible Markup Language (XML), Hypertext Transfer Protocol (HTTP) and web service. Furthermore, the eSeal Middleware provides loggers to connect to public information sharing platforms (i.e. EPCglobal Network) to share the real-time supply chain information with logistics partners and related parties.

3.3 Public Key Infrastructure (PKI)

To enhance the data protection mechanism in the current data exchange process, PKI technology is employed to protect the real-time information captured in the RFID-enabled track and track infrastructure. Such a cryptography approach provides the advantage of delivering accurate supply chain information with immediate notification when there is any abnormality in the transportation process. Figure 4 shows the authentication mechanism for the container delivery from the point of origin to the point of destination. Before the shipment of the container, specific shipping information (e.g. shipper's information, receiver's

information, and delivery path) is identified and recorded in an XML format to form an electronic shipping document (Figure 5). In order to enhance data confidentiality (i.e. the information stored in the tag can only be read by the authorized party or user), a randomized one-off blind pubic key [38] is generated.

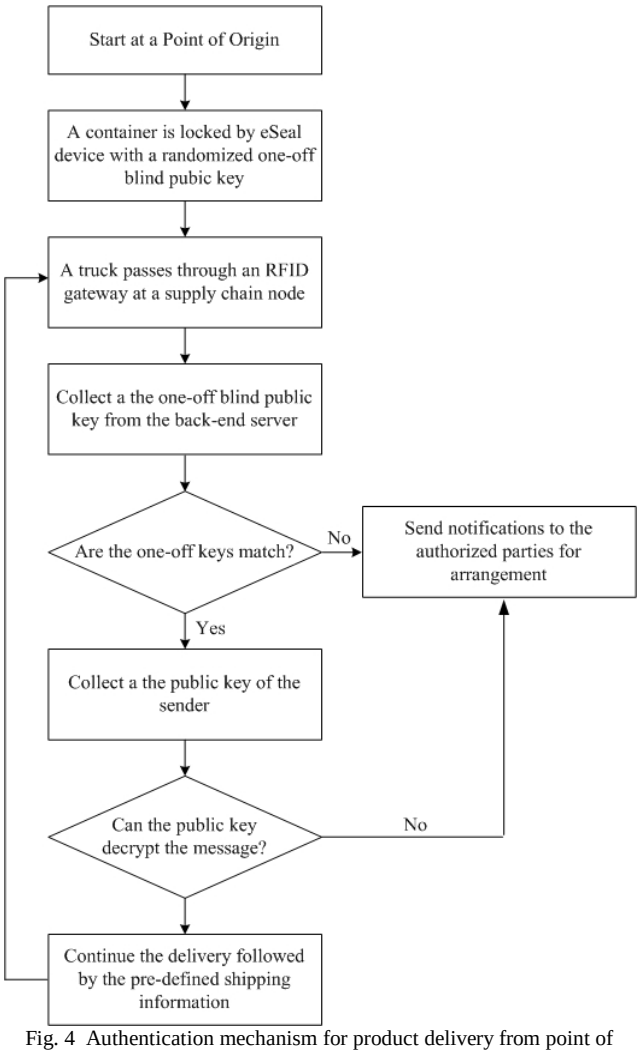


Fig. 4 Authentication mechanism for product delivery from point of origin to point of destination

Encryption Phase: Upon the completion of the setup of the electronic shipping document and the one-off key, an encryption based on the asymmetric key cryptography is employed to encode these two elements; and all these encrypted information are recorded into the reserve memory of RFID tag. Assume that there are four nodes (supply chain parties) in the track and trace infrastructure as described in Figure 5, the encryption procedures first start by encrypting the electronic shipping document by the private key of the manufacturer (i.e. $PrKey_{origin}$), the public key of the node C (i.e. $PuKey_C$) and the one-off key (i.e. $Key_{one\text{-}off}$) to form the first digital envelope (1st-DE). Then, the algorithm continues to encrypt the previous encrypted message (i.e. 1st-DE) by the private key of the manufacturer (i.e. $PrKey_{origin}$), the public key of the node B (i.e. $PuKey_C$) and one-off key (i.e. $Key_{one\text{-}off}$) to form the second digital envelope (2nd-DE). Next, the previous double encrypted message (i.e. 2nd-DE) is encrypted by private key of the manufacturer (i.e. $PrKey_{origin}$), the public key of the node A (i.e. $PuKey_A$) and the one-off key (i.e. $Key_{one\text{-}off}$) to form the third digital envelope (i.e. 3rd-DE). Finally, the triple encrypted message (i.e. 3rd-DE) is stored to the eSeal device. Such backward Nth digital envelope (Nth-DE) approach (i.e. "backward", in here, means that the algorithm encrypts the message from the end party to the first party and "N" is depending on the number of node in the delivery) can ensure the containers delivering in a right and efficient way. The encryption algorithm is depicted in Figure 6.

Decryption Phase: In contrast to the encryption phase, a decryption based on the asymmetric key cryptography is employed to decode the digital envelope. In this stage, each node is responsible to decrypt the corresponding envelope designed in the electronic shipping document. Take node A as an example, the algorithm starts by enquiring the one-off key (i.e. $Key_{one\text{-}off}$) from the back-end system and hence decrypting the 3rd-DE by the pubic key of the manufacturer (i.e. $PuKey_{origin}$), the private key of the node A (i.e. *PriKeyA*) and $Key_{one\text{-}off}$. If decryption is success (i.e. the keys are matched), the algorithm will update the encrypted message and the containers can keep going to next station. However, if a container goes a wrong way, there is a mismatch in the asymmetric key infrastructure. In other words, the digital envelope cannot be opened and alert message are prompted to the back-end system and immediate actions can then be taken. The decryption algorithm is depicted in Figure 7.

4. Performance Evaluation

A simulation is conducted to analyze the performance of PKI-TTN scheme in terms of key generation time, security and scalability. The simulation is operated on a Windows-based computer in Intel Core 2 Duo2.80 GHz with 2GB memory. The main objective of this simulation is to investigate the key generation time against the network size (i.e. the amount of network parties). The network size is set to 3, 6, 9, 12, and 15 in which the range can cover most of the real situation in today's supply chain network (i.e. the normal party involved in a track and trace infrastructure is 5 to 10). The average time taken to jointly generate the digital envelope is hence measured. Table 1 shows the time for key generations in terms of different network size. The generation time increases directly

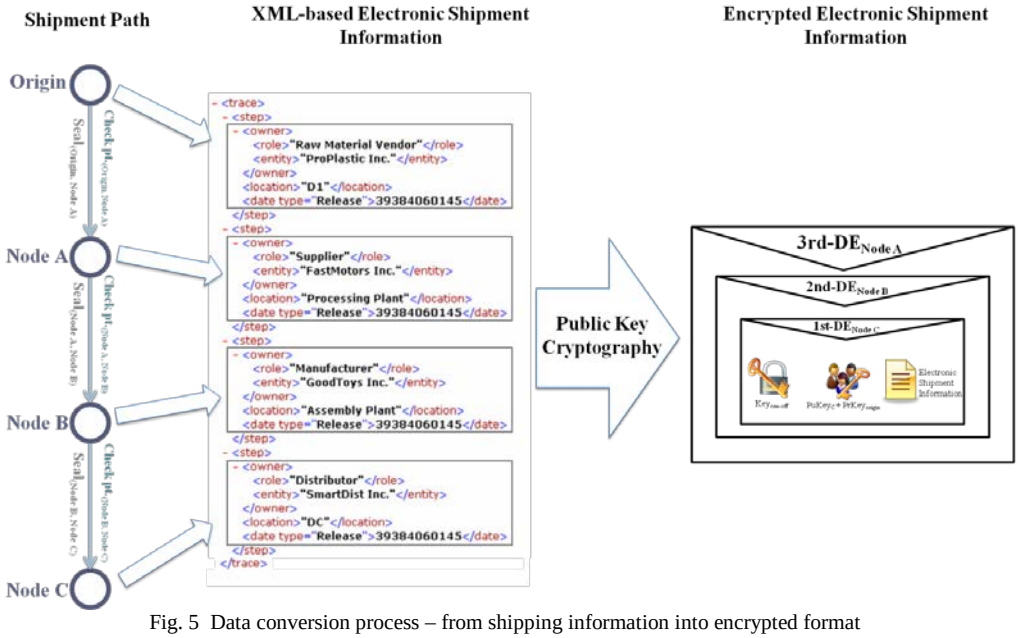


Fig. 5 Data conversion process – from shipping information into encrypted format

```
public Envelope pack(LinkedList<Site> siteList, Site shipper, Seal seal) throws Exception
        {
                Envelope existingEnvelope = null;
                /* from final destination to original source */
                for(int i = siteList.size() - 1; i >= 0; i--)
                {
                        Site site = siteList.get(i);
                        Envelope envelope = new Envelope();
                        envelope.setCheckPoint(site.getSiteId());
                        if(existingEnvelope != null)
                        {
                                envelope.setInterEnvelope(existingEnveloep);
                        }
                        envelope.encrypt(shipper.getPrivateKey(), site.getPublicKey(), seal.getOneOffKey());
                        existingEnvelope = envelope;
                }
                return existingEnvelope;
        }
```

Fig. 6 Pseudo code of encryption algorithm

```
public Envelope unpack(Envelope envelope, Site shipper, Site currentSite, Seal seal) throws Exception
        {
                envelope.decrypt(shipper.getPublicKey(), currentSite.getPrivateKey(), seal.getOneOffKey());

                if(envelope.getCheckPoint() == currentSite.getSiteId())
                {
                        /* if "envelope.getInterEnvelope()" returns "null" value, then it means the current site is
                           final destination */
                        return envelope.getInterEnvelope();
                }
                else
                {
                        throw new Exception("Invalid Check Point!");
                }
        }
```

Fig. 7 Pseudo code of decryption algorithm

proportional to the amount of network parties. The result is acceptable as it requires at most 30 seconds to generate the key set for the eSeal device.

Table 1: Key generation time against different network sizes

Network Size	Time (s)
3	0.39
6	3.88
9	11.28
12	15.73
15	32.55

In addition to the system performance, security analysis of PKI-TTN is also discussed in this section. As mentioned before, confidentiality and integrity are two important security elements to be ensured in the data protection of the current eSeal technology. In PKI-TTN scheme, the privacy of the electronic shipment information is protected by encrypting data streams as well as the messages via the asymmetric encryption. This action can prevent any unauthorized disclosure of information because only an intended recipient can get the corrected key set (i.e. the sender's public key and the eSeal one-off key) to decrypt the message. On the other hand, the integrity can be ensured by the PKI-TTN scheme in forms of the digital signature. With the property of cryptographic hash algorithm and signature algorithm, any change in the input message leads to any unpredictable change in the output message can be detected. In this case, if the data has changed, the signature is failure to verify, and hence the loss of integrity will be obvious to the recipient.

5. Conclusions

In recent years, critical issues, including theft of products, product counterfeit, cross border smuggling and invisible supply chain information, are penetrating in the current track and trace infrastructure. In cope with these challenges, RFID-tagged seals are proposed to securely protect high value of products delivered from a manufacturer to retailers. However, another problem related to the data protection raises in the RFID communication process. Thus, a framework for an integration of eSeal technology with public key cryptography is introduced in this paper. The information stored in RFID tags is securely protected by electronic envelope method with asymmetric encryption and decryption algorithms. Together with the electronic shipment information, a PKI-based Track and Trace Network (PKI-TTN) for cross-boundary container security is proposed to improve efficiency of product delivery, processing time at each supply chain node, and detection of cross-border smuggling in supply chain management.

Acknowledgments

The authors would also like to express their sincere thanks to the Research Committee of the Hong Kong Polytechnic University for providing the financial support for this research work (Project No.: 1-BB7Q).

References

[1] Organisation for Economic Co-operation and Development, Container Transport Security Across Modes, Paris: OECD, 2005.
[2] J. Zhang, and C. Zhang, "Smart Container Security: the E-seal with RFID Technology", Modern Applied Science, Vol. 1, No. 3, 2007, pp. 16-18.
[3] D. Mullen, "The Application of RFID Technology in a Port", Port Technology International, Vol. 22, 2004, pp. 181-182.
[4] F. Rizzo, M. Barboni, L. Faggion, G. Azzalin, and M. Sironi, "Improved Security for Commercial Container Transports Using an Innovative Active RFID System", Journal of Network and Computer Applications, Vol. 34, No. 3, 2011, pp. 846-852.
[5] E.W.T. Ngai, T.C.E. Cheng, S. Au, and K. Lai, "Mobile Commerce Integrated with RFID Technology in a Container Depot", Decision Support Systems, Vol. 43, No. 1, 2007, pp. 62-76.
[6] S.L. Ting, S.K. Kwok, A.H.C. Tsang, and G.T.S. Ho, "The Study on Using Passive RFID Tags for Indoor Positioning", International Journal of Engineering Business Management, Vol. 3, No. 1, 2011, pp. 9-15.
[7] J.P.T. Mo, and W. Lorchirachoonkul, "Design of RFID Cloud Services in a Low Bandwidth Network Environment", International Journal of Engineering Business Management, Vol. 3, No. 1, 2011, pp. 38-43.
[8] E.W.T. Ngai, K.K.L. Moon, F.J. Riggins, and C.Y. Yi, "RFID Research: An Academic Literature Review (1995–2005) and Future Research Directions", International Journal of Production Economics, Vol. 112, No. 2, 2008, pp. 510-520.
[9] S.K. Kwok, P.H. Ng, and K.L. Choy, "Development of an RFID-based Intelligent e-Seal System for Container and Physical Asset Management", Annual Journal of IIE(HK), Vol. 28, 2008, pp. 70-81.
[10] D. Friedlos, "Taiwan customs officials adopt RFID-enabled container seals", RFID Journal, 2001, Available at: http://www.rfidjournal.com/article/view/4727.
[11] B.L. Dos Santos, and L.S. Smith, "RFID in the Supply Chain: Panacea or Pandora's Box?", Communications of the ACM, Vol. 51, No. 10, 2008, pp. 127-131.
[12] L.W. Ferreira Chaves, and F. Kerschbaum, "Security and Privacy in Track and Trace Infrastructures", in D. Bouca, and A. Gafagnao (Eds.), Agent-Based Computing (pp. 109-122), New York: Nova Science Publishers, 2010.
[13] R. Celeste, and B.A. Cusack, "EPCglobal Standards in the Pharmaceutical Industry: Toward a Safe and Secure Supply Chain", Journal of Pharmacy Practice, Vol. 19, No. 4, 2006, pp. 244-249.
[14] S.K. Kwok, S.L. Ting, Albert H.C. Tsang, and C.F. Cheung, "A Counterfeit Network Analyzer Based on RFID

and EPC", Industrial Management & Data Systems, Vol. 110, No. 7, 2010, pp.1018-1037.

[15] L. Castro, and S. Fosso Wamba, "An Inside Look at RFID Technology", Journal of Technology Management & innovation, Vol. 2, No. 1, 2007,pp. 128-141.

[16] C. Adams, and S. Lloyd, Understanding PKI: Concepts, Standards, and Deployment Considerations, Boston: Addison-Wesley, 2003.

[17] H. Liping, and S. Lei, "Research on Trust Model of PKI", in Proceedings of the International Conference on Intelligent Computation Technology and Automation (ICICTA), 2011, Vol. 1, pp. 232-235.

[18] S. Wilson, "A Novel Application of PKI Smartcards to Anonymise Health Identifiers", in Proceedings of the AusCERT 2005 Asia Pacific Information Technology Security Conference, 2005.

[19] J. Dankers, T. Garefalakis, R. Schaffelhofer, and T. Wright, "Public Key Infrastructure in Mobile Systems", Electronics & Communication Engineering Journal, Vol. 14, No. 5, 2002, pp. 180-190.

[20] R. Brandner, M. van der Haak, M. Hartmann, R. Haux, and P. Schmücker, "Electronic Signature for Medical Documents - Integration and Evaluation of a Public Key Infrastructure in Hospitals", Methods of Information in Medicine, Vol. 41, No. 4, 2002, pp. 321-330.

[21] P.S. Tsilingiris, H.N. Psaraftis, and D.V. Lyridis, "RFID-enabled Innovative Solutions Promote Container Security", in Proceedings of the Annual International Symposium on Maritime Safety, Security and Environmental Protection, 2007.

[22] S. Bastia, "Next Generation Technologies to Combat Counterfeiting of Electronic Components," IEEE Transactions on Components and Packaging Technologies, Vol. 25, No. 1, 2002, pp. 175-176.

[23] F. Armenio, H. Barthel, L. Burstein, P. Dietrich, J. Duker, J. Garrett, B. Hogan, O. Ryaboy, S. Sarma, J. Schmidt, K.K. Suen, K. Traub, and J. Williams, "The EPCglobal Architecture Framework", EPCglobal Inc., 2007.

[24] C. Swedberg, "RFID Seals Provide Border Security in Eastern Europe", RFID Journal, 2008, Available at http://www.rfidjournal.com/article/articleprint/4032/-1/1.

[25] T. Yeh, Y. Wang, T. Kuo, and S. Wang, "Securing RFID Systems Conforming to EPC Class 1 Generation 2 Standard", Expert Systems with Applications, Vol. 37, No. 12, 2010, pp. 7678-7683.

[26] S.A. Brands, Rethinking Public Key Infrastructures and Digital Certificates: Building in Privacy, Cambridge, Mass.: MIT Press, 2000.

[27] B.M. Kim, K.Y. Choi, and D.H. Lee, "Disaster Coverable PKI Model Utilizing the Existing PKI Structure", in Proceedings of the Workshops On the Move to Meaningful Internet Systems, 2006.

[28] N.C. Wu, M.A. Nystrom, T.R. Lin, and H.C. Yu, "Challenges to Global RFID Adoption", Technovation, Vol. 26, No. 12, 2006, pp. 1317-1323.

[29] L. Batina, J. Guajardo, T. Kerins, N. Mentens, P. Tuyls, and I. Verbauwhede, "Public-Key Cryptography for RFID-Tags", in Proceedings of the Fifth IEEE International Conference on Pervasive Computing and Communications Workshops, 2007, pp.217-222.

[30] A. Wallstabe, and H. Pohl, "Implementing high-level Counterfeit Security using RFID and PKI", in Proceedings of the Third European Workshop on RFID Systems and Technologies, 2007, pp. 1-5.

[31] Y. Ham, N. Kim, C. Pyo, and J. Chung, "A Study on Establishment of Secure RFID Network Using DNS Security Extension", in Proceedings of the Asia-Pacific Conference on Communications, 2005, pp. 525-529.

[32] H. Deng, and W. Deng, "Identity Authentication in RFID Based Logistics-Customs Clearance Service Platform", in Proceedings of the Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2009, pp. 604-607.

[33] F. Armknecht, L. Chen, A. Sadeghi, and C. Wachsmann, "Anonymous Authentication for RFID Systems", in Proceedings of the Sixth Workshop of RFID Security, 2010.

[34] S. Vaudenay, "RFID Privacy Based on Public-Key Cryptography", in Proceedings of the Ninth International Conference of the Information Security and Cryptology, 2006.

[35] P. Lei, F. Claret-Tournier, C. Chatwin, and R. Young, A Secure Mobile Track and Trace System for Anti-counterfeiting, in Proceedings of the IEEE International Conference on e-Technology, e-Commerce and e-Service Hong Kong, 2005, pp. 686-689.

[36] EPCglobal, Inc., "RFID Implementation Cookbook", 2006, Available at: http://www.epcglobalinc.org/what/cookbook.

[37] K.H.M. Wong, P.C.L. Hui, and A.C.K. Chan, "Cryptography and Authentication on RFID Passive Tags for Apparel Products", Computers in Industry, Vol. 57, No. 4, 2006, pp. 342-349.

[38] J. Zhang, and X. Liu, "On the Security of ID-Based One-Off Blind Public Key", in Proceedings of the International Conference on Internet Technology and Applications, 2010, pp. 1-4.

Adaboost Ensemble with Genetic Algorithm Post Optimization for Intrusion Detection

Hany M. Harb[1] and Abeer S. Desuky[2]

[1]Computers and Systems Engineering Dept., Faculty of Eng., Azhar University
Cairo, Egypt

[2]Mathematics Dept., Faculty of Science, Azhar University
Cairo, Egypt

Abstract

This paper presents a fast learning algorithm using AdaBoost ensemble with simple genetic algorithms (GAs) for intrusion detection systems. Unlike traditional approaches using AdaBoost algorithms, it proposed a Genetic Algorithm post optimization procedure for the found classifiers and their coefficients removing the redundancy classifiers which cause higher error rates and leading to shorter final classifiers and a speedup of classification. This approach has been implemented and tested on the NSL-KDD dataset and its experimental results show that the method reduces the complexity of computation, while maintaining the high detection accuracy. Moreover, the method improves the processing time, so it is especially appealing for the real-time processing of the intrusion detection system.

Keywords: *Intrusion Detection; AdaBoost; Genetic Algorithm; Feature Selection; Classification; NSL-KDD dataset.*

1. Introduction

With the development of the Internet, the information security threat is becoming one of the most crucial problems. Reliable connections, information integrity and privacy on computer systems and networks are demanded more intensively nowadays than ever before. Therefore, intrusion detection systems (IDS) are used to monitor systems during their lifetime and used to detect possible attacks against them.

IDS are usually classified in two categories: misuse and anomaly [1]. A misuse or knowledge-based IDS aims at detecting the occurrence of states or action sequences that have been previously identified to be an intrusion. Thus, in this kind of IDS, attacks must be known and described a priori and IDS are usually unable to deal with new or unknown attacks. Alternatively, an anomaly or behavior-based IDS assumes that an intrusion can be detected by observing deviations from a normal or expected behavior of a monitored entity. The valid behavior is extracted from previous reference information about the system. The IDS later compares the extracted model with the current activity and raises an alert each time that a certain degree of divergence from the original model is observed.

An effective Intrusion Detection system needs to limit false positives—incorrectly identifying an attack when there is none. At the same time it needs to be effective at catching attacks. False alarms are distracting and so they reduce the effectiveness of an Intrusion Detection system. Hence, the basic target of IDS research is reducing the false positives rate while maintaining high detection rate.

One of the main problems with IDS is the overhead, which can become prohibitively high. The detection response time is another major problem in the IDS. Computer networks have a dynamic nature in a sense that information and data within them are continuously changing. Therefore, detecting an intrusion accurately and promptly is crucial especially in real time IDS.

Most of the research works are aimed to introduce the most time efficient methodologies [2]. Our research goal is to improve the AdaBoost (adaptive boosting) algorithm to be suitable for the real time implementation.

In this paper we tried to construct an intrusion detection approach with a low computational complexity, a high detection accuracy, and efficiency in the real time implementation. In this correspondence, we apply the AdaBoost algorithm with Genetic algorithm to intrusion detection.

2. Related work

M. Panda and M. R. Patra [3] reported a work on the subject of intrusion detection for the anomaly detection. Authors in this paper have proposed a framework of NIDS based on Naïve Bayes algorithm. When compared their approach to the neural network based approach, their approach achieved higher detection rate which is about 95%.

S. Teng et al. in [4] proposed cooperative network intrusion detection Based on Fuzzy SVMs. They firstly preprocessed the data. Then the fuzzy membership function is introduced into SVM. Three types of detecting agents are generated according to TCP, UDP and ICMP protocol. Finally, using the KDD CUP 1999 data set, a comparison between single FSVM and multi FSVM showed that, with cooperative detection based on multi FSVMs, the accuracy rate is 91.2101%, and with single FSVM, the accuracy rate is 82.5682%.

In [5], Bankovic´ Z et al., presented an improvement to network security using genetic algorithm approach. In this work they have realized a misuse detection system based on genetic algorithm (GA) approach. To be able to process network data in real time, they have deployed principal component analysis (PCA) to extract the most important features of the data and then to keep the high level of detection rates of attacks while speeding up the processing of the data.

In [6], Weiming Hu et al. proposed an intrusion detection algorithm based on the AdaBoost algorithm. In the algorithm, decision stumps are used as weak classifiers. Their results showed 90.88% detection rate and 1.79 % false alarm rate and in comparison with other algorithms, the algorithm gave lower computational complexity and error rates.

3. The Proposed Method: GA-AdaBoost

Ideally, an IDS should be fast, simple, and accurate, while at the same time being complete. It should detect all attacks with little performance penalty. In our Work we aim at constructing a fast intrusion detection approach with a low computational complexity to be suitable for use in real time applications. To keep a high accuracy, in this correspondence, we apply the AdaBoost with Genetic algorithm to intrusion detection. The motivations for applying the AdaBoost with Genetic algorithm are as it follows:

- The AdaBoost algorithm is one of the most popular machine learning algorithms. It has been applied to many pattern recognition problems, such as the face recognition. However, the application of the AdaBoost algorithm to intrusion detection has not been explored so far [6].

- AdaBoost is a sequential forward search procedure using the greedy selection strategy. Because of its greedy character, neither the found weak classifiers nor their coefficients are optimal [7]. Genetic Algorithm, proposed as a post optimization procedure for the found classifiers and their coefficients, removes the redundancy classifiers and leads to shorter final classifiers and a speedup of classification.

To process network data in real time and perform efficient intrusion detection, we need to extract the most important pieces of information that can be deployed for efficient detection of network attacks. We have deployed an alternative way proposed in [8] to reduce the dimension of the used data. According to the obtained results, we have selected sixteen features out of forty one used to describe each connection of KDD dataset [14, 15]. Our results confirm maintenance of high accuracy while using lower dimension of data. The other benefit is that data processing and the decision making whether a connection is an attack are performed much faster.

3.1 AdaBoost

The AdaBoost algorithm is a kind of boosting algorithms which were proposed by Freund and Schapire [9]. Fig. 1 is a generalized version of the AdaBoost algorithm for binary classification problems.

Given:
$S=\{(x_1, y_1),(x_2, y_2),...,(x_l, y_l)\}$ With $x_i \in X$
And $y_i \in \{-1, +1\}$.

Initialize the distribution:
$D_l^{(i)} = 1/l,\ i = 1,\ 2,\ ...,l.$
For t = 1, 2, …, T:
Train the weak learner using the distribution
$D_t^{(i)}$, $i = 1,\ 2,\ ...,l.$
Get the weak hypothesis $c_t : X \to R.$
Update:
$D_{t+1}^{(i)} = D_t^{(i)} \exp(-\alpha_t y_i c_t(x_i))/Z_t,\ i = 1,\ 2,\ ...,\ l,$
where Z_t is a normalization factor
($D_t^{(i)}$ is still a distribution)
and $\alpha_t = \frac{1}{2} \ln\frac{1-\varepsilon_t}{\varepsilon_t}$ with
$\varepsilon_t = \sum_{i=1}^{l} D_t^{(i)}[y_i \neq c_t(x_i)].$
Output the final hypothesis:
$C(X) = sign(\sum_{t=1}^{T} \alpha_t c_t(X)).$

Fig. 1 A generalized version of the AdaBoost algorithm.

AdaBoost algorithm is used to boost the classification performance of a weak learner. It does this by combining a collection of weak classification functions to form a stronger classifier. AdaBoost combines iteratively the weak classifiers by taking into account a weight distribution on the training samples such that more weight is attributed to samples misclassified by the previous iterations. The final strong classifier takes the form of a perceptron, a weighted combination of weak classifiers followed by a threshold. As in Fig. 1 the algorithm takes as input a training set (x_1, y_1), (x_2, y_2), …, (x_l, y_l) where each x_i belongs to some domain or instance space X, and each label y_i is in some label set Y. We iterate T rounds of AdaBoost training where T is the number of weak classifiers c_t and ensemble weights α_t are yielded by learning to constitute the final strong classifiers [9] [10].

The weak classifier is the core of an AdaBoost algorithm. In our AdaBoost-based algorithm for intrusion detection, classification and regression tree (CART) algorithm - proposed by Breiman et al. [11] - is used as weak classifiers.

3.2 Genetic Algorithm Overview

Genetic algorithms (GA) are search algorithms based on the principles of natural selection and genetics. The bases of genetic algorithm approach are given by Holland [12] and it has been deployed to solve wide range of problems.

GA operates on a population of potential solutions applying the principle of survival of the fittest to produce better and better approximations to the solution of the problem that GA is trying to solve. At each generation, a new set of approximations is created by the process of selecting individuals according to their level of fitness value in the problem domain and breeding them together using the operators borrowed from the genetic process performed in nature, i.e. crossover and mutation. This process leads to the evolution of populations of individuals that are better adapted to their environment than the individuals that they were created from, just as it happens in natural adaptation [13].

3.3 Overview of GA-AdaBoost Algorithm

According to the model of the boosted classifier $C(X) = sign(\sum_{t=1}^{T} \alpha_t c_t(X))$, the final strong classifier could be regarded as the weight combination of weak classifiers $\{c_1, c_2, \ldots, c_T\}$.

Each weak classifier c_t will be determined after the boosting training.

As AdaBoost is a sequential forward search procedure using the greedy selection strategy, neither the found weak classifiers nor their coefficients are optimal. Moreover, the strong classifier comprises more weak classifiers requiring more time to evaluate and more memory to occupy. This affects the performance of IDS and slows down the detection process. To address this issue, we propose an approach based on Genetic Algorithm. This algorithm selects some weak classifiers to constitute an ensemble (final strong classifier). The weak classifiers are selected according to some evolving weights that could characterize the fitness of the included weak classifiers in the ensemble. The approach procedure is summarized in Fig. 2.

Given:

$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$ With $x_i \in X$

And $y_i \in \{-1, +1\}$.

Initialize the distribution:

$D_l^{(i)} = 1/l, \ i = 1, 2, \ldots, l.$

For t = 1, 2, …, T:

Train the weak learner using the distribution

$D_t^{(i)}, \ i = 1, 2, \ldots, l.$

Get the weak hypothesis $g_t : X \to R.$

Update:

$D_{t+1}^{(i)} = D_t^{(i)} \exp(-\alpha_t y_i g_t(x_i)) / Z_t, \ i = 1, 2, \ldots, l,$

where Z_t is a normalization factor ($D_t^{(i)}$ is still a distribution)

and $\alpha_t = \frac{1}{2} \ln \frac{1-\varepsilon_t}{\varepsilon_t}$ with

$\varepsilon_t = \sum_{i=1}^{l} D_t^{(i)} [y_i \neq g_t(x_i)].$

Output the final hypothesis:

$G(X) = sign(\sum_{t=1}^{T} \alpha_t G_t(X)).$

- Generate a population of bit strings b.
- Evolve population where the fitness of each individual is measured as

$f(b) = w_1 * (1 - l/T) + w_2 * (1/E_b^s)$

- b* is the evolved best individual.
- Use vector b* to update -reduce- the classifiers with their weights.
- Generate a new population of reduced weight vectors w.
- Evolve population where the fitness of each individual is measured as $f(w) = 1/E_w^s$
- w* is the evolved best individual.

Outbut: ensamble G*

$G^*(X) = sign(\sum^{T} w^* G_t(X))$

Fig. 2 Post-Optimization procedure of a given boosted strong classifier based on Genetic algorithms.

The proposed Algorithm is depicted in Fig. 3 which consists of the following three phases:

- Pre-processing and Features Extraction Phase:
  In this phase, randomly two separated training and testing data sets are selected from the NSL-KDD dataset, the symbolic features are converted into numerical ones, and the most suitable features are selected.
- AdaBoost Training and Testing Phase:
  Training subphase: In this phase AdaBoost is trained using the training dataset. It iterates T rounds of AdaBoost training, where T is the number of weak classifiers c_t and ensemble weights α_t are yielded by learning to constitute the final strong classifiers.
  Testing subphase: the performance of the system with the testing data set is measured.
- Post Optimization Procedure Based on Genetic Algorithm: This phase is implemented in two steps:

Step 1: Initialize population, each individual in the evolving population is a vector b composed of bit string (b_1, b_2, …, b_T) denoting the weak classifiers. These weak classifiers constitute the strong classifier. The b_i=1 denotes the i^{th} weak classifier appearance in the ensemble while b_i=0 denotes its absence. In order to evaluate the goodness of the individuals, a validation data set is employed. Let E_b^s be the validation error of the ensemble corresponding to the individual b on the validation set s. It is obvious that E_b^s expresses the goodness of b in the way that the smaller E_b^s is, the better b is. Furthermore, let $l = \sum_{i=1}^{T} b_i$ be the number of the selected weak classifiers, where T is the total number of weak classifiers. It is preferable that fewer weak classifiers give a correct prediction for a given object rather than more weak classifiers. Therefore, $f(b) = w_1 * (1 - l/T) + w_2 * (1/E_b^s)$ is used as the fitness function, where w1 and w2 are fitness weights that can be adjusted to balance the efficiency, mode size and accuracy of classifiers.

Step 2: In this step we use the resulted vector b to update - reduce- the number of classifiers and their corresponding weights and then a new population is initialized. Each individual in this population is a vector $w = (w_1, w_2, \ldots, w_T)$ which denotes the corresponding weights of the weak classifiers. As in step one let E_w^s be the validation error of the ensemble corresponding to the individual w on the validation set s and $f(w) = 1/E_w^s$ is used as the fitness function. The procedure is summarized in Fig. 2.

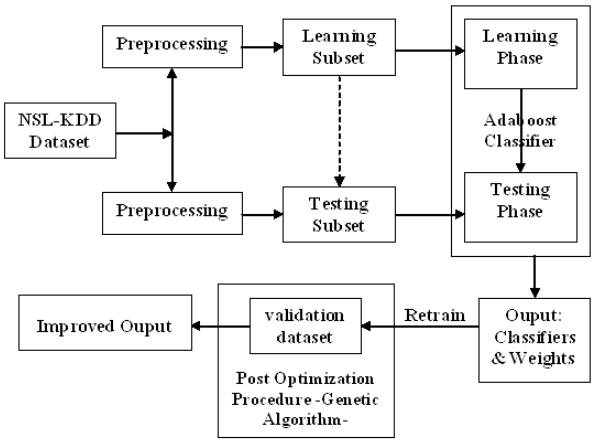


Fig. 3 The Proposed Model Structure

4. Experimental Results

We have run our experiments using Matlab 7, on a system with a 1.86GHZ Intel(R) Celeron(R) M processor and 512MB of RAM running Microsoft Windows XP Professional (SP2).

In our experiment, we used NSL-KDD data set [14]. It has solved some of the inherent problems of the KDDcup99 [15]. It is considered as a standard benchmark for intrusion detection algorithms evaluation. The training dataset of NSL-KDD is similar to KDDcup99 consisting of approximately 1,074,992 single connection vectors each of which contains 41 features. We used NSL-KDD full training dataset that contains 25192 connection records with one target value or labeled class either normal or attack.

The AdaBoost algorithm and the weak learner employed by our algorithm is implemented by the GML AdaBoost Matlab Toolbox developed by Alexander Vezhnevets [16]. The parameters used for evolution were: 0.05 crossover rate and 0.05 mutation rate, and the population was initialized randomly. In our algorithm, the GA terminates if there is no found better individual within the next 100 generations or if the validation error is equal to zero. From the experiment results shown in Table 1, we can see that the numbers of weak classifiers of the boosted strong classifier trained by standard AdaBoost are reduced by about 42% due to the Genetic Algorithms

optimization. Moreover, as the weak classifiers reduction, the average classification time (measured in seconds) of the boosted strong classifier with GA-optimization is about 600% faster. The accuracy of the classification is also slightly increased by 0.64, as also shown in Fig. 4 and Fig. 5. The results along with the comparison to other existing methods using NSL-KDD Dataset are shown in Table 2.

Table 1 Comparison of The Boosted Classifiers Without and With Post-Optimization

AdaBoost classifiers	*Our classifiers*	*AdaBoost time*	*Our time*	*AdaBoost accuracy*	*Our accuracy*
20	17	6.81	0.08	97.00	97.89
40	23	7.19	0.13	98.36	98.59
60	40	6.08	0.20	98.48	99.02
80	46	6.69	0.22	98.97	99.14
100	67	6.94	0.27	99.21	99.32
120	71	6.77	0.30	99.41	99.48
140	89	6.91	0.44	99.48	99.38
160	81	6.52	0.45	99.51	99.55
180	104	7.02	0.52	99.54	99.44
200	109	7.39	0.50	99.53	99.57
Average		6.83	0.31	98.95	99.14

Table 2 Detection Accuracy Comparison of Machine learning algorithms using NSL-KDD Dataset

Classifier	*Detection Accuracy(%)*	*Classification time(seconds)*
Discriminative Multinomial Naïve Bayes [17]	96.5	1.11
Discriminative Multinomial Naïve Bayes +PCA [17]	94.84	118.36
AdaBoost [18]	90.31	***
Decision Trees (J48) [14]	81.05	***
AdaBoost + GA (proposed)	99.57	0.5

*** indicates data not provided by the authors in their paper.

5. Conclusion

The goal of a network-based intrusion detection system (IDS) is to distinguish the attacks on the Internet from normal use of the Internet. It is an indispensable part of the information security system. Due to the variety of network behaviors and the rapid development of attack fashions, it is necessary to develop machine-learning-based intrusion detection algorithms with high detection rates and low false-alarm rates and detection time. We have proposed an AdaBoost ensemble with Genetic algorithm for intrusion detection. The method has effectively improved results of the boosted classifier. The experimental results have demonstrated that a given boosted classifier with our post optimization based on GA (Genetic Algorithm) has fewer weak classifiers and an increase of the classification accuracy and speed which is important for real time network applications.

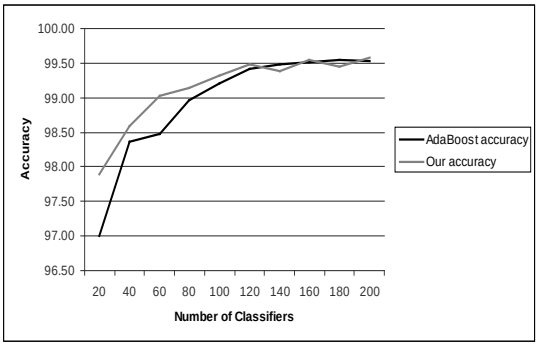


Fig. 4 AdaBoost Accuracy and Post-Optimization Accuracy.

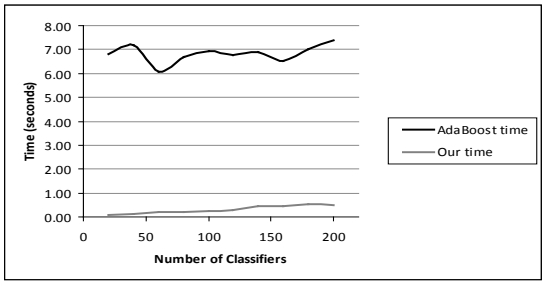


Fig. 5 AdaBoost Time and Post-Optimization Time.

References

[1] Ian Stewart, "A Modified Genetic Algorithm and Switch-Based Neural Network Model Applied to Misuse-Based Intrusion Detection", M.s. thesis, School of Computing In conformity, Queen's University, Kingston, Ontario, Canada, February 2009.

[2] S. X. Wu and W. Banzhaf, "The use of computational intelligence in intrusion detection systems: A review," Applied Soft Computing, vol. 10, 2010, pp. 1–35.

[3] Mrutyunjaya Panda and Manas Ranjan Patra, "Network Intrusion Detection Using Naïve Bayes", IJCSNS

International Journal of Computer Science and Network Security, VOL.7, No.12, 2007, pp. 258-263.
[4] Shaohua Teng, Hongle Du, Naiqi Wu, Wei Zhang, Jiangyi Su, "A Cooperative Network Intrusion Detection Based on Fuzzy SVMs", Journal Of Networks, VOL. 5, NO. 4, 2010, pp. 475-483.
[5] Bankovic, Z., Moya, José M., Araujo, A., Bojanic, S., and Nieto-Taladriz, "Improving network security using genetic algorithm approach", Computers & Electrical Engineering, Vol.33, Issue 5-6, 2007, pp. 438-451.
[6] Weiming Hu, Wei Hu, and Steve Maybank, "AdaBoost-Based Algorithm for Network Intrusion Detection", IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 38, No. 2, 2008, pp. 577-583.
[7] Zhang Z, Li S Z, Zhang H, "Real-time multi-view face detection", Fifth IEEE International Conference on Automatic Face and Gesture Recognition, Washington D.C, USA, 2002, pp. 142-147.
[8] Hany M. Harb, Afaf A. Zaghrot, Mohamed A. Gomaa and Abeer S. Desuky, "Selecting Optimal Subset of Features for Intrusion Detection Systems", Advances in Computational Sciences and echnology, Research India Publications, Volume 4 Number 2, 2011, pp. 179-192.
[9] Y. Freund and R. E. Schapire, “Experiments with a new boosting algorithm,” in Proceedings of the 13th International Conference on Machine Learning, Bari, Italy, 1996, pp. 148-156.
[10] Zhou Z-H, WU J, Tang W, "Ensembling neural networks: many could be better than all", Artificial Intelligence, 2002, pp. 239-263.
[11] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and Regression Trees", Chapman and Hall, New York, USA, 1984.
[12] Holland J. "Adaptation in natural and artificial system. Ann Arbor", The University of Michigan Press, 1975.
[13] B. Abdullah, I. Abd-alghafar, Gouda I. Salama and A. Abd-alhafez, " Performance Evaluation of a Genetic Algorithm Based Approach to Network Intrusion Detection System", 13th International Conference on Aerospace Sciences & Aviation Technology, Asat- 13, May 26 – 28, 2009, pp. 1-17.
[14] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali, A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", proceeding of IEEE symposium on computational Intelligence in security and defence application, 2009.
[15] KDD Cup 1999 Data Set, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.htm
[16] A. Vezhnevets, GML AdaBoost Matlab Toolbox [http://research. graphicon.ru/], Graphics and Media Laboratory, Computer Science Department, Moscow State University, Moscow, Russian Federation.2006.
[17] Panda, M., Abraham, A., Patra, M.R., "Discriminative multinomial Naïve Bayes for network intrusion detection", Information Assurance and Security (IAS), 2010 Sixth International Conference, 2010, pp. 5-10.
[18] V.P. Kshirsagar and Dharmaraj R. Patil, "Application of Variant of AdaBoost based Machine Learning Algorithm in Network Intrusion Detection", International Journal of Computer Science and Security (IJCSS), Vol. 4, Issue.2, 2010, pp. 1-6.

Multi-objective Numeric Association Rules Mining via Ant Colony Optimization for Continuous Domains without Specifying Minimum Support and Minimum Confidence

Parisa Moslehi[1], Behrouz Minaei Bidgoli[2], Mahdi Nasiri[3], Afshin Salajegheh[4]

[1] Computer Department, Islamic Azad University South-Tehran Branch Tehran, Tehran, Iran

[2] Computer Department, Iran University of science and Technology Tehran, Tehran, Iran

[3] Computer Department, Iran University of science and Technology Tehran, Tehran, Iran

[4] Computer Department, Islamic Azad University South-Tehran Branch Tehran, Tehran, Iran

Abstract

Currently, all search algorithms which use discretization of numeric attributes for numeric association rule mining, work in the way that the original distribution of the numeric attributes will be lost. This issue leads to loss of information, so that the association rules which are generated through this process are not precise and accurate. Based on this fact, algorithms which can natively handle numeric attributes would be interesting. Since association rule mining can be considered as a multi-objective problem, rather than a single objective one, a new multi-objective algorithm for numeric association rule mining is presented in this paper, using Ant Colony Optimization for Continuous domains (ACO_R). This algorithm mines numeric association rules without any need to specify minimum support and minimum confidence, in one step. In order to do this we modified ACO_R for generating rules. The results show that we have more precise and accurate rules after applying this algorithm and the number of rules is more than the ones resulted from previous works.

Keywords- *Ant Colony Optimization for Continuous Domains, Numeric association rules mining, Multi objective association rules mining*

1. Introduction

Data mining is the most instrumental tool in discovering knowledge from transactions [1][2][3]. Also data mining is known as an integral part of knowledge discovery in databases (KDD). Transactional database refers to the collection of transaction records, which in most cases are sales records. Data mining on transactional database focuses on the mining of association rules, finding the correlation between items in transaction records. Most of the association rule algorithms are based on methods proposed by Agrawal, Imielinski, and Swami [4] and Agrawal and Srikant [5], which are Apriori [4], SETM [6], AIS [4] and Pincer search [7] etc. Neither the rules with numeric attributes nor the rules in the form of $I_8I_{10}I_{12} \rightarrow I_4I_5I_9$ can be discovered by these methods [8]. That is why numeric association rule mining algorithms have been proposed.

On the other hand while we use these methods we must specify minimum support and minimum confidence. This issue makes these methods dependent on datasets and they must execute several times. We do not require specifying support and confidence thresholds in our method; also we extract best rules in one execution of the algorithm. Previous methods mine association rules in two steps. First they find frequent itemsets and then extract association rules from them. In a numeric association rule, attributes can be Boolean, numeric or categorical. Since, mining numeric association rules is a hard optimization problem rather than being a simple discretization one, algorithms that natively handle numeric attributes usually perform better.

In recent years, the swarm intelligence paradigm received widespread attention in research. Two main algorithms of which that are popular swarm intelligence metaheuristics for data mining, are Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO). Swarm Intelligence is based on social behavior that can be observed in nature, such as ant colonies, flocks of birds, fish schools and bee hives, where a number of individuals with limited capabilities are able to come to intelligent solutions for complex problems [9].

Recently, a continuous version of ACO metaheuristic has been proposed for tackling continuous attributes and solving continuous optimization problems in [10] and [11] by Socha.

There is not any study which uses ACO_R for mining numeric association rules. In this paper we propose a method using ant colony optimization for continuous domain (ACO_R) and extract best rules with high support and confidence without any need to specify minimum support and minimum confidence thresholds.

The rest of this paper is organized as follows. In section 2, a brief explanation of Ant Colony optimization algorithm for continuous domain (ACO_R) is presented. In section 3, numeric association rule mining and in section 4 related works are discussed. The proposed multi objective association rule mining algorithm via ACO_R is presented in section 5. Experimental setup and results is presented in section 6. Finally we conclude with a summary in section 7.

2. Ant Colony Optimization for continuous domains (ACO_R)

While ACO uses a discrete probability distribution for choosing a solution, ACO_R uses a probability density function (PDF) and samples it. A Gaussian function is used as PDF in ACO_R.

In ACO a pheromone table is used to store pheromone information. ACO_R uses a solution archive of size k in order to describe the pheromone distribution over the search space. Here, k is the number of complete solutions to the problem. Considering the solution archive as a matrix, each entry is called s_j^i where i=1,2,…,n is the number of dimensions and j=1,2,…,k is the number of rows.

First the archive is initialized with k random solutions. These solutions are ranked based on their quality. The weight ω_j of a complete solution S_j is calculated according to its rank:

$$\omega_j = \frac{1}{qk\sqrt{2\pi}} e^{\frac{-(j-1)^2}{2q^2k^2}} \quad (1)$$

where q is a parameter of the algorithm. The effect of q is that, if it is small the best-ranked solutions are strongly preferred and when it is large, the probability becomes more uniform [11].

Each ant chooses a solution from the archive probabilistically for building its own solution. The probability of choosing S_j by an ant is:

$$p_j = \frac{\omega_j}{\sum_{r=1}^{k} \omega_r} \quad (2)$$

After choosing a solution s_j^i in the archive, each ant samples a Gaussian function:

$$P(x) = g(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (3)$$

where μ and σ are the mean and standard deviation of the Gaussian function respectively. If an ant chooses a solution s_j^i then the value of s_j^i is assigned to μ, and the standard deviation is assigned as follows:

$$\sigma \leftarrow \xi \sum_{r=1}^{k} \frac{|s_r^i - s_j^i|}{k-1} \quad (4)$$

where ξ is a parameter of the algorithm that has the same effect as the pheromone evaporation parameter in ACO algorithm. The higher the value of ξ, the lower the convergence speed of the algorithm [11].

3. Numeric association rule mining

In a basket market transactions where transactions are a list of items purchased by a customer, the knowledge that association rules give us are something like: "70% of customers who buy A also buy B". The applications of association rules are in discovering customer buying patterns for cross-marketing and attached mailing applications, catalog design, product placement, customer segmentation, etc., based on their buying patterns [12].

Given a set of transactions, the problem of mining association rules is to find all association rules that have support and confidence greater than the user-specified minimum support and minimum confidence respectively. Association rules can be boolean or numeric. Numeric association rules can have some numeric attributes, like age, height and etc. they also can have categorical attributes like gender, brand, and etc. numeric attributes need to be dicretized in order to transform the problem into a Boolean one, before mining association rules.

An example of a numeric association rule in an employee database is like this [13]:
"Age ∈ [25,36] ∧ Sex=Male → Salary ∈ [2000,2400] ∧ Have-Car=Yes"
(Support=4%, Confidence=80%)
In this numeric association rule, "Age ∈ [25,36] ∧ Sex=Male" is antecedent and "Salary ∈ [2000,2400] ∧ Have-Car=Yes" is consequent part. This numeric association rule states that "4% (support) of the employees are males aged between 25 and 36 and earning a salary of between $2.000 and $2.400 and have a car", while "80 % (confidence) of males aged between 25 and 36 are earning a salary of between $2.000 and $2.400 and have a car".

In a transaction database the support and confidence of a rule is calculated by following equations [14]:

$$Support, s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \quad (5)$$

$$Confidence, c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \quad (6)$$

where N is the total number of transactions. $\sigma(X \cup Y)$ and $\sigma(X)$ is the frequency of occurrence of the itemset X and $X \cup Y$ respectively, which is called support count. Support determines how often a rule is satisfied in the transaction, and confidence determines how often items in Y appear in transactions that contain X [14].

4. Related works

There has been proposed many numeric association rule mining algorithms. Each of these algorithms use a method to deal with numeric attributes.

Sirkant and Agrawal in 1996 [15], proposed an algorithm for mining association rules in large relational tables containing both quantitative and categorical attributes, by partitioning numeric attributes into a number of intervals. Fukuda et al in 1996 [16], proposed an algorithm for mining optimized association rules for numeric attributes which uses computational geometry for computing optimized ranges.

Miller and Yang in 1997 [17], proposed an algorithm for mining association rules over interval data using Birch, an adaptive clustering algorithm. Lent and Swami in 1997 [18], proposed an algorithm for the problem of clustering two-dimensional association rules in large databases. They used a geometric-based algorithm, BitOp for clustering.

The idea of using an evolutionary algorithm (EA) for mining only frequent sets was applied in Mata et al.(2002) [19]. Aumannn and Lindell in 2003 [20], proposed a statistical theory for quantitative association rules, based on the distribution of values of the quantitative attributes. Kaya and Alhajj in 2005 [2], proposed a genetic clustering method. Ke et al in 2006 [21], proposed an information theoretic approach to quantitative association rule mining. They used discretizing numeric attributes and constructing a graph based on mutual information of attributes.

David and Yanhong in 2007 [22], proposed a fuzzy weighted association rule mining algorithm by transforming numeric and categorical data into fuzzy values. Alatas and Akin in 2007 [23], proposed a multi-objective differential evolution algorithm for mining numeric association rules, later they proposed a rough particle swarm optimization algorithm and presented its applications in data mining especially numeric association rule mining problems. Their algorithm had some improvements in performance and precision in comparison with previous one in [13]. Also they proposed a genetic algorithm for automated mining of both positive and negative quantitative association rules in [24]. Qodmanan et al. in 2010 [8], proposed a multi-objective genetic algorithm for association rule mining and proposed a method without taking into account of minimum support and minimum confidence. Nasiri et al. in 2010 [25], proposed a multi-objective numeric association rule mining algorithm using simulated annealing.

Also, there are some researches which applied ACO for combinatorial optimization problems to association rule mining. The first data mining algorithm using ACO, Ant-Miner, was proposed by Parpinelli et al. in 2002 [26], as a classification algorithm, Since then, there have been proposed many data mining algorithms using ACO in field of clustering and classification. For the first time, Kuo et al in 2007 [27], proposed an algorithm for mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan. Atabaki in [28], used Ant System for mining association rules in a distributed system.

These algorithms mine Boolean association rules. In this paper we use ACO_R which is an extension of ACO to continuous domain for mining numeric association rules.

5. ACO_R in numeric association rule mining

5.1. Solution representation

Since, ACO_R uses a Gaussian function as a PDF, for dealing with continuous variables; we got this idea and used the concept of Gaussian functions in our process of generating intervals. Based on this concept, each solution member in archive is assumed to be the central point of a Gaussian normal distribution. The structure of the solution archive is slightly modified in order to store the standard deviation of each solution member in the archive to store the interval information of numeric attributes and specify whether each attribute is in the antecedent of a rule or the consequent of it.

In ACO_R, the ants move through the archive and choose one row of it based on its associated weight (ω), which is calculated through equation (1). Then they construct a new solution by sampling the Gaussian function g of the selected solution's values of each dimension.

Each numeric attribute is considered as one dimension of the solution archive which has three parts in such a way that each complete solution is considered as a numeric association rule. The structure of each complete solution in proposed algorithm is shown in Fig. 1 where n is the number of attributes of database being mined.

	1			……..		n		
S_j	AC_j^1	s_j^1	σ_j^1			AC_j^n	s_j^n	σ_j^n

Fig.1. The structure of a complete solution in proposed algorithm

The first part of each complete solution represents the antecedent or consequent of the rule and can take values between 0 and 1like the work in [13]. If the first part of the decision variable takes values between $0.00 \leq AC_j \leq 0.33$, it means that this item will be in the antecedent of

the rule and if $0.33 < AC_j \leq 0.66$, this item will be in the consequent of the rule. If $0.66 < AC_j \leq 1.00$, it means that this item will not be involved in the rule. All decision variables which are in the interval 0.00 and 0.33 will form the antecedent of the rule while decision variables which have values in the interval between 0.33 and 0.66 will form the consequent of the rule.

The second part represents the value of the solution and the third part represents the standard deviation of that solution which is used to build the intervals of numeric data. Our numeric association rule mining algorithm is described in the following.

5.2. Objective function

In ACO_R algorithm, our goal is to optimize a function which is called objective function. The mined rules have to acquire high support and confidence. Our algorithm has been designed to find the intervals of each numeric attribute that conform an interesting rule, by the use of Gaussian functions, in such a way that the objective function itself is the one that decides the frequency and the length of the intervals. The objective function is shown in Eq. (7)

$$Objective = \alpha 1 \times \text{Support} + \alpha 2 \times \text{Confidence} - \alpha 3 \text{ Interestingness} - \alpha 4 \times Int \quad (7)$$

This objective function has four parts. The first part can be considered as support of the rule that is statistical significance of an association rule. The second part can be considered as confidence value. The third part is used for number of attributes in the rule. This parameter rewards the shorter rules with a smaller number of attributes. By interestingness measure, readability, comprehensibility, and ease of understanding that are important in data mining are increased. It is known that larger rules are more likely to contain redundant or unimportant information, and this can obscure the basic components that make the rule successful and efficiently processable. The last part of the objective function is used to penalize the amplitude of the intervals that conform the itemset and rule. In this way, between two solutions that cover the same number of records and have the same number of attributes, the one whose intervals are smaller gives the best information. *Int* has been computed as shown in Eq. (8)

$$Int = \sum_{i=0}^{n} \frac{(UB_i - LB_i)}{max\ bound_i - min\ bound_i} \quad (8)$$

Where n is the number of attributes in a rule, and *max bound* and *min bound* are Maximum and minimum allowable values of an attribute in a database, and UB_i and LB_i are upper bound value and lower bound value of an attribute in a rule. Since in our proposed algorithm the notion of Gaussian functions is used, these values can be acquired by adding a coefficient of standard deviation to the value of a solution s_j^i, in fact the upper bound and lower bound of the intervals are calculated by the equation (9):

$$UB_i = s_j^i + \alpha_5 \sigma \quad \text{and} \quad LB_i = s_j^i - \alpha_5 \sigma \quad (9)$$

This way, the distribution of the original data is kept. α_1, α_2, α_3, α_4 and α_5 are user specified parameters and one might increase or decrease the effects of parts of objective function.

5.3. Rule generation

In ACO_R pheromone table is replaced by solution archive, to keep track of the solutions and the pheromone update procedure adds a number of new solutions, each of which is generated by one ant, and eliminates the same number of bad solutions from the archive after ranking its solutions. This way, the best solutions are always on top of the solution archive, so the best solution in each execution of the ACO_R can be considered as a rule, as the structure of a complete solution in proposed algorithm represents a rule. To obtain more rules the final user can execute the algorithm as many times as it is required. Also he can consider a number of top solutions from the archive as the best ranking rules of each execution.

5.4. Initialization and parameter control

First the solution archive is filled by some uniform random data, based on a range that is defined by the user in run-time; this range is usually selected in the way that covers the upper bound and lower bound of the numeric attribute value in the database. Furthermore, as this range has a vital effect on solutions of the algorithm, it can be selected in a way to focus on some particular parts of its range. Then the weights are calculated according to the equation (1) and one solution is selected probabilistically according to equation (2).

After that, the vector of standard deviations is calculated according to the equation (4), considering s_r^i as a solution member and the central point of intervals, and k as the size of the solution archive. Choosing a proper value for ξ, affects the ability of the algorithm to find the correct solutions [11]. Fig. (2) shows the effect of ξ on generated intervals that will be used by ACO_R process.

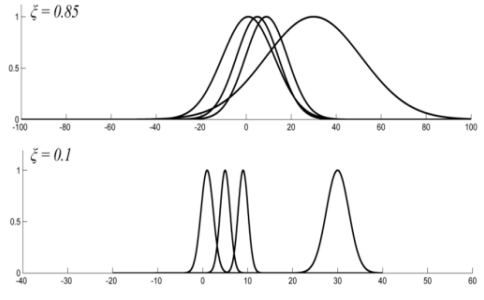


Fig. (2) The impact of ξ on generated intervals

In this work, a fixed and predefined value for ξ is used, but if its value changes dynamically, it will result in a more efficient algorithm.

5. Experimental setup and results

First the proposed algorithm was evaluated in a synthetic database with a size of 1,000 records formed by 4 numeric attributes like previous works in [13] and [25]. All of the domains of values were set to [0, 100]. The values were uniformly distributed in attributes in such a way that they were grouped in pre-determined sets as shown in Table 1.These sets are completely separated from each other. This distribution of the values was completely arbitrary. Some intervals had small size and others have larger size. Support and confidence values for these sets were 25 and 100%, respectively. Other values outside these sets were distributed in such a way that no other better rules than these rules exist.

Table 1: Predetermined sets

sets
$A_1 \in [1-10] \wedge A_2 \in [15-30]$
$A_1 \in [15-45] \wedge A_3 \in [60-75]$
$A_2 \in [65-90] \wedge A_4 \in [15-45]$
$A_3 \in [80-100] \wedge A_4 \in [80-100]$

The aim of the proposed algorithm was to optimize the objective function and mining the numeric association rules which result in optimum values for the objective function. This test was performed in order to show the accuracy of the association rules found by the proposed algorithm.

In Table 2 the association rules found by the proposed algorithm, are shown. The resulted rules have high support and confidence and fall into the pre-determined sets. The algorithm is database-independent, since it does not rely upon support/confidence thresholds which are hard to choose for each database. No other rules with higher support and confidence can be found in case of specifying the minimum support and minimum confidence thresholds. Also, the intervals of the attributes are open and have a proper length because of using the Gaussian functions for building the intervals. The proposed algorithm is able to automatically find all the rules without relying upon the minimum support and the minimum confidence thresholds and the intervals of the numeric attributes are exact and accurate.

Table 2: The association rules found by the proposed algorithm

Association Rules	Support (%)	Confidence (%)
$A_3 \in$(79/44-100/02) $\rightarrow A_4 \in$(79/78-100/17)	25	100
$A_2 \in$(14/92-30/13) $\rightarrow A_1 \in$(0/67-10/31)	25	100
$A_1 \in$(0/95-10/08) $\rightarrow A_2 \in$(14/34-30/98)	25	100
$A_1 \in$(14/94-45/01) $\rightarrow A_3 \in$(59/91-75/02)	25	100
$A_4 \in$(79/45-100/48) $\rightarrow A_3 \in$(79/21-100/05)	25	100
$A_3 \in$(59/66-75/65) $\rightarrow A_1 \in$(14/77-45/40)	25	100
$A_2 \in$(64/12-90/62) $\rightarrow A_4 \in$(14/16-45/39)	25	100
$A_4 \in$(14/40-45/41) $\rightarrow A_2 \in$(64/28-90/12)	25	100

To test the efficiency of the proposed algorithm, it has been executed on noisy synthetic database like the work in [13] and [25].The noise is applied in the way that a percentage p of records exists that is not fulfilled in the pre established interval of the second item. For example, for the first set there is a percentage p of records that are distributed in the ranges [0–14] or [31–100]. This test was performed with three levels of noise to show the ability of finding correct intervals even when they contain a percentage of noise. The experimental results have been shown in Table 3. It can be seen that the algorithm is able to find the ranges of the intervals which are almost exactly adjusted to the pre-determined ones. This shows that the proposed algorithm is able to overcome certain levels of noise among the tested data.

Table 3: The results of mining the synthetic data with noise by the proposed algorithm

Noise level (%)	Association rules	Support (%)	Confidence (%)
4	$A_1 \in$(0.60-10.75) $\rightarrow A_2 \in$(14.87-30.31)	24	97
	$A_2 \in$(14.56-30.55) $\rightarrow A_1 \in$(0.22-10.38)	24	97
	$A_1 \in$(14.99-45.24) $\rightarrow A_3 \in$(59.01-75.56)	24	97
	$A_3 \in$(59.93-75.28) $\rightarrow A_1 \in$(14.13-45.23)	24	97
	$A_2 \in$(64.95-90.76) $\rightarrow A_4 \in$(14.34-45.72)	24	97
	$A_4 \in$(45.57-14.86) $\rightarrow A_2 \in$(64.46-90.51)	24	97
	$A_3 \in$(79.96-100.99) $\rightarrow A_4 \in$(79.10-100.91)	24	97
	$A_4 \in$(79.70-100.57) $\rightarrow A_3 \in$(79.57-101.44)	24	97
6	$A_1 \in$(0.87-10.42) $\rightarrow A_2 \in$(14.81-30.36)	23.5	96
	$A_2 \in$(14.66-30.40) $\rightarrow A_1 \in$(0.15-11.05)	23.5	96
	$A_1 \in$(15.48-47.35) $\rightarrow A_3 \in$(59.42-75.27)	23.5	95
	$A_3 \in$(61.93-75.93) $\rightarrow A_1 \in$(14.95-47.59)	21	96
	$A_2 \in$(64.98-86.41) $\rightarrow A_4 \in$(14.91-45.17)	21	97
	$A_4 \in$(14.38-45.00) $\rightarrow A_2 \in$(64.36-90.48)	23.5	95
	$A_3 \in$(79.91-100.22) $\rightarrow A_4 \in$(79.40-100.84)	23.5	96
	$A_4 \in$(79.44-100.40) $\rightarrow A_3 \in$(79.55-100.91)	23.5	96
8	$A_1 \in$(1.97-10.03) $\rightarrow A_2 \in$(14.19-31.27)	21	95
	$A_2 \in$(14.91-31.27) $\rightarrow A_1 \in$(1.69-10.02)	21	95
	$A_1 \in$(14.60-45.50) $\rightarrow A_3 \in$(59.70-76.10)	23	95
	$A_3 \in$(60.92-76.14) $\rightarrow A_1 \in$(14.60-45.28)	21	95
	$A_2 \in$(64.97-90.85) $\rightarrow A_4 \in$(14.16-45.02)	23	95
	$A_4 \in$(16.18-45.09) $\rightarrow A_2 \in$(64.28-90.50)	21	94
	$A_3 \in$(79.61-100.04) $\rightarrow A_4 \in$(78.78-100.20)	23	94
	$A_4 \in$(79.88-100.88) $\rightarrow A_3 \in$(79.68-100.71)	23	95

Multi objective association rules mining using ACO for continuous domains algorithm was also evaluated in five public domain databases: Basketball, Bolts, Pollution, Quake, and Sleep in order to compare with related works. These databases are available from Bilkent University

Function Approximation Repository [29]. The proposed algorithm is stochastic so, it has fluctuations in different runs. In order to get a better result, the user needs to execute several trials of the algorithm to get the result with the best solutions.

Most of the numeric association rule mining algorithms need to build the intervals of numeric attributes before mining process. They discretize numeric attributes manually and then start mining association rules from them. So the proposed algorithm is compared to five evolutionary computation-based algorithms in the literature that discretize numerical attributes and search for association rules simultaneously.

Table 4 shows the number of records and the number of numeric attributes for each database. Table 5 shows the mean number of different rules and the mean of confidence value, which can be considered as strength of the rule, found by previous works and the proposed algorithm. The experimental comparison in terms of number of rules and confidence values has been performed because the algorithms RPSOA [13], the Genetic Association Rule Mining algorithm [24] and SA [25] find directly numeric association rules without finding frequent itemsets and search for numeric intervals while mining association rules in one step. For the algorithm proposed in [24], the population size was set to 100 and it has been modified to find only positive ARs. It can be shown that, number of rules found by the proposed algorithm is more than that reported by [13], [24] and [25], in all databases. The most effective aspect of the proposed algorithm is the way it builds numeric intervals and that's why it is able to find more different rules. The results obtained in these domains seem to indicate that the proposed algorithm is competitive with the other algorithms in terms of confidence values and number of rules.

Table 4: The number of records and numeric attributes in each database

Sleep		Quake	
8	Number of attributes	4	Number of attributes
57	Number of records	2178	Number of records
%2	Missing values	0	Missing values
Pollution		**Bolts**	
16	Number of attributes	8	Number of attributes
60	Number of records	40	Number of records
0	Missing values	0	Missing values
Basketball			
5	Number of attributes		
96	Number of records		
0	Missing values		

Table 5: The mean number of different rules and confidence value compared with other works

Database	Number of rules			
	ACO_R	SA	RPSOA	GA
Basketball	37	12	33.8	33.8
Bolts	42	3.1	39.0	39.0
Pollution	49	3.5	41.2	41.2
Quake	58	7.5	43.8	43.8
Sleep	41	3.1	32.8	32.8
Database	**Confidence (%)**			
	ACO_R	SA	RPSOA	GA
Basketball	78	93	60±2.8	60±1.2
Bolts	89	80	60±2.0	65±1.9
Pollution	78	75	66±4.7	68±4.8
Quake	85	73	63±2.8	62±5.1
Sleep	84	78	64±2.8	64±2.3

Table 6 shows the comparison of obtained results from the proposed algorithm, RPSOA [13], the Genetic Association Rule Mining algorithm proposed in [24], the work proposed in [19] and SA [25] in terms of support which refers to the usefulness of the rule, and size which refers to number of attributes in the rule. The GAR algorithm uses an EA for mining only frequent itemsets. That is why; comparisons about the values according to the rules cannot be made [13]. The value of the column "Support (%)" indicates the mean of support, while the value of the column "Size" shows the mean number of attributes contained in the rules.

Table 6: The comparison of obtained results from the proposed algorithm in terms of support and size

Database	Support (%)				
	ACO_R	SA	RPSOA	GA	GAR
Basketball	45	42	36.44	32.21	36.69
Bolts	48	41	28.48	27.04	25.97
Pollution	54	41	43.85	38.95	46.55
Quake	60	45	38.74	36.96	38.65
Sleep	53	46	36.52	37.25	35.91
Database	**Size**				
	ACO_R	SA	RPSOA	GA	GAR
Basketball	2.7	2.58	3.21	3.21	3.38
Bolts	3.88	2.71	5.14	5.14	5.29
Pollution	6.71	2.3	6.46	6.21	7.32
Quake	2.32	2.02	2.22	2.10	2.33
Sleep	4.1	2.01	4.19	4.19	4.21

Multi objective association rule mining via ACO_R algorithm has found rules with high values of support in all databases. The size values obtained from this algorithm are smaller than the values obtained from the

GAR in five out of five databases and they are smaller than the values obtained from the GA algorithm proposed in [19] in three out of five databases. The size of the intervals is based on the attributes' standard deviations and is so flexible and controllable. There is no need to refine their size by decreasing the length of them.

Another point is that, by tuning up the parameters of the algorithm many different rules can be discovered which have lower support and confidence or have 100% confidence and lower support. Using a method which can tune the parameter ξ will result in even more different rules.

6. Conclusion

Ant colony optimization for continuous domain (ACO_R) is a new metaheuristic approach. This study proposed a new algorithm that uses the notion of Gaussian functions and the modification of the solution archive of ACO_R in order to build numeric intervals and search for association rules in one step without specifying minimum support and minimum confidence.
The lower bound and upper bound values are introduced by adding and subtracting a coefficient of a partial solution's standard deviation to/from its value. The proposed algorithm has been used in data mining within databases that can take numeric attributes and has given satisfactory results in its first applications. This algorithm seems to provide useful extensions for practical applications specially while using feature selection before applying the algorithm, since ant algorithms cannot distinguish variable correlations. But still despite of this fact, the ACO_R algorithm is able to find accurate and exact rules with reasonable interval lengths for numeric attributes because of the way of building intervals using Gaussian functions. Proposing an efficient way of tuning the algorithm's parameters which would result in better solutions and another ant algorithm which can handle both numeric and categorical attributes may be presented as further works.

References

[1] Chen C.-H., Hong T.-P., and Tseng V.S., *A Cluster-Based Fuzzy-Genetic Mining Approach for Association Rules and Membership Functions*, IEEE International Conference on Fuzzy Systems, pp. 1411 - 1416, 2006.

[2] Kayaa M., Alhajj R., *Genetic algorithm based framework for mining fuzzy association rules*, Fuzzy Sets and Systems , Vol. 152, No. 3, pp. 587-601, 2005.

[3] Tsay Y. J., Chiang J. Y., *CBAR: an efficient method for mining association rules*, Knowledge] Based Systems, Vol. 18, No. 2-3, pp. 99-105, 2005.

[4] Agrawal R., Imielinski T., and Swami, A., *Mining association rules between sets of items in large databases,* In proceedings of ACM SIGMOD conference on management of data, pp. 207-206, 1993.

[5] Agrawal R., Sirkant R., *Fast algorithms for mining association rules,* In proceedings of the 20th international conference on very large databases, Santiago, Chile, 1994.

[6] Houtsma A., Swami M., *Set-oriented mining of association rules,* Research Report, 1993.

[7] Lin D. I., Kedem Z. M., *Pincer-search: An efficient algorithm for discovering the maximal frequent set,* In Proceedings of sixth European conference on extending database technology, 1998.

[8] Qodmanan H. R., Nasiri M., and Minaei-Bidgoli B., *Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence*, Expert Systems with applications, Vol. 38, No. 1, pp. 288-298, 2010.

[9] Martens, D., Baesens, B., and Fawcett, T., *Editorial Survey: Swarm Intelligence for Data Mining,* Machine Learning, Vol. 82, No. 1, pp. 1-42, 2010.

[10] Socha K., *ACO for Continuous and Mixed-Variable Optimization,* Ant Colony Optimization and Swarm Intelligence, Computer Science, Vol. 3172, pp. 53-61, 2004.

[11] Socha K., *Ant Colony Optimization For Continuous and Mixed-Variable Domains*, Ph.D. Thesis, Universit´e Libre de Bruxelles, Brussels, Belgium, 2008.

[12] Rungswang A., Tangpong P., Laohawee T., and Khampachua T., *Novel Query Expansion Technique using Apriori Algorithm.* In proceedings of TREC' 1999.

[13] Alatas B., Erhan A., *Rough particle swarm optimization and its applications in data mining,* Soft Computing – A Fusion of Foundations, Methodologies and Applications, Vol.12, No. 12, pp. 1205-1218, 2008.

[14] Tan P.-N., Steinbach M., and Kumar V., *Introduction to Data Mining*, Pearson International Edition Pearson Addison Wesley, 2006.

[15] Srikant R., Agrawal R., *Mining quantitative association rules in large relational tables*, In: Proceedings of ACM SIGMOD international conference on Management of data, Vol. 25, No. 2, pp. 1-12, 1996.

[16]] Fukuda T., Yasuhiko M., Sinichi M., Tokuyama T. Mining *optimized association rules for numeric attributes*. In: Proceedings of ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems. New York, pp. 182–191, 1996.

[17] Miller R.J., Yang Y., *Association rules over interval data*, In: Proceedings of ACM SIGMOD international conference on management of data, vol. 29, No. 2, pp. 452–461, 1997.

[18] Lent B., Swami A., and Widom J., *Clustering association rules*, In: Proceedings of IEEE international conference on data engineering, pp. 220-231, 2002.

[19] Mata J, Alvarez JL, Riquelme JC, *Discovering numeric association rules via evolutionary algorithm.* In: Sixth Pacific–Asia conference on knowledge discovery and data mining PAKDD-02 (LNAI), Taiwan 2336, pp 40–51, 2002.

[20] Aumann Y., Lindell Y., *A statistical theory for quantitative association rules*, Journal of Intelligent Information Systems, Vol. 20, No. 3, pp.255–283, 2003.

[21] Ke K., Cheng J., and Ng W., *An information-theoretic approach to quantitative association rule mining*, Journal Knowledge and Information Systems, Vol. 16, No. 2, pp. 112-114, 2008.

[22] David L. O., Yanhong L., *Mining Fuzzy Weighted Association Rules,* In: 40th Annual Hawaii international conference on system, sciences HICSS, pp 53–62, 2007.

[23] Alatas B., Akin E., Karci A., *MODENAR: multi-objective differential evolution algorithm for mining numeric association rules*, Appl Soft Comput. doi:10.1016/j.asoc.2007.05.003, 2007

[24] Alatas B, Akin E, *An efficient genetic algorithm for automated mining of both positive and negative quantitative association rules*. Soft Comput 10(3):230–237, 2006.

[25] Nasiri M., Taghavi L., Minaei Bidgoli B., *Multi-Objective Rule Mining using Simulated Annealing Algorithm*, Journal of Convergence Information Technology, Vol. 5, No. 1, 2010.

[26] Parpinelli R.S., Lopes H.S., and Freitas A.A., *Data mining with an ant colony optimization algorithm,* Evolutionary Computation, IEEE Transactions, pp. 321 – 332, Vol. 6, No. 4, 2002.

[27] Kuo R J., Shih C.W., *Association rule mining through the ant colony system for national health insurance research database in Taiwan.*. In Journal of Computers and Mathematics with Applications, pp. 1303-1318, 2007.

[28] Atabaki G., Kangavari M., *Mining association rules in Distributed Environment through Ant Colony Optimization Algorithm,* M.Sc thesis (in Persian), Iran University of Science and Technology, 2009.

[29] Guvenir H.A., Uysal I., Bilkent University Function Approximation Repository. http://funapp.cs.bilkent.edu.tr, 2000.

A New Color Feature Extraction Method Based on Dynamic Color Distribution Entropy of Neighborhoods

Fatemeh Alamdar[1] and MohammadReza Keyvanpour[2]

[1] Department of Computer Engineering, Alzahra University
Tehran, Iran

[2] Department of Computer Engineering, Alzahra University
Tehran, Iran

Abstract

One of the important requirements in image retrieval, indexing, classification, clustering and etc. is extracting efficient features from images. The color feature is one of the most widely used visual features. Use of color histogram is the most common way for representing color feature. One of disadvantage of the color histogram is that it does not take the color spatial distribution into consideration. In this paper dynamic color distribution entropy of neighborhoods method based on color distribution entropy is presented, which effectively describes the spatial information of colors. The image retrieval results in compare to improved color distribution entropy show the acceptable efficiency of this approach.

Keywords: *color feature, color histogram, annular color histogram, color distribution entropy, dynamic color distribution entropy of neighborhoods, image retrieval*.

1. Introduction

In recent decades, digital technology progress results in unprecedented growth in production of digital images. Therefore development of effective automatic techniques for image sets organization and management is required so that one can search, retrieval and categorize the images more convenient. Feature extraction is the basis of these automatic techniques. Color, texture and shape are the most common visually features[1]. These features are independent of specific domain and can used in general systems of retrieval images[2]. The color feature is the first and one of the most widely used visual features in image retrieval and indexing[3]. The most important advantages of color feature are power of representing visual content of images, simple extracting color information of images and high efficiency, relatively power in separating images from each other[4], relatively robust to background complication and independent of image size and orientation[5,6,2].

The color histogram method introduced in [7] has shown to be very effective and simple to implement. Use of color histogram is the most common way for representing color feature[8]. Despite of some drawbacks, color histogram had been used in many researches and great efforts were done for overcoming its weakness [9-14]. One of disadvantage of the color histogram method is that it is not robust to significant appearance changes because it does not include any spatial information[10]. Several schemes including spatial information have been proposed. Pass et al. [15] suggested classifying each pixel as coherent or no coherent based on whether the pixel and its neighbors have similar colors. Then, a split histogram called color coherence vector (CCV) is used to represent this classification for each color in an image. Huang[11] proposed a color correlograms method, which collects statistics of the co-occurrence of two colors some distances apart. A simplification of this feature is the autocorrelogram, which only captures the spatial correlation between identical colors. In [11-14] respectively introduced annular color histogram, spatial-chromatic histogram (SCH) and geostat to describe how pixels of identical color are distributed in the image. Sun et al. [10] propose a color distribution entropy (CDE) method, which takes account of the correlation of the color spatial distribution in an image. This feature is based on annular color histogram that draws some concentric circles from images and then the annular color histogram is calculated by counting the pixels of every color bin inside every circle. The number of circles is a predefined constant and for every image is the same regardless of its content.

In this paper we introduce a dynamic color distribution entropy of neighborhoods (D-CDEN) method, which similar to CDE describes the spatial information of an image. Instead of drawing concentric circles, D-CDEN takes account of images' content by attending to neighborhoods of pixels for every color bin. The number of extracted neighborhoods is different for every color

bins. In addition, predefining this number does not require in this approach, also the color indexing results in image clustering and retrieval is much better. The results are demonstrated by image retrieval and D-CDEN in compare to I-CDE show the efficiency of this approach.

The rest of the paper is organized as follows. A briefly review on CDE and I-CDE is presented in Section 2. Section 3 describes the proposed feature extraction base on D-CDEN. Section 4 details the similarity measurement. Experimental results are demonstrated in Section 5. Finally, a conclusion is given in Section 6.

2. Color Distribution Entropy

CDE descriptor was proposed in [10]. This descriptor expresses the color spatial information of an image. This descriptor based on the NSDH (Normalized Spatial Distribution Histogram) and information entropy was defined. NSDH was derived from Annular Color Histogram[12]. In Annular Color Histogram which introduced by Rao et al., suppose $\boldsymbol{A}_i$ be the set of pixels with color bin i of an image and $|\boldsymbol{A}_i|$ be the number of elements in $\boldsymbol{A}_i$. Let $\boldsymbol{C}_i$ be the centroid and r_i be the radius of color bin i which are defined in [12]. With $\boldsymbol{C}_i$ as the center and with jr_i / N as the radius for each $1 \le j \le N$, N concentric circles can be drawn. Let $|\boldsymbol{A}_{ij}|$ be the count of the pixels of color bin i inside circle j. Then the annular color histogram can be written as $(|\boldsymbol{A}_{i1}|,|\boldsymbol{A}_{i2}|,...,|\boldsymbol{A}_{iN}|)$. This is illustrated in Fig. 1. Based on the Annular Color Histogram, the NSDH is given in Eq. (1).

$$\begin{aligned} \boldsymbol{P}_i &= (P_{i1}, P_{i2}, ..., P_{iN}) \\ P_{ij} &= |\boldsymbol{A}_{ij}| / |\boldsymbol{A}_i| \end{aligned} \quad (1)$$

The E_i defined as CDE of color bin i, was defined as

$$E_i(\boldsymbol{P}_i) = \sum_{j=1}^{N} P_{ij} \log_2(P_{ij}) \quad (2)$$

where $\boldsymbol{P}_i$ is the Normalized Spatial Distribution Histogram.

This equation shows the dispersive degree of the pixel patches of a color bin in an image. Large E_i means the distribution of the pixels is dispersed, otherwise the distribution is compact. Then the CDE index for an image can be written as $(h_1, E_1, ...h_i, E_i, ..., h_n, E_n)$, where h_i is the histogram of color bin i, E_i is the CDE of color bin i and n is the number of bins.

The improved CDE (I-CDE) was defined as

$$E_i(\boldsymbol{P}_i) = -g(\boldsymbol{P}_i)\sum_{j=1}^{N} f(j) P_{ij} \log_2(P_{ij}) \quad (3)$$

$$f(j) = 1 + \frac{j}{N} \quad (4)$$

$$g(\boldsymbol{P}_i) = 1 + \frac{A(\boldsymbol{P}_i)}{N} \quad (5)$$

$$A(\boldsymbol{P}_i) = \sum_{j=1}^{N} (P_{ij} \times j) \quad (6)$$

$f(i)$ is the weight function which denotes the different contribution of each annular circle to the CDE. $g(\boldsymbol{P}_i)$ is the weight function using Histogram Area($A(\boldsymbol{P}_i)$) defined as Eq. (6). $g(\boldsymbol{P}_i)$ effectively removes the influence of symmetrical property of entropy. More details could be found in [10].

The CDE and I-CDE similarity measurement of image I_q and I_t was defined as[10]

$$d(I_q, I_t) = \sum_{i=1}^{n} \min(h_i^{I_q}, h_i^{I_t}) \times \frac{\min(E_i^{I_q}, E_i^{I_t})}{\max(E_i^{I_q}, E_i^{I_t})} \quad (7)$$

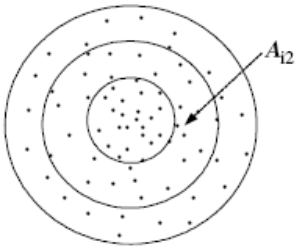


Fig 1: Annular Color Histogram[10]

2.1 Footnotes

Footnotes should be typed in singled-line spacing at the bottom of the page and column where it is cited. Footnotes should be rare.

3. Feature extraction based on Dynamic Color Distribution Entropy of Neighborhoods

D-CDEN method is based on CDE and effectively describes the spatial information of colors. In CDE, two images are considered similar when distributions of the pixels of color bins are the same but layout and neighborhoods of color pixels can be not the same, so this distribution may be similar in different images. D-CDEN method takes account of images' content and instead of drawing N concentric circles in CDE attends to

neighborhoods of pixels for every color bin of image color histogram.

3.1 Neighborhoods extraction

For extracting neighborhoods for every color bin i, an image matrix is scanned rows by rows from left to right, up to down. Because of this kind of scanning, only neighborhoods of right and up adjacent pixels of current pixel had been identified and regarded. If none of them is in the same color bin, this pixel is in the new neighborhood; but in the other cases, if the current pixel is in the same color bin of the right, 135° diagonal or up adjacent pixels, it is assigned to its neighborhood. It is illustrated in Fig. 2. If the middle pixel is the current pixel, the 1-8 pixels are its neighbors (Fig. 2(a)). Because this pixel and pixel 2 are in the same color bin, it is in the neighborhoods of pixel 2. In Fig. 2(b) the neighborhoods which were detected up to current pixels are determined by different numbers.

A problem may be appear in the neighborhoods' specifying when the current pixel is in the same color bin of both right and up adjacent pixels but their neighborhoods are different. In this case these two neighborhoods are merged. Fig. 3 shows this problem. Pixel 0 is the current pixel and is not in the same color bin of 1,2,3 pixels (Fig. 3(a)) and Fig. 3(b) shows detected neighborhoods up to now, so a new neighborhood is defined for current pixel. In continues of scan, when the last pixel is the current pixel, this problem occur (Fig. 3(c)). Fig. 3(d) shows the final detected neighborhoods when neighborhood 1, 7 had been merged.

2	3	4
1	0	5
8	7	6

(a)

1	2	2
3	1	4
3	5	5

(b)

Fig. 2: (a) neighbor pixels of pixel 0. (b) extracted neighborhoods for image (a)

2	3	4		
1	0	5		
8	7	6		

(a)

1	1	1	2	3
4	5	5	1	3
6				

(b)

1	1	1	2	3
4	5	5	1	3
6	7	8	8	1
6	9	7	7	

(c)

1	1	1	2	3
4	5	5	1	3
6	1	8	8	1
6	9	1	1	1

(d)

Fig. 3: (a) neighbor pixels of pixel 0. (b,c,d) extracted neighborhoods for image (a)

Extracted neighborhoods for a real image are shown in Fig. 4(b). In this figure, neighborhoods of a color bin have the same color.

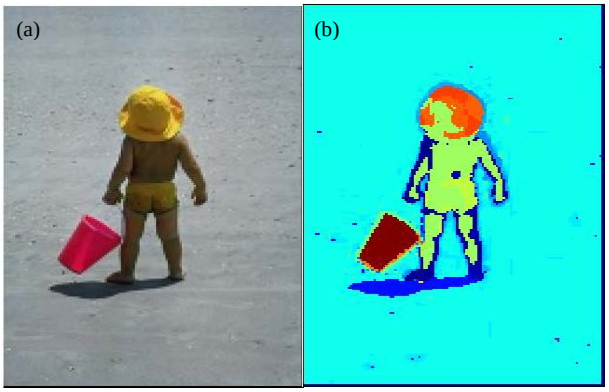


Fig. 4: (a) original image. (b) neighborhoods of image (a).

3.2 Dynamic Color Distribution Entropy of Neighborhoods

Like CDE, Normalized Spatial Distribution Histogram is defined as:

$$\boldsymbol{P}'_i = (P'_{i1}, P'_{i2}, ..., P'_{inb_i})$$
$$\text{and } P'_{ij} = |\boldsymbol{A}'_{ij}| / |\boldsymbol{A}'_i|$$
$$\text{for } 1 \le j \le nb_i \qquad (8)$$

nb_i is the number of extracted neighborhoods for color bin i and is different for every color bins and $\boldsymbol{A}'_i$ is the set of pixels with color bin i of an image and $|\boldsymbol{A}'_{ij}|$ is the count of the pixels of neighborhood j for color bin i.

The D-CDEN is defined as Eq. (2) by replacing N with nb_i and $\boldsymbol{P}_i$ with $\boldsymbol{P}'_i$, so the D-CDEN index for an image is written as $(h_1, E'_1, ...h_i, E'_i, ..., h_n, E'_n)$, where E'_i is the D-CDEN of color bin i.

3.3 Feature extraction

Firstly, before extraction of neighborhoods, the images are resized into 128x128 pixels, because it makes noises removed and small neighborhoods reduced especially in cluttered scene images. Then for every image, neighborhoods are extracted and nb_i is calculated according the previous sub-sections. Afterwards the images are indexed by D-CDEN descriptor and in HSV color space. The color space is uniformly quantized into 8 levels of hue, 2 levels of saturation and value giving a total of 32 bins.

4. Similarity measurement

In this case, we introduce a dissimilarity measurement based on vector space retrieval model (VSM), which was used in [16].

In [16], two problems of CDE similarity measurement have been mentioned. First problem is that two color bins are similar or two color bins are not similar but just have the same histogram, so different images with the same histogram are considered similar. Another problem is that the same $d(I_q, I_t)$ can be produced by a number of different sets. In order to overcome these problems, the similarity measurement was done by using vector space retrieval model (VSM). VSM measure is as follows[17]:

$$\cos\theta_H(I_q, I_t) = \frac{\sum_{i=1}^{n} h_i^{I_q} \times h_i^{I_t}}{\sqrt{\sum_{i=1}^{n} h_i^{I_q\,2}} \times \sqrt{\sum_{i=1}^{n} h_i^{I_t\,2}}} \quad (9)$$

$$\cos\theta_E(I_q, I_t) = \frac{\sum_{i=1}^{n} E_i^{I_q} \times E_i^{I_t}}{\sqrt{\sum_{i=1}^{n} E_i^{I_q\,2}} \times \sqrt{\sum_{i=1}^{n} E_i^{I_t\,2}}} \quad (10)$$

$$d'(I_q, I_t) = 2 - (\cos\theta_H(I_q, I_t) + \cos\theta_E(I_q, I_t)) \quad (11)$$

where $\cos\theta_H(I_q, I_t)$ and $\cos\theta_E(I_q, I_t)$ are the color histogram similarity and color distribution entropy similarity between two images I_q and I_t in vector space retrieval model, and $d'(I_q, I_t)$ is the dissimilarity of two images.

For measuring the image dissimilarity, we use the following distance:

$$d''(I_q, I_t) = 3 - (\cos\theta_H(I_q, I_t) + \cos\theta_{E'}(I_q, I_t) + \cos\theta_N(I_q, I_t)) \quad (12)$$

where $\cos\theta_H(I_q, I_t)$ and $\cos\theta_E(I_q, I_t)$ are calculated by Eq. (9) and (10) by replacing E with E' and $\cos\theta_N(I_q, I_t)$ is computed as:

$$\cos\theta_N(I_q, I_t) = \frac{\sum_{i=1}^{n} Nb_i^{I_q} \times Nb_i^{I_t}}{\sqrt{\sum_{i=1}^{n} Nb_i^{I_q\,2}} \times \sqrt{\sum_{i=1}^{n} Nb_i^{I_t\,2}}} \quad (13)$$

where Nb_i is the normalized number of neighborhoods for color bin i for an image, which is defined as:

$$Nb_i = \frac{nb_i}{\sum_{j=1}^{n} nb_j} \quad (14)$$

5. Experimental Results

In this section, the results of D-CDEN are demonstrated by image retrieval and these results are compared with I-CDE. Experiments were carried out by using two databases. The first is SIMPLIcity[1][18] database of 1000 images which included 10 categories of Africa people, Beach, Buildings, Buses, Dinosaurs, Elephants, Flowers, Horses, Mountains and Food. Every category contains 100 images. Fig. 5 shows the different types of images in this experimental database. Each of the images is with the size of 256×384 or 384×256 pixels.

Secondly we used 70 categories of Caltech101 [2] [19] database. Categories contain of 6384 images of different objects. There are about 40 to 800 images in each object category. Size of each image is roughly 200×300 or 300×200 pixels. Some of the different types of images in this database are shown in Fig. 6.

For comparing, an image query was chosen and the retrieval results of using on D-CDEN, I-CDE has been calculated and for specifying the number of circles in I-CDE, we averaged over nb_i s mean ($\frac{1}{n}\sum_{i=1}^{n} nb_i$) for all of images. The used image dissimilarity measurements for D-CDEN and I-CDE are d'' and d' respectively. The retrieval accuracy was measured in terms of the Recall, Precision. The Precision rate and Recall rate are defined as follows:

[1] - available at http://wang.ist.psu.edu/~jwang/test1.tar
[2] - available at http://www.vision.caltech.edu/Image_Datasets/Caltech101/101_ObjectCategories.tar.gz

Fig. 5: Different types of image in SIMPLIcity database

Fig. 6: Different types of image in Caltech101 database

$$\text{Precision} = \frac{r}{Nr}$$
$$\text{Recall} = \frac{r}{Ni} \qquad (15)$$

where r is the number of relevant images selected, Nr is the total number of retrieved images and Ni is the total number of similar images in the database. Results show the superiority of D-CDEN.

For comparing in SIMPLIcity database, every 100 of the categories was chosen as a query image. Fig. 7 is the Recall and Precision graph of results averaged over 100 images in Buildings(7(a)), Buses(7(b)), Flowers(7(c)).

In Caltech101 database, 50 images were selected randomly as query images. Fig. 8 is the Recall and Precision graph of results averaged over 50 random selected queries in 3 times.

Because of attending to neighborhoods of color pixels and the number of them, D-CDEN has better results in both databases.

In Fig. 9, the top-left one (9(a)) is the query image from SIMPLIcity database and the top ten retrieval results of query are sorted by similarity from left-to-right and top-to-down sequence by using of D-CDEN method(9(b)) and I-CDE method(9(b)). The same results for Caltech101 database are shown in Fig. 10.

6. Conclusions

In this paper, a color features extraction method based on dynamic color distribution entropy of neighborhoods was expressed. D-CDEN method measures the spatial relation of colors in an image and takes account of images' content by neighborhoods extraction of pixels for every color bin of image color histogram.

In this work we introduce a new dissimilarity measuring to demonstrating results by image retrieval and these results are compared with I-CDE. Experiments were carried out by using two databases of 1000 and 6384 images. These experiments show the acceptable efficiency of this approach.

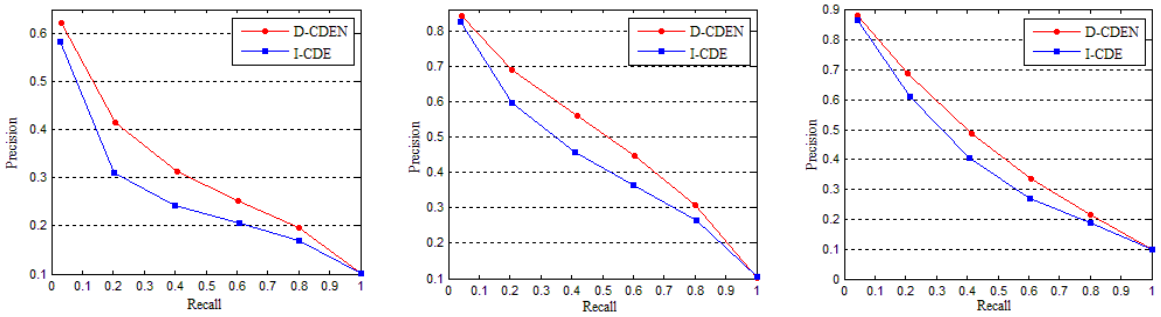


Fig. 7: Precision/Recall graph

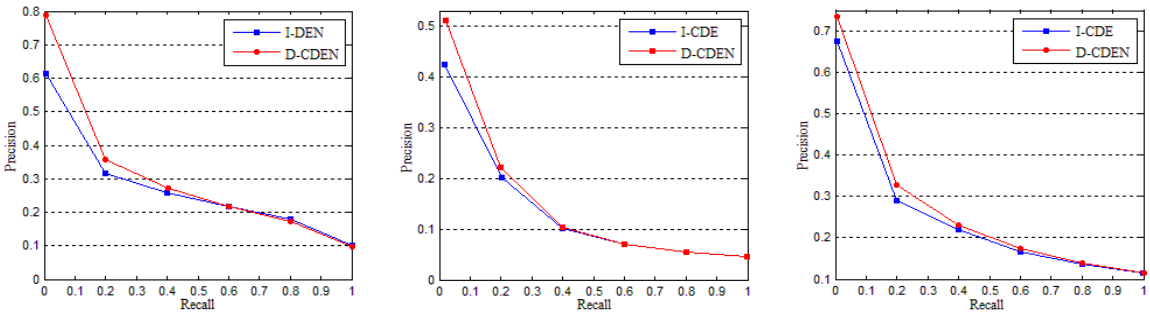


Fig. 8: Precision/Recall graph

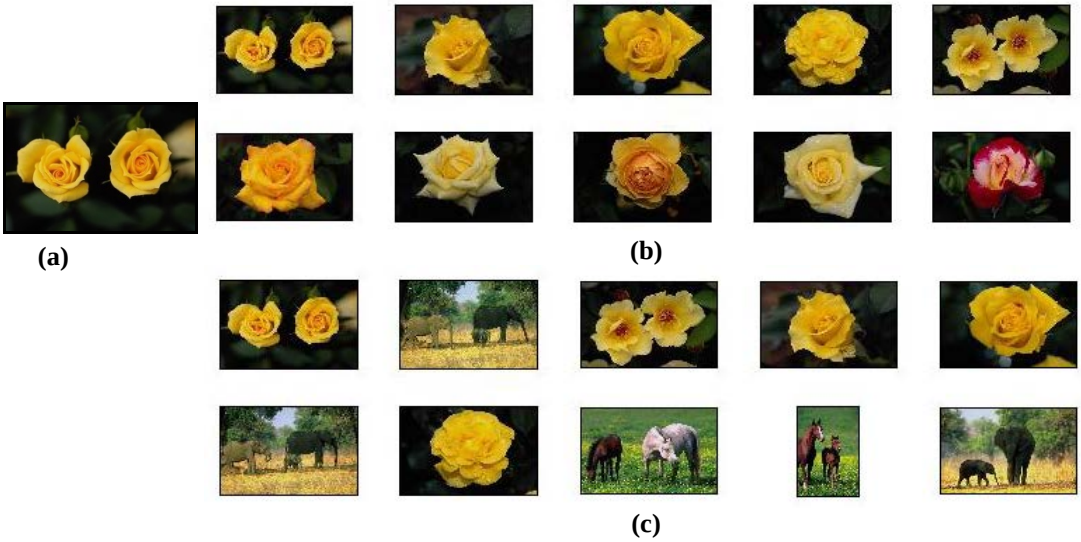


Fig. 9: (a) Query image from SIMPLIcity database. (b) Query results of D-CDEN method. (c) Query results of I-CDE method

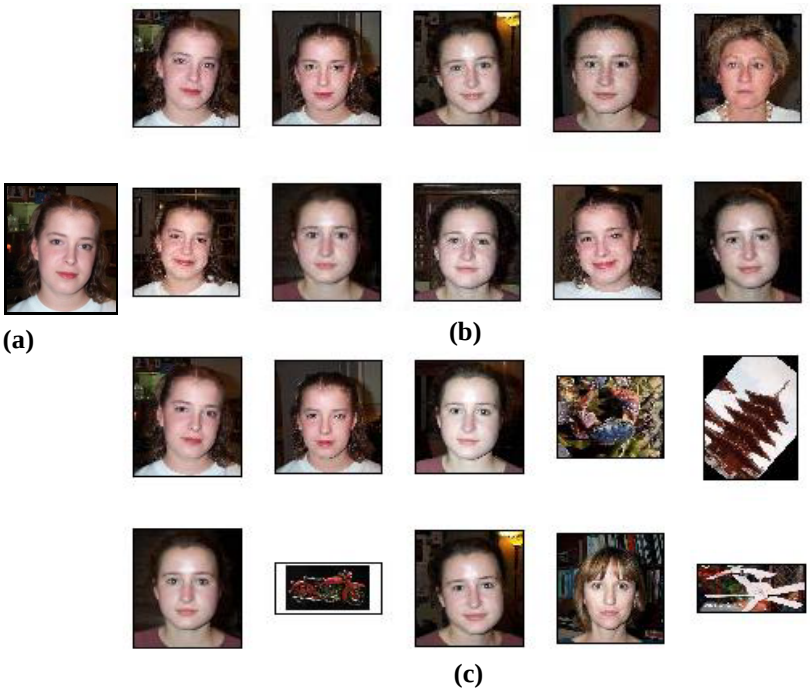


Fig. 10: (a) Query image from Caltech101 database. (b) Query results of D-CDEN method. (c) Query results of I-CDE method

Acknowledgments

This work was supported in part by the Iran Telecommunication Research Center, ITRC.

References

[1] L. Hove, "Improving Content Based Image Retrieval Systems with a Thesaurus for Shapes", M.S. Thesis, Institute for Information and Media Sciences, University of Bergen, Bergen, Norway, 2004.

[2] M. Keyvanpour, N. Moghadam Charkari, "Image Retrieval Using Hybrid Visual Features", in ICEE, Tehran, Iran, 2008, pp. 62-67.

[3] R. S. Torres, A. X. Falcao, "Content-Based Image Retrieval: Theory and Applications", Revista de Informática Teórica e Aplicada, Vol. 13, No.2, 2006, pp.161-185.

[4] S. Panchanathan, Y. Park, et al., "The Role of Color in Content-Based Image Retrieval", in ICIP'00, Canada, 2000, Vol. 1, pp. 517-520.

[5] Y. Rui and T.S. Huang, "Image retrieval: Current techniques promising directions and open issues", Journal of Visual Communication and Image Representation, Vol.10, 1999, pp. 39-62.

[6] J. Zhang, W. Hsu, M. Lee, "An Information driven Framework for Image Mining", in Proc of 12th International Conference on Database and Expert Systems Applications, Munich, Germany, 2001, pp. 232-242.

[7] M. J. Swain, D. H. Ballard, "Color indexing, International Journal of Computer Vision", Vol. 7, No. 1, 1991, pp. 11-32.

[8] L. Tran, "Efficient Image Retrieval with Statistical Color Descriptors", Ph.D. Thesis, Department of Science and Technology, Linkoping University, Linkoping, Sweden, 2003.

[9] R. Datta, D. Joshi, J. Li, and J.Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age", ACM Computing Surveys, Vol. 40, No. 2, 2008, pp. 1-60.

[10] J. Sun, X. Zhang, J. Cui and L. Zhou, "Image retrieval based on color distribution entropy", Pattern Recognition Letters, Vol. 27, No. 10, 2006, pp. 1122-1126.

[11] J. Huang, "Image indexing using color correlograms", in IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Juan., 1997, pp. 762–768.

[12] A. B. Rao, R. K. Srihari, and Z. F. Zhang, "Spatial color histogram for content-based retrieval, Tools with artificial intelligence", in Proceedings of 11th IEEE International Conference, 1999, pp. 183-186.

[13] L. Cinque, S. Levialdi, K. Olsen, et al., "Color-based image retrieval using spatial-chromatic histogram", in IEEE International Conference on Multimedia Computing and Systems, 1999, Vol. 2, pp. 969-973.

[14] S. Lim, G.J. Lu, "Spatial statistics for content based image retrieval", in International Conference on Information Technology: Computers and Communications, Las Vegas, Nevada, 2003, pp. 28-30.

[15] G. Pass, R. Zabih, J. Miller, "Comparing images using color coherence vectors", in ACM 4th International Conference on Multimedia, Boston, Massachusetts, United States, 1996, pp. 65-73.

[16] G. Lui, B. Lee, "A Color-based Clustering Approach for Web Image Search Results", in International Conference on Convergence and Hybrid Information Technology, Korea, 2009, pp. 481-484.

[17] G. Lu, multimedia database management systems, USA: Artech house publisher, 1999, pp. 81-82.

[18] J.Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: semantic sensitive integrated matching for picture libraries", IEEE Trans. Pattern Anal. Machine Intell., Vol. 23, No. 9, 2001, pp. 947-963.

[19] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories", CVPR 2004, Workshop on Generative-Model Based Vision, 2004.

A Review of Congestion Control Algorithm for Event-Driven Safety Messages in Vehicular Networks

Mohamad Yusof Darus[1] and Kamalrulnizam Abu Bakar[2]

**[1] Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, 40450 Malaysia
603-55211139**

**[2] Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia, 81310 UTM Skudai, Malaysia
607-5532382**

Abstract

Congestion control algorithm in Vehicular Networks (VANETs) has been extensively studied. However, most of congestion control algorithms are not directly applicable to uni-priority of event-driven safety messages. The event-driven safety messages have stringent requirement on delay and reliability. The uni-priority of event-driven safety messages are caused by the traffic of the same priority, typically the warning messages of safety applications from different transmitters. The uni-priority messages should be schedule before the node starts the transmitting process. In dense network, a large number of vehicles broadcast a beacon messages at a high number of frequency. Then the Control Channel (CCH) easily congested. It's very important to keep the CCH channel free from congestion to ensure timely and reliable delivery of event-driven safety messages [9, 20]. Hence, this study takes a closer look at existing congestion control algorithms to solve congestion problems because it affects the performance of safety messages. The study further exposes the weaknesses and advantages of some of these congestion control algorithms which can assist researchers to tackle the inherent problems of congestions in VANETs. This paper also concludes with a planned future research for disseminating uni-priority of event-driven safety messages while solving congestion problems.

Keywords: *VANETs, Event-Driven Safety Message, Control Channel, Uni-Priority, IEEE 802.11p*

1. Introduction

VANETs are composed of vehicles equipped with advanced wireless communication devices and self-organized networks built up from moving vehicles. The VANETs tends to operate without any infrastructure or legacy client and server communication. Each vehicle equipped with communication devices will be a node in the VANETs and allow to receive and send other messages through the wireless communication channels. This network will provide wide variety of services such as Intelligent Transportation System (ITS). The safety application is one of the most crucial application in ITS. For example, if a vehicle detects road accident, it will inform other neighboring vehicles about this road accident. The safety messages must to be delivered to each neighboring node with almost no delays.

The safety messages can be categorized into two categories; beacon and event-driven messages. The vehicles are supposed to issue beacon messages periodically to announce other vehicles about their situations such as speed, positioning and direction. These periodic messages are used by neighboring vehicles to become aware of their surrounding and to avoid potential dangers [6, 20]. The event-driven safety messages are generated when an abnormal condition or an imminent danger is detected, and disseminated within a certain area with high priority [20]. The event-driven safety messages should be delivered to neighboring node with high reliability and limit time. A single delayed or lost message could result in loss of life [8, 20].

The Federal Communication Commission (FCC) allocated a frequency spectrum for VANETs wireless communication. The Commission then established Dedicated Short Range Communications (DSRC) Service in 2003. The DSRC is a communication service that uses

the 5.850-5.925 GHz band for the use of public safety and private applications [5].

In order to provide DSRC for VANETs communication the IEEE is currently working on the IEEE 802.11p or Wireless Access in Vehicular Environments (WAVE) standard [14]. The original 802.11 protocols are not suitable for VANETs because of high vehicular mobility, faster topological changes, requirements of high reliability and low latency for safety applications. The DSRC was designed into multi-channel system. The FCC divided the spectrum into seven each with 10 to 20 MHz channels which six were identified as Service Channels (SCH), and one was identified as the Control Channel (CCH), as shown in fig. 1. The CCH channel is used for safety messages, but non-safety services and WAVE-mode short messages are expected to be provided in the six service channels [10, 13].

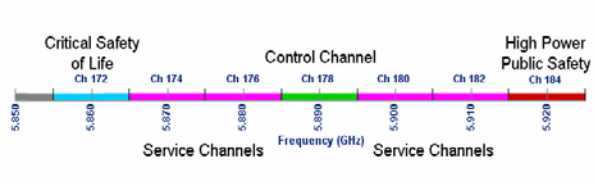


Fig. 1 The Seven Channels of DSRC

In dense network, a large number of vehicles broadcast beacon messages at a high frequency, the CCH channel will easily congested. The periodic messages are broadcast may lead to broadcast storm/blind flooding problem in VANETs [6, 16, 18]. It is very important to keep the CCH channel free from congestion in order to ensure timely and reliable delivery of event-driven safety messages [19, 20]. In order to avoid congestion of CCH channel and delay of event-driven safety message, a reliable and efficient congestion control algorithm is needed [9, 14, 20].

The purpose of this study is to reveal the strong and weak points of some of these congestion control algorithm so that researchers can come up with broader algorithm to tackle the inherent problems of congestions in VAENTs. This study focused on uni-priority of event-driven safety message congestion. The uni-priority message is caused by the traffic of the same priority, typically the warning messages of safety applications from different transmitters. According to [19], if there are many nodes with the same priority to transmit, the collisions may occur. Furthermore, in real life various reactions from drivers will happen, its will generate multiple event-driven safety messages.

The rest of this paper is structured as follows. The rest part of the paper is organized as follows: In section 2, network congestions in wireless ad hoc networks is shortly reviewed. Section 3 discusses various congestion control algorithms and some implementation strategies in VANETs. Section 4 concludes the paper with outlooks on the future work.

2. Wireless Ad Hoc Network Congestion

Wireless ad hoc network is a decentralized wireless network. The wireless ad hoc network does not rely on a pre-existing infrastructure, such as routers in wired networks or access points in managed wireless networks. Instead, each node participates in routing by forwarding data for other nodes, and so the determination of which nodes forward data is made dynamically based on the network connectivity. Every node in wireless ad hoc network can become aware of the presence of other nodes within its range. The wireless ad hoc networks can be further classified by their application such as Mobile Ad Hoc Networks (MANET), Wireless Mesh Networks (WMN), Wireless Sensor Networks (WSN) and Vehicular Networks (VANETs).

Wireless ad hoc is prone to network congestion due to the mobility of nodes, synchronization difficulties in self-coordination, and the limited capacity of the wireless channels [22, 3]. Therefore, node in wireless ad hoc may experience low throughput and long latency under the circumstance of network congestion.

One of the important aspects in wireless ad hoc networks is to maintain the efficiency network operation while preventing degradation of wireless channels communication [3, 24]. They were proposed the congestion control algorithm as solution. The major goal of congestion control mechanism is simply to use the network as efficiently as possible by attaining the highest possible throughput while maintaining a low loss ratio and small delay [17].

3. Congestion Control Algorithms in VANETs

In VANETs, many of studies focused to control the load of wireless channels over congestion control algorithms such as [7, 9, 11, 14]. The main purpose of congestion control algorithm is to control the load of traffic conditions and the performance of wireless communication channels will be increased.

Research in [11] was proposed congestion control algorithm focused on comfort applications such as browsing Internet. However, our study concern about

dissemination of the safety messages especially event-driven safety messages. The successfully dissemination of event-driven safety message is very crucial and can save our life. If a vehicle detects dangerous stuff such as a sharp object fallen from a construction truck on the road, it will notify other vehicles behind to avoid the object. The safety message must to be delivered to others node with high reliability and without delays.

Researches in [9, 14] concentrated on safety messages. However research in [9] focused only on the performance of the Emergency Electronic Brake Light with Forwarding (EEBL-F) safety application. This congestion control algorithm should be testing on other event-driven safety applications such as pre-crash sensing and lane change warning. Research in [14] also focused on safety messages but this research didn't separate safety messages into beacon and event-driven safety messages. As mentioned, the event-driven safety message is most important and should be delivering on time with high reliability.

Some of researchers consider the utility of packets as an important part in congestion control algorithm such as [11, 14]. Research in [11] proposed a novel concept for utility-based congestion control and packet forwarding in VANETs. This protocol called as decentralized Utility-Based Packet Forwarding and Congestion Control (UBPFCC) is implemented on top of the IEEE 802.11 MAC protocol. The congestion control algorithm uses an application specific utility function and encodes the quantitative utility information in each transmitted data packet in a transparent way for all users within a local environment. A decentralized algorithm then calculates the average utility value of each individual node based on the utility of its data packets and assigns a share of the available data rate proportional to the relative priority. This congestion control algorithm evaluated priority message based on utility and packet size, it will reduced performance of disseminating of event-driven safety messages.

Research in [14] applied message utility one of dynamic factor, according to the number of its retransmitting by the neighborhood. For example, if a node X has to send message A but at the same time node X receives the same message A sent by another node. The node X should change the message A based on dynamic factor. The higher priority message is given to the smaller covered zone. Furthermore research in [14] evaluated priority with based on other factors such as node speed and message validity. The result showed that the delay of event-driven safety message is 50 ms in the worst scenario. This result is critical because pre-crash sensing safety application message need to disseminate to adjacent nodes within 20 ms.

Furthermore research in [14] also developed a congestion control algorithm and then adapt dynamic priorities-based scheduling. The purpose of dynamic priorities-based scheduling is to ensure high priority packets to be sent first without delay while medium and low priority packets will be rescheduling. In fact with EDCA, the message has been given based on its content such as critical or not.

Research in [20] proposed transmit power control in their framework of congestion control algorithm. Research on this area has been increase researchers to study especially from [16]. The purpose of transmit power control to maximize energy consumption and connectivity for point-to-point communications. Generally, a higher data rate usually requires a higher transmit power from a sender, thus may cause a higher interference to other nodes. The congestion control via dynamic transmit power control, which are usually periodical one-hop broadcast messages, can restrict the channel usage level and dynamically reserve a fraction of bandwidth for the safety application. The original idea is to control the transmit power of low priority messages and keep the transmit power of the highest priority traffic [16]. Research in [9] also proposed congestion control via dynamic transmit power control. They adjusted the transmit power for all packet types and study the impacts of transmit power control on the congestion problem in VANET. With increased transmit power level, the IEEE 802.11p physical layers (PHY) is able to provide communications within a distance from 100m to 1km in vehicular environments. This congestion control algorithm need places equipped with Road Site Units (RSUs).

Researches in [9, 20] were proposed smart/efficient rebroadcast scheme algorithms to prevent the congestion channels problem by limiting the forwarded packets. The blindly broadcasting beacon messages will causes a lot of redundancy packets and lead the broadcast storm problem. Smart rebroadcasting scheme from [20] operates only vehicles on the same lane and located behind the accident vehicle will forward the event-driven safety message. The vehicles only forward event-driven safety message after their successful reception of this event-driven safety message from front vehicles. In real scenarios, when accident happened it's also involve other lanes. The researchers in [20] also proposed efficient rebroadcasting in their concepts and framework of congestion control. The efficient rebroadcasting will reduce the transmission rate with minimum overhead.

Some of congestion detection methods are introduced in congestion control algorithm such as event-driven detection, measurement-based detection and MAC blocking.

3.1 Event-Driven Detection Method

The event-driven detection method was adapted in [9, 14]. With this event-driven detection method, each node applies the brute force queue freezing for all MAC transmission queues except for the safety queue with the highest priority. For example, when a node detects event-driven safety message either generated at its own application layer or received from another device, it will launch the congestion control immediately to guarantee the delivering of event-driven safety message [9].

3.2 Measurement-Based Detection Method

In the measurement-based detection method, each device periodically senses the channel based on the predefined thresholds such as channel usage level [9], number messages queue [14] and channel occupancy time [7]. The predefined threshold play important role in the performance of the wireless network by monitor and detect congestion of communication channels. The predefined threshold will be measure by metrics above. For example in [9] applied channel usage level as threshold. With this method, each device periodically senses the channel usage level, and detects the congestion whenever the measured channel usage level exceeds the predefined threshold. However research in [7] was set channel occupancy time as predefined threshold. If channel occupancy time measured at a node in CCH channel is longer than a given predefine threshold, all beacon messages will be blocked immediately. Research in [14] set a queue length in SCHs channels as threshold. If the queue length of comfort applications exceeds a predefine threshold, congestion is indicated and the preceding node is notified in order to decrease its transmission rate. To ensure performance of event-driven safety message, they should control CCH communication channel compared to SCHs channels.

3.3 MAC Blocking Detection Method

The MAC blocking detection mechanism is used for immediate and aggressive control of beacon message transmissions to mitigate congestion, and also adaptive traffic rate control is used for congestion avoidance. The MAC blocking detection method applied in [7]. For example, if MAC blocking happens at a node due to excessively long channel occupancy time, the channel is considered as congested for event-driven safety messages. The congestion signal is sent to the application layer, triggering traffic rate control actions.

4. Conclusions and Future Works

In this paper, we tried to expose the strong and weak points of some of these existing congestion control algorithms in VANETs. We conclude that many of these of congestion control algorithms are found will solved congestions problems in VANETs. However, most of proposed congestion control algorithms not focusing on event-driven safety message. Furthermore these congestion control algorithms are not addressed on uni-priority of event-driven safety message congestion. In real situation, various reactions from drivers will generate multiple event-driven safety messages. The uni-priority message of event-driven safety messages also generated from different transmitters. The nodes with same high priority packets need to schedule before start transmitting process.

In future work, we will propose framework for congestion control for disseminating uni-priority of event-driven safety messages. We also plan to verify and evaluate performance of our proposed congestion control algorithm using network simulator such as NS-2.

References

[1] A. Kajackas, A. Vindašius, et al. “Inter-Vehicle Communication: Emergency Message Delay Distributions”, Journal of Electronics and Electrical Engineering, No. 86, page 33-38.

[2] A. Vinel, S. Andreev, et al. “Estimation of A Successful Beacon Reception Probability In Vehicular Ad-Hoc Networks”, In Proceedings of the 2009 International Conference on Wireless Communications and Mobile Computing, 2008, pp. 416-420

[3] C. Lochert, B. Scheuermann, and M. Mauve, “A Survey on Congestion Control for Mobile Ad-Hoc Networks”, Wiley Wireless Communications and Mobile Computing 7 (5), 2007, pp. 655–676,

[4] C.Y. Yang, and S.C. Lo, “Street Broadcast with Smart Relay for Emergency Messages in VANET”. In Proceeding 24th International Conference on Advanced Information Networking and Applications Workshops, 2010.

[5] D. Jiang, and L. Delgrossi, “IEEE 802.11p: Towards an International Standard for Wireless Access in Vehicular Environments”. In Proceeding IEEE Vehicular Technology Conference, 2008 pp. 2036 – 2040

[6] F.J. Martinez, P. Manzoni, and J.M. Barrios, “Assessing the Feasibility of a VANET Driver Warning System”, In

Proceedings of the 4th ACM workshop on Performance Monitoring and Measurement of Heterogeneous Wireless and Wired Networks, New York, 2009

[7] J. He, H.C. Chen, et al. "Adaptive Congestion Control for DSRC Vehicle Networks", IEEE Communications Letters, Vol. 14, No. 2, February 2010.

[8] J. Peng, and L. Cheng, "A Distributed MAC Scheme for Emergency Message Dissemination in Vehicular Ad Hoc Networks." IEEE Transactions on Vehicular Technology 56 (6): 8, 2007.

[9] J. Zang, L. Stibor, et al. "Congestion Control in Wireless Networks for Vehicular Safety Applications", In Proceeding The 8th European Wireless Conference, Paris, France. 2007, pp.7

[10] K.M. Tony, P.L. Kenneth, et al. "Multichannel Medium Access Control for Dedicated Short-Range Communications", IEEE Transactions on Vehicular Technology 58(1): 17. 2009.

[11] L. Wischhof, and H. Rohling, "Congestion Control in Vehicular Ad Hoc Networks", In Proceeding of IEEE International Conference on Vehicular Electronics and Safety, Germany, 2005, pp. 58-63

[12] L. Zhou, G. Cui, et al. "NPPB: A Broadcast Scheme in Dense VANETs", Information Technology Journal, 2010, pp. 247-256.

[13] M. Amadeo, C. Campolo, et al. "A WAVE-Compliant MAC Protocol to Support Vehicle-to-Infrastructure Non-Safety Applications", In Proceeding IEEE International Conference on Communications (IEEE ICC 2009). Dresden, Germany, 2009, pp. 1-6

[14] M.S. Bouassida, and M. Shawky "On the Congestion Control within VANET." 1st IFIP Wireless Days, 2008, IEEE Explore.

[15] M. Li, W. Lou, et al. "OppCast: Opportunistic Broadcast of Warning Messages in VANETs with Unreliable Links. Mobile Ad Hoc and Sensor Systems", In Proceeding IEEE 6th International Conference (MASS '09), 2009, China, pp. 534 - 543

[16] M. Torrent-Moreno, D. Jiang, et al. "Broadcast Reception Rates and Effects of Priority Access in 802.11Based Vehicular Ad-Hoc Networks", International Conference on Mobile Computing and Networking USA, 2004, pp. 10-18

[17] M. Welzl "Network Congestion Control-Managing Internet Traffic", John Wiley & Sons Ltd. Chichester, West Sussex, England, 2005.

[18] N. Wisitpongphan, O.K. Tonguz, et al. "Broadcast Storm Mitigation Techniques in Vehicular Ad Hoc Networks", IEEE Wireless Communications, 2007, Vol.14. pp. 84 - 94

[19] W. Lee, H. Lee, et al. "Packet Forwarding Based on Reachability Information for VANETs", Journal of Information Networking, Towards Ubiquitous Networking and Services. SpringerLink, 2008, page 305–314.

[20] W. Zhang, A. Festag, et al. "Congestion Control for Safety Messages in VANETs: Concepts and Framework", In Proceeding 8th Conference on ITS Telecommunications (ITST), Thailand, 2008, pp. 199-203

[21] X. Yang, J. Liu, and F. Zhao, "A Vehicle-to-vehicle Communication Protocol for Cooperative Collision Warning", In Proceeding of the 1st Annual International Conference on Mobile and Ubiquitous Systems Networking and Services, IEEE Computer Society, Massachusetts, USA, 2004, pp: 114-123.

[22] Y. C. Hu and D. B. Johnson, "Exploiting Congestion Information in Network and Higher Layer Protocols in Multihop Wireless Ad Hoc Networks", In Proceedings of the 24th International Conference on Distributed Computing Systems (ICDCS'04), 2004, pp. 301-310

[23] Y. Mylonas, M. Lestas, et al. "Speed Adaptive Probabilistic Flooding in Cooperative Emergency Warning", In Proceeding ACM International Conference, on Wireless Internet (WICON'08), Hawaii, USA, 2008.

[24] Y. Yu, and G. B. Giannakis, "Cross-Layer Congestion and Contention Control for Wireless Ad Hoc Networks" IEEE Transactions On Wireless Communications, Vol. 7, No. 1, Jan 2008.

Mohamad Yusof Darus is currently a PhD student in the Department of Computer Systems and Communications of the Faculty of Computer Science and Information Systems at the Universiti Teknologi Malaysia. He obtained M.Sc. Information Technology from Universiti Utara Malaysia (Malaysia) in 2003. He has been involved in lots of academic research since then; presently he is a member of Pervasive Computing Research Group at UTM, while his research interest Vehicular Networks (VANETs). He has published in many national and international learned journals.

Kamalrulnizam Abu Bakar obtained his PhD degree from Aston University (Birmingham, UK) in 2004. Currently, he is an Associate Professor in Computer Science at Universiti Teknologi Malaysia (Malaysia) and member of the "Pervasive Computing" research group. He involves in several research projects and is the referee for many scientific journals and conferences. His specialization includes mobile and wireless computing, information security and grid computing.

RGWSN: Presenting a genetic-based routing algorithm to reduce energy consumption in wireless sensor network

Arash Ghorbannia Delavar[1], Amir Abbas Baradaran[2] and Javad Artin[3]

[1] Department of Computer Engineering and Information Technology, Payam Noor University Tehran, Iran

[2] Department of Computer Engineering and Information Technology, Payam Noor University Tehran, Iran

[3] Department of Computer Engineering and Information Technology, Payam Noor University Tehran, Iran

Abstract

In this paper, a genetic- based routing algorithm to reduce energy consumption in sensor networks is presented. This method, by regarding distance hybrid parameters, energy and density, has created a fitness function which has optimal conditions compared to previous parameters. In the proposed algorithm, a new technique to select several probable cluster heads(CHs) in each area, has been used. The results of simulation indicate that the proposed algorithm has increased network's lifetime, compared to former techniques. Also, we have proved , in the new technique, that the number of alive nodes at the end of each round has increased , compared to previous techniques, and that is the result of proper selection of CH in each area.

Keywords: *Genetic algorithm, Wireless sensor network, Routing, Reduce energy consumption.*

1. Introduction

Wireless sensor networks are increasingly used in military surveillance and civilian usages[1]. These networks are composed of hundreds or thousands sensors which receive data from environment and send them to a Base Station(BS)[2]. One important problem in wireless sensor networks, is creating an effective routing protocol. Generally this kind of network has some constraints in calculation potency, storage capacity, energy, etc .To reduce energy consumption and to increase network's lifetime are the most considerable problems in sensor networks[2,3]. Receiving environment data by nodes and sending them to BS can lead to run out of node's energy and then threaten network's lifetime[4]. One of the techniques used to find the best way of receiving and sending data by nodes is using genetic algorithms(GA)[5].In the routing algorithms other than GA, the problem of optimum consumption of energy is point to point, which causes ignorance of other points, but in GA all points in all phases of running an algorithm can be included, and this leads to good results. GA is a multi-purpose ,optimal search inspired by genetic theory and natural selection . The problems using GA are as a coded chromosome including several genes[5]. The solution is shown by a group of chromosomes related to a population .When the algorithm is repeated, current chromosomes take a genetic operation which includes selection, crossover and mutation. This operation results in the appearance of next generation. These processes are running till finding an optimum, and certain solution. In this research, we try to present a genetic- based routing algorithm to reduce energy consumption of supply source in sensor networks. Routine work is comparing to previous algorithms.

2. Related Works

One of the most important clustering algorithms to LEACH which is based on the rounds each of which includes setup and steady phases. Each node in every round can be or cannot be a CH[6]. Being a CH or not being a CH is based on the following threshold[6]:

$$T(n)=\begin{cases}\dfrac{p}{1-p*(r \mod \frac{1}{p})} & if \quad n\in G,\\ 0 & otherwise\end{cases} \qquad (1)$$

in which :
P: CH decision percentage(percentage of being CH)
R: current round

G: set of nodes not changed to CH in 1/p of the recent round

One other technique to optimize energy consumption in sensor networks, based on genetics, was presented by Ming, Shiyuan Jin, Annie S.Wu and Zhou[7]. In this technique, sensor network is clustered by GA. Some other genetic – based techniques have been presented by Zhou, Jingcheng Ouyang, Jianming Zhang, Yaping Lin, Cuihong[8].

3. Genetic Algorithm

Genetic algorithm (GA) is an optimization technique inspired by nature which operates as numerical, direct and random search .GA is based on repeat and the primary principles thereof are obtained from Genetics .GA , due to using nature, has some basic differences from other searching methods , as follow[5]:

1. GA operates by bit strands called chromosomes, each of which represents all sets of variables, while most search techniques include special and independent variables.
2. GA searches by random selection that leads to non-optimum points, in each round, to be included in next processes.
3. GA, in each repeat, considers some points of search space. So it is less likely to converge into a local maximum.

Each process of repeat in GA is called a generation and a series of solutions or answers is called population .GA starts with some initial points generated randomly or selectively. These points are called Initial Population. After the generation of initial population , genetic operators operate on initial population and create new population. Generating new population is done by fitness function, i.e. fitness function operates on initial population and the new population is generated. Initiating next generation is by means of new populatuion . These processes are run through various generations to obtain the best solution[9].

Generally, GA includes the following stages[5,9]:

1. **Initial stage**: In this stage an n-chromosomes population is generated randomly.
2. **Fitness**: In this stage, for all existing n- chromosomes, a fitness value is defined.
3. Generation of new population through following stages[9]:

A. **Selection**: In this stage the two chromosomes having more fitness are selected as parents. Selection procedures are random. Some current methods of selection includes: wheel roulette selection, sequential selection, competitive selection, Boltzman selection and etc.

B. **Crossover**: In this stage, two parents generated in selection stage bring new children. Generally, crossover is a process in which the old generation of chromosomes crossover and a new generation of chromosomes is generated.

Some current crossover methods are: the combination of single- point, the combination of two- point, the combination of multi-point and uniform composition

C. **Mutation** :In this stage a child having mutation condition mutates. After this stage, children are decoded and compared to fitness function. If , regarding fitness function , the conditions are not optimal , new children in initial population are used and the algorithm proceeds. In this stage, generated chromosomes are considered as initial population and the answers having low fitness are omitted and the algorithm proceeds with n chromosomes . We can refer to Binary and True mutations as two of the most considerable mutation methods .It is noteworthy that to do the above stages needs using an encoding system.

Under the following conditions, we can put an end to running GA[5]:

1. Algorithm reaches a fixed number of generations.
2. No improvement is obtained while algorithm proceeds.
3. The average value of fitness function reaches a certain measure as per several repeats.
4. The greatest fitness rate is gained for children.
5. A combination of the above conditions happen.

4. Network Model

Network model has the following characteristics[6]:

1. All nodes of sensor and BS are motionless and after being established, nodes can not be added or omitted.
2. The base energy of nodes is different.
3.The nodes of sensor are informed of situation , i.e to do this, they need hardwares like GPS.

5. Radio Model

The sensors consume energy while receiving and transmitting data[6,10]. The standard radio model used in WSN, uses free space and multi- path fading model which depend on the distance between transmitter and receiver. This distance is the shortest crossover distance, $d_{crossover}$. Transmit power equals to[6,11,12,13] :

$$p_r(d) = \frac{p_t G_t G_r \lambda^2}{(4\pi d)^2} \qquad (2)$$

In which:

p_t : transmit power

G_t : gain of transmit antenna

λ : carrier signal's wave-length (in meter)

When the receiver s distance is longer than $d_{crossover}$

transmit power equals to[11]:

$$p_r(d) = \frac{p_t G_t G_r h_t^{\,2} h_r^{\,2}}{(d)^4} \qquad (3)$$

h_t : transmit antena height in meter

h_r : receiver antena height in meter

To transmit n-bit message in a d- meter distance , radio energy consumption equals to[6,11]:

$$E_{TX}(n,d) = n(E_{elect} + \in_{fs} d^2) \quad d < d_{crossover}$$
$$E_{TX}(n,d) = n(E_{elect} + \in_{mp} d^4) \quad d \geq d_{crossover} \qquad (4)$$

To receive an n-bit message radio energy equals to[11,13]:

$$E_{rx}(n) = nE_{elect} \qquad (5)$$

$\in_{fs}$ and $\in_{mp}$ are parameters which depend on sensitivity(intelligence) of receiver and noise's shape and E_{elect} is the electric energy depending on such factors as digital code, modulation. Filtering , etc[11].

6. The New Presented Algorithm RGWSN

The new proposed algorithm is based on the rounds each of which includes two phases: steady phase and setup phase. In setup phase, the area where nodes are distributed, are networked and optimum CHs are specified. In the second phase or steady phase, data are transmitted from normal nodes to CHs and from CHs to BS (Base Station). Set up phase starts by an initial manage from BS, including nodes' location and initial energy. Then some nodes having a higher fitness shall be selected and attended in GA to detect optimum nodes as CHs, for data transmission in steady phase to do this, the environment in which nodes are distributed, is divided to separate areas called grides. The selection of nodes in each grid is based on their distance from the nodes' gravity center of each grid. The nodes closer to gravity center in each grid, are selected as initial population, to attend in GA, i.e. initial population includes the nodes that are likely to be optimum CH. (for example if the environment is divided into 8 areas and 3 nodes are selected from each area, the initial population has 24 bits. In this algorithm we used binary coding system. After creating initial population, to select optimum nodes, regarding initial population, we create some random populations and by genetic operators of selection, combination and mutation, we created new children, randomly and binarily, from current population.(For example if we have 50 populations, after binary operation of genetic operators, we have 100 populations.). In the proposed algorithm we use the combination of single-point and binary mutation. For example, if two parents are as P2:10001111, P1=01000011, P=3, the following children are generated.
F1=110011
F2=011110
After running some generations, by applying fitness function on generated populations, the best population is selected, i.e. the population having the lowest fitness (the least average energy of entire network) is selected. For this proposed approach, fitness function equals average energy consumed by entire network. Fitness function is calculated based on the model proposed by Heinzleman. Heinzleman has stated, in a model, that each node, to transmit L bits of data in a distance d from itself, will use energy E_s .

$$E_s = LE_{elect} + L \in_{fs} d^2 \quad d < d_0$$
$$E_s = LE_{elect} + L \in_{mp} d^4 \quad d \geq d_0 \qquad (6)$$

In which:

d_0 : The shortest crossover distance

E_{elect} : The Energy required for activation of electric circuits.

$\in_{mp}$, $\in_{fs}$: parameters related to receiver's sensitivity and noise shape.

In the proposed algorithm and set up phase, we calculate fitness value for final population bits and we have supposed bits 0 as representing common nodes and bits 1 as representing CHs. Fitness Function(The energy consumed by entire network) equals:

$$E = E_1 + E_2 + E_3 + E_4 \qquad (7)$$

E_1 : Energy necessary to transfer from normal node to CH

E_2 : Energy necessary to receive CH data from normal node

E_3 : Aggregation energy

E_4 : Energy necessary to transfer from CH to BS

Which we have:

$$E_1 = LE_{elect} + L \in_{fs} d^2_{distoch} \quad d_{distoch} < d_0$$
$$oR \qquad (8)$$
$$E_1 = LE_{elect} + L \in_{mp} d^4_{distoch} \quad d_{distoch} \geq d_0$$

$d_{distoch}$: The distance between common node and CH

L: bits number

$$E_2 = LE_{elect} \times N\_Common \qquad (9)$$

N_Common : common node number

$$E_3 = LE_{ag} \times N\_ch \qquad (10)$$

E_{ag} : aggregation energy

N_ch :CHs number (bits 1)

$$E_4 = LE_{elect} + L \in_{fs} d^2_{distoBs} \quad d_{distoBs} < d_0$$

$$oR \qquad (11)$$

$$E_4 = LE_{elect} + L \in_{mp} d^4_{distoBs} \quad d_{distoBs} \geq d_0$$

$d_{distoBS}$: distance from CH to BS

In set up phase, after applying fitness function on final population and specifying common nodes and CHs, E_1 for common nodes (bits 0) and E_2, E_3, E_4 for bits 1(CHs) are calculated. At last in steady phase, based on transition ,The energy of nodes reduces. Chart 1 shows flowchart of the proposed algorithm.

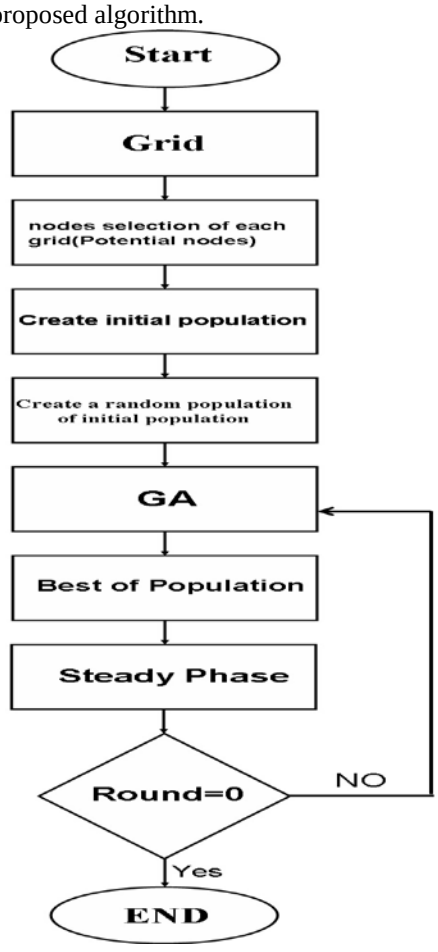


CHART 1 . FLOWCHART OF THE PROPOSED ALGORITHM

7. Simulation

Analysis of the proposed algorithm is done by MATLAB software. The number of alive nodes at the end of each round, the number of grids ,the number of selected nodes in every grid have been considered to generate initial population. Also, the energy of initial nodes, are random measures from 0.2 to 0.5. Other parameters used in simulation are as follow:

1. The nodes are randomly placed in a squared environment.
2. BS position is variable.
3. E_{elect} : 50nj/bit
4. $\in_{fs}$: 10pj/bit/m2
5. $\in_{mp}$: 0.0013pj/bit/m4
6. E_{ag} : 5nj/bit/signal
7. $d_0 = \sqrt{\frac{\in_{fs}}{\in_{mp}}}$

The proposed method is compared to GSAGA,RCSDN,LEACH methods. Figure 1 shows alive nodes at the end of 1000 rounds and figure 2 shows fitness function for the proposed method. Also table 1 shows simulation parameters.

TABLE I. SIMULATION PARAMETERS

Parameter	Value
Network size	100*100 m
Base station location	50,50 m
Initial energy for node	rand [0.2,0.5] J
E_{elec}	50nJ/bit
ε_{fs}	10pj/bit/ m2
ε_{mp}	0.0013pj/bit/m4
Data aggregation energy	5nj/bit/signal
Nodes number	100
Grids Number	6
Nodes number of each grid	5
d_0	87m

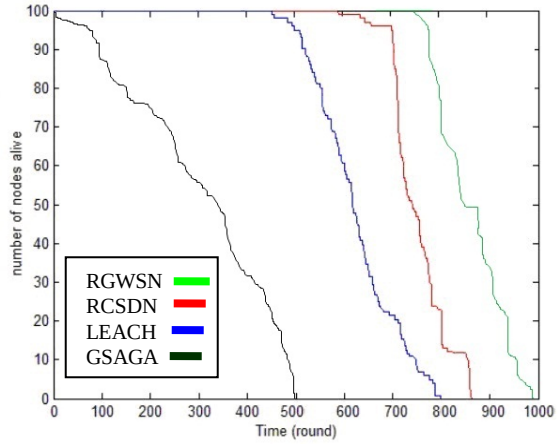


Figure 1. Total number of alive nodes in the RGWSN, LEACH,GSAGA,RCSDN

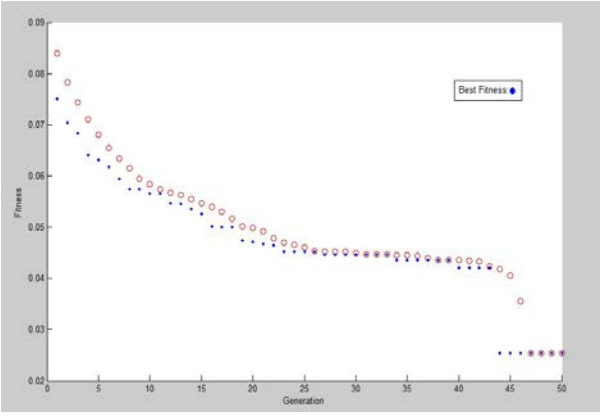


Figure 2. The energy consumed by entire network

As shown by diagrams, after running 1000 rounds, the number of alive nodes in the proposed approach is more than that of GSAGA ,RCSDN ,LEACH .As a result, the network lifetime is longer.

8. Conclusion

In this research, a new method to transmit data from normal nodes to CH and from CH to BS in sensor networks, is presented. The selection of an optimum CH has an effective role in increasing a sensor network's lifetime. By means of multi-simulations, we have shown that the proposed algorithm is different from other presented algorithms, by virtue of reducing energy consumption.

References

[1] Hongjuan Li, Kai Lin, Keqiu Li," Energy-efficient and high-accuracy secure data aggregation in wireless sensor networks",in: School of Computer Science and Technology, Dalian University of Technology, No. 2, Linggong Road, Dalian 116024, China

[2] A.H. Mohajerzadeh,M.H.Yaghmaee,H.S.Yazdi,A.A.Rezaee," A Fair Routing Protocol Using Generic Utility Based Approach in wireless sensor networks ",in: 9781-4244-3941-6/09/$25.00 ©2009 IEEE

[3] CHENG Hong-bing, YANG Geng," NHRPA: a novel hierarchical routing protocol algorithm for wireless sensor networks",in: Journal of China Universities of Posts and Telecommunications , September 2008, 15(3): 75–81

[4] Ataul Bari , Shamsul Wazed, Arunita Jaekel, Subir Bandyopadhyay," A genetic algorithm based approach for energy efficient routing in two-tiered sensor networks",in: School of computer Science, University of Windsor, 401 Sunset Avenue, Windsor, ON, Canada N9B 3P4, April 2008

[5] S.Yussof,R.Z. Razali,O.H.See," A Parallel Genetic Algorithm for Shortest Path Routing Problem",in: 2009 International Conference on Future Computer and Communication, 978-0-7695-3591-3/09 $25.00 © 2009 IEEE DOI 10.1109/ICFCC.2009.36

[6] A.G. Delavar,J.Artin,M.M.Tajari," RCSDN : a Distributed Balanced Routing Algorithm with Optimized Cluster Distribution",in: ICSAP 2011,3rd International Conference onSignal Acquisition And Processing , 26-28, February, 2011, Singapore

[7] Shiyuan Jin, Ming Zhou, Annie S. Wu," Sensor Network Optimization Using a Genetic Algorithm",in: School of EECS University of Central Florida Orlando, FL 32816

[8] Jianming Zhang,Yaping Lin,Cuihong Zhou,Jingcheng Ouyang," Optimal Model for Energy-Efficient Clustering in Wireless Sensor Networks Using Global Simulated Annealing Genetic Algorithm"in: 978-0-7695-3505-0/08 $25.00 © 2008 IEEE DOI 10.1109/IITA.Workshops.2008.40

[9] V.PURUSHOTHAMREDDY,G.MICHAEL,M.UMAMAH ESHWARI,"Coarse-Grained ParallelGeneticAlgorithm to solve the Shortest Path Routing problem using Genetic operators"in: V.Purushotham Reddy et al./ Indian Journal of Computer Science and Engineering (IJCSE), ISSN : 0976-5166

[10] Wang, Q., Yang, W. "Energy consumption model for power management in wireless sensor networks. " In 4th Annual IEEE communications society conference on sensor, mesh and ad hoc communications and network (SECON 2007)

[11] W.R. Heinzelman, A. Chandrakasan, H. Balakrishnan, "Energy efficient communication protocol for wireless sensor networks", in:Proceedings of the 33rd Hawaii International Conference on System Science, vol. 2, 2000.

[12] T. Rappaport, "Wireless Communications": Principles & Practice,New Jersey, Prentice Hall, 1996.

[13] A.G. Delavar,J.Artin,M.M.Tajari," PRWSN: A Hybrid Routing Algorithm with Special Parameters in Wireless Sensor Network", in: A. Özcan, J. Zizka, and D. Nagamalai (Eds.): WiMo/CoNeCo 2011, CCIS 162, pp. 145–158, 2011. © Springer-Verlag Berlin Heidelberg 2011

Computer Science, Payam Noor University, Tehran, IRAN. He is also the Director of Virtual University and Multimedia Training Department of Payam Noor University in IRAN.

Dr.Arash Ghorbannia Delavar is currently editor of many computer science journals in IRAN. His research interests are in the areas of computer networks, microprocessors, data mining, Information Technology, and E-Learning.

Amir Abbas Baradaran received the BS , in 2008 and Now, department of Computer Engineering and Information Technology in Payam Noor University, Tehran, IRAN. His research interests include wireless communication, network computer and Genetic algorithm.

Javad Artin received the BS, in 2007 and now, he is a Student the MS degree in the department of Computer Engineering and Information Technology in Payam Noor University, Tehran, IRAN. His research interests include computer networks, wireless communication, Genetic algorithm, and Fuzzy logic.

Clustering Web Access Patterns Based on learning Automata

Babak Anari [1], Mohammad Reza Meybodi [2] and Zohreh Anari[3]

**[1] Computer Engineering Department, Islamic Azad University
Shabestar, 5381915351, Iran**

**[2] Computer Engineering and Information Technology Department, Amirkabir University of Technology
Tehran, 5381915351, Iran**

**[3] Computer Engineering Department, Islamic Azad University
Shabestar, 5381915351, Iran**

Abstract

The interest of web users can be revealed by the visited web pages and time duration on these web pages during their surfing. In this paper a new method based on Learning Automata for clustering web access patterns is proposed. At the first step of the proposed algorithm, each web access pattern from web logs is transformed into a weight vector using the learning automata. In the second step a primitive clustering is performed to group weight vectors into a number of clusters. Finally, the weighted Fuzzy c-means approach is developed to deal with the results of the second step. Our experiments on a large real data set show that the method is efficient and practical for web mining applications.

Keywords: ***Web access patterns, Clustering, Learning automata, Distributed learning automata, Time duration.***

1. Introduction

The problem of clustering web access patterns is a part of a larger work of web usage mining which is the application of data mining techniques to discover usage patterns from Web data typically collected by web servers in large logs [9]. Clustering web access patterns is an important step in studying the characteristics of web surfers. The patterns from web data are non-numerical, thus Runkler and Beadek [7] proposed relational clustering method to group non-numerical web data. Some other researchers tried to explore clustering by another soft computing technique, rough theory. Wu [10] proposed a two-layer evolutionary clustering algorithm to group web access patterns from web logs. Shi [8] applied rough k-means clustering method in fuzzy environment to group web access patterns from web logs. De [3] tries to use rough approximation to cluster web transactions.

In this paper we propose a new algorithm based on learning automata to group web access patterns from web logs. First each pattern is converted into a weight vector using learning automata. Then the learning automaton is used to group all the weight vectors into a number of clusters. And the set of centers of these clusters is further clustered by the weighted Fuzzy c-means. The rest of the paper is organized as follows: In section 2 the learning automata and the distributed learning automata will be explained. In section 3 the proposed algorithm to cluster web access patterns is proposed. In section 4 the effectiveness of the method is demonstrated. Finally we conclude in section 5.

2. Learning Aautomata

An automaton can be regarded as an abstract model that has finite number of actions. This action is applied to the selected action of automata. The random environment evaluates the applied action and gives a grade to the selected action of automata. The response from environment (i.e. grade of action) is used by automata to select its next action. By continuing this process, the automaton learns to select an action with the best grade. The learning algorithm is used by automata to determine the selection of next action from the response of environment. Figure 1 shows the relationship between the environment and the learning automata [6].

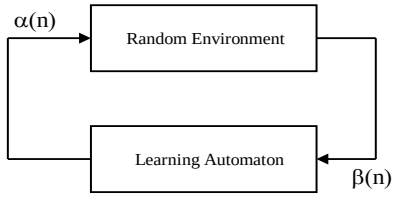


Figure1: The relationship between learning automata and the environment

2.1 Environment

The environment can be shown by E ≡{α, β, c} in which α ={α_1,α_2,….,α_r} represents a finite action / output set,

$\beta=\{\beta_1, \beta_2,\ldots, \beta_m\}$ represents an input / response set, and $c=\{c_1, c_2,\ldots,c_r\}$ is the set of penalty probabilities, where each element c_i corresponds to one action α_i of the set α. The output (action) α_n of the automaton belongs to the set α, and it is applied to the environment at time t = n.

2.2 Learning Automata with Variable Structure

Variable structure learning automata is represented by $< \beta, \alpha, T, p >$, where $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_r\}$ is a set of actions. $\beta = \{0, 1\}$ is the set of inputs from the environment; where 0 represents a reward and 1 represents a penalty, $p(n+1) = T[\alpha(n), \beta(n), p(n)]$ is learning algorithm and defines the method of updating the action probabilities on receiving an input from the random environment. $p = \{p_1(n), p_2(n)\ldots p_r(n)\}$ is the action probability vector, where $p_i(n)$ represents the probability of choosing action α_i at time n. In these kinds of automata, if the action of α_i is chosen in the n^{th} stage and receive the desirable response from the environment, the probability of $p_i(n)$ increases and the other probabilities decreases and in undesirable response, the - probability of $p_i(n)$ decreases and the other probabilities increase. The following algorithm is one of the simplest learning schemes for updating action probabilities, and is defined as follows:

$$p_i(n+1) = p_i(n) + a[1 - p_i(n)] \qquad \forall j \quad j \neq i$$
$$p_j(n+1) = (1-a)\,p_j(n) \qquad (1)$$

a) Desirable response

$$p_i(n+1) = (1-b)\,p_i(n)$$
$$\forall j \quad j \neq i \quad p_j(n+1) = \frac{b}{r-1} + (1-b)\,p_j(n) \qquad (2)$$

b) Undesirable response

As seen from the definition, the parameter a is associated with reward response, and the parameter b with penalty response. According to the values of a and b we can consider three scheme. If the learning parameters a and b are equals, the scheme called reward penalty (L_{R-P})When b is less than a, we call it linear reward epsilon penalty ($L_{R\varepsilon P}$) scheme. When b equals to zero, we call it as linear reward inaction (L_{R-I}) scheme. For more information about the theory and applications of learning automata, refer to [2, 7, 8].

2.3 Distributed Learning Automata (DLA)

A distributed learning automaton (DLA) is a network of learning automata which collectively cooperates to solve a particular problem. In DLA, the number of actions for any automaton in the network is equal to the number of outgoing edges from that automaton. When an automaton selects one of its actions, another automaton on the other end of edge corresponding to the selected action will be activated. At any time only one automaton in the network will be active. Formally, a DLA with n learning automata can be defined by a graph (V, E).Where $V= \{LA_1, LA_2\ldots LA_n\}$ is the set of automata and $E \subset V \times V$ is the set of edges in the graph in which an edge (LA_i, LA_j) corresponds to action α_j of automata LA_i. Figure 2 shows the network of distributed learning automata. The action probability vector for automaton LA_i is shown by $p^i = \{p_1^i, p_2^i, \ldots, p_m^i\}$ where p_m^i denotes the probability of selecting action α_m , that is, edge (LA_i, LA_m). The choice of action α_m^i by LA_i activates LA_m. r_i shows the number of actions done by LA_i automata.

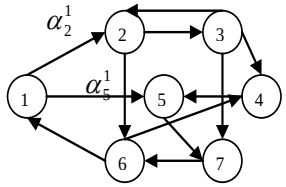


Figure 2: Network of distributed learning automata

3. Clustering Web Access Patterns based on Learning Automata

The proposed algorithm includes three steps: At the first step of the proposed algorithm, each web access pattern from web logs is transformed into a weight vector using the learning automata. In the second step we put each weight vector in the nearest cluster using the learning automata. By doing this, a primitive clustering is performed on the web access patterns and the primitive centers of clusters are determined. Finally, these primitive clusters which have no or several access patterns are used by weighted c-means clustering algorithm and on the basis of the weight of each cluster which has been determined according to the number of web access patterns in each cluster and also the centers of clusters which have been determined by the learning automata are clustered. Finally, the final clusters are determined from the primitive clusters.

3.1 Characterizing Web Access Pattern as a Weight Vector

Suppose there are m users and user transactions where $s_i (1 \le i \le m)$ representing the unique surfing behavior of the i^{th} user. Let $W = \{Url_1, Url_2, \ldots, Url_n\}$ is the union set of distinct n web pages visited by users, $U = \{(Url_1, t_{1_1}), \ldots, (Url_1, t_{1_g}), \ldots, (Url_n, t_{n_1}), \ldots, (Url_n, t_{n_h})\}$ be the union set of $s_i (1 \le i \le m)$, where g is the number of time

duration on web page Url_1, h is the number of all time durations on web page. Url_n, n is the number of all different web pages visited by users. Let T_{max} be the maximum time duration on web pages. Web pages and users play the role of stochastic environment for the existing learning automata in DLA. In the proposed method for each web page like P_k the learning automata of LA_k is considered. Suppose $p^k = \{p_1^k, p_2^k, \dots, p_n^k\}$ is the action probability vector for LA_k automata has been assigned to web page of P_k and $p_m^k(n)$ is the probability of choosing action α_m in learning automata of LA_k at the n^{th} time. If the user moves from page P_k to page P_m ($P_k \rightarrow P_m$) learning automata of LA_k updates its action probability vector based on learning algorithm. We can represent each web access pattern $s_i \in S(1 \le i \le m)$ by a weight vector as $W_i =< w_{i1}, w_{i2}, \dots, w_{ik}, \dots, w_{in} >$ where w_{ik} shows that the i^{th} user has been visited the web page of k. The main steps of the algorithm for determining the action probability of each automaton are provided as follows:

Step1. Create a DLA according to web pages structure.
Step2. Do Eq.3 for each learning automata such as LA_k where$(1 \le k \le n)$,

$$p_i^k(n) = \begin{cases} \frac{1}{r} & i \ne k \\ 0 & i = k \end{cases} \qquad (1 \le i \le n) \qquad (3)$$

Step3. Do step3-1 for every user access pattern in the log file

Step3-1 If the user moves from page D_k to page D_m $(k \ne m)$ and t_{value} is the time duration on D_m then update the action probability vector for LA_k automata based on Equations of 4, 5, 6 and 7 respectively.

$$a_m^k(n+1) = \frac{t_{value} \times p_m^k(n)}{T_{\max}} \qquad (4)$$

$$p_m^k(n+1) = p_m^k(n) + a_m^k[1 - p_m^k(n)] \qquad (5)$$

$$p_j^k(n+1) = (1 - a_m^k)\, p_j^k(n) \; j \ne m \forall j \qquad (6)$$

$$w_{ik} = \begin{cases} p_m^k & \text{If the user moves from page } D_k \text{ to page } D_m, (k \ne m) \\ 0 & \text{Otherwise,} \quad (1 \le k \le n) \end{cases} \qquad (7)$$

By increasing the number of access frequency to each web page and spending more time duration on each web page the amount of probability increases.
For example let $S = \{s_1, s_2, \dots, s_6\}$ be the set of user transactions and W= {A, B, C, D, E, F, G, H} be the union set of distinct web pages visited by all users, then 10 learning automata are considered (for each web page one automata and also one automata for the exit and start pages). Suppose that start and exit are the two pages that the user has entered and left from the site respectively. Also we consider 10 actions for each automaton according to equation 1. If the browsing sequence of a user is as $Start \rightarrow A \rightarrow B \rightarrow C \rightarrow D \rightarrow Exit$ and the time duration for visited pages as $(A, 30), (B, 42), (C, 118), (D, 91)$ and $T_{max} = 120sec$ then the action vector of each automata that is $(Start, A, B, C, D, E, F, G, H, Exit)$ is done based on equation (3) ,(4) and (5) is updated and the probability vector for each web page which is computed by learning automata as follows:

Start=(0,0.136,0.108,0.108,0.108,0.108,0.108,0.108,0.108,0.108)

A=(0.10625,0,0.15,0.10625,0.10625,0.10625,0.10625,0.10625,0.10625, 0.10625,0.10625)

B=(0.09875,0.09875,0,0.21,0.09875,0.09875,0.09875,0.09875,0.09875, 0.09875)

C=(0.101625,0.101625,0.101625,0,0.187,0.101625,0.101625,0.101625, 0.101625,0.101625)

D=(1/9, 1/9, 1/9, 1/9,0, 1/9, 1/9, 1/9, 1/9, 1/9)

Exit=(1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9,0)

p_k^j is the probability of choosing action α_k in learning automata LA_j . The value of p_k^j is computed in the first step of algorithm. For example, if the user access pattern for s_1 like is $Start \rightarrow A \rightarrow B \rightarrow C \rightarrow D \rightarrow Exit$ and the action probability values are like the previous example then $W_1 =< 0.136, 0.15, 0.21, 0.187, 0, 0, 0, 0, 0 >$.

3.2 Clustering Web Access Patterns by DLA

By increasing the number of access frequency to each web page and spending more time duration on each web page the amount of probability increases. In the second step of algorithm, we use DLA to group web access patterns into a number of clusters. In the proposed method for each web page like P_i $(1 \le i \le N)$ the learning automata LA_i is considered. Also we consider N be the number of clusters. The center of each cluster like $L(1 \le L \le N)$, can be denoted as $P_L = [p_L^i, p_L^{i+1}, p_L^{i+2}, \dots, p_L^m]$ $(1 \le L \le N)$. We suppose P_L^i is the probability of choosing action α_L in the learning automata of LA_i. The main steps of the algorithm are provided as follows:

Input: Web access patterns S
Output: N clustering centers
Step1. Create a DLA according to web pages structure and initialize action probability vector of each LA based on Eq.3
Step2.
Repeat
for L=1 to number of clusters **do**

// P_L^i is the probability of choosing action α_L in the learning automata of LA_i

// P_L is the cluster vector($1 \le L \le N$)

$$P_L = [p_L^i, p_L^{i+1}, p_L^{i+2}, \dots, p_L^m]$$

end for
for k =1 to number of users **do**
for L=1 to number of cluster vector

Compare each user weight vector W_k with vector P_L so that this Equation is satisfied.

$$\|W_k - P_L\|^2 =_L^{min} \|W_k - P_L\|^2$$

end for
end for

for i=1 to m //m is the number of automata

Enable action ith of LA_i according to the following equation:

$$p_L^i = p_L^i + a[1 - p_L]$$
$$p_m^i = (1-a)p_m^i \; \forall m \;\; L \neq m$$

end for
Until (no noticeable changes in the cluster vector)

The above algorithm classifies all user access patterns into N clusters, where the lth clusters is defined as $U^L = s_k \in S: \|W_k - P_L\|^2 ==_L^{min} \|E[W_k] - P_L\|^2$ where $(1 \le k \le N)$ and $(1 \le L \le N)$.

3.3 Clustering Centers of Clusters by Weighted c-means

In this step, the set of clustering centers $P_L(1 \le L \le N)$ generated in algorithm3.2 is further clustered based on the weighted c-means. Since $U^L(1 \le L \le N)$ includes different number of web access patterns, different weight is assigned to different $U^L(1 \le L \le N)$. The weight of $U^L(1 \le L \le N)$ is defined as follows:

$$w_i = \frac{\aleph(U^i)}{\sum_{j=1}^N \aleph(U^i)} \qquad (8)$$

Weighted c-means is applied to group $P_L(1 \le L \le N)$ into c different nonempty subsets. The main steps are described as follows:

Input: clustering center $P_L(1 \le L \le N)$ generated by the second step of algorithm
Output: Clustering results $(c\ clusters)$
Step1. Assign initial means $v_i(1 \le i \le c)$
Step2. According to the following membership function, assign each pattern P_L into the nearest cluster

(9)

$$u_{il} = \frac{1}{\sum_{j=1}^c \left(\frac{d(P_L, v_i)}{d(P_L, v_j)}\right)^{\frac{2}{m-1}}}$$

in which u_{il} is the membership degree of the P_L belonging to ith, $m \in (1, \infty)$ is a fuzzy parameter.
Step3. Recompute $v_i(1 \le i \le c)$ according to the following equation:

$$v_i = \frac{\sum_{L=1}^N w_L (u_{il})^m v_l}{\sum_{j=1}^N w_L (u_{il})^m} \qquad (10)$$

Step4. Repeat step 2 to step 3 until the following objective function convergence, i., there are no more new assignment.

$$J_m(M, V) = \sum_{l=1}^N w_L (u_{il})^m \sum_{i=1}^c d(P_L, v_i) \qquad (11)$$

In which $M = \{[u_{il}], 1 \le i \le c, 1 \le l \le N\}$ is a clustering matrix, $V = \{[v_i], 1 \le i \le c\}$ is the set of final clustering centers. After the algorithm 3.3 is executed, each $P_L(1 \le L \le N)$ belongs to c clusters according to different degrees. Each cluster center $v_i(1 \le i \le c)$. Every web access pattern s_i in U^l belongs to c clusters by the same degree with P_L.

4. An Experiment

Several preprocessing tasks have to be performed prior to clustering web access patterns from web logs. In this experiment, we just need data cleaning and simple session identification. 15,534 web access patterns are extracted from an information resource after a web log is downloaded [11]. Assume these web access patterns are grouped into 5 clusters, the clustering results are shown in Table 1. The same data is clustered into groups by other algorithms. In this paper, we compute the optimal number of clusters in terms of the Davies-Bouldin cluster validity index [2], which is a function of the ratio of the sum of within-cluster distance to between-cluster separation. The optimal clustering method for c clusters minimizes

$$DB = \frac{1}{c}\sum_{k=1}^c \max_{l \neq k} \left\{\frac{S(C_k) + S(C_l)}{d(C_k, C_l)}\right\}, \qquad (12)$$

Where $S(C_k)$ is the within-cluster and $d(C_k, C_l)$ is the between-cluster separation.

Table 1. Clustering result

Cluster number	The included web access patterns
Cluster 1	4567
Cluster 2	3478
Cluster 3	3589

Cluster 4	1897
Cluster 5	2003

In order to evaluate the results of clustering algorithms, the experiments have been done and the results of the proposed algorithm are compared to the different algorithms. Table 2 shows the Davies-Bouldin cluster validity index between our approach with other algorithms.

Table 2. Comparison Davies-Bouldin Index (DB) with Other Algorithms

Clustering algorithm	Davies-Bouldin Index
LVQ	0.697
DLA based approach	0.574
Fuzzy c-means	0.411
LVQ+fuzzyc-means[10]	0.335
Our Approach	0.232

Table 2 shows the DB criterion comparison between the proposed algorithms to different algorithms. As it is in the table the factor of DB in the proposed approach is lower than other algorithms. This shows that the proposed algorithm in the clustering of web access patterns has a higher proficiency. If only the first step of proposed algorithm (DLA based approach) is used, it won't have appropriate proficiency.

5. Conclusion

The visited web page and the time duration on it reflect the interest of web users. In this paper, each web access pattern from web logs is transformed into a weight vector using the learning automata then we put each weight vector in the nearest cluster using the learning automata. By doing this, a primitive clustering is performed on the web access patterns and the primitive centers of clusters are determined. Finally, these primitive clusters which have no or several access patterns are used by Fuzzy weighted c-means clustering algorithm and on the basis of the weight of each cluster which has been determined according to the number of web access patterns in each cluster and also the centers of clusters which have been determined by the learning automata are clustered. In order to avoid disadvantage of single DLA based approach or LVQ or Fuzzy weighted c-means, a hybrid approach based on DLA and Fuzzy weighted c-means is proposed to cluster web access patterns from web logs. Using this approach, the surfing behaviors of web users can be more quickly and exactly disclosed which is useful to build adaptive web server and design personalized service according to users' surfing behaviors.

Acknowledgement

This work has been supported by a grant from the Islamic Azad University, Shabestar Branch with Number: 5195489020400

References

[1] H. Beigy, and M. R. Meybodi , "A Learning Automata based Algorithms for Determination of Minimum Number of Hidden Unites for Three Layers Neural Networks", Journal of Amirkabir , Vol. 48, No. 4, 2002, pp. 957-974.
[2] J. Bezdek, and N. Pal, "Some New Indexes for Cluster Validity", IEEE Transactions on Systems, Man, and Cybernetics, Part-B, 1998, Vol. 28, pp. 957-974.
[3] S. De, and P. Krishna, "Clustering Web Transactions Using Rough Approximation", Fuzzy Sets and Systems, 2004 , Vol. 148, pp. 131-138.
[4] M. R. Meybodi, and H. Beigy, "A Note on Learning Automata based Schemes for Adaptation of BP Parameters", Journal of Neuro Computing, Vol. 48, 2002, pp. 957-974.
[5] M. R. Meybodi, and S. Lakshmivarahan, "On A class of Learning Algorithms Which have Symmetric Behavior under Success and Failure", Lecture Notes in Statistics, Berlin: Springer Verlag, 1984, pp. 145-155.
[6] K. S. Narendra, and M. A. L., Thathachar, Learning Automata: An introduction, Prentice Hall, 1989.
[7] T. Runkler, and J. Beadek, "Web Mining with Relational Clustering". International Journal of Approximate Reasoning, Vol. 32, 2003, pp. 217-236.
[8] P. Shi, "An Efficient Approach for Clustering Web Access Patterns from Web Logs", International Journal of Advanced Science and Technology, Vol. 5, 2009, pp. 1- 13.
[9] J. Srivastava, R. Cooley, M. Deshpande, and P.-N., Tan,. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, ACMSIGKDD Explorations, 2000.
[10] R. Wu, "Clustering Web Access Patterns based on Hybrid Approach", Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008, Vol. 5, pp.52-56.
[11] http://ita.ee.lbl.gov/html/traces.html.

Babak Anari received the B.Sc. degrees in computer engineering from Azad University of Shabestar, Iran in 2002, and M.Sc. degrees from Azad University of Arak, Iran in 2007, respectively. He has some research papers in web mining and learning automata field.

Mohammad Reza Meybodi received the B.Sc. and M.Sc. degrees in economics from Shahid Beheshti University, Tehran, Iran, in 1973 and 1977, respectively, and the M.S. and Ph.D. degrees incomputer science from Oklahoma University, Norman, in 1980 and 1983, respectively. He is currently a Full Professor with Department of Computer Engineering and Information Technology, Amirkabir University of Technology, Tehran. Prior to his current position, he was an Assistant Professor with Western Michigan University, Kalamazoo, from 1983 to 1985 and an Associate Professor with Ohio University, Athens, from 1985 to 1991. He has many research papers in learning systems and in International Proceedings or in International Journals. He has many research papers in learning systems in International Proceedings or in International Journals.His research interests include learning systems, parallel algorithms, soft computing, and software development.

Zohreh Anari received the B.Sc. and M.Sc. degrees in computer engineering from Azad University of Shabestar, Iran in 2003 and

2009, respectively. She has some research papers in fuzzy web mining field. Her current research interests include learning automata, web mining and soft computing.

Improving the User Query for the Boolean Model Using Genetic Algorithms

Mohammad Othman Nassar[1], Feras Al Mashagba[2], and Eman Al Mashagba[3]

[1] Computer Information Systems, Amman Arab University, Amman, 11942, Jordan

[2] Computer Information Systems, Amman Arab University, Amman, 11942, Jordan

[3] Computer Information Systems, Irbid Private University, Irbid, 22110, Jordan

Abstract

The Use of genetic algorithms in the Information retrieval (IR) area, especially in optimizing a user query in Arabic data collections is presented in this paper. Very little research has been carried out on Arabic text collections. Boolean model have been used in this research. To optimize the query using GA we used different fitness functions, different mutation strategies to find which is the best strategy and fitness function that can be used with Boolean model when the data collection is the Arabic language. Our results show that the best GA strategy for the Boolean model is the GA (M2, Precision) method.

Keywords: *information retrieval, Boolean model, query optimization, genetic algorithms.*

1. Introduction

The resource discovery problem is concerned with how to find information interest among the vast and growing amount of information available, this resource discovery problem is one of the most pressing issues with the explosive growth of the Internet [7]. Information retrieval (IR) can be defined broadly as the study of how to determine and retrieve from a corpus of stored information the portions which are responsive to particular information needs [1]. The major information retrieval models includes: the vector space model, Boolean model, Fuzzy sets model and the probabilistic retrieval model. These models are used to find the similarity between the query and the documents in order to retrieve the documents that reflect the query. The similarity then used to evaluate the effectiveness of IR system using two measures: Precision which is a ratio that compares the number of relevant documents found to the total number of returned documents [8], and Recall which is the system's ability to retrieve all related documents of a query [2].

The problem with the IR models is that it may converge to a result that is only locally optimal, which means it may lead to form a query that is better than the original form but significantly poorer than another undetected form, so Genetic Algorithm (GA) can be used to solve this problem. A (GA) is an adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetics [3]. The Genetic algorithm (GA) approach has gained importance and popularity, as evident in the number of studies that have used it to improve different optimization procedures to be able to find a global solution in many problems.

In this paper, we will work on Boolean IR model to optimize the user query using different genetic algorithms settings (different mutation techniques, different fitness functions). As a test bed; we are going to use an Arabic data collection which was presented for the first time by [24]; this data set is composed from 242 documents and 59 queries, the correct answer for each query (relevant documents) is also known in advanced.

Arabic is the official language of over than twenty one Arab countries, and it is the religious language of more than one billion Muslims around the world. The Arabic language is unique and difficult language; the difficulty comes from several sources; amongst them: it differs syntactically, morphologically, and semantically from other Indo-European languages [13]. Compared to English, Arabic language is more sparsed, which means that for the same text length English words are repeated more often than Arabic words [14, 15]. Sparseness may negatively affect the retrieval quality in Arabic language because Arabic terms will get less weight compared to English. In written Arabic, most letters take many forms of writing. Also, there is a punctuation associated with some letters that may change the meaning of two identical words. Finally; comparing to English roots, Arabic roots are more complex. The same Arabic root, depending on the context, may be derived from multiple Arabic words.

Finally, we can say that the uniqueness and the special properties for the Arabic language, its differences from the English and the other languages, and the lack of similar studies in the literature was our motivator to conduct a deep and rich comparative study that apply different Genetic algorithm (GA) strategies using different mutation techniques and different fitness functions on the output of traditional IR system based on Boolean model in order to improve the user query.

2. Previous Studies

Query optimization is an active research area in IR, many studies have been conducted in this area based on English data collections [4,8,9,10,11,12,16,17,18,19,20,21,22,23]. Vaclav S, Dusan H [4] deals with Genetic algorithms to optimize the Boolean query in information retrieval system based on English data collection, in this study the authors used three different mutation criteria, they found that GA improves the performance compared to traditional approach, and the improvement is different from mutation criteria to another. Masaharu et al. [8] employed a few number of query terms and concept categories with Boolean expressions; they use only the words that exist in the original query for reformulating the Boolean query. Morgan and Kilgour employ GAs to choose search terms from a thesaurus and dictionary [12]. Unlike [8, 12]; in our study we used terms not only from the original query; but also from the retrieved documents. The authors in [9, 10, 11] examine GAs for information retrieval and they suggested new crossover and mutation operators, all of them used English data collections.

Other contributions towards evolutionary optimization of search queries were introduced by Kraft et al. [18]; they used genetic programming to optimize Boolean search queries only, and based on English data collection. Cordn et al. [19] introduced MOGA-P, an algorithm to deal with search query optimization as a multi-objective optimization problem and compared their approach with several other methods including Kraft's. Yoshioka and Haraguchi [20] introduced query reformulation interface to transform Boolean search queries into more efficient search expressions. Finally the researchers in [23] investigate evolutionary algorithms as a tool for the optimization of user queries and seek for its good settings.

Using GA to improve the performance of Arabic information system is rare in the literature. In [17] the researchers used Genetic Algorithms to improve performance of Arabic information retrieval system, which based on vector space model.

3. Boolean model

Retrieval systems based on Boolean logic have long served as the cornerstone of the commercial document retrieval system market and remain very important because of the relative simplicity of the query language and the ease with which it can be understood and implemented [5]. The most common use for a Boolean expression is to state what characteristics must be present in material to be retrieved in a system that retrieves and presents to users bibliographic records or full-text. A second use of Boolean expressions, likely to increase in importance over the next decade, is in rules incorporated into document and email filtering systems. Boolean expressions typically use three operators: AND, OR, and NOT.

4. Genetic Algorithms (GA)

A GA is an adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetics [3]. The basic concept of GA is designed to simulate processes in natural systems necessary for evolution. As such they represent an intelligent exploitation of a random search within a defined search space to solve a problem. GAs exploits the idea of the survival of the fittest and an interbreeding population to create a novel and innovative search strategy. A population of strings, representing solutions to a specified problem, is maintained by the GA. The GA then iteratively creates new populations from the old by ranking the strings and interbreeding the fittest to create new strings, which are hopefully closer to the optimum solution to the problem at hand. So in each generation, the GA creates a set of strings from the bits and pieces of the previous strings. The idea of survival of the fittest is of great importance to genetic algorithms. GAs use what is termed as a fitness function in order to select the fittest string that will be used to create new, and conceivably better, populations of strings. The only thing that the fitness function must do is to rank the strings in some way by producing the fitness value. These values are then used to select the fittest strings. The GA algorithm flowchart is illustrated in Figure 1.

Genetic algorithm operations can be used to generate new and better generations. As shown in Figure 1 the genetic algorithm operations include:

A. Reproduction: the selection of the fittest individuals based on the fitness function.

B. Crossover: is the exchange of genes between two individual chromosomes that are reproducing. In one point cross over [3] a chunk of connected

genes will be swapped between two chromosomes.

C. Mutation: is the process of randomly altering the genes in a particular chromosome. There are two types of mutation:

1) Point mutation: in which a single gene is changed.
2) Chromosomal mutation: where some number of genes is changed completely.

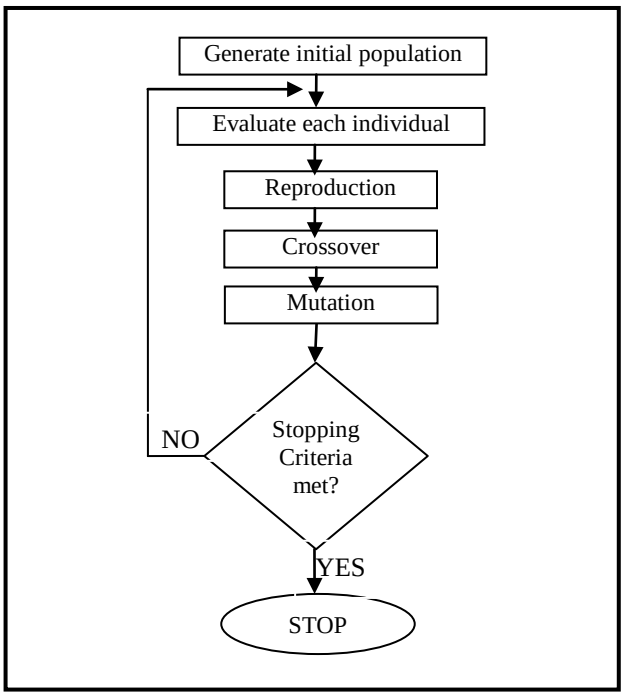


Fig. 1 Flowchart for Typical Genetic Algorithm.

5. Experiment (GA)

In this study we used IR system based on Boolean model and Fuzzy set model that was built and implemented by Hanandeh [6] to handle the 242 Arabic abstracts collected from the Proceedings of the Saudi Arabian National Conference [24]. The study was conducted as following:

1) Select the highest 15 terms frequency from the top 10 documents retrieved using the original query used by Hanandeh [6].
2) Construct 10 Queries from the selected terms.
3) Represent queries as a tree and calculate the fitness function which is either precision or recall for each query.
4) Select the best two queries.
5) Perform Crossover (one point crossover is used).
6) Perform Mutation (three different mutation techniques are used; for more details see the next section).
7) Update Population by replacing the new two queries with the worst two queries of the 10 Queries selected in step 2.
8) Go to step 3.

In order to use GA a set of parameters must be determined, these parameters are:

1) Number of generation: the number of iteration can be determined by predefined scale of accepted error, or can be defined before the GA starts. In this study the number of iterations used is 75 iterations.
2) Fitness Function Operator: Fitness function is a performance measure or reward function which evaluates how each solution is good. In this study precision and recall are used as two fitness functions.

$$\mathrm{Re}call = \frac{\sum_d [rd \times fd]}{\sum_d [rd]}$$

$$\Pr ecision = \frac{\alpha \sum_d [rd \times fd]}{\sum_d [rd]} + \frac{\beta \sum_d [rd \times fd]}{\sum_d [rd]}$$

Where rd is the number of relevance documents and fd is the number of retrieved document and α and β are arbitrary weights. In this study and based on previous studies [32] the value of α, β used is 0.25, 1.0 respectively.

3) Selection operator: In this study we used a single point crossover strategy with crossover probability Pc =0.8. The best two individuals with best fitness values are chosen from a population, and represented as trees. When the one point crossover is applied (i.e. if Random number < Probability of crossover) the two trees will exchange sub tree between them.

4) Mutation operators: In this experiment the mutation operator works as the most important operator for the learning of query. Each node from the new offsprings may be mutated; that depends on mutation probability (pm=0.2), this

mutation is applied if random number is less than probability of mutation. Different types of mutations are used in this study:

a) Mutation on Boolean operator: randomly exchanging one operator to another.
b) Mutation on term node (leaf node): in Boolean model one term is selected randomly from the offspring and replace by any other one from the terms in a given collection of documents. But in fuzzy model the term is not replaced, only the term weight is changed.
c) Mutation by inserting or deleting operator between two nodes in the offsprings.

As a result we create six different GA strategies for the Boolean and fuzzy models, those strategies are as following:

1) GA(M1,Precision): GA that use mutation on the operator and the precision as a Fitness Function.
2) GA(M2,Precision): GA that use Mutation on the term node (leaf node) and the precision as a Fitness Function.
3) GA(M3,Precision): GA that use Mutation by inserting or deleting operator between two nodes and the precision as a Fitness Function.
4) GA(M1,Recall): GA that use Mutation on the operator and the recall as a Fitness Function.
5) GA(M2,Recall): GA that use Mutation on the term node (leaf node) and the recall as a Fitness Function.
6) GA(M3,Recall): GA that use Mutation by inserting or deleting operator between two nodes and the recall as a Fitness Function.

6. Experiment Results

The results for the GA strategies based on Boolean model are shown in Table 1, Table 2. From those tables we notice that GA(M2,Precision), GA(M3,Precision) and GA(M2,Recall) give a high improvement than user query while GA(M1,Precision), GA(M1,Recall) and GA(M3,Recall) gives a low improvement than user query. We can also notice that GA(M2,Precision) gives the highest improvement over the user query in the Boolean model. The results for our experiments can be improved for the Boolean model by increasing the number of iterations for the GA, in one hand increasing the number of iterations will improve the performance, but in the other hand this will lead to increase the run time.

Table 1: Results when Precision was used as a Fitness Function in the Boolean Model.

Recall	Traditional	GA(M1)	GA(M2)	GA(M3)
0.1	0.156	0.161	0.169	0.146
0.2	0.162	0.162	0.173	0.187
0,3	0.166	0.167	0.179	0.174
0.4	0.178	0.169	0.191	0.172
0.5	0.188	0.178	0.203	0.182
0.6	0.221	0.213	0.232	0.219
0.7	0.223	0.219	0.243	0.225
0.8	0.241	0.236	0.256	0.233
0.9	0.245	0.239	0.265	0.239
Average	0.19777	0.19377	0.21233	0.19744

Table 2: Results when Recall was used as a Fitness Function in the Boolean Model.

Recall	Traditional	GA(M1)	GA(M2)	GA(M3)
0.1	0.156	0.152	0.166	0.145
0.2	0.162	0.159	0.169	0.157
0.3	0.166	0.167	0.173	0.168
0.4	0.178	0.176	0.187	0.176
0.5	0.188	0.187	0.194	0.179
0.6	0.221	0.219	0.228	0.211
0.7	0.223	0.226	0.238	0.216
0.8	0.241	0.245	0.251	0.232
0.9	0.245	0.243	0.261	0.241
Average	0.197778	0.19711	0.20744	0.19166

References

[1] Baeza-Yates, and Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, 1999.
[2] J. Carlberger, H. Dalianis, M. Hassel, O. Knutsson, "Improving Precision in Information Retrieval for Swedish using Stemming", In the Proceedings of NoDaLiDa-01 - 13th Nordic Conference on Computational Linguistics, May 21-22, 2001, Uppsala, Sweden.
[3] Goldberg, D. E., Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, 1989.
[4] Vaclav, S., and Dusan, H., "Using Genetic Algorithms for Boolean Queries Optimization", Ninth IASTED International Conference on Internet and Multimedia Systems and Application, ISBN 0-88986-510-8, 2005.
[5] Salton, G., McGill,M., Introduction to Modern Information Retrieval, McGraw-Hill, New-York, 1983.
[6] Hananda E, "Evaluation of Different Information Retrieval models and Different indexing methods on Arabic Documents", Phd Thesis, ARAB Academy, 2008.
[7] Yuwono, B., and Lee, D. L "WISE: A World Wide Web Resource Database System," IEEE Transaction on Knowledge and Data Engineering, ISSN: 1041-4347, Volume: 8 Issue:4, pp. 548-554, 1996.
[8] Masaharu, Y., and Makoto, H, "An Appropriate Boolean Query Reformulation Interface for Information Retrieval Based on Adaptive Generalization", International Conference on Data Engineering, pp. 148-153, 2005.

[9] M. Boughanem, C. Chrisment, and L. Tamine, “On using genetic algorithms for multimodal relevance optimization in information retrieval”, Journal of the American Society for Information Science and Technology, 53(11), pp. 934–942, 2002.
[10] J. T. Horng, and C. C. Yeh, “Applying genetic algorithms to query optimization in document retrieval”, Information Processing and Management, 36(5), pp. 737–759, 2000.
[11] D. Vrajitoru, “Crossover improvement for the genetic algorithm in information retrieval”, Information Processing& Management, 34(4), pp. 405–415, 1998.
[12] J. Morgan, and A. Kilgour, “Personalising on-line information retrieval support with a genetic algorithm”. In A. Moscardini, & P. Smith (Eds.), PolyModel 16: Applications of artificial intelligence, pp. 142–149, 1996.
[13] Khoja, S., "APT:Arabic part-of-speech tagger", proceedings of the student workshop at second meeting of north American chapter of Association for Copmputational Linguistics (NAACL2001), Pittsburgh, Pennsylvania, pp. 20-26, 2001.
[14] yahaya, A., "on the Complexity of the initial stage of Arabic text processing", First Great Lakes Computer Science Conference, Kalamazoo, Michigan, USA, October, 1989.
[15] Goweder, A., De Roeck, A., "Assessment of a Significant Arabic Corpus", Arabic Natural Language Processing Workshop (ACL2001), Toulouse, France. Downloaded from: (http://www.elsnet.org/acl2001 arabic.html).
[16] Kushchu, I., "Web-Based Evolutionary And Adaptive Information Retrieval", Evolutionary Computation, IEEE Transactions, Volume: 9, Issue: 2, ISSN: 1089-778X, pp. 117 – 125, 2005.
[17] Bassam Al-Shargabi, Islam Amro, and Ghassan Kanaan, "Exploit Genetic Algorithm to Enhance Arabic Information Retrieval", 3rd International Conference on Arabic Language Processing (CITALA’09), Rabat, Morocco, May 4-5, pp. 37-41, 2009.
[18] D. H. Kraft, F. E. Petry, B. P. Buckles, and T. Sadasivan, "Genetic Algorithms for Query Optimization in Information Retrieval: Relevance Feedback", In E. Sanchez, T. Shibata, and L.A. Zadeh, editors, Genetic Algorithms and Fuzzy Logic Systems Soft Computing Perspectives, Singapore, pp. 155-173, 1997.
[19] Oscar Cordn, Flix de Moya, and Carmen Zarco, "Fuzzy logic and multiobjective evolutionary algorithms as soft computing tools for persistent query learning in text retrieval environments", In IEEE International Conference on Fuzzy Systems 2004, ISBN: 0-7803-8353-2, pp. 571-576, Budapest, Hungary, 2004.
[20] Masaharu Yoshioka, and Makoto Haraguchi, "An Appropriate Boolean Query Reformulation Interface for Information Retrieval Based on Adaptive Generalization", WIRI '05 Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration, ISBN: 0-7695-2414-1, pages 145-150, 2005.
[21] Owais, S., Kromer, P., and Snasel, V., "Implementing GP on Optimizing Boolean and Extended Boolean Queries in IRs With Respect to Users Profiles", ISBN: 1-4244-0271-9 , pp. 412 – 417, 2006 .
[22] Simon, P., and Sathya, S.S., "Genetic algorithm for information retrieval", International Conference on Intelligent Agent & Multi-Agent Systems (IAMA 2009) , ISBN: 978-1-4244-4710-7, pp. 1 – 6, 2009.
[23] Kromer, P., Snásel, V., Platos, J., and Abraham, A., "Evolutionary improvement of search queries and its parameters", 2010 10th International Conference on Hybrid Intelligent Systems (HIS), pp. 147 - 152 ISBN: 978-1-4244-7363-2, 2010.
[24] I. Hmedi, and G. Kanaan and M. Evens, "design and implementation of automatic indexing for information retrieval with Arabic documents", Journal of American society for information science, Volume 48 Issue 10, pp. 867-881, 1997.

First Author Dr. Mohammad Othman Nassar is Assistant Professor at the Computer Information Systems department, faculty of computer sciences and informatics, Amman Arab University, Amman, Jordan. He worked as Assistant Professor at the Computer Information Systems department in the Arab Academy for Banking & Financial Sciences University. He holds a Ph.D. in Computer Information Systems from the Arab Academy for Banking & Financial Sciences University, Jordan. His research interests include information retrieval, supply chain management, reengineering, outsourcing, and security. Dr. Nassar has published more than 12 articles in these fields in various journals and international conferences. He is included in the Panel of referees of "International Journal of Modeling and Optimization" and in the "International Journal of Computer Theory and Engineering", he was reviewer in the 2011 3rd International Conference on Machine Learning and Computing, also he is currently reviewer in A collection of open access journals called (academic journals).

Second Author Dr. Feras Fares Al-mashakba is Assistant Professor at the Computer Information Systems department, faculty of computer sciences and informatics, Amman Arab University, Amman, Jordan. He holds a Ph.D. in Computer Information Systems from the Arab Academy for Banking & Financial Sciences University, Jordan. His research interests include information retrieval, artificial intelligence, M-commerce.

Third Author Dr. Eman Fares Al-mashakba is Assistant Professor at the Computer Information Systems department, faculty of computer sciences and informatics, Irbed University, Irbed, Jordan. She holds a Ph.D. in Computer Information Systems from the Arab Academy for Banking & Financial Sciences University, Jordan. His research interests include information retrieval, Security, E-learning and image processing.

A New Algorithm Based Entropic Threshold for Edge Detection in Images

Mohamed A. El-Sayed

CS dept, Faculty of Computers and Information Systems , Taif Univesity, 21974 Taif, KSA
Mathematics department, Faculty of Science, Fayoum University, 63514 Fayoum, Egypt

Abstract

Edge detection is one of the most critical tasks in automatic image analysis. There exists no universal edge detection method which works well under all conditions. This paper shows the new approach based on the one of the most efficient techniques for edge detection, which is entropy-based thresholding. The main advantages of the proposed method are its robustness and its flexibility. We present experimental results for this method, and compare results of the algorithm against several leading edge detection methods, such as Canny, LOG, and Sobel. Experimental results demonstrate that the proposed method achieves better result than some classic methods and the quality of the edge detector of the output images is robust and decrease the computation time.

Keywords: *Segmentation, Edge detection, Clustering, Entropy, Thresholding, Measures of information.*

1. Introduction

Edge detection is an important field in image processing. It can be used in many applications such as segmentation, registration, feature extraction, and identification of objects in a scene. An effective edge detector reduces a large amount of data but still keeps most of the important feature of the image. Edge detection refers to the process of locating sharp discontinuities in an image. These discontinuities originate from different scene features such as discontinuities in depth, discontinuities in surface orientation, and changes in material properties and variations in scene illumination. [6]

Many operators have been introduced in the literature, for example Roberts, Sobel and Prewitt [2, 3, 5 , 8, 12, 15]. Edges are mostly detected using either the first derivatives, called gradient, or the second derivatives, called Laplacien. Laplacien is more sensitive to noise since it uses more information because of the nature of the second derivatives.

Most of the classical methods for edge detection based on the derivative of the pixels of the original image are Gradient operators, Laplacian and Laplacian of Gaussian (LOG) operators [14].

Gradient based edge detection methods, such as Roberts, Sobel and Prewitts, have used two 2-D linear filters to process vertical edges and horizontal edges separately to approximate first-order derivative of pixel values of the image. Marr and Hildreth achieved this by using the Laplacian of a Gaussian (LOG) function as a filter [10]. The paper [11] classified and comparative studies of edge detection algorithms are presented. Experimental results prove that Boie-Cox, Shen- Castan and Canny operators are better than Laplacian of Gaussian (LOG), while LOG is better than Prewitt and Sobel in case of noisy image.

The paper [6] used 2-D gamma distribution, the experiment showed that the proposed method obtained very good results but with a big time complexity due to the big number of constructed masks.

To solve these problems, the study proposed a novel approach based on information theory, which is entropy-based thresholding. The proposed method is decrease the computation time. The results were very good compared with the well-known Sobel gradient [7] and Canny [4] gradient results.

This paper is organized as follows: in Section 2 presents some fundamental concepts of the mathematical setting of the threshold selection. Section 3, we describe the proposed method of edge detection, and we describe the proposed algorithm. In Section 4, we report the effectiveness of our method when applied to some real-world and synthetic images, and compare results of the algorithm against several leading edge detection methods, such as Canny, LOG, and Sobel method. At last conclusion of this paper will be drawn in Section 5.

2. Image thresholding

The set of all source symbol probabilities is denoted by P, $P= \{p_1, p_2, p_3, ..., p_k\}$. This set of probabilities must satisfy the condition $\sum_{i=1}^{k} p_i = 1$, $0 \leq p_i \leq 1$. The average

information per source output, denoted $S(Z)$ [9], Shannon entropy may be described as:

$$S(Z) = -\sum_{i=1}^{k} p_i \ ln(p_i) \qquad (1)$$

being k the total number of states. If we consider that a system can be decomposed in two statistical independent subsystems A and B, the Shannon entropy has the extensive property (additivity) $S(A+B) = S(A) + S(B).$, this formalism has been shown to be restricted to the Boltzmann-Gibbs-Shannon (BGS) statistics.

However, for non-extensive systems, some kind of extension appears to become necessary. Tsallis [13] has proposed a generalization of the BGS statistics which is useful for describing the thermo statistical properties of non-extensive systems. It is based on a generalized entropic form,

$$S_q = \frac{1}{q-1}\left(1 - \sum_{i=1}^{k} p_i^q\right) \qquad (2)$$

where the real number q is a entropic index that characterizes the degree of non-extensivity. This expression recovers to BGS entropy in the limit $q \to 1$. Tsallis entropy has a non-extensive property for statistical independent systems, defined by the following rule [1] :

$$S_q(A+B) = S_q(A) + S_q(B) + (1-q)\ .S_q(A)\ .S_q(B). \qquad (3)$$

Similarities between Boltzmann-Gibbs and Shannon entropy forms give a basis for possibility of generalization of the Shannon's entropy to the Information Theory. This generalization can be extended to image processing areas, specifically for the image segmentation, applying Tsallis entropy to threshold images, which have non-additive information content.

Let $f(x, y)$ be the gray value of the pixel located at the point (x, y). In a digital image $\{ f(x,y) \mid x \in \{1,2,...,M\}, y \in \{1,2,...,N\}\}$ of size $M \times N$, let the histogram be $h(a)$ for $a \in \{0,1,2,..., 255\}$ with f as the amplitude (brightness) of the image at the real coordinate position (x, y). For the sake of convenience, we denote the set of all gray levels $\{0,1,2,..., 255\}$ as G. Global threshold selection methods usually use the gray level histogram of the image. The optimal threshold t^* is determined by optimizing a suitable criterion function obtained from the gray level distribution of the image and some other features of the image.

Let t be a threshold value and $B = \{b_0, b_1\}$ be a pair of binary gray levels with $\{b_0, b_1\} \in G$. Typically b_0 and b_1 are taken to be 0 and 1, respectively. The result of thresholding an image function $f(x, y)$ at gray level t is a binary function $f_t(x, y)$ such that $f_t(x,y) = b_0$ if $f_t(x,y) \le t$ otherwise, $f_t(x,y) = b_1$. In general, a thresholding method determines the value t^* of t based on a certain criterion function. If t^* is determined solely from the gray level of each pixel, the thresholding method is point dependent [9].

Let $p_i = p_1, p_2, \ldots, p_k$ be the probability distribution for an image with k gray-levels. From this distribution, we derive two probability distributions, one for the object (class A) and the other for the background (class B), given by:

$$p_A : \frac{p_1}{P_A}, \frac{p_2}{P_A}, \ldots, \frac{p_t}{P_A}, \qquad p_B : \frac{p_{t+1}}{P_B}, \frac{p_{t+2}}{P_B}, \ldots, \frac{p_k}{P_B} \qquad (4)$$

and where

$$P_A = \sum_{i=1}^{t} p_i \quad , \quad P_B = \sum_{i=t+1}^{k} p_i \qquad (5)$$

The Tsallis entropy of order q for each distribution is defined as:

$$S_q^A(t) = \frac{1}{q-1}\left(1 - \sum_{i=1}^{t} p_A^q\right), \quad \text{and} \quad S_q^B(t) = \frac{1}{q-1}\left(1 - \sum_{i=t+1}^{k} p_B^q\right) \qquad (6)$$

The Tsallis entropy $S_q(t)$ is parametrically dependent upon the threshold value t for the foreground and background. It is formulated as the sum each entropy, allowing the pseudo -additive property, defined in equation (2). We try to maximize the information measure between the two classes (object and background). When $S_q(t)$ is maximized, the luminance level t that maximizes the function is considered to be the optimum threshold value [4].

$$t^*(q) = Arg \max_{t \in G}[S_q^A(t) + S_q^B(t) + (1-q).S_q^A(t).S_q^B(t)]. \qquad (7)$$

In the proposed scheme, first create a binary image by choosing a suitable threshold value using Tsallis entropy. The technique consists of treating each pixel of the original image and creating a new image, such that $f_t(x,y) = 0$ if $f_t(x,y) \le t^*(q)$ otherwise, $f_t(x,y) = 1$ for every $x \in \{1,2,...,M\}$, $y \in \{1,2,...,N\}$.

When $q \to 1$, the threshold value in Equation (2), equals to the same value found by Shannon's method. Thus this proposed method includes Shannon's method as a special case. The following expression can be used as a criterion function to obtain the optimal threshold at $q \to 1$.

$$t^*(1) = Arg \max_{t \in G}[S^A(t) + S^B(t)]. \quad (8)$$

The *Threshold* procedure to select suitable threshold value t^* and q can now be described as follows:

Procedure *Threshold*,
Input: A digital grayscale image A of size $M \times N$.
Output: The suitable threshold value t^* of A, for $q \geq 0$.
Begin
1. Let $f(x, y)$ be the original gray value of the pixel at the point (x, y), $x=1..M$, $y=1..N$.
2. Calculate the probability distribution $0 \leq p_i \leq 255$.
3. For all $t \in \{0,1,\ldots,255\}$,
   i. Apply Equations (4 and 5) to calculate P_A, P_B, p_A and p_B .
   ii. Apply Equation (7) to calculate optimum threshold value t^* .

End.

3. The edge detection

We will use the usual masks for detecting the edges. A spatial filter mask may be defined as a matrix w of size $m \times n$. Assume that $m=2\alpha+1$ and $n=2\beta+1$, where α, β are nonzero positive integers. For this purpose, smallest meaningful size of the mask is 3×3. Such mask coefficients, showing coordinate arrangement as Figure 1.a .

$w(-1,-1)$	$w(-1,0)$	$w(-1,1)$
$w(0,-1)$	$w(0,0)$	$w(0,1)$
$w(1,-1)$	$w(1,0)$	$w(1,1)$

Fig 1.a

$f(x-1, y-1)$	$f(x-1,y)$	$f(x-1, y+1)$
$f(x, y-1)$	$f(x, y)$	$f(x, y+1)$
$f(x+1, y-1)$	$f(x+1,y)$	$f(x+1, y+1)$

Fig. 1.b

1	1	1
1	×	1
1	1	1

Fig. 1.c

Image region under the above mask is shown as Figure 1.b. In order to edge detection, firstly classification of all pixels that satisfy the criterion of homogeneousness, and detection of all pixels on the borders between different homogeneous areas. In the proposed scheme, first create a binary image by choosing a suitable threshold value using Tsallis entropy. Window is applied on the binary image. Set all window coefficients equal to 1 except centre, centre equal to × as shown in Figure 1.c.

Move the window on the whole binary image and find the probability of each central pixel of image under the window. Then, the entropy of each central pixel of image under the window is calculated as $S(CPix) = -p_c \ln(p_c)$.

Table 1: p and S of central under window

P	***S***	***P***	***S***
1/9	0.2441	6/9	0.2703
2/9	0.3342	7/9	0.1955
3/9	0.3662	8/9	0.1047
4/9	0.3604	9/9	0.0
5/9	0.3265	-	-

where, p_c is the probability of central pixel $CPix$ of binary image under the window. When the probability of central pixel, p_c =1, then the entropy of this pixel is zero. Thus, if the gray level of all pixels under the window homogeneous, p_c= 1 and S =0. In this case, the central pixel is not an edge pixel. Other possibilities of entropy of central pixel under window are shown in Table 1.
In cases p_c= 8/9, and p_c=7/9 , the diversity for gray level of pixels under the window is low. So, in these cases, central pixel is not an edge pixel. In remaining cases, $p_c \leq 6/9$, the diversity for gray level of pixels under the window is high.

The complete ***EdgeDetection*** **Procedure** can now be described as follows:

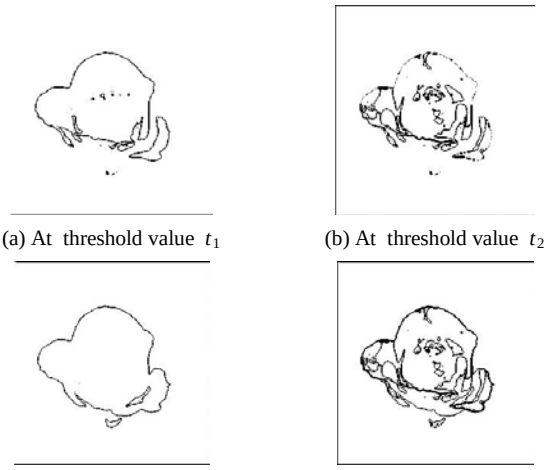
Procedure *EdgeDetection*;
Input: A grayscale image A of size $M \times N$ and t^* .
Output: The edge detection image g of A.
Begin
Step 1: Create a binary image: For all x, y,
If $f(x, y) \leq t^*$ then $f(x, y) = 0$ Else $f(x, y) = 1$.
Step 2: Create a mask, w, with 3×3, $a = (m-1)/2$ and $b = (n-1)/2$. (see Figure 1)
Step 3: Create an M×N output image, g: For all x and y, Set $g(x, y) = f(x, y)$.
Step 4: Checking for edge pixels:
For all $y \in \{ b+1, \ldots, N-b\}$, and $x \in \{ a+1, \ldots, M-a\}$,
sum = 0;
For all $k \in \{ -b, \ldots, b\}$, and $j \in \{ -a, \ldots, a\}$,
If ($f(x, y) = f(x+j, y+k)$) Then $sum = sum+1$.
If (sum >6) Then $g(x,y)=0$ Else $g(x,y)=1$.

(see Table 1)
End Procedure.

The steps of proposed algorithm are as follows:
Step1: We use Shannon entropy, the equation (8), to find the global threshold value (t_1). The image is segmented by t_1 into two parts, the object and the background. See Figure 2.

(a) Original image (b) Part1 (c) Part2

Fig. 2. Original image , and its parts, Part1 and Part2.



(a) At threshold value t_1 (b) At threshold value t_2

(c) At threshold value t_3 (d) final edge image.

Fig. 3 Edge images of original image , its parts, Part1 and Part2 and final output of edge image

Step2: We use Tsallis entropy, the equation (7) , q=0.5, Since, we can write the Equation (6) as:

$$S_{0.5}^{A}(t) = 2\sum_{i=1}^{t}\left|\sqrt{p_A}\right| - 2 \ , \quad \text{and} \ S_{0.5}^{B}(t) = 2\sum_{i=t+1}^{k}\left|\sqrt{p_B}\right| - 2 \tag{9}$$

Therefore, we have

$$t^{*}(0.5) = Arg\max_{t\in G}[(\sum_{i=1}^{t}\left|\sqrt{p_A}\right|)(\sum_{i=t+1}^{k}\left|\sqrt{p_B}\right|) - 1]. \tag{10}$$

Applying the equation (10), to find the locals threshold values (t_2) and (t_3) of Part1 and Part2, respectively.

Step3: Applying *EdgeDetection* Procedure with threshold values t_1, t_2 and t_3. See Figure 3 .a-c

Step4: Merge the resultant images of step 3 in final output edge image. See Figure 3.d

In order to reduce the run time of the proposed algorithm, we make the following steps:
Firstly, the run time of arithmetic operations is very much on the M×N big digital image, I , and its two separated regions, Part1 and Part2. We are use the linear array p (probability distribution) rather than I , for segmentation operation, and threshold values computation t_1, t_2 and t_3.
Secondly, rather than we are create many binary matrices f and apply the edge detector procedure for each region individually, then merge the resultant images into one. We are create one binary matrix f according to threshold values t_1, t_2 and t_3 together, then apply the edge detector procedure one time. This modifications will reduce the run time of computations.
The following procedures summarize the proposed technique for calculating the optimal threshold values and the edge detector. The above procedures can be done together in the following MATLAB program:

MainProgram.m file

```
I=imread('Lena.tif');
[M,N]=size(I);
p = zeros(256,3);
for ii=1:256 p(ii,1)=ii-1; end;
p(:,2) = imhist(I);
p (p(:,2)==0,:) = []; % remove zero entries in p
% Calling Shannon procedure, return t1 value and
its location in p
[T1,Loc]=Shannon(p);
% Calling Tsallis procedure of Part1
pLow= p(1:Loc,:);          T2= Tsallis_Sqrt(pLow);
% Calling Tsallis procedure of Part2
pHigh=p(Loc+1:size(p),:); T3=Tsallis_Sqrt(pHigh);
% Cerate binary matrices  f
f=zeros(M,N);
for i=1:M;
  for j=1:N;
     if ((I(i,j)>= T2)&(I(i,j)<T1))|(I(i,j)>= T3)
         f(i,j)=1; end;
   end;
end
% Calling EdgeDetector procedure, return edge
detection image.
[g]= EdgeDetector(f);
figure;
imshow(g);
```

Shannon.m file

```
function [T,Loc]=Shannon(p)
p(:,3) = p(:,2)./ sum(p(:,2)); % normalize p so
that sum(p) is one.
[M1, N1]=size(p); Max1= 0;
for t=1 : size(p)
    PA= sum(p(1:t,3));
    PB= 1-PA ;
    p1=p(1:t,3)./PA; % p1 is i probability in PA
    p2=p(t+1:M1,3)./PB; %p2 is i prob. in PB
    Sa= -sum(p1.*log2(p1));
    Sb= -sum(p2.*log2(p2));
```

```
    Sab= Sa + Sb;
    if(Sab>Max1)  T=p(t,1); Loc=t;Max1=Sab; end;
end;
```

Tsallis_Sqrt.m file

```
function [T]=Tsallis_Sqrt(p)
p(:,3) = p(:,2)./ sum(p(:,2));
[M1, N1]=size(p); Max1= 0;
for t=1 : M1
    PA= sum(p(1:t,3));
    PB= 1-PA ;
    p1= p(1:t ,3)./PA;
    p2= p(t+1 :M1 ,3)./PB;
    Tab=sum(sqrt(p1))*sum (sqrt(p2))-1;
    if ( Tab > Max1 ) T=p(t,1); Max1=Tab; end;
end;
```

EdgeDetector.m file

```
function [g]=EdgeDetector(f);
[M,N]=size(f);
g=zeros(M,N);
for y=2 : N-1;
  for x=2 : M-1;
     sum1=0;
     for k=-1:1;
          for j=-1:1;
            if(f(x,y)==f(x+j,y+k))
              sum1=sum1+1;
            end;
          end;
     end;
     if (sum1<=6)  g(x,y)=1;  end;
  end;
end;
```

4. Results and Discussions

In order to test the method proposed in this paper and compare with the other edge detectors, common gray level test images with different resolutions and sizes are detected by Canny, LOG, and Sobel and the proposed method respectively. The performance of the proposed scheme is evaluated through the simulation results using MATLAB. Prior to the application of this algorithm, no pre-processing was done on the tested images.

As the algorithm has two main phases – global and local enhancement phase of the threshold values and detection phase, we present the results of implementation on these images separately. Here, we have used in addition to the original gray level function $f(x, y)$, a function $g(x, y)$ that is the average gray level value in a 3×3 neighborhood around the pixel (x, y).

The proposed algorithm used the good characters of each Shannon entropy and Tsallis entropy, together, to calculate the global and local threshold values. Hence, we ensure that the proposed algorithm done better than the algorithms that based on Shannon entropy or Tsallis entropy separately.

Though the performance of the proposed entropic edge detector excels as a shape and detail detector, it is fraught with some drawbacks. It fails to provide all thinned edges. The presence of thick edges at some locations needs to be addressed by the proper choice of parameter q. The weak edges are not eliminated but for some applications, these may be required. This detector has another distinctive feature, i.e. it retains the texture of the original image. This feature can be utilized for the identification of fingerprints, where the ridges may have different intensities. As the success of the edge detection depends on these parameters, we are experimenting on several images to come up with a useful selection guideline with $0<q<1$.

We run the Canny, LOG, and Sobel methods and the proposed algorithm 10 times for each image with different sizes. As shown in Figures 4-7, The charts of the test images and the average of run time for the classical methods and proposed scheme. It has been observed that the proposed edge detector works effectively for different gray scale digital images as compare to the run time of Canny method.

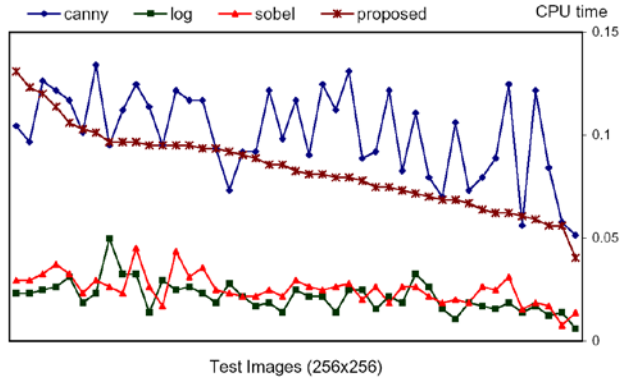


Fig. 4. CPU time of Canny, LOG, Sobel, and proposed method with 256×256 pixel test images

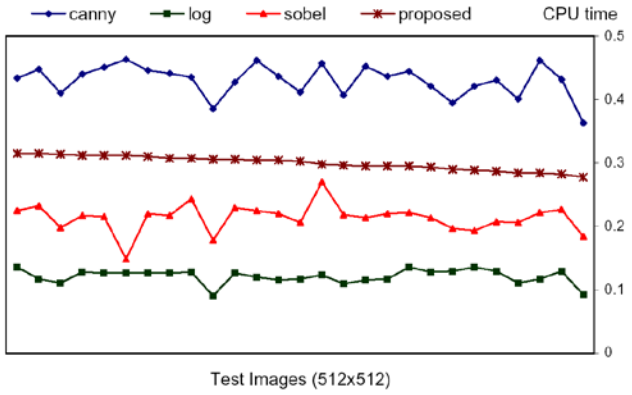


Fig. 5. CPU time of Canny, LOG, Sobel, and proposed method with 512×512 pixel test images

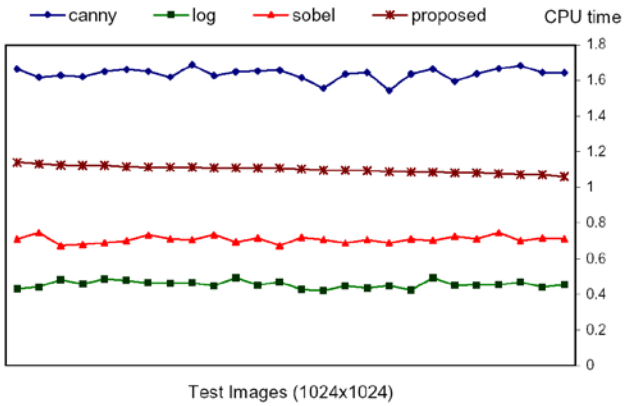


Fig. 6. CPU time of Canny, LOG, Sobel, and proposed method with 1024×1024 pixel test images

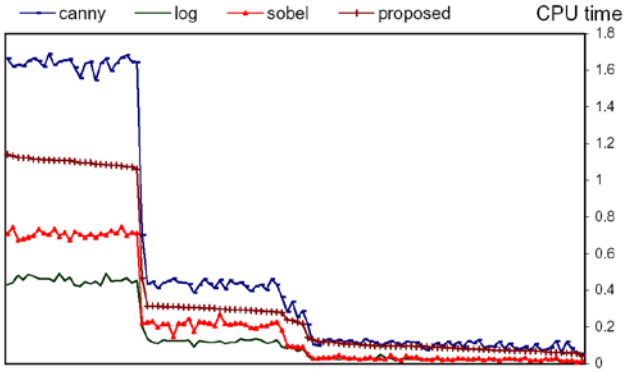


Fig. 7. CPU time of Canny, LOG, Sobel, and proposed method with different size test images

Some selected results of edge detections for these test images using the classical methods and proposed scheme are shown in Figures 8-12. From the results; it has again been observed that the proposed method works well as compare to the previous methods, LOG and Sobel (with default parameters in MATLAB).

5. Conclusion

The hybrid entropic edge detector presented in this paper uses both Shannon entropy and Tsallis entropy, together. It is already pointed out in the introduction that the traditional methods give rise to the exponential increment of computational time. However, the proposed method is decrease the computation time with generate high quality of edge detection. Experiment results have demonstrated that the proposed scheme for edge detection works satisfactorily for different gray level digital images. Another benefit comes from easy implementation of this method.

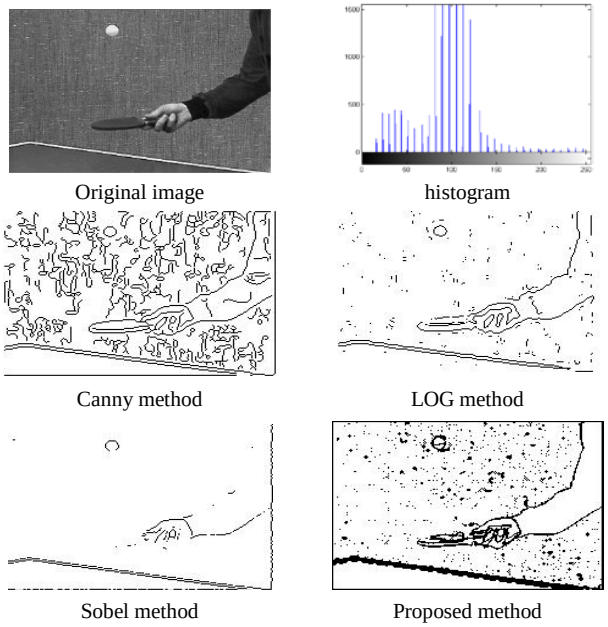


Fig. 8. Sport image with 224×153 pixel .

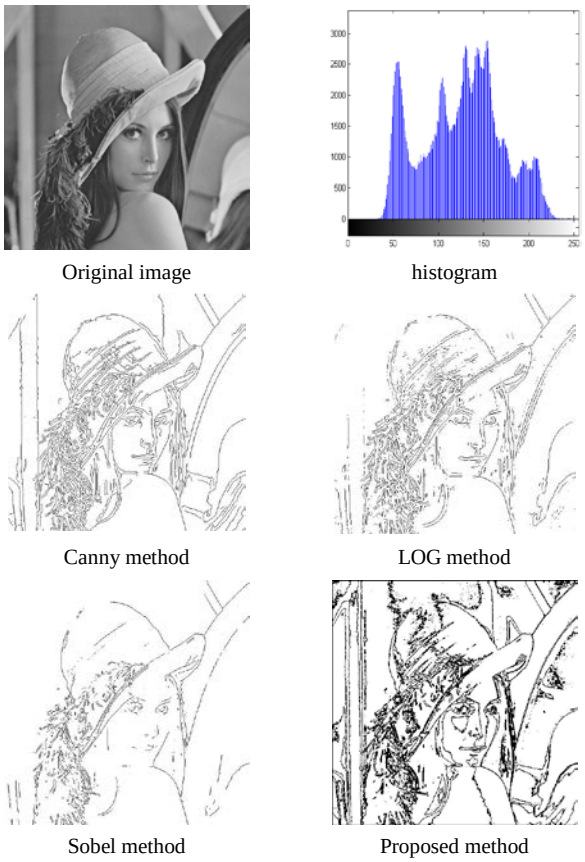


Fig. 9. Lena image with 512×512 pixel.

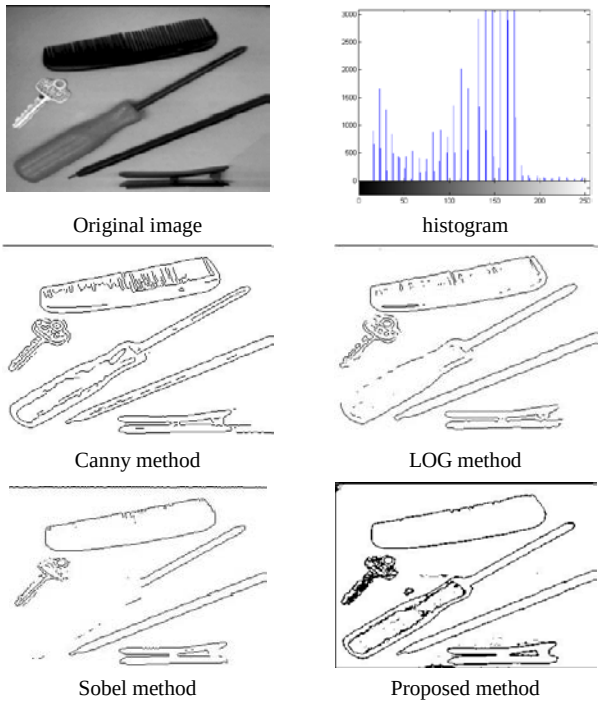


Fig. 10. Tools image with 322×228 pixel.

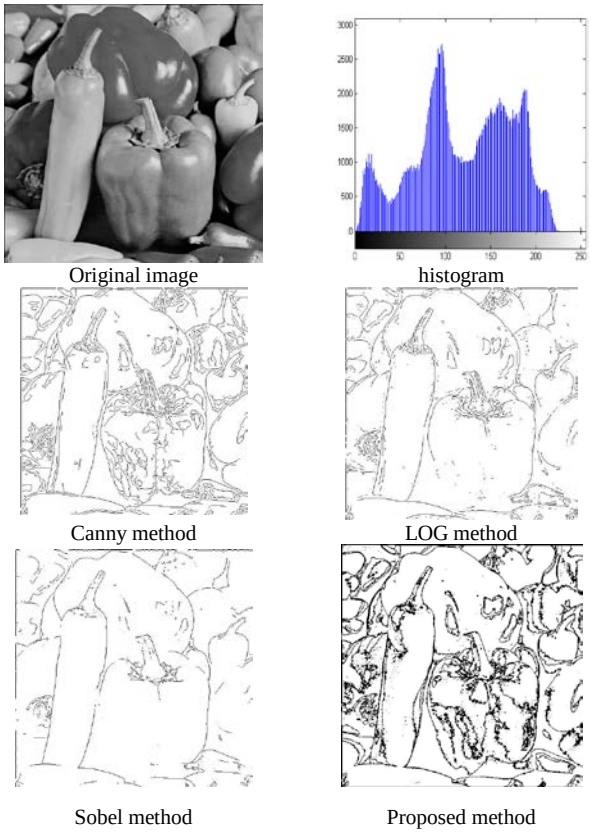


Fig. 11. Peppers image with 512×512 pixel.

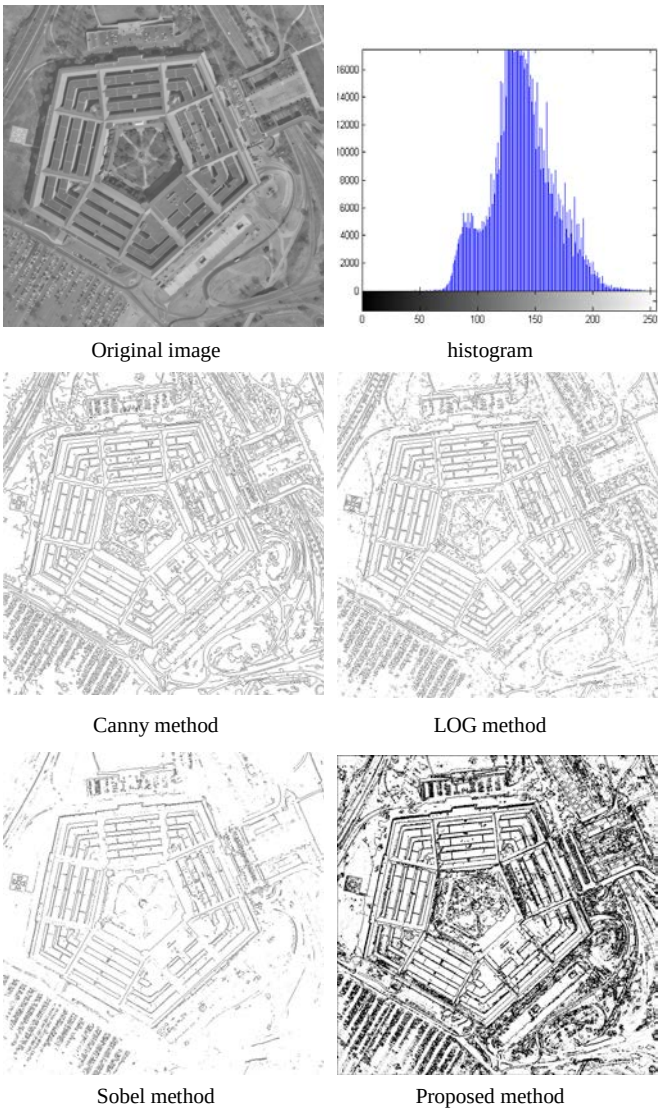


Fig. 12, pentagon image with 1024×1024 pixel.

References

[1] M. P. de Albuquerque, I. A. Esquef , A.R. Gesualdi Mello, "Image Thresholding Using Tsallis Entropy." Pattern Recognition Letters 25, 2004, pp. 1059 – 1065.

[2] V. Aurich, and J. Weule, "Nonlinear Gaussian filters performing edge preserving diffusion. ", Proceeding of the 17th Deutsche Arbeitsgemeinschaft für Mustererkennung (DAGM) Symposium, Sept. 13-15, Bielefeld, Germany, Springer-Verlag, 1995, pp. 538-545 .

[3] M. Basu, "A Gaussian derivative model for edge enhancement.", Patt. Recog., 27:1451-1461, 1994.

[4] J. Canny, "A computational approach to edge detection.", IEEE Trans. Patt. Anal. Mach. Intell., 8, 1986, pp. 679-698.

[5] G. Deng, and L.W. Cahill, "An adaptive Gaussian filter for noise reduction and edge detection.", Proceeding of the IEEE Nuclear Science Symposium and Medical Imaging Conference, Oct. 31-Nov. 6, IEEE Xplore Press, San Francisco, CA., USA, 1993, pp. 1615-1619.

[6] A. El-Zaart, "A Novel Method for Edge Detection Using 2 Dimensional Gamma Distribution", Journal of Computer Science 6 (2), 2010 , pp. 199-204,.

[7] R.C. Gonzalez, and R.E. Woods, "Digital Image Processing.", 3rd Edn., Prentice Hall, New Jersey, USA. ISBN: 9780131687288, 2008, pp. 954.

[8] C. Kang, and W. Wang, "A novel edge detection method based on the maximizing objective function.", Pattern. Recog., 40, 2007, pp. 609-618.

[9] F. Luthon, M. Lievin and F. Faux, "On the use of entropy power for threshold selection." Int. J. Signal Proc., 84, 2004, pp. 1789-1804.

[10] B. Mitra, "Gaussian Based Edge Detection Methods- A Survey ". IEEE Trans. on Systems, Manand Cybernetics , 32, 2002, pp. 252-260.

[11] M. Roushdy, "Comparative Study of Edge Detection Algorithms Applying on the Grayscale Noisy Image Using Morphological Filter", GVIP, Special Issue on Edge Detection, 2007, pp. 51-59.

[12] J. Siuzdak, "A single filter for edge detection.", Pattern Recog., 31, 1998 , pp.1681-1686.

[13] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," J. Stat. Phys., 52,1988, pp. 479–487.

[14] M. Wang and Y. Shuyuan, "A Hybrid Genetic Algorithm Based Edge Detection Method for SAR Image", In: IEEE Proceedings of the Radar Conference'05 May 9-12, 2005, pp. 1503-506.

[15] Q. Zhu, "Efficient evaluations of edge connectivity and width uniformity.", Image Vis. Comput., 14. 1996, pp.21-34.

Biography

Mohamed A. El-Sayed received his B.Sc. from Faculty of Science (Maths & CS in June 1994) at Minia University - Egypt. His M.Sc. from Faculty of Science (Maths\CS) at South Valley University in 2002. Ph.D. degrees in Computer Science from Faculty of Science at Minia University in 2007. His research interests include image processing, computer graphic and graphs drawing. He has published several international journals and Conference papers in the above area. He is working in Mathematics department, Faculty of Science, Fayoum University, Egypt. Currently, he is an Assistant professor of Computer Science at Faculty of Computers and Information Science , Taif Univesity, KSA.

An Enhanced Two-Stage Impulse Noise Removal Technique based on Fast ANFIS and Fuzzy Decision

V. Saradhadevi[1] and Dr.V.Sundaram[2]

[1]Research Scholar, Karpagam University, Coimbatore, India.
[2] Director of MCA , Karpagam Engineering College, Coimbatore, India.

Abstract

Image enhancement plays a vital role in various applications. There are many techniques to remove the noise from the image and produce the clear visual of the image. Moreover, there are several filters and image smoothing techniques available in the literature. All these available techniques have certain limitations. Recently, neural networks are found to be a very efficient tool for image enhancement. A novel two-stage noise removal technique for image enhancement and noise removal is proposed in this paper. In noise removal stage, Adaptive Neuro-Fuzzy Inference System (ANFIS) with a Modified Levenberg-Marquardt training algorithm was used to eliminate the impulse noise. The usage of Modified Levenberg-Marquardt training algorithm will reduce the execution time. In the image enhancement stage, the fuzzy decision rules inspired by the Human Visual System (HVS) are used to categorize the image pixels into human perception sensitive class and nonsensitive class, and to enhance the quality of the image. The Hyper trapezoidal fuzzy membership function is used in the proposed technique. In order to improve the sensitive regions with higher visual quality, a Neural Network (NN) is proposed. The experiment is conducted with standard image. It is observed from the experimental result that the proposed FANFIS shows significant performance when compared to existing methods.

Keywords--- *Fuzzy Decision, Impulse Noise, Peak Signal to Noise Ratio (PSNR), Modified Levenberg-Marquardt Training Algorithm, Adaptive Neuro-Fuzzy Inference System*

1. Introduction

IMAGES are usually affected by means of impulse noise because of noisy sensors or channel transmission mistakes. The main aim of noise reduction is to smother the noise, and also probably to safeguard the sharpness of edge and feature information. The nonlinear filtering method—standard median (SM) [9], [10] filter—according to order statistic, has been discussed to be usually better than linear filtering in reducing the impulse noise. Conversely, the median filter is inclined to blur fine details and demolish edges while smoothing out the impulse noise. For producing better result, the median filter has been altered in several manners. Those were expected to raise the signal preservation but relatively reduce the noise reduction capability, applying these algorithms altogether.

In the majority of applications, it is very essential to suppress the impulse noise [11] from image because the performances of subsequent image processing techniques are strictly reliable on the accomplishment of image noise removal process [12]. Conversely, this is very difficult in any image processing technique since the restoration filter must not alter the useful data in the image and conserve image data and texture during noise removal. The existing noise removal filters generally have the demerits of inducing unwanted distortions and blurring effects into the resulted image during noise removal phase.

So a novel approach is required for noise removal and image enhancement. Two-stage techniques that integrate noise identification and image enhancement have been proposed to eliminate the noise and keep the detail information well [13].

Since neural networks (NNs) have the ability to learn from examples, and fuzzy systems have the ability to deal with uncertainty, they also have a growing number of applications in image noise removal in the past few years [16]. Moreover, fuzzy techniques are also used to detect impulse noise. These methods exhibit relatively better performance but require more computation and memory cost. It is desired to improve the quality of noise removal and reduce the time consumption at the same time.

A new two-stage noise removal technique to deal with impulse noise is proposed in this paper. An Adaptive Neuro-Fuzzy Inference System is designed for fast and accurate noise detection such that various widespread densities of noisy pixels can be distinguished from the detail edge pixels well. The proposed ANFIS uses Modified Levenberg-Marquardt Training Algorithm for reducing the execution time. After suppressing the impulse noise, the image quality enhancement is applied to compensate the corrupted pixels to enhance the visual quality of the resultant images. It consists of fuzzy decision rules based on the Human Visual System (HVS) for image analysis and an NN for image quality enhancement. If a noise-corrupted pixel is in the perception sensitive region, the proposed NN module is applied to this pixel for further quality compensation.

2. Related Works

Schulte et al., [1] proposed a fuzzy two-step filter for impulse noise reduction from color images. A novel method for suppressing impulse noise [4] from digital images is provided in this paper, in which a fuzzy detection process is followed by an iterative fuzzy filtering method [7]. The filter proposed by author is called as fuzzy two-step color filter.

Sun et al., [2] provided an impulse noise image filter using fuzzy sets. The successful use of fuzzy set theory

performance on many domains, together with the increasing requirement for processing digital images, have been the main intentions following the efforts concentrated on fuzzy sets [5, 6]. Ibrahim et al., [3] given a simple adaptive median filter for the removal of impulse noise from highly corrupted images.

A novel impulse noise removal approach based on wavelet neural network is applied to restore digital images corrupted by impulse noise. Initially, wavelet neural network is applied to identify the noise-pixels and differentiate it from noise-free pixels. Then, the noise-pixels are categorized further by equivalent threshold and assigned the coefficient. Ultimately, the median filter is combined with the coefficient for the output. The proposed approach effectively eliminates the impulse noise while preserving more fine details. Visual evaluation and detailed statistical analysis show that the proposed technique is very significant than the conventional filters.

3. System Architecture

Optimal noise removal should delete the visible noise as cleanly as possible and maintain the detail information and natural appearance to obtain a natural-looking image. In order to remove the impulse noise cleanly from input images without blurring the edge, the proposed system is divided into two stages.

1. Impulse Noise Removal

2. Image Enhancement

In impulse noise removal stage, the impulse noise is removed without affecting too much detail information, and then, the image quality enhancement is applied to compensate the edge sharpness in the second stage. The two-level NN noise removal process is shown in Figure 1. Inside the first level, only the noisy pixels identified by the NN detection are processed with the 3 × 3 median filter. The second-level noise removal process is used to detect and remove the misclassified and the detected but un-removed noise pixels in the first-level noise removal process with an adaptive median filter.

The 3X3 window is used in this stage to get the features equivalent to the pixel P (O, O) for noise detection.

Figure 2 shows the schematic block diagram of the second stage image quality enhancement system. The proposed approach contains a fuzzy decision module, an angle evaluation module, and an adaptive compensation module. A fuzzy decision module based on the HVS categories each reference pixel O (0, 0) as sensible delineated edge or not. Depending on this category, the proposed adaptive NN compensation module is applied to the sensible delineated edge region. When the adaptive NN compensation is activated, the angle evaluation section will estimate the dominant orientation of the original image present in the sliding block as the input data of the proposed NN. The 4 × 4 window is applied at this stage to get the features equivalent to the pixel O(0, 0) for HVS-based image compensation.

The weighted compensation of O (0, 0) is applied to the noise-corrupted pixel F(m, n) at the position (m, n) in the sensible delineated edge region and can be presented as

$$F(m,n) = \sum_{i=-1}^{2} \sum_{j=-1}^{2} O(i,j) W_{\theta}(i,j) \quad (1)$$

where W_{θ} is derived from an NN after offline training. The NN is trained according to the edge angle of the reference image pixel to obtain the corresponding weights.

4. Proposed Impulse Noise Removal

4.1. Impulse Noise Model

Impulse noise is when the pixels are randomly failed and replaced by other values in an image. The image model containing impulse noise can be described as follows:

$$X_{ij} = \begin{cases} N_{ij}, & with\ probability\ p \\ S_{ij}, & with\ probability\ 1-p \end{cases} \quad (2)$$

where S_{ij} represents the noiseless image pixel and N_{ij} represents the noise substituting for the Original Pixel (OP).With the noise ratio p, only p percent of the pixels in the image are replaced and others keep noise uncorrupted. In a variety of impulse noise models for images, fixed- and random-valued impulse noises are mostly discussed. Fixed-valued impulse noise, known as the “salt-and-pepper” noise, is made up of corrupted pixels whose values are replaced with values equal to the maximum or minimum (255 or 0) of the allowable range with equal probability (p/2). The random-valued impulse noise is made up of corrupted pixels whose values are replaced by random values uniformly distributed in the range within [0, 255]. In this paper, both fixed and random-valued impulse noises are adopted as the noise model to test the system robustness.

4.2. NN for Noise Detection

As the residual noise greatly affects human perception, exact noise detection is an important factor for the noise removal.

A NN with high accuracy and ability of dealing with various noisy images is proposed for noise detection. It is a 3-layer NN with one hidden layer. The input layer contains three nodes equivalent to the Gray-level Difference (GD), Average Background Difference (ABD), and Accumulation Complexity Difference (ACD) in the 3 × 3 sliding window. The second layer is the hidden layer that contains six nodes, and the bipolar sigmoid function is applied as the activation function. The weighting vectors between the first and second layers, and between the second and third layers, are denoted as S and R, respectively. The output layer contains one node that denotes the identified attribution of the pixel: “noise” or “non-noise,” and moreover the bipolar sigmoid function is

also used as the activation function. The three features in the input layer are discussed as follows.

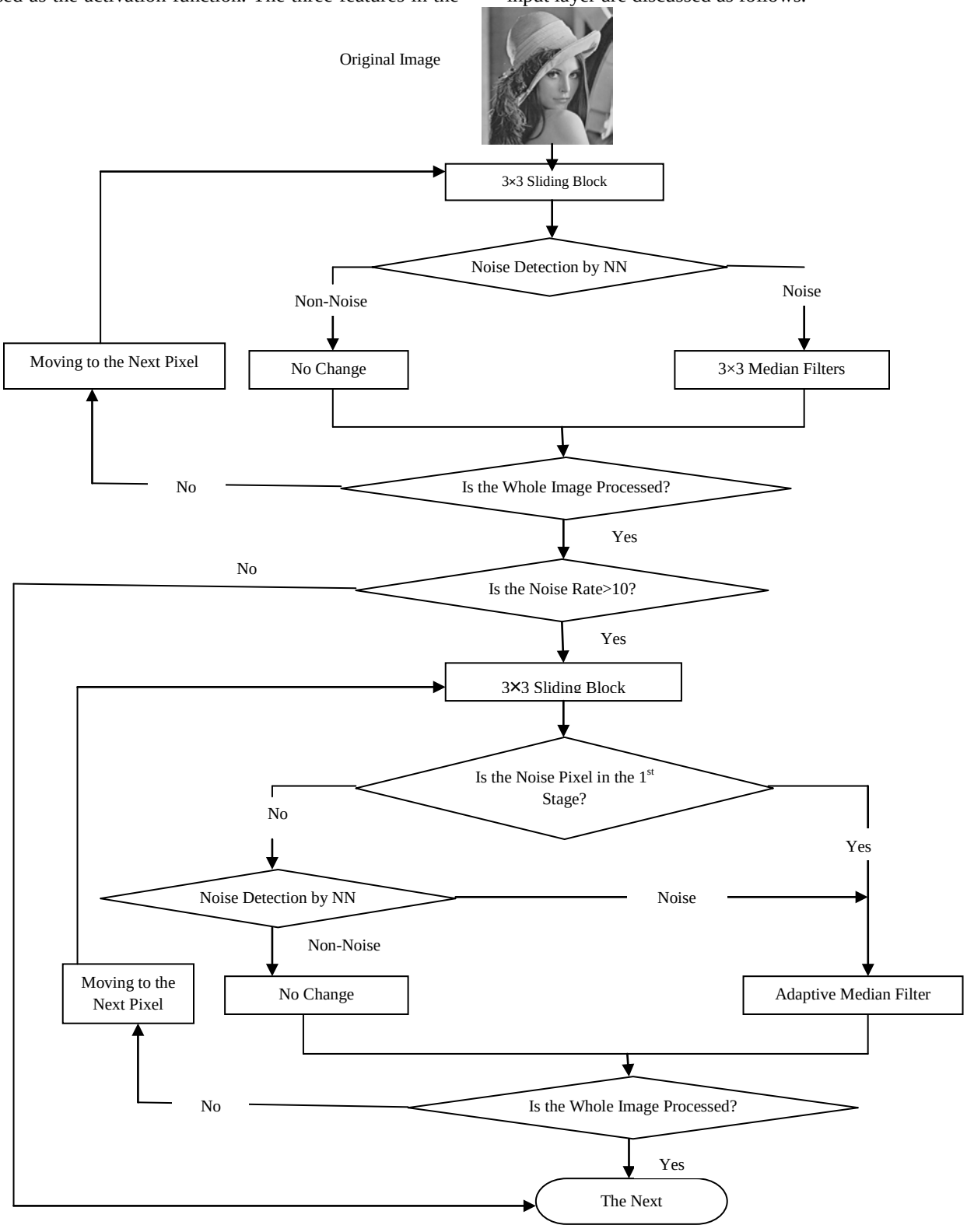


Figure 1: Procedure diagram of the two-level impulse noise removal

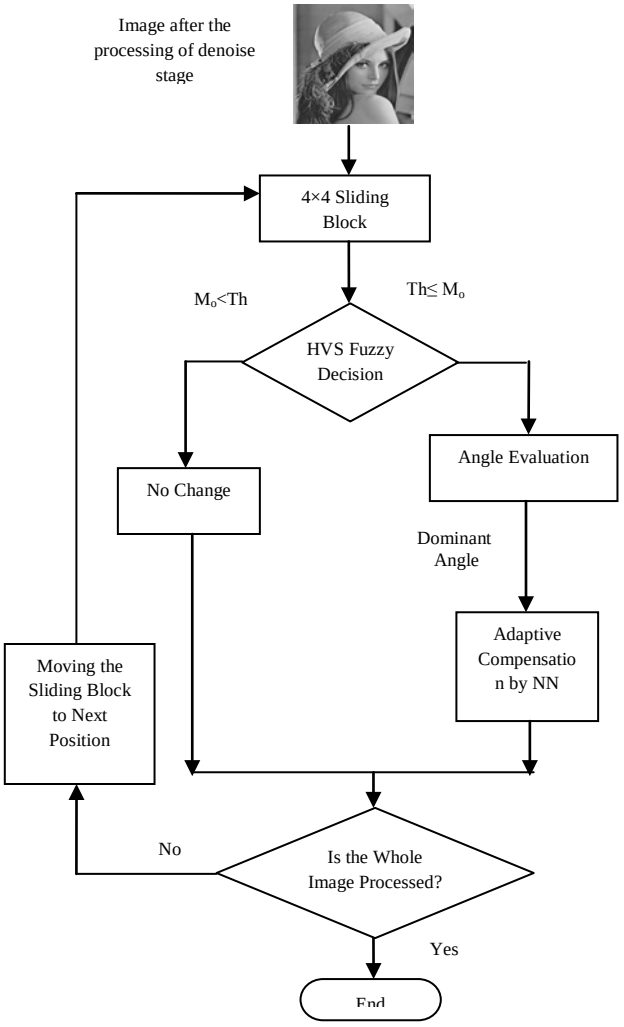


Figure 2: Procedure diagram of the image quality enhancement

1) Gray-Level Difference (GD): The GD represents the accumulated variations between the central pixel for identification and each surrounding local pixel. It is defined by

$$GD = \sum_{i=-1}^{1} \sum_{\substack{j=-1 \\ (i,j)\neq(0,0)}}^{1} |P(0,0) - P(i,j)| \qquad (3)$$

where P(0, 0) is the reference pixel and P(i, j) is the surrounding local pixel.

The GD feature is considered to identify the noise over a flat area. It is expected that the corrupted pixels would yield much bigger differences as compared with the uncorrupted pixels.

2) Average Background Difference (ABD): The surrounding pixels averaged as the Background Luminance (BL) of the sliding block which is then compared with the central pixel. This is an assistant feature to detect the noise. This feature, called the ABD, denoting the overall average variation with the central pixel in the block, is defined by

$$ABD = \left| P(0,0) - \frac{\sum_{i=-1}^{1} \sum_{\substack{j=-1 \\ (i,j)\neq(0,0)}}^{1} P(i,j)}{8} \right| \qquad (4)$$

The corrupted pixels provide bigger differences as compared with the clean ones. For the pixels in the texture area, the GD value is large but the ABD feature will be small.

3) Accumulation Complexity Difference (ACD): Accumulating the difference between each pixel in the 3 × 3 sliding block and its four neighboring pixels as defined next shows the structure information of the block

$$ACD = \sum_{i=-1}^{1} \sum_{j=-1}^{1} |4 \times P(i,j) - P(i-1,j) - P(i+1,j) - P(i,j-1) - P(i,j+1)| \qquad (5)$$

In the edge area, the summation is lower than that in the noise-pixel area, though the GD difference might be similar. So, it provides an assistant feature between the edge and noise pixels.

The neural network can be replaced by Adaptive Neuro-Fuzzy Inference System for better detection of noise. ANFIS is explained as below:

Architecture of ANFIS

The ANFIS is a framework of adaptive technique to assist learning and adaptation. This kind of framework formulates the ANFIS modeling highly organized and not as much of dependent on specialist involvement. To illustrate the ANFIS architecture, two fuzzy if-then rules according to first order Sugeno model are considered:

$$Rule\ 1: If(x\ is\ A_1) and\ (y\ is\ B_1) then\ (f_1 = p_1x + q_1y + r_1)$$

$$Rule\ 2: If(x\ is\ A_2) and\ (y\ is\ B_2) then\ (f_2 = p_2x + q_2y + r_2)$$

where x and y are nothing but the inputs, Ai and Bi represents the fuzzy sets, fi represents the outputs inside the fuzzy region represented by the fuzzy rule, pi, qi and ri indicates the design parameters that are identified while performing training process. The ANFIS architecture to execute these two rules is represented in figure 2, in which a circle represents a fixed node and a square represents an adaptive node.

In the first layer, every node are adaptive nodes. The outputs of first layer are the fuzzy membership grade of the inputs that are represented by:

$$O_i^1 = \mu_{A_i}(x)\ \ i = 1,2 \qquad (6)$$

$$O_i^1 = \mu_{B_{i-2}}(y)\ \ i = 3,4 \qquad (7)$$

where $\mu_{A_i}(x), \beta_{B_{i-2}}(y)$, can accept any fuzzy membership function. For example, if the bell shaped membership function is employed, $\mu_{A_i}(x)$ is represented by:

$$\mu_{A_i}(x) = \frac{1}{1 + \left\{\left(\frac{x - c_i}{a_i}\right)\right\}^{b_i}} \qquad (8)$$

where ai, bi and ci represents the parameters of the membership function, controlling the bell shaped functions consequently.

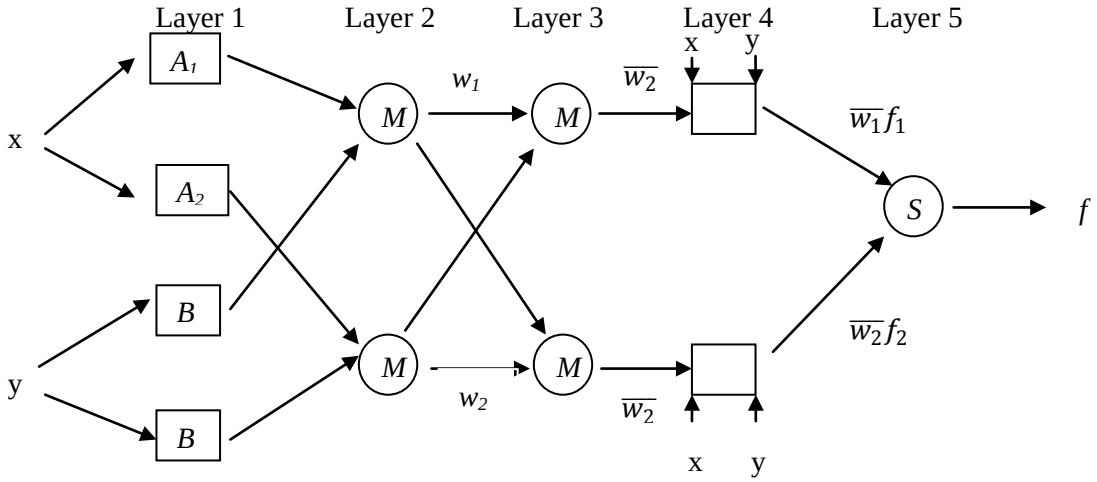


Figure 2: ANFIS Architecture

In layer 2, the nodes are fixed nodes. These nodes are labeled with M, representing that they carry out as a simple multiplier. The outputs of this layer can be indicated by:

$$O_i^2 = w_i = \mu_{A_i}(x)\mu_{B_i}(y) \;\; i = 1,2 \qquad (9)$$

which are called as firing strengths of the rules.

The nodes are fixed in layer 3 as well. They are labeled with N, representing that they are engaged in a normalization function to the firing strengths from the earlier layer.

The outputs of this layer can be indicated as:

$$O_i^3 = \overline{w}_i = \frac{w_i}{w_1 + w_2} \qquad i = 1,2 \qquad (10)$$

which are called as normalized firing strengths.

In layer 4, all the nodes are adaptive nodes. The output of the every node in this layer is merely the product of the normalized firing strength and a first order polynomial. Therefore, the outputs of this layer are provided by:

$$O_i^4 = \overline{w}_i f_i = \overline{w}_i(p_i x + q_i y + r_i) \qquad i = 1,2 \qquad (11)$$

In layer 5, there exists only one single fixed node labeled with S. This node carries out the operation like summation of every incoming signal. Therefore, the overall output of the model is provided by:

$$O_i^5 = \sum_{i=1}^{2} \overline{w}_i f_i = \frac{\sum_{i=1}^{2} w_i f_i}{w_1 + w_2} \qquad (12)$$

It can be noted that layer 1 and the layer 4 are adaptive layers. Layer 1contains three modifiable parameters such as ai, bi, ci that is associated with the input membership functions.

These parameters are called as premise parameters. In layer 4, there exists three modifiable parameters as well such as {pi, qi, ri}, related to the first order polynomial. These parameters are called consequent parameters.

Learning algorithm of ANFIS

The intention of the learning algorithm is to adjust all the modifiable parameters such as{ai, bi, ci} and {pi, qi, ri}, for the purpose of matching the ANFIS output with the training data.

If the parameters such as ai, bi and ci of the membership function are unchanging, the outcome of the ANFIS model can be given by:

$$f = \frac{w_1}{w_1 + w_2} f_1 + \frac{w_2}{w_1 + w_2} f_2 \qquad (13)$$

Substituting Eq. (5) into Eq. (8) yields:

$$f = \overline{w}_1 f_1 + \overline{w}_2 f_2 \qquad (14)$$

Substituting the fuzzy if-then rules into Eq. (15), it becomes:

$$f = \overline{w}_1(p_1 x + q_1 y + r_1) + \overline{w}_2(p_2 x + q_2 y + r_2) \qquad (15)$$

After rearrangement, the output can be expressed as:

$$f = (\overline{w}_1 x)p_1 + (\overline{w}_1 y)q_1 + (\overline{w}_1)r_1 + (\overline{w}_2 x)p_2 + (\overline{w}_2 y)q_2 + (\overline{w}_2)r_2 \qquad (16)$$

which is a linear arrangement of the adjustable resulting parameters such as p1, q1, r1, p2, q2 and r2. The least squares technique can be utilized to detect the optimal values of these parameters without difficulty. If the basis parameters are not adjustable, the search space becomes larger and leads to considering more time for convergence. A hybrid algorithm merging the least squares technique and the gradient descent technique is utilized in order to solve this difficulty. The hybrid algorithm consists of a forward pass and a backward pass. The least squares technique which acts as a forward pass is utilized in order to determine the resulting parameters with the premise parameters not changed. Once the optimal consequent parameters are determined, the backward pass begins straight away. The gradient descent technique which acts as a backward pass is utilized to fine-tune the premise parameters equivalent to the fuzzy sets in the input domain. The outcome of the ANFIS is determined by using the resulting parameters identified in the forward pass. The output error is utilized to alter the premise parameters with the help of standard backpropagation method. It has been confirmed that this hybrid technique is very proficient in training the ANFIS. Learning can be fast up in ANFIS using Modified Levenberg-Marquardt algorithm

Modified Levenberg-Marquardt algorithm

A Modified Levenberg-Marquardt algorithm is used for training the neural network.

Considering performance index is $F(w) = e^T e$ using the Newton method we have as:

$$W_{K+1} = W_K - A_K^{-1}.g_K$$

$$A_k = \nabla^2 F(w)|_{w=w_k}$$

$$g_k = \nabla F(w)|_{w=w_k}$$

$$[\nabla F(w)]_j = \frac{\partial F(w)}{\partial w_j} = 2\sum_{i=1}^{N} e_i(w).\frac{\partial e_i(w)}{\partial w_j}$$

The gradient can write as:

$$\nabla F(x) = 2J^T e(w)$$

Where

$$J(w) = \begin{bmatrix} \frac{\partial e_{11}}{\partial w_1} & \frac{\partial e_{11}}{\partial w_2} \cdots & \frac{\partial e_{11}}{\partial w_N} \\ \frac{\partial e_{21}}{\partial w_1} & \frac{\partial e_{21}}{\partial w_2} \cdots & \frac{\partial e_{21}}{\partial w_N} \\ & \vdots & \\ \frac{\partial e_{KP}}{\partial w_1} & \frac{\partial e_{KP}}{\partial w_2} \cdots & \frac{\partial e_{KP}}{\partial w_N} \end{bmatrix}$$

$J(w)$ is called the Jacobian matrix.

Next we want to find the Hessian matrix. The k, j elements of Hessian matrix yields as

$$[\nabla^2 F(w)]_{k,j} = \frac{\partial^2 F(w)}{\partial w_k \partial w_j}$$
$$= 2\sum_{i=1}^{N} \left\{ \frac{\partial e_i(w)}{\partial w_k} \frac{\partial e_i(w)}{\partial w_j} + e_i(w).\frac{\partial^2 e_i(w)}{\partial w_k \partial w_j} \right\}$$

The Hessian matrix can then be expressed as follows

$$\nabla^2 F(w) = 2J^T(W).J(W) + S(W)$$

$$S(w) = \sum_{i=1}^{N} e_i(w).\nabla^2 e_i(w)$$

If S(w) is small assumed, the Hessian matrix can be approximated as

$$\nabla^2 F(w) \cong 2J^T(w)J(w)$$

$$W_{k+1} = W_k - [2J^T(w_k)\cdot J(w_k)]^{-1} 2J^T(w_k)e(w_k)$$

$$\cong W_k - [J^T(w_k)\cdot J(w_k)]^{-1} J^T(w_k)e(w_k)$$

The advantage of Gauss-Newton is that it does not require calculation of second derivatives.
There is a problem the Gauss-Newton method is the matrix $H = J^T J$ may not be invertible. This can be overcome by using the following modification.
Hessian matrix can be written as

$$G = H + \mu I$$

Suppose that the eigenvalues and eigenvectors of H are $\{\lambda_1, \lambda_2, \ldots\ldots, \lambda_n\}$ and $\{z_1, z_2, \ldots\ldots, z_n\}$.Then:

$$Gz_i = [H + \mu I]z_i$$
$$= Hz_i + \mu z_i$$
$$= \lambda_i z_i + \mu z_i$$
$$= (\lambda_i + \mu)z_i$$

Therefore the eigenvectors of G are the same as the eigenvectors of H, and the eigen values of G are $(\lambda_i + \mu)$. The matrix G is positive definite by increasing μ until $(\lambda_i + \mu) > 0$ for all i therefore the matrix will be invertible.
This leads to Levenberg-Marquardt algorithm:

$$w_{k+1} = w_k - [J^T(w_k)J(w_k) + \mu I]^{-1} J^T(w_k)e(w_k)$$

$$\Delta w_k = [J^T(w_k)J(w_k) + \mu I]^{-1} J^T(w_k)e(w_k)$$

As known, learning parameter, μ is illustrator of steps of actual output movement to desired output. In the standard LM method, μ is a constant number. This paper modifies LM method using μ as

$$\mu = 0.01 e^T e$$

Where e is a $k \times 1$ matrix therefore $e^T e$ is a 1×1 therefore $[J^T J + \mu I]$ is invertible.
Therefore, if actual output is far than desired output or similarly, errors are large so, it converges to desired output with large steps.
Likewise, when measurement of error is small then, actual output approaches to desired output with soft steps. Therefore error oscillation reduces greatly.

4.3. Noise Removal Algorithm

After the first level, the image noise density is calculated to decide whether the second level is necessary or not by the precise detection procedure. By the experiments, it is observed that when the noise density is below 10%, only a one-level noise removal process is enough. More residual noises will occur when the noise density increases. In this case, the second-level noise removal process is essential to detect and remove the residual noises.
As the local features may influence the correctness of the detection part and the median filter may still retain certain noises, the residual noise pixels are detected and removed

with an adaptive median filter in the second level. If there are more than 30% noisy pixels in this image, it is identified as a highly corrupted region and the 5 × 5 median filter is applied for processing. Otherwise, the 3 × 3 median filter is used to process the noisy pixel. The proposed adaptive two-level noise removal technique is very efficient to suppress the impulse noise as well as to preserve the sharpness of edges and detail information.

5. Proposed Image Quality Enhancement

The conventional median filtering techniques have the limitation of blurring details and cause artifacts around edges. In order to compensate the edge sharpness, image quality enhancement is applied to the modified pixels. As the first stage has eliminated the visible noise, the second stage focuses the image enhancement on the edge region. For image analysis, the properties of the HVS are used to acquire the features of images. Thus, region which would worth quality enhancement is realized, since human eyes would be usually more sensitive to this region. For sensitive regions, an adaptive NN is used to enhance the visual quality to match the characteristics of human visual perception.

5.1. HVS-Directed Image Analysis

A novel fuzzy decision system motivated by the HVS is proposed to categorize the image into human perception sensitive and nonsensitive regions. There are three input variables: Visibility Degree (VD); Structural Degree (SD); and Complexity Degree (CD), and one Output Variable (Mo) in the proposed fuzzy decision system.

Visibility Degree (VD): The capability of human eyes to identify the magnitude difference between an object and its background depends on the BL. Figure 3 shows the actual visibility thresholds called JND corresponding to different BLs, and they were verified by a subjective experiment [15]. The experiments were conducted in a dark room and a square area was located in the center of a flat field of constant gray level. Through varying the amplitude of the object, the visibility threshold for each gray level was determined when the object was just noticeable.

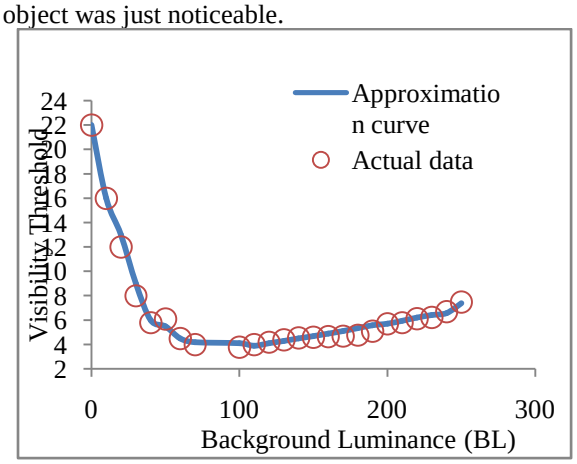


Figure 3: Visibility thresholds corresponding to different BLs

It is observed from figure 3 that the visibility threshold is lower when the BL is within the interval from 70 to 150, and the visibility threshold will increase if the BL becomes darker or brighter away from this interval. In addition, a high visibility threshold will occur when the BL is in a very dark region.

BL is the average luminance of the sliding block proposed to approximate the actual BL and can be calculated by

$$BL = \frac{1}{23}\sum_{i=-1}^{2}\sum_{j=-1}^{2} O(i,j) \times B(i,j) \quad (22)$$

$$B(i,j) = \begin{bmatrix} 2 & 2 & 2 & 1 \\ 2 & 0 & 2 & 1 \\ 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad (23)$$

and the denominator 23 in (22) is the weighted sum of all elements in (23) for normalization. The weighting coefficients of B decrease as the corresponding distance away from the reference pixel increases to estimate the average BL. Feature D is the difference between the maximum and minimum pixel values in the sliding block and can be calculated by

$$D = \max\big(o(i,j)\big) - \min\big(o(i,j)\big) \quad (24)$$

A nonlinear function V (BL) is also designed to approximate the relation between the visibility threshold and BL (as figure. 3), and can be represented as

$$V(BL) = 20.66e^{-0.03BL} + e^{0.008BL} \quad (25)$$

The parameter of 20.66 is obtained by substituting 0 for BL in the nonlinear approximation equation by setting the coefficient of e0.008 BL to be 1.

The first input variable of the fuzzy decision system, VD, is defined as the difference between D and V (BL) and can be represented as

$$VD = D - V(BL) \quad (26)$$

If VD > 0, it means the magnitude difference between the object and its background exceeds the visibility threshold and the object is sensible. Otherwise, this object is not sensible.

The other two input variables, SD and CD, are used to indicate whether the pixels in the sliding block own the edge structure.

Structural Degree (SD): SD shows if the sliding block is a high contrast region, and the pixels in the block can be evidently partitioned into two clusters. It is calculated by

$$SD = \frac{\left|\max\big(O(i,j)\big) - \mathrm{mean}\big(O(i,j)\big) - \left[\mathrm{mean}\big(O(i,j)\big) - \min\big(o(i,j)\big)\right]\right|}{\max\big(O(i,j)\big) - \min\big(O(i,j)\big)} \quad (27)$$

Where

$$\mathrm{mean}\big(O(i,j)\big) = \frac{1}{16}\sum_{i=-1}^{2}\sum_{j=-1}^{2} O(i,j) \quad (28)$$

Equation (11) can be expressed as $|\sigma 1 - \sigma 2|/(\sigma 1 + \sigma 2)$. So, the SD has been normalized to [0, 1] and this rule can also be applied to images with a different intensity range. If SD is small and σ2 and σ1 are close, it means the pixels in the block can be partitioned into two even clusters. The block may contain edge or texture structure. On the contrary, if SD

is a large value, $0 < |\sigma_1 - \sigma_2|$, it means the pixel number of one cluster and that of the other cluster are not even; thus, the block may contain noise.

Complexity Degree (CD):

In these two plots, pixel numbers of the two clusters are the same. Hence, the SD values equivalent to these two structures are close. As the proposed NN is used to compensate the sensitive regions, a CD input variable based on the differential process is used to tell the delineated edge structure from the texture structure. It is calculated by

$$CD = \sum_{i=-1}^{2}\sum_{j=-1}^{2}\big|4O'(i,j) - [O'(i+1,j) + O'(i-1,j) + O'(i,j-1) + O'(i,j+1)]\big| \quad (29)$$

Where $O'(i,j)$ is the binarized version of $O(i,j)$. Assuming mean (O) is the mean gray value of the sliding block, $O'(i,j)$ is defined as

$$O'(i,j) = \begin{cases} 1, & if\ O(i,j) \geq mean(O) \\ 0, & otherwise \end{cases} \quad (30)$$

In (29), each pixel in the 4 × 4 sliding block takes the 4-directional local gradient operation and the CD is the summation of the 16 local gradient values. If the CD is a large value, it means the block may contain texture structure. On the contrary, if the CD is a small value, the block may contain delineated edge structure.

Hyper Trapezoidal Fuzzy Membership Function

The proposed system uses the Hyper Trapezoidal fuzzy Membership function.

Hyper trapezoidal membership functions are defined by prototype points and a crispness factor. In a fuzzy partitioning of an N-dimensional space, let each fuzzy set, S_i be defined by a prototype point, λ_i. Furthermore, let the partitioning of the space also be parameterized by a crispness factor, σ . The prototype point, λ_i has a degree of membership in set, S_i , of $\mu_i(\lambda_i) = 1$ and a degree of membership in set S_j, of $\mu_j(\lambda_j) = 1$ where $j \neq i$.

The crispness factor, 0≤σ≤1, determines how much ambiguity exists between the sets of the partitioning. For σ =1, no fuzziness exists between the sets and the partitioning is equivalent to a minimum distance classifier. For fuzzy sets σ <1. One way to define the crispness factor is using Figure 4 and equation (31).

$$\sigma = \frac{2\alpha}{d} \quad (31)$$

The crispness factor establishes fuzzy space between the prototype points. The prototype points are selected as ideal representatives of each fuzzy set. Then, the designer's selection of σ specifies the ratio of α and d.

The next step in the derivation is the definition of a suitable distance measure relating the distance from the crisp input to two prototype points. This distance measure is a ratio of the distance between two prototype points, and the difference in the distances from the crisp input to the two prototype points. For fuzzy sets S_i and S_j, with prototype points λ_i and λ_j, and a crisp input $\square\Lambda$, that distance measure is

$$\rho_{i|j}(\Lambda) = \frac{|\vec{v}_i|^2 - |\vec{v}_j|^2}{|\vec{v}_{ji}|^2} \quad (32)$$

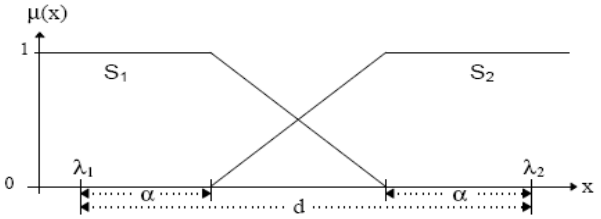


Figure 4: Defining Crispness of a partitioning

Where $\vec{v}_{ji}$ is a vector from λ_i to λ_j; $\vec{v}_i$ is a vector from λ_i to Λ. This distance measure is used to determine if the crisp input Λ lies completely in fuzzy set i, or completely in fuzzy set j, or in the fuzzy region between the two sets.

The third step in the derivation of hyper trapezoidal membership functions is determining the degree of membership that L has in set i, given that set j is the only other set in the partition. Suppose fuzzy sets i and j are the only two sets defined in an N-dimensional space. Using the distance measure of equation (32), that degree of membership is

$$\mu_{i|j}(\Lambda) = \begin{cases} 0; & \rho_{i|j}(\Lambda) \geq 1-\sigma \\ 1; & \rho_{i|j}(\Lambda) \leq \sigma - 1 \\ \dfrac{\vec{v}_{ji} \cdot \vec{v}_j - \frac{\sigma}{2}|\vec{v}_{ji}|^2}{(1-\sigma)|\vec{v}_{ji}|^2}; & otherwise \end{cases} \quad (33)$$

For the first case in equation (33), Λ lies completely in fuzzy set j. For the second case, Λ lies completely in fuzzy set i. The third case is the case of Λ being in the transition from set i to set j.

The numerical value of Mo after defuzzification is compared with a threshold value, Th, where Th is preferably set as the value 5 by experiments. When Mo ≥ Th, the adaptive NN compensation module with angle evaluation would be chosen; otherwise, the OP value would be used.

5.2. Angle Evaluation

As Mo ≥ Th, the fuzzy system identifies the reference pixel as sensible delineated edge and the trained adaptive NN model is chosen for quality enhancement according to its corresponding edge angle. The angle evaluation is performed to determine the dominant orientation of the sliding block. When the orientation angle of O (i, j) denoted as A (i, j) is computed, the luminance values of the OPs nearby O (i, j) are used for the following computations:

$$Dx(i,j) = O(i-1,j-1) + 2O(i-1,j) + O(i-1,j+1) - (O(i+1,j-1) + 2O(i,j+1) + O(i+1,j+1)) \quad (34)$$

$$Dy(i,j) = O(i-1,j-1) + 2O(i,j-1) + O(i+1,j-1) - (O(i-1,j+1) + 2O(i,j+1) + O(i+1,j+1)) \quad (35)$$

$$A(i,j) = -\frac{180}{\pi}\left[\tan^{-1}\left(\frac{Dy(i,j)}{Dx(i,j)}\right)\right] \quad (36)$$

Where $-1 \le i \le 2$ *and* $-1 \le j \le 2$.

The obtained angle of each pixel in the sliding window is quantized into eight quantization sectors such as θ = 22.5 × k (in degrees), where k = 0, 1, . . . , 7. Assuming θ is the quantized angle for most pixels in the window; it is regarded as the dominant orientation of the reference edge pixel. The corresponding weighting coefficient Wθ derived from the offline training NN is adopted for compensation filtering.

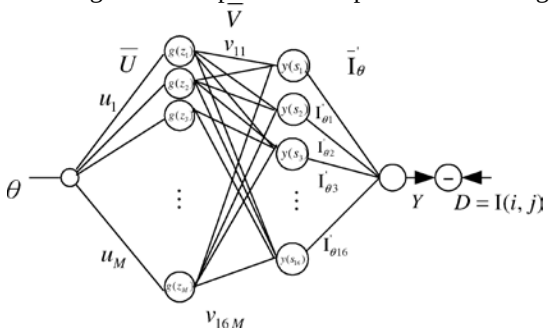


Figure 5: Proposed feed forward NN for image quality enhancement

5.3. NN-Based Image Compensation

The function of the proposed NN is to obtain the weights Wθ defined in (1), where θ represents the quantized dominant orientation of the reference pixel. Thus, the proposed NN is used to obtain eight sets of weighting matrices through training.

Each weighting matrix W_θ can be represented as

$$W_\theta(i,j) = \begin{bmatrix} w_{-1-1} & w_{-10} & w_{-11} & w_{-12} \\ w_{0-1} & w_{00} & w_{01} & w_{02} \\ w_{1-1} & w_{10} & w_{11} & w_{12} \\ w_{2-1} & w_{20} & w_{21} & w_{22} \end{bmatrix} \quad (37)$$

In order to use supervised learning algorithms to train the proposed NN, several clean image portions with dominant orientation are used as training patterns. Assuming a clean image portion is denoted as I, the noise-corrupted version of I has been processed by the proposed noise removal method in the first stage and the filtered result is denoted as I'. According to figure 5, let $I'(i,j)$ be the reference pixel, where O (0, 0) = $I'(i,j)$, and it is classified as an edge pixel with dominant orientation θ after angle evaluation. The input of the NN can be defined as IP = θ and the network output is the compensated pixel value of $I'(i,j)$. The pixel value of I (i, j) obtained from the clean original image is used as the desired output of the NN for training.

6. Experimental Results

For experimenting the proposed filtering technique, several 512 X 512 grayscale images affected by the noise with noise occurrence of 1% to 50% is considered. The result of the proposed filter is compared with several existing filters such as median filter and two stage filter with different window size. The quantitative measures used for comparison is the Peak Signal-to-Noise Ratio (PSNR) between the original and restored images and average execution time. PSNR value is evaluated by using the following equation:

$$PSNR = 10log_{10}\left(\frac{\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}255^2}{\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}[I(i,j)-Y(i,j)]^2}\right) \quad (38)$$

Where {I(i, j)} and {Y (i, j)} are the original and restored images, respectively.

Table 1 provides the comparison of the proposed filter with median filter (3X3 window size), median filter (5X5 window size) and two stage filter. The experimentation is performed at different noise level such as 1%, 5%, 10%, 20%, 40% and 50%.

From the table 1, it can be observed that the PSNR value resulted in 5% noise affected image is 29.6 for the median filter of window size 3 X 3, 30.2 for the median filter of window size 5 X 5, 34.7 for the two stage filter whereas it is higher for the Two Stage filter with hyper trapezoidal fuzzy membership function and Modified LM i.e., 38.2 and for the proposed ANFIS technique the resulted PSNR value is 41.2 . When the image is affected by higher noise i.e., 50%, other filters results only less PSNR value i.e., 13.6 for the median filter of window size 3 X 3, 14.3 for the median filter of window size 5 X 5, 28.5 for the two stage filter whereas it is higher for the Two Stage filter with hyper trapezoidal fuzzy membership function and Modified LM i.e., 33.9 and for the proposed ANFIS technique the resulted PSNR value is 36.7. When the overall PSNR is considered, the proposed filter shows better PSNR values when compared to the conventional filters.

Table 1: Comparative results in PSNR of different filtering methods for various percentages of noise (Lena image)

Noise Ratio / Filter	1%	5%	10%	20%	40%	50%
Median Filter (3X3)	31.3	29.6	25.9	23.2	20.9	13.6
Median Filter (5X5)	32.5	30.2	27.1	25.1	21.3	14.3
Two Stage filter	36.8	34.7	33.2	32.4	31.2	28.5
Two Stage filter with hyper trapezoidal fuzzy membership function and Modified LM	39.7	38.2	37.6	35.2	34.8	33.9
Proposed Two Stage Noise Removal using FANFIS	41.2	40.5	40.1	39.4	37.5	36.7

Table 2 shows the execution time taken for the proposed filter and the different existing image filters with the noise rate as 80%. From the table, it can be observed that execution time required for the median filter with 3 X 3 window size is 14 seconds, median filter with 5 X 5 window size is 16

seconds, two stage filter is 13 seconds, whereas, the execution time required by the two stage technique is 9 seconds and finally for the proposed FANFIS, only 6 seconds are needed. This clearly indicates that the overall execution time required by the proposed filter is lesser when compared to the existing filters.

Table 2: Comparative results in average execution time of different filtering methods for Lena image corrupted image by 80% salt and pepper noise

Filters	Median Filter (3X3)	Median Filter (5X5)	Two Stage filter with LM training Algorithm	Two Stage filter with hyper trapezoidal fuzzy membership function and Modified LM	Proposed Two Stage Noise Removal using FANFIS
Time (seconds)	14	16	13	9	6

7. Conclusion

A novel two-stage noise removal technique is proposed in this paper. In the first stage, a two level noise removal procedure with NN-based noise detection was applied to remove the noise. In the second stage, a fuzzy decision rule inspired by the HVS was proposed to categorize pixels of the image into human perception sensitive and nonsensitive classes. A Fast Adaptive Neuro-Fuzzy Inference System is proposed to enhance the sensitive regions to perform better visual quality. Moreover, the hyper trapezoidal member function is used which provides significance performance. The proposed technique is experimented with 512 X 512 grayscale image with the noise occurrence of 1% to 50%. The PSNR value obtained for the proposed technique is higher when compared to the existing filtering techniques. Moreover, the execution time taken by the proposed approach is very significant when compared with the existing approaches. Thus the proposed technique is very efficient when compared with the conventional methods in perceptual image quality, and it can provide a quite a stable performance over a wide variety of images with various noise densities.

8. References

[1] Schulte, S., De Witte, V., Nachtegael, M., Van der Weken, D. and Kerre, E.E., "Fuzzy Two-Step Filter for Impulse Noise Reduction From Color Images", IEEE Transactions on Image Processing, Vol. 15, No. 11, Pp. 3567 – 3578, 2006

[2] Sun Zhong-gui, Chen Jie and Meng Guang-wu, "An Impulse Noise Image Filter Using Fuzzy Sets", International Symposiums on Information Processing (ISIP), Pp. 183 – 186, 2008.

[3] Ibrahim, H., Kong, N.S.P. and Theam Foo Ng, "Simple adaptive median filter for the removal of impulse noise from highly corrupted images", IEEE Transactions on Consumer Electronics, Vol. 54, No. 4, Pp. 1920 - 1927, 2008.

[4] Abreu, E., Lightstone, M., Mitra, S.K. and Arakawa, K., "A New Efficient Approach for the Removal of Impulse Noise from Highly Corrupted Images", IEEE Transaction on Image Processing, Vol. 5, No. 6, Pp. 1012-1025, 1996.

[5] Russo, F. and Ramponi, G., "A Fuzzy Filter for Images Corrupted by Impulse Noise", IEEE Signal Processing Letters, Vol. 3, No. 6, Pp. 168-170, 1996.

[6] Choi, Y.S. and Krishnapuram, R., "A Robust Approach to Image Enhancement Based on Fuzzy Logic", IEEE Transaction on Image Processing, Vol. 6, No. 6, Pp. 808-825, 1997.

[7] Boskovitz, V. and Guterman, H., "An Adaptive Neuro-Fuzzy System for Automatic Image Segmentation and Edge Detection", IEEE Transactions on Fuzzy Systems, Vol. 10, No. 2, Pp. 247-262, 2002.

[8] Chao Deng; Ji Yu An; “An Impulse Noise Removal Based on a Wavelet Neural Network”, Second International Conference on Information and Computing Science, ICIC '09, Volume 2, pages 71-74, 2009.

[9] J. B. Bednar and T. L. Watt, “Alpha-trimmed means and their relationship to median filters,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 1, pp. 145–153, Feb. 1984.

[10] D. A. F. Florencio and R.W. Schafer, “Decision-based median filter using local signal statistics,” in *Proc. SPIE Symp. Vis. Commun. Image Process.*, 1994, vol. 2308, pp. 268–275.

[11] S. J. Ko and Y. H. Lee, “Center weighted median filters and their applications to image enhancement,” *IEEE Trans. Circuits Syst.*, vol. 38, no. 9, pp. 984–993, Sep. 1991.

[12] T. Chen and H. R. Wu, “Impulse noise removal by multi-state median filtering,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, pp. 2183–2186, Jun. 2000.

[13] X. Li and M. Orchard, “True edge-preserving filtering for impulse noise removal,” in presented at the 34th Asilomar Conf. Signals, Syst. Comput., Pacific Grove CA, Oct. 2000.

[14] D. Zhang and Z.Wang, “Impulse noise detection and removal using fuzzy techniques,” *Electron. Lett.*, vol. 33, no. 5, pp. 378–379, Feb. 1997.

[15] C. H. Chou and Y. C. Li, “A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile,” *IEEE Trans. Fuzzy Syst.*, vol. 3, no. 3, pp. 467–476, Dec. 1995.

[16] P. Civicioglu, “Using uncorrupted neighborhoods of the pixels for impulsive noise suppression with ANFIS,” *IEEE Trans. Image Process.*, vol. 16, no. 3, pp. 759–773, Mar. 2007.

Comparative Analysis of Congestion Control Algorithms Using ns-2

Sanjeev Patel[1], P. K. Gupta[2], Arjun Garg[3], Prateek Mehrotra[4] and Manish Chhabra[5]

[1] Deptt. of Computer Sc. & Engg, Jaypee Institute of Information Technology, Noida, UttarPradesh, 201307, India

[2] Deptt. of Computer Sc. & Engg, Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh, 173215, India

[3] Deptt. of Computer Sc. & Engg, Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh, 173215, India

[4] Deptt. of Electronics & Communication Engg., Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh, 173215, India

[4] Deptt. of Electronics & Communication Engg., Thapar University, Patiala, Punjab, 147004, India

Abstract

In order to curtail the escalating packet loss rates caused by an exponential increase in network traffic, active queue management techniques such as Random Early Detection (RED) have come into picture. Flow Random Early Drop (FRED) keeps state based on instantaneous queue occupancy of a given flow. FRED protects fragile flows by deterministically accepting flows from low bandwidth connections and fixes several shortcomings of RED by computing queue length during both arrival and departure of the packet. Stochastic Fair Queuing (SFQ) ensures fair access to network resources and prevents a busty flow from consuming more than its fair share. In case of (Random Exponential Marking) REM, the key idea is to decouple congestion measure from performance measure (loss, queue length or delay). Stabilized RED (SRED) is another approach of detecting nonresponsive flows. In this paper, we have shown a comparative analysis of throughput, delay and queue length for the various congestion control algorithms RED, SFQ and REM. We also included the comparative analysis of loss rate having different bandwidth for these algorithms.

Keywords:*Stochastic Fair Queing (SFQ), Random Early Detection (RED), Random Exponential Marking (REM), First In First Out (FIFO), Throughput, Delay, Queue length, Loss rate, and Utilization.*

1. Introduction

SFQ (Stochastic Fair Queuing) is a class of queue scheduling disciplines that are designed to allocate a pretty large number of separate FIFO queues [1]. Increasing the number of queues to a large extent helps to achieve fairness. RED queue management aims at alleviating this problem by detecting incipient congestion in advance and communicating the same to the end-hosts, allowing them to trim down their transmission rates before queues begin to overflow and packets start dropping. For this, RED maintains an exponentially weighted moving average of the queue length which it used as a congestion detection mechanism [2]. In order to be efficient, RED must ensure that congestion notification is conveyed at a rate which sufficiently suppresses the transmitting sources without underutilizing the link. RED must also ensure that the queue is configured with enough buffer space to hold an applied load greater than the link capacity from the time when congestion detection occurs to the time when the applied load reduces at the bottleneck link in response to the notification regarding congestion. FRED proposes to transform RED mechanisms to provide fairness by using per-active-flow accounting to make different dropping decisions for connections with different bandwidth usages [3]. When a flow persistently occupies a considerable amount of the queue's buffer space, it is identified and restrained to a smaller buffer space. Severity of congestion is indicated by queue lengths in various queue management algorithms. This inherent problem can be dealt by a fundamentally different active queue management algorithm, called BLUE [4]. BLUE has been shown to perform significantly better than RED both in terms of packet loss rates and buffer size requirements in the network. If buffer overflow causes the queue to

recurrently drop packets, BLUE increments the marking probability, thus augmenting the rate at which congestion notification is sent back [4].

REM is an active queue management scheme that aims to achieve both high utilization and negligible loss and delay in a simple and scalable manner [5]. While congestion measure indicates excess demand for bandwidth and must track the number of users, performance measure, independently of the number of users, should be stabilized around their targets. The first idea of REM [5] attempts to match user rates to network capacity while clearing buffers, irrespective of number of users. The second idea embeds the sum of link prices (congestion measures), summed over all the routers in the path of the user to the end-to-end marking (or dropping) probability [5]. Number of active flows shares a linear relationship with number of different flows in the buffer. We simulated the network configuration having higher delay and lower bandwidth at the main bottleneck link [6]. In this paper, we used ns-2 network simulator. The structure of ns-2 and performance metrics considered in this paper has been given in section 2. In section 3, we described about network configuration and network parameters used in the simulation. This section also deals with implementation and result analysis observed in this simulation. Last section concludes with discussion of future work.

2. Performance Metrics

The complete NS class hierarchy has been shown in Figure 1. Now in the queue object of the hierarchy there are only two active queue management algorithm, RED and Drop tail. Now the other algorithm discussed in the section of network congestion control are implement under this object only. Queue serves as the parent of all these algorithms, and they are appended as a child to this element in the hierarchy. Similarly other can also be deployed in ns2 architecture and make them run [7].

2.1. Analysig Trace File

When the ns is run, the trace of each event can be stored in a trace file. While tracing into an output ASCII file, the trace is organized in 12 fields as shown in the following figure. Class hierarchy of ns2 and the description of each field is shown by their name and an sample example of trace file as given below in Figure 1 and Figure 2.

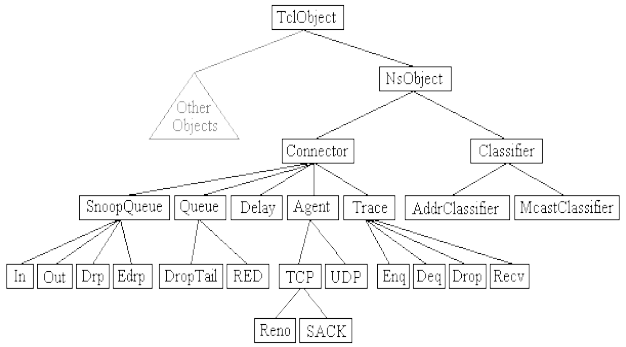


Fig. 1 Class Hierarchy

event	time	from node	to node	pkt type	pkt size	flags	fid	src addr	dst addr	seq num	pkt id

```
r : receive (at to_node)
+ : enqueue (at queue)              src_addr : node.port (3.0)
- : dequeue (at queue)              dst_addr : node.port (0.0)
d : drop    (at queue)
```

```
r 1.3556 3 2 ack 40 ------- 1 3.0 0.0 15 201
+ 1.3556 2 0 ack 40 ------- 1 3.0 0.0 15 201
- 1.3556 2 0 ack 40 ------- 1 3.0 0.0 15 201
r 1.35576 0 2 tcp 1000 ------- 1 0.0 3.0 29 199
+ 1.35576 2 3 tcp 1000 ------- 1 0.0 3.0 29 199
d 1.35576 2 3 tcp 1000 ------- 1 0.0 3.0 29 199
+ 1.356 1 2 cbr 1000 ------- 2 1.0 3.1 157 207
- 1.356 1 2 cbr 1000 ------- 2 1.0 3.1 157 207
```

Fig. 2 Trace files structure

2.2 Packet Loss

Packets can be lost in a network because they may be dropped when a queue in the network node overflows. The amount of packet loss during the steady state is another important property of a congestion control scheme. The larger the value of packet loss, the more difficult it is for transport-layer protocols to maintain high bandwidths, the sensitivity to loss of individual packets, as well as to frequency and patterns of loss among longer packet sequences is strongly dependent on the application itself. This characteristic can be specified in a number of different ways, including loss rate, loss patterns, loss free seconds, and conditional loss probability. In this paper, we considered that packet loss would occur only due to the dropping of the packets. There is no loss due to other means.

2.3 Throughput

This is the main performance measure characteristic, and most widely used. This measure how soon the receiver is able to get a certain amount of data send by the sender. It is determined as the ratio of the total data received to the end to end delay. Throughput is an important factor which directly impacts the network performance.

2.4 Delay

Delay is the time elapsed while a packet travels from one point (e.g., source premise or network ingress) to another (e.g., destination premise or network degrees). The larger the value of delay, the more difficult it is for transport layer protocols to maintain high bandwidths. This characteristic can be specified in a number of different ways, including average delay, variance of delay (jitter), and delay bound. In this paper, we calculated end to end delay

2.5 Queue Length

A queueing system in networks can be described as packets arriving for service, waiting for service if it is not immediate, and if having waited for service, leaving the system after being served. Thus queue length is very important characteristic to determine that how well the active queue management of the congestion control algorithm has been working.

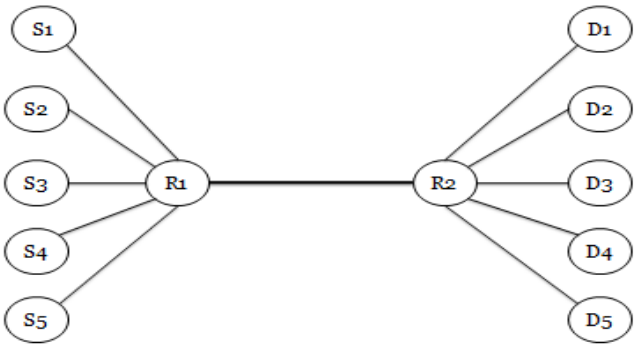


Fig. 3 Simulation Scenario

3. Simulations And Results

In this section, we discussed about network configuration used over the network simulator ns2 to simulate the three algorithms RED, SFQ and REM and after that we analyzed about the results obtained from our simulations. The algorithms compared here are first deployed into the ns2 architecture then following simulation scenario has been generated to compare their performance on the simulation setting as shown in Figure3.

3.1 Simulation Scenario

There are five nodes at each side of the bottleneck link. Here five nodes are acting as a TCP source and five nodes are acting as a TCP sink so that both routers are applying the congestion control algorithm. There is two- way traffic in the system. We consider the network scenario as shown in Figure 3. We simulate this network on ns2 for different AQM algorithms RED, SFQ and REM for same network parameters as given in Table 1 except to the bottleneck link. We simulated these three algorithms RED, SFQ, and REM on the same bottleneck link R_1R_2. Firstly we consider the bottleneck link to 5Mbps for each considered AQM algorithm. We considered a fixed packet size of 2 KB and buffer capacity of 4KB throughout the simulation. Round trip delay for each link has been displayed in Table 1. So it could be concluded from the Table 1 that minimum end to end delay should be larger than 60 ms. Our simulation has been observed over the period of 100 seconds. Whole simulation has been observed over small buffer capacity of 4KB.

3.2 Analysis of Loss Rate

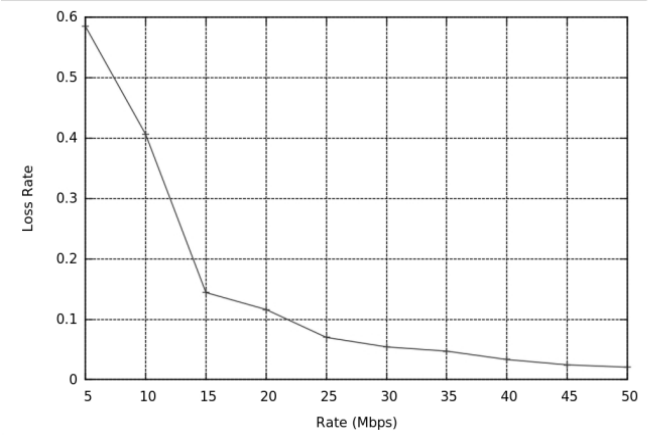
Figure 4 shows about the loss rate occurred in RED, SFQ, and REM respectively. In our simulation, we vary the bandwidth of the bottleneck link as given in Figure 4 for each algorithm RED, SFQ, and REM. It has been observed that loss rate smoothly decreased as we are increasing the bandwidth of bottleneck link in case of RED. We got the drastic change in loss rate at 15 Mbps in case of SFQ because of unfairness achieved at this bandwidth. It has been concluded that SFQ and REM could achieved higher loss rate at higher bandwidth at some specific bandwidth but it could not be happen. It has been reflected more in case of SFQ. But RED shows smooth decrease in loss rate over increase in bandwidth.

3.3 Analysis of Throughput

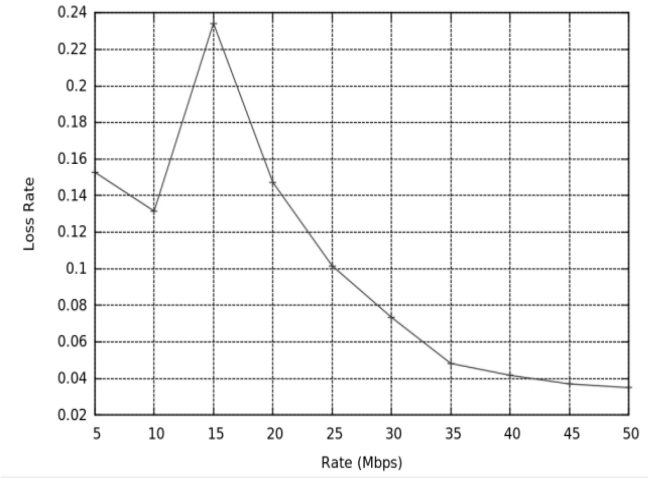
It has been observed that REM had a best throughput and RED had least throughput among all these three algorithms for the simulation achieved at 5 Mbps of bandwidth. Figure 5 show that REM gets the good result and RED gets the poor result. It could be observed one point on throughput graph whenever smooth growth in throughput has been broken. It indicated about a starting point when dropping of packet took place. This achieved point in each algorithm has a same ratio as compared to their maximum achieved throughput.

Table 1:.Parameters for simulation

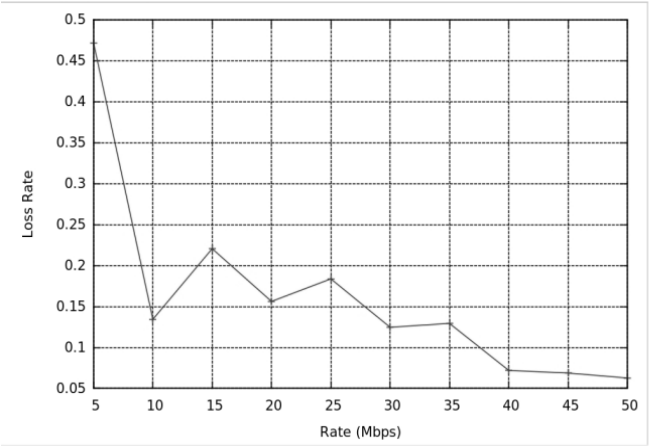
Link	*RTT (ms)*	*Rate (Mbps)*	*Protocol*
S1 R1	10	100	Drop tail
S2 R1	10	100	Drop tail
S3 R1	10	100	Drop tail
S4 R1	10	100	Drop tail
S5 R1	10	100	Drop tail
R1R2	40	10	RED / SFQ /REM
R2D1	10	100	Drop tail
R2D2	10	100	Drop tail
R2D3	10	100	Drop tail
R2D4	10	100	Drop tail
R2D5	10	100	Drop tail



(a) RED

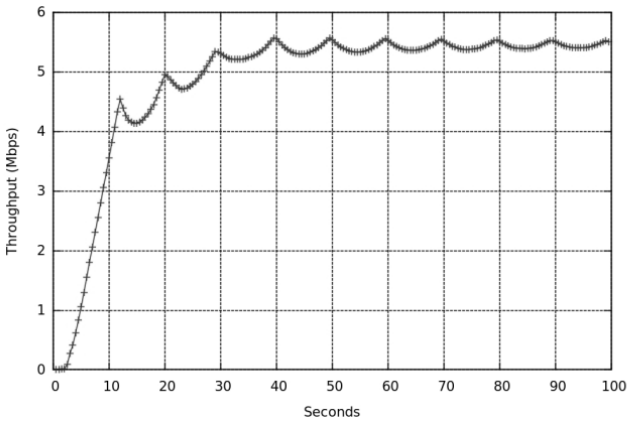


(b) SFQ

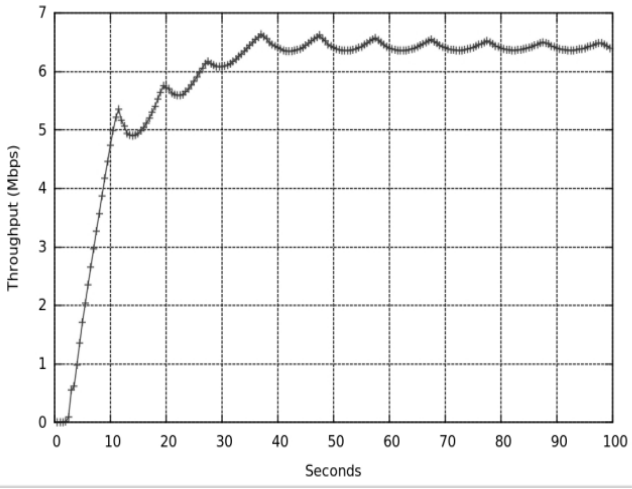


(c)REM

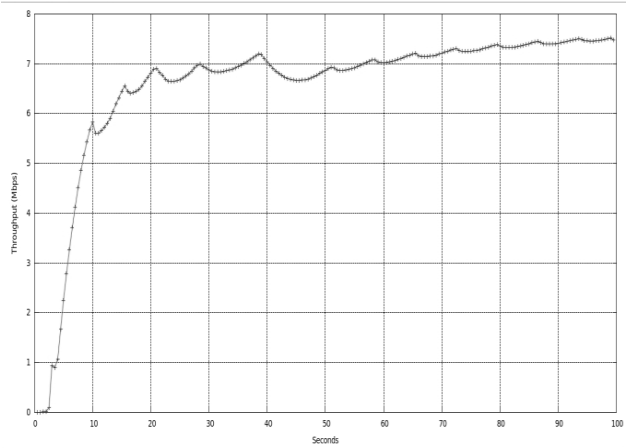
Fig. 4 Loss Rate for various algorithms (a) RED, (b) SFQ, (c) REM



(a)RED

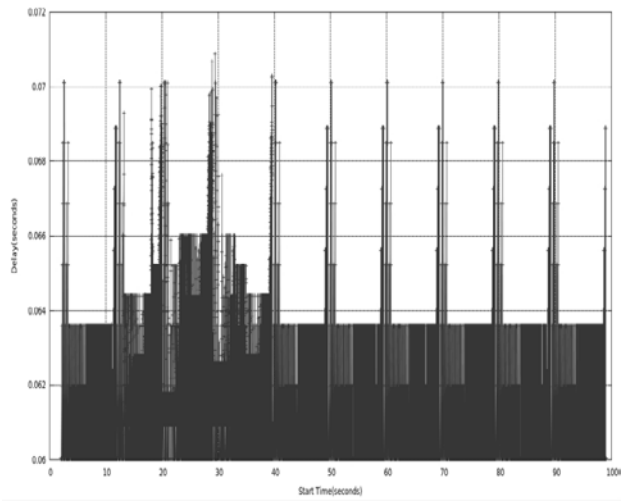


(b)SFQ

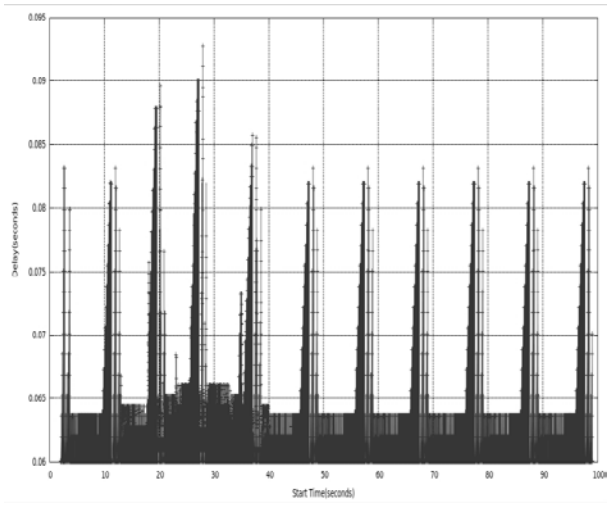


(c) REM

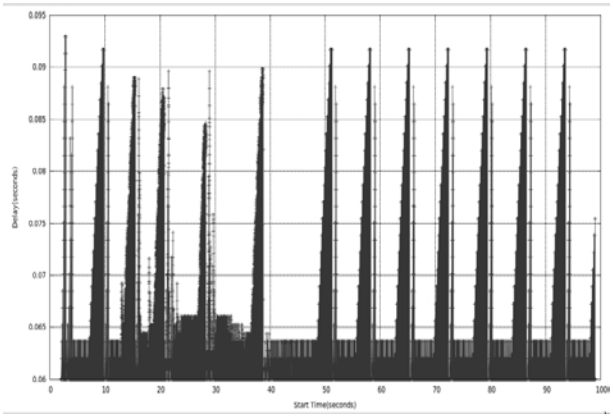
Fig. 5 Throughput Diagram for various algorithms (a) RED, (b) SFQ, (c) REM



(a) RED

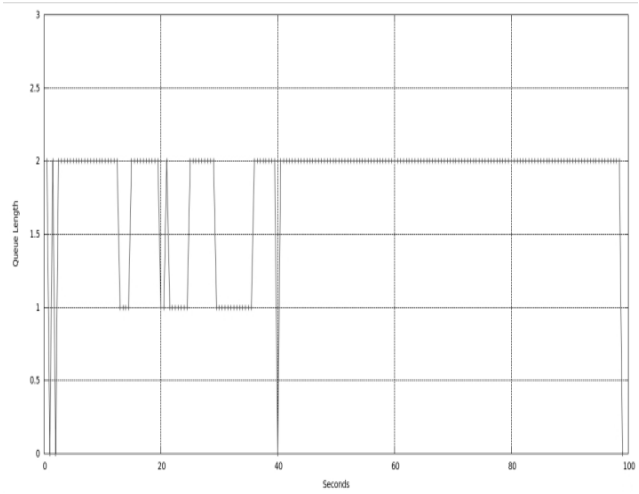


(b)SFQ

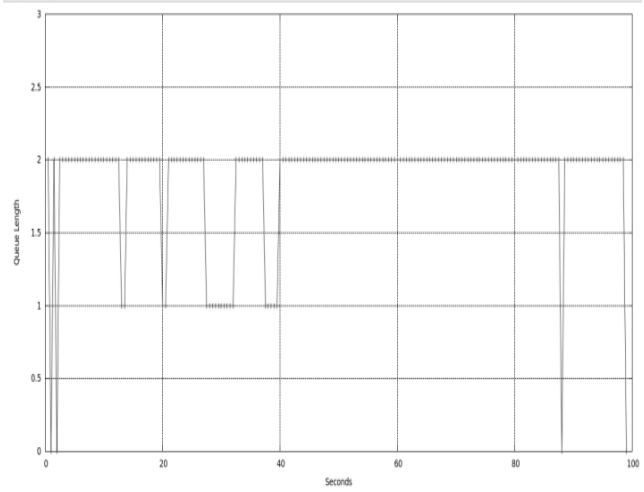


(c)REM

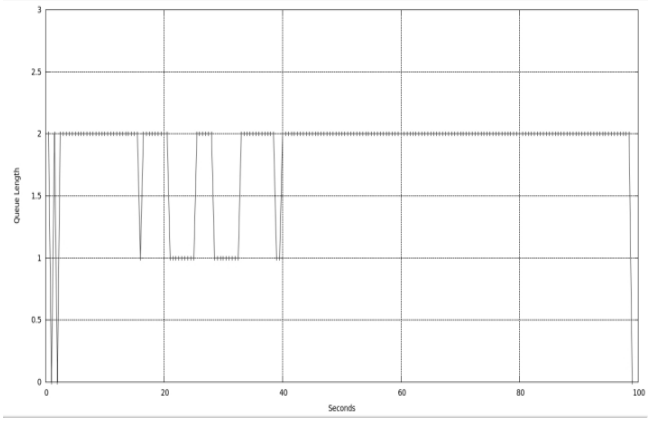
Fig. 6 Delay Diagram for various algorithms (a) RED, (b) SFQ, (c) REM



(a)RED



(b)SFQ



(c)REM

Fig. 7 Queue Length Diagram for various Algorithms (a) RED, (b) SFQ, (c) REM

Table 2 Comparative results

Performance Metrics		RED	SFQ	REM
Queue length	Max.	2	2	2
	Min.	0	0	0
Throughput	Max.	5.53	6.64	7.51
	Min.	0	0	0
Delay	Max.	67.25	90.01	92.96
	Min.	60.03	60.03	60.03
Send Packets		37157	42554	49117
Lost Packets		151	56	66
Average Loss Ratio (%)		0.4064	0.1316	0.1344
Utilization (%)		59.45	68.08	78.58

Table 3 Ranking of the different algorithms

Algorithm	Delay	Queue Length	Throughput	Loss Rate
RED	A	A	C	C
SFQ	B	B	B	A
REM	C	C	A	B

3.4 Analysis of Delay

Figure 6 plots the actual response time for each packet achieved in RED, SFQ, and REM. It has been observed from Table 2 that minimum delay occurred in each algorithm is same but maximum delay achieved in REM. Therefore we could conclude that each algorithm would get a same response time provided congestion has been observed because queueing delay would be same for each algorithm if there is no congestion in network.

3.5 Analysis of Queue length

Here we did not achieve much difference in queue length between these algorithms because at most two packets could be allowed to enter into queue due to the small buffer capacity. REM achieved queue length of two packets for a longer time as shown in Figure 7.

4. Future Work And Conclusion

In this paper we address the problems with existing congestion control algorithms and we tried to show about various performance parameters of RED, SFQ, and REM for our considered network configurations. We have calculated the different performance parameters for each algorithm of considered network configuration as given in Figure 3 and Table 1. We calculated the total number of packets sent over the bottleneck link R_1R_2 and total number of packets lost during the simulation over the period of 100 seconds. SFQ has a minimum average loss ratio and RED has a maximum loss ratio. Now actual number of bytes transmitted over the bottleneck link R_1R_2 could be computed termed as utilization has been shown in Table 2. It has been observed that performance parameters are varying according to the algorithms. RED achieved the best result in terms of the delay but in terms of throughput, loss ratio, and utilization REM shows the best results. If we would provide the equal weightage to each performance parameter then we could conclude that REM would be the better one among all three algorithms considered in our simulation. Ranking for each performance parameter has been displayed in Table 3 as A indicates a higher ranking and ranking decrease up to C.

For future work, we plan to extend the simulation for the new algorithm which would comprise all the advantage of each algorithm. There would be hybridization of RED, SFQ, and REM to provide the better results.

References

[1] P.E. Meckenney, “Stochastic Fair Queuing”, In proceedings of INFCOM, vol. 2, pp. 733-740, June 1990.

[2] S. Floyd and V. Jacobson. Random early detection gateways for congestion avoidance. IEEE/ACM Trans. On Netwoking, vol.1, no.4, pp.397–413, August 1993.

[3] D.Lin and R. Morris, “Dynamics of random early detection,”, ACM SIGCOMM Computer Communication Review, vol. 27,no.4, pp.127-137, October 1997.

[4] W. Feng, K. G. Shin,, D. D. Kandlur, , and D. Saha,, “The BLUE Active Queue Management Algorithms,” IEEE/ACM Transaction on Networking, vol. 10, no. 4, pp.513 – 528, August 2002.

[5] S.Athuraliya, Victor H. Li, Steven H. Low and Qinghe YinK. Elissa, “REM: Active Queue Management,” *IEEE Network*, vol. 15 ,no.3, pp. 48 – 53, August 2002.

[6] S. Patel, P. Gupta and G. Singh, “ Performance Measure of Drop tail and RED,” In proceedings of *ICECT* , pp. 35 –38, June 2010.

[7] Network Simulator Manual, http:www.isi.edu/nsam/ns/index.html.

Bidirectional Agglomerative Hierarchical Clustering using AVL Tree Algorithm

Hussain Abu-Dalbouh[1] and Norita Md Norwawi[2]

[1] Faculty of Science & Technology
University Sains Islam Malaysia (USIM)
Bandar Baru Nilai, 71800 Nilai
Negeri Sembilan
Malaysia

[2] Faculty of Science & Technology
University Sains Islam Malaysia (USIM)
Bandar Baru Nilai, 71800 Nilai
Negeri Sembilan
Malaysia

Abstract

The hierarchy is often used to infer knowledge from groups of items and relations in varying granularities. Hierarchical clustering algorithms take an input of pairwise data-item similarities and output a hierarchy of the data-items. This paper presents Bidirectional agglomerative hierarchical clustering to create a hierarchy bottom-up, by iteratively merging the closest pair of data-items into one cluster. The result is a rooted AVL tree. The *n* leafs correspond to input data-items (singleton clusters) needs to *n/2* or *n/2+1* steps to merge into one cluster, correspond to groupings of items in coarser granularities climbing towards the root. As observed from the time complexity and number of steps need to cluster all data points into one cluster perspective, the performance of the bidirectional agglomerative algorithm using AVL tree is better than the current agglomerative algorithms. The experiment analysis results indicate that the improved algorithm has a higher efficiency than previous methods.

Keywords: Hierarchical, Clustering, Bidirectional algorithm, agglomerative**,** AVL tree

1. Introduction

Recently, dramatic increases in the amount of information or data are being stored in electronic format. This accumulation of data has taken place at an explosive rate [1]. It has been estimated that the amount of information in the world doubles every 20 months and the size and number of databases are increasing even faster [1]. These have transformed societies into one that strongly depends on information and knowledge. Huge volumes of data, that have accumulated and generated, contains important information. These databases contain not only known information, but also new knowledge as well and are not easy to be extracted and understood.

This essentially requires the development of reliable and scalable analysis procedures to extract the hidden rules, technology and dramatic growth in applications such as internet search, digital imaging, and video surveillance have created many high-volume and high dimensional data sets. It is estimated that the digital universe consumed approximately 281 exabytes in 2007, and it is projected to be 10 times that size by 2011 (1 exabyte is 1018 bytes or 1,000,000 terabytes). Many domains started collecting and sorting data from different sources and the massive amounts of information from many fields, such as math, biology, medical science, business, banking, engineering, education, medical and DNA technology, have led to the accumulation of tremendous amounts of data. However, traditional clustering algorithms become computationally expensive when the data set to be clustered is large. Clustering is an area where the analysis of large data sets becomes a problem. Analyzing large data sets via traditional methods has moved from being tedious, to being highly computational cost.

2. Proposed Algorithm for Bidirectional Agglomerative Hierarchical Clustering using AVL tree in the case of single-linkage clustering method

In this section, in depth discussion is presented on how bidirectional algorithm using AVL tree works in the case of single-linkage clustering. The algorithm is an agglomerative scheme that erases nodes in the tree as old clusters are merged into new ones.

The clustering are assigned sequence numbers 0,1,....., (*n/2*), (*n/2*) +1 and L(*k*) is the level of the *k*th clustering. A cluster with sequence number *m* is denoted (*m*) and the proximity between clusters (*r*) and (*s*) is denoted d [(*r*),(*s*)]. The algorithm composed of the following steps:

BIDIRECTIONAL AGGLOMERATIVE HIERARCHICAL CLUSTERING USING AVL TREE ALGORITHM

1. *Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.*
2. *Arrange the pairs from minimum distance (similarities) to the maximum distance.*
3. *Count how many pairs, say (n) pairs.(If n >=3) do,*
   {

3.1 *Find the median/root*

3.2 *Divide the pairs in two sides according the median. Say left side and right side.*

3.3 *Find the least dissimilar pair of clusters in the left and right current clustering, say pair (r), (s), according to*
d[(r),(s)] = min d[(i),(j)]
where the minimum is over all pairs of clusters in the current clustering.

3.4 *Check if left and right side have at least one similar object(element)then merge it together in one cluster, and find minimum is over all pairs of clusters in the current clustering.*
   }
   Else
   {

3.5 *Find the least dissimilar pair of clusters in the left and right current clustering, say pair (r), (s), according to*
d[(r),(s)] = min d[(i),(j)]
where the minimum is over all pairs of clusters in the current clustering.
   }

4. *Increment the sequence number: m = m +1. (In both sides) Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to*
L(m) = d[(r),(s)]

5. *Update the tree, T, by deleting the nodes corresponding to clusters (r) and (s) and adding a node corresponding to the newly formed cluster. The proximity between the new cluster, denoted (r,s) and old cluster (k) is defined in this way:*
d[(k), (r,s)] = min d[(k),(r)], d[(k),(s)].
6. *If all objects are in one cluster, stop. Else, go to step 2.*

Fig. 1: Pseudo code of the bidirectional agglomerative hierarchical clustering using AVL tree algorithm

2.1 Complexity of bidirectional agglomerative hierarchical clustering using AVL tree algorithm

Initially each of the *n* objects to be clustered is in a cluster by itself, in step 1 of each loop iteration the Tree *T* has nodes for each of the m remaining clusters. The number of clusters decreases by one for step 8 and 9. When step 9 completes, the revised tree T has a nodes for each of the (*m-1*) remaining clusters. Table 1 shows the complexity to the major steps of the algorithm.

Table 1: Paradigmatic bidirectional agglomerative hierarchical clustering using AVL tree algorithm

Algorithm	***Time complexity***
1. *Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.*	
2. *Arrange the pairs from minimum distance (similarities) to the maximum distance.*	
3. *Count how many pairs, say (n) pairs. .(If n >=3) do,*	O(1)
3.1 *Find the median/root.*	O(1)
3.2 *Divide the pairs in two sides according the median. Say left side and Right side.*	O(1)
3.3 *Find the least dissimilar pair of clusters in the left and right current clustering, say pair (r), (s), according to d[(r),(s)] = min d[(i),(j)] where the minimum is over all pairs of clusters in the current clustering.*	O(1)
3.4 *Check if left and right side have at least one similar object (element) then merge it together in one cluster, and find minimum is over all pairs of clusters in the current clustering.* *Else;* 3.5 *Find the least dissimilar pair of clusters in the left and right current clustering, say pair (r), (s), according to d[(r),(s)] = min d[(i),(j)] where the minimum is over all pairs of clusters in the current clustering.*	O(1) O(1)
4. *Increment the sequence number: m = m +1. (In both sides) Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to: L(m) = d[(r),(s)]*	O(1)
5. *Update the tree, T, by deleting the nodes corresponding to clusters (r) and (s) and adding a node corresponding to the*	O(logn)

	newly formed cluster. The proximity between the new cluster, denoted (r,s) and old cluster (k) is defined in this way: *d[(k), (r,s)] = min d[(k),(r)], d[(k),(s)].*	
6.	*If all objects are in one cluster, stop. Else, go to step 2.*	

The complexity = Max { O(1), O(1), O(1), O(1), O(1), O(1), O(1), O(logn)} = O(logn).

In bidirectional agglomerative clustering using AVL tree, the distance of each cluster to all other clusters, and at each step the number of clusters decreases by one. Considering bidirectional *agglomerative* hierarchical clustering using AVL tree in the case of single-linkage clustering, if the number of objects are *n*, there are *n/2* (*in the best case*) or *n/2+1* (*in the worst case*) levels. Each level involves finding a minimum from tree *T* with time complexity O(1). Merging two clusters into a single cluster need O(1) then updating the proximity tree, *T*, by deleting the nodes corresponding to clusters need O(*logn*).

3. Related Works

Clustering is considered as an unsupervised classification process that means no predefined classes [4-6]. Clustering large data sets of high dimensionality has always been a serious challenge for clustering algorithms. Clustering of large datasets can be very difficult with the available clustering algorithms mainly due to the time complexity. Hierarchical methods rely on a distance function to measure the similarity between clusters. These methods do not scale well with the number of data objects. Their computational complexity is usually O(n^2). [7-9].

To solve the complexity problem, many improved algorithms are proposed [10-12]. That aimed to improve performance, some of these partition algorithms. It was chosen to reduce the distance calculation process. Like the method based on the k-d tree structure and pruning function proposed by [10], P-CLUSTER, the parallel clustering algorithm utilizes three kinds of pruning methods proposed by [13] and the parallel algorithm based on the k-d tree structure proposed by [14].

According to [15] by reducing distance or similarity calculation, the algorithm does not guarantee accuracy. [15] use parallel computing, assign the distance computing to show different nodes in a distributed environment, which improved the efficiency and ensure the effectiveness.

There are many recently developed representative of hierarchical clustering algorithm found in the literature that attempted and proposed for handling large data sets and to overcome the complexity time such as: i) Agglomerative Nesting (AGNES) [16] it with O(n^2) time complexity. ii) Divisive Analysis (DIANA) [16] it with O(n^2) time complexity. iii) Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) [17] it with O(n) time complexity. iv) Clustering Using REpresentatives (CURE) [18] it with O(n^2) time complexity, v) RObust Clustering using links (ROCK) [19] it with O(O(n^2)+nmmma +n2logn)) time complexity.

4. An Example

The following discussion on bidirectional agglomerative hierarchical clustering using AVL tree algorithm is based on a simple example of distances in kilometers between some Malaysian states. The method used is single-linkage. There are eleven data points: JHR, KED, KTN, MLK, NSN, PHG, PRK, PLS, PNG, SGR, and TRG.

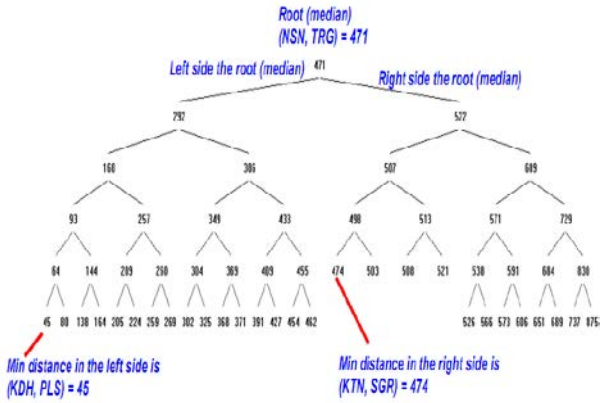


Fig. 2: First cluster in AVL tree

Based on Figure 2 the median/ root is NSN and TRG, at distance 471. The *Min* Left Side pair of states is KDH and PLS, at distance 45. These are merged into a single cluster called "KDH/PLS" and the Right Side pair of states is KTN and SGR, at distance 474. The level of the new cluster is L (KED/PLS) = 45, L (KTN/SGR) = 474, and the new sequence number is m = 1. In addition the left side and right side elements did not have any same element. Therefore no need to merge the left side cluster and right side cluster into a one cluster. The cluster, side, sequence number, elements and distances are shown in Table 2.

Table 2: First cluster

Cluster	***Side***	***Sequence No***	***Element***	***Distance***
KDH/PLS	Left side	1	KED and PLS	45
KTN/SGR	Right side		KTN and SGR	474

Then compute the distance from this new compound object in the left side to all other objects, and compute the distance from this new compound object in the right side to

all other objects. In single link clustering the rule is that the distance from the compound object to another object is *MIN* distance from any member of the cluster to the outside object. Therefore the distance from "KED/PLS" to JHR is chosen to be 830, which is the distance from "KED" to "JHR", and so on. The distances between this cluster and the remaining elements in the distance matrix are computed as shown below.

d (KED/PLS) JHR = *min* [*d* (KED, JHR), *d* (PLS,JHR)] = *d* (KED,JHR) = 830.
d (KED/PLS) KTN = *min* [*d* (KED, KTN), *d* (PLS, KTN)] = *d* (KED, KTN) = 409.
d (KED/PLS) MLK = *min* [*d* (KED, MLK), *d* (PLS, MLK)] = *d* (KED, MLK) = 609.
d (KED/PLS) NSN = *min* [*d* (KED, NSN), *d* (PLS, NSN)] = *d* (KED, NSN) = 526.
d (KED/PLS) PHG = *min* [*d* (KED, PHG), *d* (PLS, PHG)] = *d* (KED, PHG) = 684.
d (KED/PLS) PRK = *min* [*d* (KED, PRK), *d* (PLS, PRK)] = *d* (KED, PRK) = 257.
d (KED/PLS) PNG = *min* [*d* (KED, PNG), *d* (PLS, PNG)] = *d* (KED, PNG) = 93.
d (KED/PLS) SGR = *min* [*d* (KED, SGR), *d* (PLS, SGR)] = *d* (KED, SGR) = 462.
d (KED/PLS) TRG = *min* [*d* (KED, TRG), *d* (PLS, TRG)] = *d* (KED, TRG) = 521.

The distances between the right side cluster and the remaining elements in the distance matrix are computed as shown below.

d (KTN/SGR) JHR = *min* [*d* (KTN, JHR), *d* (SGR,JHR)] = *d* (SGR,JHR) = 368.
d (KTN/SGR) KTN = *min* [d(KTN, KED/PLS), *d* (SGR, KED/PLS)] = d (KTN, KED/PLS) = 409.
d (KTN/SGR) MLK = *min* [*d* (KTN, MLK), *d* (SGR, MLK)] = *d* (SGR, MLK) = 144.
d (KTN/SGR) NSN = *min* [*d* (KTN, NSN), *d* (SGR, NSN)] = *d* (SGR, NSN) = 64.
d (KTN/SGR) PHG = *min* [*d* (KTN, PHG), *d* (SGR, PHG)] = *d* (SGR, PHG) = 259.
d (KTN/SGR) PRK = *min* [*d* (KTN, PRK), *d* (SGR, PRK)] = *d* (SGR, PRK) = 205.
d (KTN/SGR) PNG = *min* [*d* (KTN, PNG), *d* (SGR, PNG)] = *d* (SGR, PNG) = 369 .
d (KED/PLS) TRG = *min* [*d* (KTN, TRG), *d* (SGR, TRG)] = *d* (KTN, TRG) = 168.

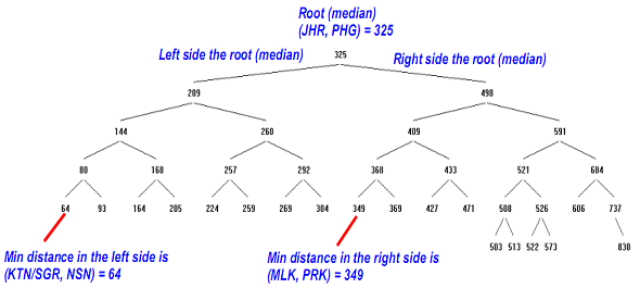


Fig. 3: Second cluster in AVL tree

Based on Figure 3 the root or the median is JHR and PHG, at distance 325. The minimum Left Side pair of states is KTN/SGR and NSN, at distance 64. These are merged into a single cluster called "KTN/SGR/NSN", and the minimum Right Side pair of states is MLK and PRK, at distance 349. The level of the new cluster is L (KTN/SGR/NSN) = 64, L (MLK/PRK) = 349, and the new sequence number is m = 2. In addition the left side and right side elements did not have any same element. Therefore no need to merge the left side cluster and right side cluster. The cluster, side, sequence number, elements and distances are shown in Table 3.

Table 3: Second cluster

Cluster	***Side***	***Sequence No***	***Element***	***Distance***
KTN/SGR/NSN	Left side	2	KTN, SGR and NSN	64
MLK/PRK	Right side		MLK and PRK	349

The distances between this cluster and the remaining elements in the distance matrix are computed as shown below.

d (KTN/SGR/NSN) JHR = *min* [*d*(KTN/SGR, JHR), *d* (NSN,JHR)] = *d*(NSN,JHR) = 304.
d (KTN/SGR/NSN) KTN = *min* [*d*(KTN/SGR, KED/PLS), *d* (NSN, KED/PLS)] = *d* (KTN, KED/PLS) = 409.
d (KTN/SGR/NSN) MLK = *min* [*d* (KTN/SGR, MLK), *d* (NSN, MLK)] = *d*(NSN, MLK) = 80.
d (KTN/SGR/NSN) PHG = *min* [*d* (KTN/SGR, PHG), *d* (NSN, PHG)] = *d*(SGR, PHG) = 259.
d (KTN/SGR/NSN) PRK = *min* [*d* (KTN/SGR, PRK), *d* (NSN, PRK)] = *d* (SGR, PRK) = 205.
d (KTN/SGR/NSN) PNG = *min* [*d* (KTN/SGR, PNG), *d* (NSN, PNG)] = *d* (SGR, PNG) = 369 .
d (KTN/PLS/NSN) TRG = *min* [*d* (KTN/SGR, TRG), *d* (NSN, TRG)] = *d* (KTN, TRG) = 168.

The distances between the right side cluster and the remaining elements in the distance matrix are computed as shown below.

d (MLK/PRK) JHR = *min* [*d*(MLK, JHR), *d* (PRK,JHR)] = *d* (MLK,JHR) = 224.
d (MLK/PRK) KTN = *min* [*d*(MLK, KED/PLS), *d* (PRK, KED/PLS)] = *d* (PRK, KED/PLS) = 257.
d (MLK/PRK) KTN/SGR/NSN = *min* [*d* (MLK, KTN/SGR/NSN), *d* (PRK, KTN/SGR/NSN)] = *d* (MLK, KTN/SGR/NSN) = 80.
d (MLK/PRK) PHG = *min* [*d* (MLK, PHG), *d* (PRK, PHG)] = *d* (MLK, PHG) = 292.
d (MLK/PRK) PNG = *min* [*d* (MLK, PNG), *d* (PRK, PNG)] = *d* (PRK, PNG) = 164 .
d (MLK/PRK) TRG = *min* [*d* (MLK, TRG), *d* (PRK, TRG)] = *d* (PRK, TRG) = 503.

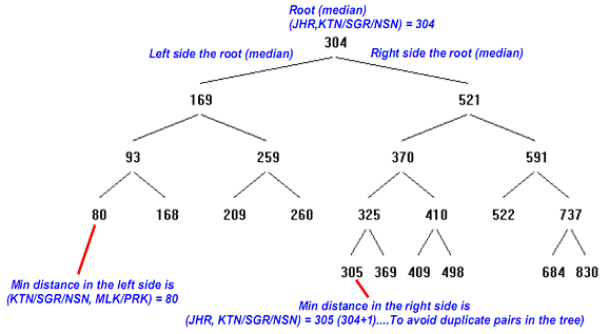


Fig. 4: Third cluster in AVL tree

Based on Figure 4 the root or the median is JHR and KTN/SGR/NSN, at distance 304. The minimum Left Side pair of states is KTN/SGR/NSN and MLK/PRK, at distance 80. These are merged into a single cluster called "KTN/SGR/NSN/MLK/PRK", and the minimum Right Side pair of states is JHR and KTN/SGR/NSN, at distance 305. The level of the new cluster is L (KTN/SGR/NSN/MLK/PRK) = 80, L (JHR/KTNSGR/NSN) = 305, and the new sequence number is m = 3. In addition the left side and right side have same element(s) in both sides. Therefore merge the left side cluster and the right side cluster in one cluster. The level of the new cluster is L (KTN/SGR/NSN/ JHR/MLK/PRK) = 385. The cluster, side, sequence number, elements and distances are shown in Table 4.

Table 4: Third cluster

Cluster	***Side***	***Sequence No***	***Element***	***Distance***
KTN/SGR/NSN/JHR/MLK/PRK	-	3	KTN, SGR, NSN, JHR, MLK, PRK	385

The distances between this cluster and the remaining elements in the distance matrix are computed as shown below.

d (KTN/SGR/NSN/JHR/MLK/PRK) JHR = *min* [*d* (KTN/SGR, JHR), *d* (NSN,JHR)] = *d* (NSN,JHR) = 304.
d (KTN/SGR/NSN/JHR/MLK/PRK) KTN = *min* [*d* (KTN/SGR, KED/PLS), *d* (NSN, KED/PLS)] = *d* (KTN, KED/PLS) = 409.
d (KTN/SGR/NSN/JHR/MLK/PRK) MLK = *min* [*d* (KTN/SGR, MLK), *d* (NSN, MLK)] = d(NSN, MLK) = 80.
d (KTN/SGR/NSN/JHR/MLK/PRK) PHG = *min* [*d* (KTN/SGR, PHG), *d* (NSN, PHG)] = *d*(SGR, PHG) = 259.
d (KTN/SGR/NSN/JHR/MLK/PRK) PRK = *min* [*d* (KTN/SGR, PRK), *d* (NSN, PRK)] = *d* (SGR, PRK) = 205.
d (KTN/SGR/NSN/JHR/MLK/PRK) PNG = *min* [*d* (KTN/SGR, PNG), *d* (NSN, PNG)] = *d* (SGR, PNG) = 369 .
d (KTN/SGR/NSN/JHR/MLK/PRK) TRG = *min* [*d* (KTN/SGR, TRG), *d* (NSN, TRG)] = *d* (KTN, TRG) = 168.

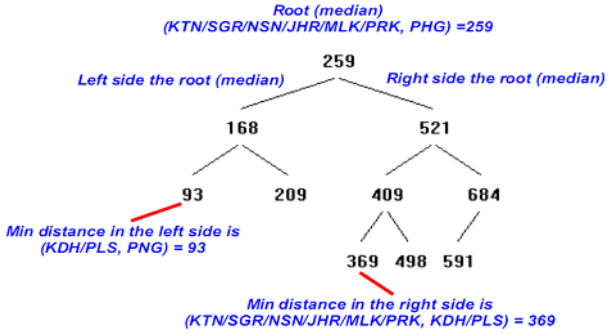


Fig. 5: Forth cluster in AVL tree

Based on Figure 5 the root or the median is KTN/SGR/NSN/JHR/MLK/PRK and PHG, at distance 259. The minimum left side pair of states is KDH/PLS and PNG, at distance 93. These are merged into a single cluster called "KDH/PLS/PNG", and the minimum Right Side pair of states is KTN/SGR/NSN/JHR/MLK/PRK and KDH/PLS, at distance 369. The level of the new cluster is L (KDH/PLS/PNG) = 93, L (KTN/SGR/NSN/JHR/MLK/PRK/KDH/PLS) = 369, and the new sequence number is m = 4. In addition the left side and right side have same element(s) in both sides. Therefore merge the left side cluster and the right side cluster in one cluster. The level of the new cluster is L (KTN/SGR/NSN/JHR/MLK/PRK/KDH/PLS/PNG) = 462. The cluster, side, sequence number, elements and distances are shown in Table 5.

Table 5: Forth cluster

Cluster	*Side*	*Sequence No*	*Element*	*Distance*
KTN/SGR/NSN/JHR/MLK/PRK/KDH/PLS/PNG	-	4	KTN, SGR, NSN, JHR, MLK, PRK, KDH, PLS, PNG	462

The distances between this cluster and the remaining elements in the distance matrix are computed as shown below.

d (KTN/SGR/NSN/JHR/MLK/PRK/KDH/PLS/PNG) PHG = *min* [*d* (KTN/SGR/NSN/JHR/MLK/PRK, PHG), *d* (KDH/PLS/PNG, PHG)] = *d* (KTN/SGR/NSN/JHR/MLK/PRK, PHG) = 259.
d (KTN/SGR/NSN/JHR/MLK/PRK/KDH/PLS/PNG) TRG = *min* [*d* (KTN/SGR/NSN/JHR/MLK/PRK, TRG), *d* (KDH/PLS/PNG, TRG)] = *d* (KTN/SGR/NSN/JHR/MLK/PRK, TRG) = 168.

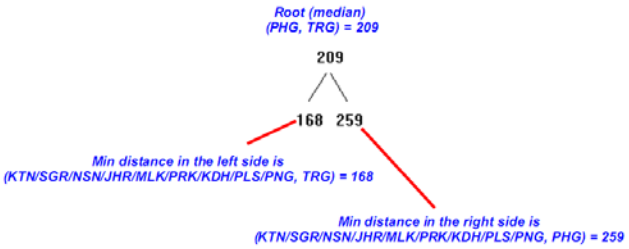


Fig. 6: Fifth cluster in AVL tree

Based on Figure 6 the minimum left side pair of states is KTN/SGR/NSN/JHR /MLK/PRK/KDH/PLS/PNG and TRG, at distance 168. These are merged into a single cluster called "KTN/SGR/NSN/JHR/MLK/PRK/KDH/PLS/PNG/TRG", and the minimum right side pair of states is KTN/SGR/NSN/JHR/MLK/PRK/KDH/PLS/PNG and PHG the level of the new cluster is L (KTN/SGR/NSN/JHR/MLK/PRK /KDH/PLS/PNG/TRG) = 168, and the new sequence number is m = 5. In addition the left side and right side have same element(s) in both sides. Therefore merge the left side cluster and the right side cluster in one cluster. The level of the new cluster is L (KTN/SGR/NSN/JHR/MLK/PRK /KDH/PLS/PNG/PHG/TRG) = 427. The cluster, side, sequence number, elements and distances are shown in Table 6.

Table 6: Fifth cluster

Cluster	*Side*	*Sequence No*	*Element*	*Distance*
KTN/SGR/NSN/JHR/MLK /PRK/KDH /PLS/PNG/ PHG/TRG	-	5	KTN, SGR, NSN, JHR, MLK, PRK, KDH, PLS, PNG,PHG,TRG	168

The final cluster is (KTN/SGR/NSN/JHR/MLK /PRK/KDH/PLS/PNG/PHG/TRG)

Table 7 shows an example of hierarchical clustering of eleven labeled points, namely JHR, KED, KTN, MLK, NSN, PHG, PRK, PLS, PNG, SGR and TRG. The example showing the following sequence of nested partitions:

Table 7: Overall cluster and sequence number

Sequence No	**Root**	**Cluster**		**Overall cluster**
		Left cluster	**Right cluster**	
m1	NSN, TRG	KDH/PLS	KTN/SGR	
m2	JHR, PHG	KTN/SGR /NSN	MLK/PRK	
m3	JHR, KTN/SGR/NSN	KTN/SGR /NSN/MLK/PRK	JHR and KTN/SGR /NSN	KTN/SGR/NSN/ JHR/MLK/PRK
m4	KTN/SGR/NSN/JHR/MLK/PRK,PHG	KDH/PLS/PNG	KTN/SGR /NSN/JHR /MLK/PRK and KDH/PLS	KTN/SGR/NSN/JHR/MLK/PRK/KDH/PLS/PNG
m5	PHG, TRG	KTN/SGR /NSN/JHR /MLK/PRK/KDH/PLS/PNG and TRG,	KTN/SGR /NSN/JHR /MLK/PRK/KDH/PLS/PNG and PHG	KTN, SGR, NSN, JHR, MLK, PRK, KDH, PLS, PNG,PHG,TRG

5. Results and Discussions

From the results of the manual analysis of applying bidirectional agglomerative hierarchical clustering using single-link method on Malaysian states example. Consider the number of objects are n, there are *n/2* (*in the best case*) or *n/2+1* (*in the worst case*) levels. Bidirectional algorithm in each level involves finding a minimum from tree *T* with time complexity O(1), then merge two clusters into a single cluster it need O(1), finally update the proximity tree, *T*, by deleting the nodes corresponding to clusters its need

O(*logn*). Therefore, this approach is more efficient in clustering huge amount of data and the performance of the bidirectional agglomerative hierarchical clustering using AVL tree algorithm is better from the bidirectional agglomerative hierarchical clustering using distance matrix and traditional agglomerative hierarchical clustering algorithms. In fact of the manual analysis on bidirectional agglomerative hierarchical clustering use distance matrix for single-link method. It is obvious *n* data point's need (*n/2*) in the best case or (*n/2+1*) in the worst case steps to merge all data points into one cluster. While the traditional agglomerative hierarchical clustering algorithms need (*n-1*) steps for merging all data points into single cluster. The bidirectional agglomerative hierarchical clustering using distance matrix and as for traditional agglomerative hierarchical clustering algorithms, the dissimilarity matrix *D* has a row and a column for each of the n elements. The overall complexity to merge all data points into single cluster is $O(n^2)$. Therefore the bidirectional use distance matrix and agglomerative hierarchical clustering method is a limitation to handle large datasets within a reasonable time and memory resources.

6. Conclusions

The complexity analysis is the algorithm performance in determining the resources such as execution time and memory usage necessary to execute it. Usually, the complexity of an algorithm is a function related to the input length/size to the number of fundamental steps. This paper proposes a hierarchical algorithm called bidirectional agglomerative hierarchical clustering algorithm based on the AVL tree by clustering the objects in left and right the median/root to enhance the complexity time of the current agglomerative hierarchical clustering algorithms and to reduce the gap between the flooding of information and the current agglomerative hierarchical clustering algorithm. One of the advantages of the proposed bidirectional agglomerative hierarchical clustering algorithm using AVL tree and that of other similar agglomerative algorithm is that, it has relatively low computational requirements. The overall complexity of the proposed algorithm is O(*logn*) and need (*n/2* or *n/2+1*) to cluster all data points in one cluster whereas the previous algorithm is $O(n^2)$ and need (*n-1*) steps to cluster all data points into one cluster.

Appendix

Table 8: Original distances

	JHR	***KDH***	***KTN***	***MLK***	***NSN***	***PHG***	***PRK***	***PLS***	***PNG***	***SGR***	***TRG***
JHR		830	689	224	304	325	573	875	737	368	521
KDH	830	0	409	606	526	684	257	45	93	462	521
KTN	689	409	0	609	538	371	391	454	386	474	168
MLK	224	606	609	0	80	292	349	651	513	144	508
NSN	304	526	538	80	0	259	269	571	433	64	471
PHG	325	684	371	292	259	0	427	729	591	259	209
PRK	573	257	391	349	269	427	0	302	164	205	503
PLS	875	45	454	651	571	729	302	0	138	507	566
PNG	737	93	386	513	433	591	164	138	0	369	498
SGR	368	462	474	144	64	259	205	507	369	0	455
TRG	521	521	168	508	471	209	503	566	498	455	0

Acknowledgments

The authors wish to thank Universiti Sains Islam Malaysia, Faculty of Science and Technology.

References

[1] S. Moran, Y. Hey, and K. Liu. An Empirical Framework for Automatically Selecting the Best Bayesian Classifier. Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K.

[2] S. Kanaujiya. Visual Data Mining. *Proceedings of 2nd National Conference on Challenges & Opportunities in Information Technology (COIT-2008) RIMT-IET, Mandi Gobindgarh. March 29, 2008.*

[3] J. *Gantz. F. 2008.* The diverse and exploding digital universe. Available online at: <http://www.emc.com/collateral/analyst-reports/diverse-exploding-digitaluniverse. pdf>.

[4] G. Bordogna, and G. Pasi. Hierarchical-Hyperspherical Divisive Fuzzy C-Means (H2D-FCM) Clustering for Information Retrieval. IEEE computer society 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology -Workshops.

[5] M. Halkidi, D. Gunopulos, M. Vazirgiannis, N. Kumar, and C. Domeniconi. A Clustering

Framework Based on Subjective and Objective Validity Criteria. ACM Transactions on Knowledge Discovery from Data (TKDD). Vol 1, No.4, 2008.

[6] A. Jain, Murty, P.J. Flynn. 1999. "Data Clustering: A Review", *ACM Computing Surveys*. Vol. *31*, *No.* 3. 1999, pp.264-323.

[7] C. Carpineto, S. Osiński, G. Romano, and D. Weiss. A Survey of Web Clustering Engines. *ACM* Computing Surveys. Vol. 41, No. 3, 2009.

[8] J. J. Hu, C. G. Tang, J. Peng, C. Li, C. A. Yuan, and A. L. Chen. A Clustering Algorithm Based Absorbing Nearest Neighbors. WAIM 2005, Volume 3739 of Lecture Notes in Computer Science, Springer, 2005, p.p 700-705.

[9] Ke-Bing, Z. 2007. Visual Cluster Analysis in Data Mining. (PhD Thesis). Macquarie University.

[10] K. Alsabti, S. Ranka, and V. Singh. An Efficient K-Means Clustering Algorithm, http://www.cise.ufl.edu/_ranka/, 1997.

[11] M. N. Joshi. Parallel K-Means Algorithm on Distributed Memory Multiprocessors. 2003

[12] L. Liping, and M. Zhi-Qing Meng. A method of choosing the initial cluster centers, Computer Engineering and Applications, pp.179-180.

[13] D. Judd, P. K. McKinley, and A. K. Jain. Large-Scale Parallel Data Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 8, AUGUST 1998, pp.871-876.

[14] M. Hemalatha, P. Ranjith Jebah Thangiah, K. Vivekanandan. A Distributed and Parallel Clustering Algorithm for Massive Biological Data. JCIT: Journal of Convergence Information Technology, Vol. 3, No. 4, 2008, pp. 84 -88.

[15] H. Gao, J. Jiang, L. She, and Y. Fu. A New Agglomerative Hierarchical Clustering Algorithm Implementation based on the Map Reduce Framework. International Journal of Digital Content Technology and its Applications, Vol. 4 No. 3, 2010.

[16] L. Kaufman, and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis", John While & Sons.

[17] T. Zhang , R. Ramakrishnan, and M. livny BIRCH: An efficient data clustering method for very large databases. In Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'96). 1996, Pp.103–114.

[18] S. Guha, R. Rastogi, and K. Shim. An efficient clustering algorithm for large databases. In Proceedings of SIGMOD, June 1998, pp.73–84.

[19] S. Guha, R. Rastogi, and K. Shim. ROCK: A Robust Clustering Algorithm for Categorical Attributes", In the Proceedings of the IEEE Conference on Data Engineering. 1999.

Hussain Mohammad Abu Dalbouh is a PhD student at the University Science Islam Malaysia (USIM), was born on the 26 of May 1982, his nationality Jordanian. He obtained his Bachelor's degree in Computer Information System in 2005 from the Al Yarmouk University, Jordan. He received his Master's degree in Information Technology from University Utara Malaysia (UUM) in 2009. His interest Areas: Artificial Intelligence (AI), Data Mining (DM), Visualization, Tree data structure, Data structure and algorithms.

Norita Md Norwawi is an Associate Professor at Universiti Sains Islam Malaysia. She obtained her Bachelor in Computer Science in 1987 from the University of New South Wales, Australia. She received her Master's degree in Computer Science from National University of Malaysia in 1994. In 2004, she obtained her PhD specializing in Temporal Data Mining and Multiagent System from University Utara Malaysia. As an academician, her research interests include artificial intelligence, multi-agent system, temporal data mining, text mining, knowledge mining, information security and digital Islamic application and content. Her works have been published in international conferences, journals and won awards on research and innovation competition in national and international level.

Graph based E-Government web service composition

Hajar Elmaghraoui [1], Imane Zaoui[2], Dalila Chiadmi[3] and Laila Benhlima[4]

[1] Department of Computer Science, Mohammad Vth University -Agdal, Mohammadia School of Engineers(EMI) Rabat, Agdal, Morocco

[2] Department of Computer Science, Mohammad Vth University -Agdal, Mohammadia School of Engineers(EMI) Rabat, Agdal, Morocco

[3] Department of Computer Science, Mohammad Vth University -Agdal, Mohammadia School of Engineers(EMI) Rabat, Agdal, Morocco

[4] Department of Computer Science, Mohammad Vth University -Agdal, Mohammadia School of Engineers(EMI) Rabat, Agdal, Morocco

Abstract

Nowadays, e-government has emerged as a government policy to improve the quality and efficiency of public administrations. By exploiting the potential of new information and communication technologies, government agencies are providing a wide spectrum of online services. These services are composed of several web services that comply with well defined processes. One of the big challenges is the need to optimize the composition of the elementary web services. In this paper, we present a solution for optimizing the computation effort in web service composition. Our method is based on Graph Theory. We model the semantic relationship between the involved web services through a directed graph. Then, we compute all shortest paths using for the first time, an extended version of the Floyd-Warshall algorithm.

Keywords: *Web services composition, optimization, graph theory; Floyd-Warshall, e-government.*

1. Introduction

Many countries around the world are attempting to strengthen and revitalize the quality and the efficiency of their public administration and make it more service oriented. In the last decade, electronic (e-) government has emerged as a solution to the problems of traditional public administrations such as high costs, poor quality services and corruption. According to the European Commission, e-government is "the use of Information and Communication Technologies (ICTs), in public administrations combined with organizational change and new skills in order to improve public services and democratic processes and strengthen support to public policies" [1]. Thus, ICT integration in government operations plays a crucial role in improving the quality and transparency of public services in several domains including social programs, healthcares, tax filling, voting, etc… ICT has then become crucial to a successful e-Gov program. By deploying public services via the Internet, communication is made easier for citizens and businesses, resources are federated, and finally are sped up substantially. Furthermore, adopting web services in e-Gov enables government agencies to provide value-added services through the service composition process. Web service composition allows flexible creation of new services by assembling independent and reusable service components. Traditionally, web services are composed either in a static or dynamic ways each time a user asks for a service. These methods are still tedious and very costly. Thus, one of the big challenges in e-service composition is optimizing the composition effort. By composing the most suitable web services with the lowest costs, we will certainly increase e-government efficiency, reduce considerably the response time and give more satisfying responses to users' queries. In this paper, we propose a graph based approach for optimizing web service composition. The approach is based on representing the web services semantic relationship using a directed graph built at the time of publishing. This graph is traversed to find the optimized combination of all component services that compose targeted services. The solution extends the Floyd-Warshall algorithm to reconstruct all shortest paths between all the vertices in the service graph. This operation is performed at the time of web service publishing. Indeed, by computing shortest paths before executing user queries, we optimize the composition time and costs.

The rest of this paper is organized as follows. In section 2, we present some related work. Section 3 outlines our optimization approach .In section 4, we present our web services model which is based on the graph theory. Section 5 describes our solution for optimizing web service composition, based on Floyd-Warshall algorithm. We illustrate our approach in section 6 with an example of retiring e-services. We conclude in section 7 with ongoing works.

2. Related Work

Many proposals of web services composition methods have been presented in recent years. For a detailed survey, we refer to [2, 3]. In this section, we present a brief overview of some techniques that deal with automatic web service composition. We consider only techniques that use service dependency information, graph models, and semantics. The simple idea behind dependency is that whenever a web service receives some input and returns some output, the output is somehow related or dependent on the given input. By using a graph model, the behavior of available web services is represented in terms of their input-output information, as well as semantic information about the web data. A graph is a collection of vertices or 'nodes' and a collection of edges that connect pairs of vertices. A graph may be undirected, meaning that there is no distinction between the two vertices associated with each edge, or its edges may be directed from one vertex to another. A weighted graph is a graph where each edge has a weight (some real number) associated with it. The dependency graph is used in finding a composite service to satisfy a given request.

Most of composition graph-based methods build web services dependency graphs at runtime. They use a search algorithm for traversing dependency graphs in order to compose services. The main difference between these methods is attributed to how they search the dependency graph. A*, Dijkstra, Floyd, Forward chaining, backward chaining and bidirectional search algorithms are examples of the most common search techniques.

Hashemian et al. [4] store I/O dependencies between available Web services in their dependency graph, and then build composite services by applying a graph search algorithm. In their graph, each service and I/O parameter is represented as a vertex, service's input and output are represented as incoming and outgoing edges, respectively. The authors consider only the matching and dependencies between input and output parameters without considering functional semantics, thus they cannot guarantee that the generated composite services provides the requested functionality correctly.

In [5], the authors use the backward chaining method in combination with depth first search to get the required services for a composite task. Their solution is rather abstract and does not clearly discuss execution plan generation algorithm.

Arpinar et al. [6] present an approach which not only use graphs for web service composition, but also use semantic similarity as we present in this work. They consider edges with weights and deploy a shortest-path dynamic programming algorithm based on Bellman-Ford's algorithm for computing the shortest path. For cost, the authors consider the execution time of each service and input/output similarity but they don't take into consideration the services' nonfunctional attributes.

Gekas et al. [7] develop a service composition registry as a hyperlinked graph network with no size restrictions, and dynamically analyze its structure to derive useful heuristics to guide the composition process. Services are represented in a graph network and this graph is created and explored during the composition process to find a possible path from the initial state to a final state. In order to reduce the time of searching, a set of heuristics is used. But according to the authors, creating the graph at the time of composition is very costly in term of computation and limits the applicability of graph-based approaches to the problem of web service composition.

Talantikite et al. [8] propose to pre-compute and store a network of services that are linked by their I/O parameters. The link is built by using semantic similarity functions based on ontology. They represent the service network using a graph structure. Their approach utilizes backward chaining and depth-first search algorithms to find sub-graphs that contain services to accomplish the requested task. They propose a way to select an optimal plan in case of finding more than one plan. However, they also create the graph at the time of composition which incurs substantial overhead.

In this paper, we propose a graph based solution which creates the graph at the time of publishing and therefore optimize the composition by reducing the computational effort at the time of composition. Our approach uses the Floyd-Warshall algorithm.

3. The optimization approach

Conceptually, our problem can be described as follows: "Given a set of available services, and given a goal, our aim is to automatically compose an optimal subset of services to satisfy the goal". We propose a solution based on Graph Theory. Our optimization approach consists of two fundamental pillars: i) representing the semantic relationships between all available web services using an oriented graph called SCG (Service Composition Graph), and ii) Applying a graph search algorithm in order to compute all shortest paths between all nodes. We store the algorithm results into a matrix called the Shortest Predecessor Matrix (SPM). SPM is used to reconstruct the shortest path which corresponds to an optimal automated service composition. In order to find the best combination of web services that meets the user goal, we propose to extend the Floyd Warshall graph search algorithm.

The SCG and the SPM are built at the time of publishing and updated each time new services are published or existing services are changed or removed from the repository. Thus, we avoid building the graph and applying the graph search algorithm each time a request for a service composition is made. By creating the graph and computing the shortest paths at the time of publishing, we reduce considerably the computational effort at the time of composition and thereafter the execution time of composition.

During the composition runtime, and for each query, we identify the starting and ending web services (called Vstrat and Vgoals respectively) into the SCG. Then, we extract from the SPM matrix the corresponding shortest path between VStart and Vgoals. Our optimization approach is summarized in the following process represented in Figure 1.

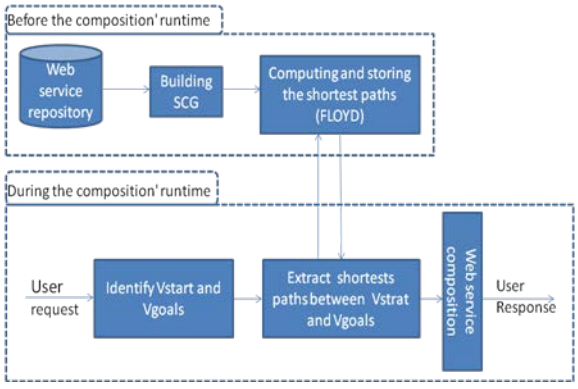


Fig.1. Optimization process

The details about this process are provided in the next sections.

4. Modeling web service composition as a directed weighted graph

In this section, we present the service composition graph (SCG).

4.1 Defining the service composition graph (SCG)

We start from a local Semantic Web Services repository, which is populated with OWL-S [9] descriptions (a popular and well-understood solution to support ontology-based Semantic Web). We will use the semantic description of the input and output parameters of the services to build a Service Composition Graph (SCG). Thus, building a SCG is based on semantic similarity between web services. In fact, for a set of web services, we need to check if two services are to be invoked in sequence during the composition process. This means that for each input of the former service, there is some output of the following service that is equivalent, more, or less general than the demanded input. In such cases, we can say that these services are semantically composable.

Formally, the SCG is a directed Graph G= (V,E,W), where V is the set of vertices representing the web services, E is the set of edges representing the semantic relationship between web services, and W is the set of edges's weights. To create the SCG, we follow the following procedure:

- For each web service WS_i in the repository, create a vertex V_i in the graph
- If two web services WS_i and WS_j represented respectively by vertices V_i and $V_j \in V$ have a semantic similarity among their inputs and outputs, then introduce an edge $V_i \rightarrow V_j$
- For each edge connecting vertices V_i and V_j, associate a weight W_{ij}

4.2 Measuring the semantic similarity

Semantic similarity stands for the degree of likeness between concepts. To compute the semantic similarities among services, we use subsumption reasoning as originally proposed by Paolucci et al. [10]. Subsumption reasoning verifies whether a concept is more general than another one. Given two web services represented by vertices V_i and V_j, this reasoning allows computing the degrees of similarity between the services using the scale: equivalent, subclass, subsumes and the following rules:

1-Exact match: If the output parameters of V_i and the input parameters of V_j are equivalent concepts.
2-Plug-in match: If the output of V_i is a sub-concept of the input of V_j (V_j subsumes V_i)
3-Subsumes match: If the input of V_j is a sub-concept of the output of V_i. (V_i subsumes V_j)
4-Fail match: No subsumption nor equivalence relation between V_i and V_j.
Thus, we associate an edge connecting vertices V_i to V_j if the degree of similarity between the outputs of V_i and the inputs of V_j is Exact, PlugIn or Subsumes.

4.3 Weighting the SCG

The weight is a key point of the model. It influences the choice of composition paths which directly affects the composition result. We associate to all the edges in the SCG a weight calculated using the semantic similarity value S between input and output parameters, and a function f(QOS) of non-functional properties of services that are cost, execution time, reliability, availability…etc .
In this paper, we propose three criteria as parameters for f(QOS) which are: cost, time and availability where:

- **Cost:** the fee that a requester has to pay for invoking the service V_i.
- **Execution time:** measures the expected delay time between the moment when V_i is invoked and when the results are received.
- **Availability:** is the probability that V_i is accessible.

Other non-functional properties can be considered in computing the weight of the edges, such as reliability, reputation, security…etc.

We calculate f(QOS) as follows:

$$f(QOS(V_i))=(\alpha*cost)+(\beta*execution\ time)+(\mu*availability) \quad (1)$$

Where α, β and μ are relative factors that can be defined by the system administrator.
The weight W_{ij} of a given edge $V_i \rightarrow V_j$ is computed as follows:

$$W_{ij} = f(QOS(V_i)) + S_{ij} \quad (2)$$

Where S_{ij} is the semantic similarity value between the input parameters of V_j and the output parameters of V_j.

5. Optimizing the web service composition

Given a user query which specifies web service's inputs and outputs, the composition problem involves automatically finding a directed acyclic graph of services from the SCG that can be composed to get the required service, when a matching service is not found. Our service composition research aims at reducing the complexity and time needed to generate and execute a composition. We also improve its efficiency by selecting the best possible services available. To achieve these optimization goals, we compute at the time of publishing all shortest paths between every pair of vertices in the SCG using the Floyd all-pairs shortest path which is a dynamic programming algorithm. Then, for each user query, we identify in the SCG the start and goal vertices based on semantic similarities between input and output parameters of the query and the SCG vertices. Thus, we calculate the shortest path between the start and goals vertices which represent the optimal services combination that meet the user needs.

5.1 Identifying start and goal vertices

Given the weighted directed graph SCG that models the web services in our repository, and when we receive a user query that requires service composition, we identify vertices and edges which represent the input and output parameters provided by the requester and we update temporarily the graph as follows:

- A starting node (V_{START}) is created and connected with all vertices (services) that contain at least one input provided by the requester;
- For each output demanded by the requester, create a goal node (V_{GOAL}) and connect it with all services providing this output.

These additional nodes are used to guide the service composition, which is based on the computation of minimum cost paths which are the shortest paths from the start node (representing the inputs) to the goal nodes (representing the outputs).

5.2 Shortest path issue

The shortest path problem has been widely used in many fields such as project planning, geographic information systems and military operations research. The classical algorithms to solve the shortest path problem are mainly Floyd algorithm and Dijkstra algorithm. Floyd algorithm is a multi-source shortest path algorithm, which is mainly used to calculate the shortest path among all nodes; whereas Dijkstra algorithm is a single-source shortest path algorithm, which is an efficient algorithm, used to calculate the shortest path from a source node to its all places nodes.

In order to reduce the overhead at runtime, we need to calculate the shortest paths for all the vertex pairs in advance and store this information for an eventual use at the time of service composition. Of course, this problem can be solved by applying Dijkstra's algorithm, but the

program will be complex and the computational overhead will be large to $O(N^4)$ [11], where N is the number of vertices. However, if we use Floyd algorithm, the computation will be simplified with a time complexity of $O(N^3)$ [11]. Floyd's algorithm is globally optimal, and the code is simple, easy to implement, and easy to integrate with other program modules [12]. In the next section, we will present a brief description of this algorithm.

Once the user request is received, we define the start node (V_{START}) and the set of goal nodes (V_{GOAL}) introduced in section 4.1, then we extract from the pre-computed list of shortest paths (already calculated using the Floyd algorithm and stored for future use), the shortest paths from the node (V_{START}) to each goal node member of (V_{GOAL}). This means that paths are found from the available input parameters, to the desired output parameters. Thus, we generate partial compositions that may have common vertices. Since our goal is to generate a single connected graph, we achieve this by eliminating duplicate paths by analyzing the intersections of the extracted paths.

5.3 Overview of Floyd- Warshall algorithm

The Floyd-Warshall algorithm [13] is an efficient Dynamic Programming [14] algorithm that computes the shortest paths between every pair of vertices in a weighted and potentially directed graph. This is arguably the easiest-to-implement algorithm for computing shortest paths in the literature [15]. The time complexity of this algorithm takes $O(N^3)$. A single execution of the algorithm will provide the lengths (summed weights) of the shortest paths between all pairs of vertices though it does not return details of the paths themselves. With some modifications of the algorithm, we create a method to reconstruct the actual shortest path between any two endpoint vertices. Path reconstruction runs in O(**N**) and thus the complexity of the algorithm is not affected. We describe below Floyd-Warshall algorithm with path reconstruction.

- Let DIST be a NxN adjacency matrix (N is the number of vertices), DIST(i,j) representing the length (or cost) of the shortest path from V_i to V_j. For each element DIST(i,j) assign a value equal to the cost of the edge going from Vi to Vj, or an infinity value if this edge does not exist.
- At each step, for each pair of vertices V_i and V_j, check if there is an intermediate vertex V_k so that the path from V_i to V_j through V_k is shorter than the one already found for V_i and V_j.
- Let NEXT be an N×N matrix that will contain the final path. The NEXT matrix is updated along with the DIST matrix such that at completion both tables are complete and accurate, and any entries which are infinite in the DIST table will be null in the NEXT table. The path from V_i to V_j is then the path from V_i to NEXT(i,j), followed by path from NEXT(i, j) to V_j.

In the next section, we present briefly the pseudo code of our Floyd-Warshall algorithm with path reconstruction.

5.4 Pseudo code of Floyd- Warshall with Path Reconstruction

```
# We assume an input graph of N vertices
# weight (i,j) is the weight of the edge from vertex Vi to Vj
: equal to infinity if such an edge does not exist and 0 if
i=j

BEGIN
For i = 1 to N
   For j = 1 to N
     #Initialization of the adjacency matrix
     DIST (i,j) = weight(i,j)
    # Initialization of the predecessor matrix
     If i! =j and there exists an edge from i to j
     Then
         NEXT (i,j) = i
      Else
         NEXT (i,j) = NIL
For k=1 to N                    # k is the intermediate vertex
   For i=1 to N
      For j=1 to N
# check if the path from i to j that goes through k is shorter
than the one already found
        If      DIST(i,k) + DIST(k,j) < DIST(i,j)
        Then
                # New shorter path length
                DIST (i,j) = DIST(i,k) + DIST(k,j)
               # The predecessor matrix
                NEXT (i,j)=k

Return DIST    # matrix of final distances
```

// GetPath(i,j) : function for path reconstruction between two vertices V_i and V_j.

```
If          DIST (i,j) = infinity
Then
            Return "no path"
IF          NEXT (i,j)= Null
Then
            Return " "    # there is an edge from Vi to Vj, with
no vertices between
Else
            Return          GetPath(i,NEXT(i,j))+NEXT(i,j)+
GetPath(NEXT(i,j),j)
END.
```

While the outcomes of our research are generic enough to be applicable to a wide range of applications, we use the area of e-government as a case study.

6. Case study

In this section, we illustrate our approach for web services composition with the example of retiring web services offered by the RCAR agency (Regime Collectif des Allocations de Retraite). The agency allows all members to access the retirement services via its web portal. The portal www.rcar.ma contains four spaces which are employer space, member space, recipient space and provider space. This categorization customizes the agency services according to different user profiles. All web services are stored in the repository. We model the relationship between them using a SCG as we have explained in section 3.

6.1 The retiring SCG

The SCG of the RCAR contains a hundred of web services. To simplify, we only present in Figure.2, a part of the SCG of the RCAR agency. Vertices are the web services described in table 1, edges represent the relationships between inputs and outputs and the weights represent the web services costs. Note that for confidentiality issues, web services and values are purely illustrative.

Table 1: Web services description

	Name	Description
V_1	GetAuth	Allows authenticating users using a login and password
V_2	FillForm	Return and fill adequate form for the demanded service
V_3	GetRegister	Allows new users to subscribe in order to get their login and passwords
V_4	SendReq	Send the user's request with the input data to the RCAR system
V_5	FillStatus	Allows users to create their status during the registration process and to update it.
V_6	SendResp	The system send the adequate response to the user's request
V_7	Unsubscribe	Allows user to unsubscribe, which automatically delete his profile and status
V_8	CheckData	Checks the user information to get the adequate response.

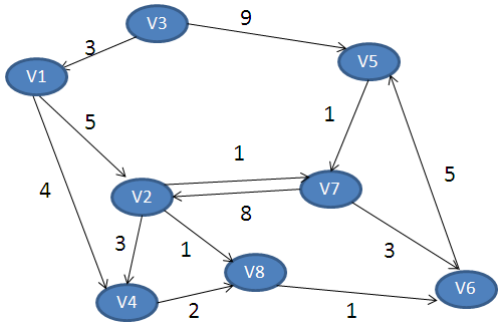


Fig.2. The retiring SCG

6.2 Shortest paths using Floyd-Warshall with path reconstruction

We apply the Floyd-Warshall algorithm with path reconstruction presented in section 4.4 to compute all shortest paths between all nodes in the SCG. The adjacency matrix, which represents the weights between all pairs of vertices in the SCG, is given bellow (3).

$$\begin{array}{c|cccccccc} & V_1 & V_2 & V_3 & V_4 & V_5 & V_6 & V_7 & V_8 \\ V_1 & 0 & 5 & \infty & 4 & \infty & \infty & \infty & \infty \\ V_2 & \infty & 0 & \infty & 3 & \infty & \infty & 1 & 1 \\ V_3 & 3 & \infty & 0 & \infty & 9 & \infty & \infty & \infty \\ V_4 & \infty & \infty & \infty & 0 & \infty & \infty & \infty & 2 \\ V_5 & \infty & \infty & \infty & \infty & 0 & \infty & 1 & \infty \\ V_6 & \infty & \infty & \infty & \infty & 5 & 0 & \infty & \infty \\ V_7 & \infty & 8 & \infty & \infty & \infty & 3 & 0 & \infty \\ V_8 & \infty & \infty & \infty & \infty & \infty & 1 & \infty & 0 \end{array} \quad (3)$$

The Shortest Predecessor Matrix (SPM) that will be used to extract all shortest paths, given bellow (4), will be stored to be manipulated at the time of service composition. Note that (V_i,V_j) holds a vertex V_k which is the direct predecessor of V_j on the least cost path between V_i and V_j

$$\begin{array}{c|cccccccc} & V_1 & V_2 & V_3 & V_4 & V_5 & V_6 & V_7 & V_8 \\ V_1 & \emptyset & V_1 & \emptyset & V_1 & V_8 & V_8 & V_2 & V_2 \\ V_2 & \emptyset & \emptyset & \emptyset & V_2 & V_7 & \emptyset & V_2 & V_2 \\ V_3 & V_1 & V_1 & \emptyset & V_1 & V_8 & V_7 & V_2 & V_2 \\ V_4 & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & V_4 \\ V_5 & \emptyset & V_7 & \emptyset & V_7 & \emptyset & V_7 & V_5 & V_7 \\ V_6 & \emptyset & V_7 & \emptyset & V_7 & V_6 & \emptyset & V_5 & V_7 \\ V_7 & \emptyset & V_7 & \emptyset & V_2 & V_6 & V_7 & \emptyset & \emptyset \\ V_8 & \emptyset & \emptyset & \emptyset & \emptyset & V_6 & V_8 & \emptyset & \emptyset \end{array} \quad (4)$$

We note also that in case of adding, deleting or modifying the services, the SCG will be updated as well as the adjacency matrix (3) and the Shortest Predecessor Matrix (SPM) (4). This update guarantees an accurate web service composition.

6.3 Delivering the retiring e-services based on the optimized web service composition.

Let us consider a member seeking to get an online certificate of his membership (V_{GOAL1}), to calculate his retirement pension (V_{GOAL2}) and to update his profile information (V_{GOAL3}). All these operations require user registration and authentication. These inputs are related to the start node (V_{START}). To get his request satisfied, several web services which are transparent for the user are involved. These web services are *V_1: GetAuth, V_4: SendReq, V_8:CheckData, and V_6:SendResp*. As explained before, the first step of our web service composition is to identify the root (V_{START}) and the targeted nodes V_{GOAL1}, V_{GOAL2} and V_{GOAL3} that answer the user request. The identification is based on comparing the inputs and outputs of the existing web services with the user's query. The result of this operation is connecting virtual vertices with a null weight to the SCG. Figure 3 illustrates this stage.

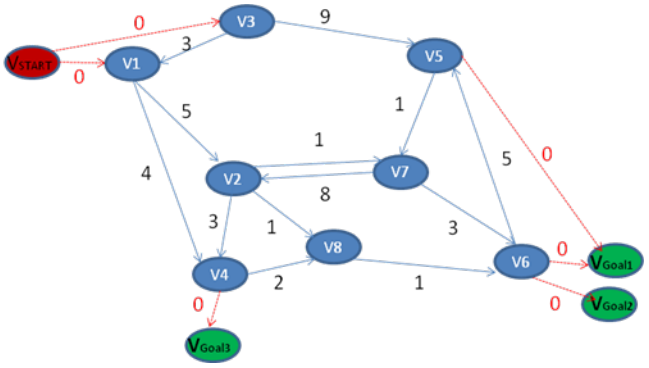


Fig.3. Identifying the START AND GOALS Vertices into the SCG

After identifying the input and output services in the SCG, which are connected to V_{START} and V_{GOAL1}, V_{GOAL2} and V_{GOAL3}, the second step of the web service composition is to extract from the matrix (2), all shortest paths between V_{START} and all V_{GOALS} . The results are given in table 2 and the optimal sub-graph of services that will meet the user needs is illustrated by figure 4.

Table 2: Shortest paths for web service composition

V_{START} *to* V_{GOAL1}	V1→V4→V8→V6→V_{GOAL1}
V_{START} *to* V_{GOAL2}	V1→V4→V8→V6→V_{GOAL2}
V_{START} *to* V_{GOAL3}	V1→V4→V_{GOAL3}

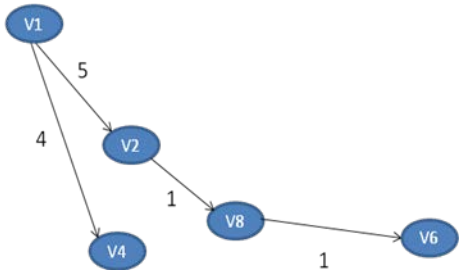


Fig.4. the optimal sub-graph of services that respond to the user request

7. Conclusion and ongoing work

Automatic composition of web services has drawn a great deal of attention recently. By composition, we mean taking advantage of currently existing web services to provide a new service that does not exist on its own. In this paper, we present a graph based approach for optimizing web service composition, which takes into consideration the semantic similarity of services computed using subsumption reasoning on their inputs and outputs. Our results include:

- Modeling the web service composition by means of a directed weighted graph, where the weight calculation takes into account the non-functional properties of services and the semantic similarity between them.
- Using Floyd algorithm to compute the shortest paths at the time of publication, in order to reduce the complexity and time needed to generate and execute a web service composition.

The implementation and evaluation of the solution proposed in this paper in real systems is the main focus of our ongoing work. We also intend to evaluate other cost policies such as reliability, reputation, security…etc.

Our near future work is mainly focusing on addressing reliability and availability of web services. Indeed, during the execution of web service composition, if one service fails or becomes unavailable, a failure recovery mechanism is needed to ensure that the running process is not interrupted and the failed service can be replaced quickly and efficiently.

References

[1] Commission of the European Communities; The Role of e-Government for Europe's Future. Brussels, 26.9.2003, COM (2003)567 final.

[2] A. Alamri et al., “Classification of the state-of-the-art dynamic web services composition”, In International Journal of Web and Grid Services , 2006, Vol. 2, pp. 148-166.
[3] J. Rao, X. Su, “A Survey of Automated Web Service Composition Methods”, In Semantic Web Services and Web Process Composition (SWSWPC'04), 2004, pp. 43-54.
[4] S. Hashemian, F. Mavaddat, “A graph-based framework for composition of stateless web services”, In Proceedings of ECOWS’06, IEEE Computer Society, Washington, DC, 2006, pp. 75–86
[5] R. Aydogan, H. Zirtiloglu, “A graph-based web service composition technique using ontological information”, 2007, Vol. 0. Los Alamitos, CA, USA: IEEE Computer Society, pp. 1154–1155.
[6] I.B. Arpinar et al., “Ontology-driven web services composition platform”. Inf. Syst. E-Business Management, 2005, 3(2):175–199.
[7] J. Gekas, M. Fasli, “Automatic Web Service Composition Based on Graph Network Analysis Metrics”, In Proceedings of the International Conference on Ontology, Databases and Applications of Semantics (ODBASE). Agia Napa, Cyprus, 2005, pp. 1571-1587.
[8] H.N. Talantikite et al., “Semantic annotations for web services discovery and composition”, Computer Standards Interfaces,31(6),1108-1117. Elsevier B.V, 2009.
[9] D. Martin et al., “OWL-S: Semantic Markup for Web Services”, http://www.w3.org/Submission/OWL-S/, 2004.
[10] M. Paolucci et al., “Semantic Matching of Web Services Capabilities”, In First International Semantic Web Conference, Sardinia, Italy, 2002, pp. 333-347.
[11] P. Krumins, http://www.catonmat.net/blog/mit-introduction-to-algorithms-part-twelve.
[12] D. Wei, “An Optimized Floyd Algorithm for the Shortest Path Problem”, Journal Of Networks, 2010, Vol 5, No 12.
[13] Cormen, T. H. et al, “Introduction to Algorithms”, MIT Press, 1990.
[14] http://en.wikipedia.org/wiki/Dynamic_programming
[15] https://vo.homelinux.org/wiki/code/FloydWarshall.

Mining Frequent Ranges of Numeric Attributes via Ant Colony Optimization for Continuous Domains without Specifying Minimum Support

Parisa Moslehi[1], Behrouz Minaei Bidgoli[2], Mahdi Nasiri[3], Erfan Nazari Fazel [4]

**[1] Computer Department, Islamic Azad University South Tehran Branch
Tehran, Tehran, Iran**

**[2] Computer Department, Iran University of science and Technology
Tehran, Tehran, Iran**

**[3] Computer Department, Iran University of science and Technology
Tehran, Tehran, Iran**

**[4] Computer Department, Islamic Azad University South Tehran Branch
Tehran, Tehran, Iran**

Abstract

Currently, all search algorithms which use discretization of numeric attributes for numeric association rule mining, work in the way that the original distribution of the numeric attributes will be lost. This issue leads to loss of information, so that the association rules which are generated through this process are not precise and accurate. Based on this fact, algorithms which can natively handle numeric attributes of a dataset would be interesting. In this paper a new approach to finding frequent intervals of numeric attributes is presented using Ant Colony Optimization for Continuous domains (ACO_R). The results show that this approach leads to more precise and accurate intervals in comparison with other approaches like discretizing into intervals with equal lengths.

Keywords- *Numeric Association Rule Mining; Preprocessing; Ant Colony Optimization; Data Mining*

1. Introduction

Data mining is the most instrumental tool in discovering knowledge from transactions [1][2][3]. Also data mining is known as an integral part of knowledge discovery in databases (KDD). Transactional database refers to the collection of transaction records, which in most cases are sales records. Data mining on transactional database focuses on the mining of association rules, finding the correlation between items in transaction records. The most important application of data mining is discovering association rules. This is one of the most important methods for pattern recognition in unsupervised systems [4]. Most of the association rule algorithms are based on methods proposed by Agrawal in [5] and [6]. The rules with numeric attributes cannot be discovered by these methods. That is why numeric association rule mining algorithms have been proposed. In a numeric association rule, attributes can be Boolean, numeric or categorical.

There has been proposed many numeric association rule mining algorithms. Each of these algorithms use a method to deal with numeric attributes while mining association rules or preprocessing numeric data.
Sirkant and Agrawal in [7], proposed an algorithm for mining association rules in large relational tables containing both quantitative and categorical attributes, by partitioning numeric attributes into a number of intervals.

Fukuda et al in [8], proposed an algorithm for mining optimized association rules for numeric attributes which uses computational geometry for computing optimized ranges.

Miller and Yang in [9], proposed an algorithm for mining association rules over interval data using Birch, an adaptive clustering algorithm.

Lent and Swami in [10], proposed an algorithm for the problem of clustering two-dimensional association rules in large databases. They used a geometric-based algorithm. BitOp for clustering.

Ke et al in [11], proposed an information theoretic approach to quantitative association rule mining. They used

discretizing numeric attributes and constructing a graph based on mutual information of attributes.

David and Yanhong in [12], proposed a fuzzy weighted association rule mining algorithm by transforming numeric and categorical data into fuzzy values.

Aumannn and Lindell in [13], proposed a statistical theory for quantitative association rules, based on the distribution of values of the quantitative attributes.

These approaches except fuzzy sets may have some drawbacks. The first problem is caused by sharp boundary between intervals which is not intuitive with respect to human perception. The algorithms either ignore or over-emphasize the elements near the boundary of the intervals. Furthermore, distinguishing the degree of membership for the interval technique without a priori knowledge is not easy [14]. Similarly, partitioning by means of fuzzy sets is not an easy task because it is hard to determine the most appropriate fuzzy sets for the numeric attribute values [15][16]. Characteristics of numeric attributes are in general unknown and it is unrealistic that the most appropriate fuzzy sets can always be provided by domain experts. That is why some researchers have proposed an evolutionary algorithm for automatically obtaining the fuzzy sets [2] for numeric attributes as a pre-processing

In recent years, the swarm intelligence paradigm received widespread attention in research. Two main algorithms of which that are popular swarm intelligence metaheuristics for data mining, are Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO). Swarm Intelligence is based on social behavior that can be observed in nature, such as ant colonies, flocks of birds, fish schools and bee hives, where a number of individuals with limited capabilities are able to come to intelligent solutions for complex problems [17].

An important insight of early research on ants' behavior was that most of the communication among individuals, or between individuals and the environment, is based on the use of chemicals produced by the ants. These chemicals are called pheromones. By sensing pheromone trails foragers can follow the path to food discovered by other ants. This collective trail-laying and trail-following behavior whereby an ant is influenced by a chemical trail left by other ants is the inspiring source of ACO. The first ACO algorithm for tackling combinatorial optimization problems was proposed by Dorigo et al in [18].

Recently, a continuous version of ACO metaheuristic has been proposed by Socha for tackling continuous attributes and solving continuous optimization problems in [19] and [20].

There is not any study which uses ACO_R for preprocessing numeric data and finding frequent ranges of them for data mining. In this paper we describe how we find frequent ranges of numeric attributes via ACO_R as the preprocessing phase of numeric association rule mining process. This leads us to have accurate and exact frequent ranges.

The rest of this paper is organized as follows. In section 2, a brief explanation of Ant Colony optimization for continuous domain (ACO_R) is discussed. In section 3, numeric association rule mining is discussed. The proposed algorithm for mining frequent ranges of numeric attributes via ACO_R is presented in section 4. Experimental setup and results are presented in section 5. Finally we conclude with a summary in section 6.

2. Ant Colony Optimization for Continuous Domain (ACO_R)

Socha in [23] proposed an extension of ACO algorithm to continuous domain for tackling continuous optimization problems. While ACO uses a discrete probability distribution for choosing a solution, ACO_R uses a probability density function (PDF) and samples it. A Gaussian function is used as PDF in ACO_R.

In ACO a pheromone table is used to store pheromone information. ACO_R uses a solution archive of size k in order to describe the pheromone distribution over the search space. Considering the solution archive as a matrix, each entry is called s_j^i where i=1,2,…,n is the number of dimensions and j=1,2,…,k is the number of complete solutions to the problem.

First the archive is initialized with k random solutions. These solutions are ranked based on their quality. The weight ω_j of solution s_j is calculated according to its rank:

$$\omega_j = \frac{1}{qk\sqrt{2\pi}} e^{\frac{-(j-1)^2}{2q^2k^2}} \quad (1)$$

Where q is a parameter of the algorithm: the effect of q is that, if it is small the best-ranked solutions are strongly preferred and when it is large, the probability becomes more uniform [21].

Each ant chooses a solution from the archive probabilistically for building its own solution. The probability of choosing s_j by an ant is:

$$p_j = \frac{\omega_j}{\sum_{r=1}^{k} \omega_r} \quad (2)$$

After choosing a solution s_j^i in the archive, each ant samples a Gaussian function:

$$P(x) = g(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (3)$$

Where μ and σ are the mean and standard deviation of the Gaussian function respectively. If an ant chooses a solution

s_j^i then the value of s_j^i is assigned to μ , and the standard deviation is assigned as follows:

$$\sigma \leftarrow \xi \sum_{r=1}^{k} \frac{|s_r^i - s_j^i|}{k-1} \quad (4)$$

Where ξ is a parameter of the algorithm that has the same effect as the pheromone evaporation parameter in ACO. The higher the value of ξ, the lower the convergence speed of the algorithm [21].

ACO_R aims to optimize a function, called objective function. So the sampled number is given to the objective function to calculate the result of it. After all the ants finished constructing their solutions, the solution archive is ranked, based on the objective function values of the solutions. This results in having the best solution on top of the archive and the worst ones on the bottom of it. Then the m worst solutions are removed from the archive, where m is the number of ants.

3. Numeric Association Rule Mining

3.1. Frequent Range

A frequent range is an interval of a numeric attribute in which most of the values of the attribute are fallen. It can have either a short length but rather cover remarkable number of values in transactions, or long length and cover quite a few numbers of values of transactions. So there is a trade-off between the length of an interval and the number of transactions which it covers.

3.2. Numeric association rule mining

An example of a numeric association rule in an employee database is as follows [14]:
Age ∈ [25, 36] ⋀ Gender=male→Salary ∈ [2000-2400] ⋀ Has-Car=Yes
(Support=4%, Confidence=80%)

In this association rule, "Age ∈ [25, 36] ⋀ Gender=male" is the antecedent of the rule and "Salary ∈ [2000-2400] ⋀ Has-Car=Yes" is the consequent of it. This rule says that, "4 percent (support) of the employees are the men whose ages range between 25 and 36 and their salaries range from 2000 to 2400 and have their own cars". While it can be said that, "80 percent (confidence) of the men between 25 and 36 years old receive salaries from 2000 to 2400 dollars and have their own cars". Here the intervals [25, 36] and [2000-2400] are frequent ranges.

In a transaction database the support and confidence of a rule is calculated by following equation [23]:

$$support, s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \; and \quad (5)$$

$$confidence, c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \quad (6)$$

where N is the total number of transactions. σ(X ∪ Y) and σ(X) is the frequency of occurrence of the itemset X and X ∪ Y respectively, which is called support count.

Support determines how often a rule is satisfied in the transaction, and confidence determines how often items in Y appear in transactions that contain X [22].

4. Proposed Work

Here we show how ACO_R is used in order to preprocess continuous data for numeric association rule mining. Since, ACO_R uses a Gaussian function as a PDF, for dealing with continuous variables; we got this idea and used the concept of Gaussian function described in section 2.1, in our process of generating frequent ranges. Based on this concept, each solution member in archive is assumed to be the central point of Gaussian normal distribution. The structure of the solution archive is slightly modified in order to store the standard deviation (σ) of each solution member in the archive, which is used to determine the boundaries of an interval of numeric attributes.

Each numeric attribute is given to ACO_R for finding its frequent ranges, so it is considered as one dimension or column of the solution archive. The structure of the modified solution archive is shown in Fig. 1.

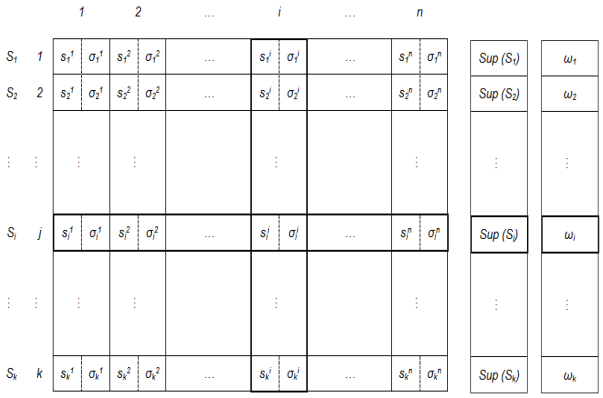


Fig.1 solution archive structure in proposed algorithm

A single dimensional solution archive is used in proposed algorithm and ACO_R is applied to each numeric attribute separately in order to find its frequent ranges. The algorithm consists of some procedures, each of which is described in the following.

4.1. Initialization( )

First the data are loaded into memory. Then the prerequisites of the objective function are prepared.

Here the objective function is the function that calculates the support of a numeric attribute. The structure of the objective function will be discussed later.

4.2. SolutionArchiveInitialization( )

In this procedure, the solution archive is filled by some uniform random data [21], based on a range that is defined by user in run-time. This range is usually selected in the way that covers the upper bound and lower bound of the numeric attribute value in the database. Furthermore, as this range has a vital effect on solutions of the algorithm, it can be selected in a way to focus on some particular parts of the numeric attribute's range. Then the weights are calculated according to the equation (1) and one solution is selected probabilistically according to equation (2).

After that, the vector of standard deviations is calculated according to the equation (4), considering s_r^i as a solution member and the central point of intervals, and k as the size of the solution archive. Choosing a proper value for ξ, affects the ability of the algorithm to find the correct solutions. Fig. 2 shows the effect of ξ on generated intervals that will be used by ACO_R.

In this work, a fixed and predefined value for ξ is used, but if its value changes dynamically, it will result in a more efficient algorithm.

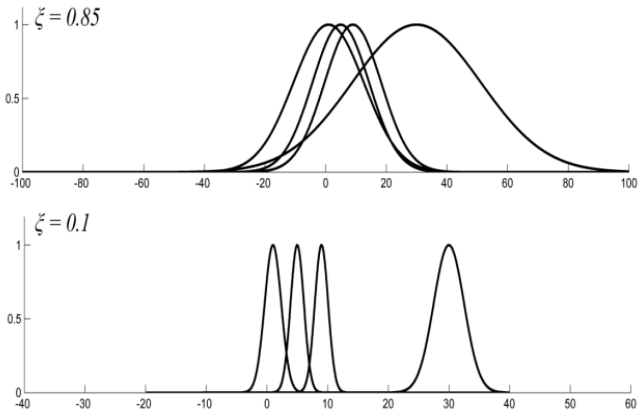


Fig. 2 The impact of ξ on generated intervals

4.3. AntBasedSolutionConstruction( )

This procedure consists of two parts:

- Constructing new solutions: In this part, the ants choose a solution according to the weights generated by Gaussian functions. They move through the archive and choose one row of it based on its associated weight (ω), which is calculated through equation (1). Then they construct a new solution by sampling the Gaussian function (g) of the selected solution.

- Calculating the objective function: Since the proposed algorithm aims to find some intervals of a solution which has an optimum support value, the objective function is the function of calculating support of sampled intervals. This function is calculated through the following equation:

$$Objective\ function = \frac{t}{N} \times \frac{1}{l} \quad (7)$$

Where t is the number of transactions in which the value of corresponding attribute falls in the interval, N is the total number of transactions, and l is the length of the interval. The impact of $\frac{1}{l}$ is that shorter intervals are prior to others, so we have more exact and accurate results.

4.4. PheromoneUpdate( )

As we said before, in ACO_R pheromone table is replaced by solution archive, so here like ACO_R, the pheromone update procedure adds a number of new solutions, each of which is generated by one ant, and eliminates the same number of bad solutions from the archive after ranking its solutions.

4.5. DaemonActions( )

Local search heuristics are not used in our proposed algorithm. However, the algorithm performance can improve using these heuristics.

All of these procedures are applied to one numeric attribute of the dataset, in parallel with the others.

5. Experimental Setup and Results

In this section, we present the experimental setup for evaluating the accuracy and precision of the results of our proposed approach. In order to do this, the resultant intervals are compared with the intervals produced by dividing the overall range of an attribute in a database into 10 intervals of equal length.

The datasets used for applying our algorithm are Basketball and Bolts which are available from Bilkent University Function Approximation Repository [23]. Table 1 shows the specifications of the both datasets. One characteristic of the proposed algorithm is that it is stochastic so that it has fluctuations in different runs. In order to get a better result, the user may execute several trials of the algorithm to tune the parameters up and get the results with the best solutions. Table 2 and Table 3 show the parameters which resulted in finding the best solutions by ACO_R.

Sigma coefficient specifies the boundary of intervals. Since our goal is to find optimum frequent ranges of numeric

attributes, and Basketball dataset contains four numeric attributes and one nominal attribute, which is defined as a class attribute, we do not apply our algorithm to this one. All of the attributes of Bolts dataset are numeric. The algorithm is implemented in C# and run on a personal computer with CPU corei5 Intel and 4G main memory. After several run-time iterations of the algorithm on each attribute of the dataset, four solutions are chosen to compare with the intervals of the same number, resulted from dividing the overall range of an attribute into 10 intervals of the same length. Table 4 shows the results.

Table 1: The dataset specifications

Basketball		Bolts	
Number of records	96	Number of records	40
Number of attributes	5	Number of attributes	8
Number of numeric attributes	4	Number of numeric attributes	8
Missing values	0	Missing values	0

Table 2: The parameters used to run the algorithm on basketball dataset

Basketball						
Parameter	m	ξ	q	k	n	Sigma coefficient
Assists-per-minute	2	5×10^{-4}	0.1	50	1	100
Height	2	2×10^{-1}	0.1	50	1	2
Time-played	2	10^{-7}	0.1	50	1	10^{7}
Age	2	10^{-6}	0.1	50	1	2×10^{6}

Table 3: The parameters used to run the algorithm on bolts dataset

Bolts						
Parameter	m	ξ	q	k	n	Sigma coefficient
Run	2	0.1	0.1	50	1	2
Speed1	2	0.09	0.1	50	1	2
Total	2	0.04	0.1	50	1	1
Speed2	2	0.005	0.1	50	1	3
Number2	2	0.009	0.1	50	1	2
Sens	2	0.009	0.1	50	1	2
Time	2	0.15	0.1	50	1	2
T20Bolt	2	0.5	0.1	50	1	2

Table 4: Frequent Ranges resuletd from ACO_R in comparison with the intervals of equal length

Basketball				
Numeric attribute	Intervals of equal length	Count	ACO_R Intervals	Count
Assists-per-minute	(0.07883,0.1082]	18	(0.0826,0.1053)	11
	(0.10826,0.1376]	14	(0.1057,0.1099)	6
	(0.22598,0.1965]	11	(0.1925,0.2561)	26
	(0.28484,0.3142]	0	-	-
Height	(172.9,177.2]	0	-	-
	(177.2,181.5]	5	(177.2450,183.6633)	9
	(181.5,185.8]	22	(181.8896,194.7184)	63
	(194.4,198.7]	21	(191.4370,197.7300)	29
Time-played	(13.143,16.206]	6	(13.2381,13.4124)	3
	(19.269,22.332]	12	(18.4885,20.1026)	6
	(22.332,25.395]	9	(19.5913,25.0038)	21
	(34.584,37.647]	14	(32.7565,38.8818)	26
Age	(25,26.5]	8	(22.3014,27.3910)	48
	(26.5,28]	24	(27.2348,28.5820)	12
	(28,29.5]	4	(28.3108,31.0539)	21
	(32.5,34]	6	(29.5503,34.9400)	29
Bolts				
Numeric attribute	Intervals of equal length	Count	ACO_R Intervals	Count
Run	(12.7,16.6]	4	(12.8865,41.0366)	28
	(−∞,4.9]	4	(0.6053,23.6242)	23
	(24.4,28.3]	4	(22.3889,39.5316)	17
	(16.6,20.5]	4	(14.4751,24.6106)	10
Speed1	(−∞,2.4]	16	(1.9979,2.3316)	16
	(3.6,4)	8	(3.9002,4.2133]	8
	(5.2,5.6]	0	-	-
	(5.6,+∞]	16	(5.7842,6.1026)	16
Total	(−∞,12]	16	(9.9542,10.0580)	16
	(18,20]	8	(19.9754,20.1676)	8
	(26,28]	0	-	-
	(28,+∞]	16	(29.8826,30.1286)	8
Speed2	(−∞,1.6]	16	(1.4908,1.5400)	16
	(1.9,2]	8	(1.9880,2.0171)	8
	(2.3,2.4]	0	-	-
	(2.4,+∞]	16	(2.4836,2.5155)	16
Number2	(−∞,0.2]	16	(-0.0165,0.0393)	16
	(0.8,1]	8	(0.9504,1.0037)	8
	(1.6,1.8]	0	-	-
	(1.8,+∞]	16	(1.9659,2.0082)	16
Sens	(−∞,1]	4	(-0.0999,0.0104)	4
	(5,6]	16	(5.8632,6.0213)	16
	(7,8]	8	(7.9190,8.0544)	8
	(9,+∞]	12	(9.9749,10.1130)	12
Time	(−∞,16.947]	18	(3.8690,20.5400)	23
	(16.947,29.954]	10	(17.1604,19.4943)	5
	(29.954,42.961]	4	(33.1563,40.7584)	3
	(94.989,107.996]	3	(99.7842,11.6971)	6
T20Bolt	(−∞,15.522]	14	(6.6923,27.9836)	25
	(31.926,40.128)	3	(32.6876,35.2207]	3
	(48.33,56.532]	0	-	-
	(56.532,64.734]	1	(56.2287,90.0640)	12

The results show that, considering the length of the intervals and the count of transactions included by them, the ones resulted from ACO_R are more precise and accurate, and there is a trade-off between the intervals' precision and accuracy (or length and count). Also, since the length of the ACO_R intervals are not equal, and the algorithm considers the support of them as the objective function, it doesn't converge to unpromising solutions.

Another important point is that we don't need to specify the minimum support threshold for each attribute, since ACO_R handles this issue as its objective function.

Furthermore, as there are some overlaps between ACO_R intervals and they are open intervals (not closed or half-bounded) because of using Gaussian function to build them, their boundaries are not sharp and separate so that we have more flexible intervals, and also the distribution of data is reflected. Another point is that sometimes ACO_R frequent ranges cover 2 or more ranges of equal length with low count.

6. Conclusion

An extension of Ant colony optimization algorithm to continuous domain called ACO_R, is a new metaheuristic for tackling continuous optimization problems. This study

proposed an algorithm which uses ACO_R and the notion of Gaussian functions used by it in order to mine frequent ranges of numeric attributes as a preprocessing step of numeric association rule mining process.

An interval resulted from our algorithm is open and reflects the distribution of the original data. ACO_R finds frequent ranges without any need to specify minimum support threshold. The algorithm is used for mining frequent ranges of a dataset which has numeric attributes and it has given satisfactory results in its application.

Our algorithm seems to provide useful extensions for practical applications since it overcomes the problems which other algorithms proposed for preprocessing numeric data cannot handle.

Mining frequent ranges by ACO_R considering other attributes in an n-dimensional solution archive may be presented as further works.

References

[1] Chen C.-H., Hong T.-P., and Tseng V.S., *A Cluster-Based Fuzzy-Genetic Mining Approach for Association Rules and Membership Functions*, IEEE International Conference on Fuzzy Systems, pp. 1411 - 1416, 2006.

[2] Kayaa M., Alhajj R., *Genetic algorithm based framework for mining fuzzy association rules*, Fuzzy Sets and Systems , Vol. 152, No. 3, pp. 587-601, 2005.

[3] Tsay Y. J., Chiang J. Y., *CBAR: an efficient method for mining association rules*, Knowledge-Based Systems, Vol. 18, No. 2-3, pp. 99-105, 2005.

[4] Qodmanan H. R., Nasiri M., and Minaei-Bidgoli B., *Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence*, Expert Systems with applications, Vol. 38, No. 1, pp. 288-298, 2010.

[5] Agrawal R., Imielinski T., and Swami, A., *Mining association rules between sets of items in large databases,* In proceedings of ACM SIGMOD conference on management of data, pp. 207-206, 1993.

[6] Agrawal R., Sirkant R., *Fast algorithms for mining association rules,* In proceedings of the 20th international conference on very large databases, Santiago, Chile, 1994.

[7] Srikant R., Agrawal R., *Mining quantitative association rules in large relational tables*, In: Proceedings of ACM SIGMOD international conference on Management of data, Vol. 25, No. 2, pp. 1-12, 1996.

[8] Fukuda T., Yasuhiko M., Sinichi M., Tokuyama T. Mining *optimized association rules for numeric attributes*. In: Proceedings of ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems. New York, pp. 182–191, 1996.

[9] Miller R.J., Yang Y., *Association rules over interval data*, In: Proceedings of ACM SIGMOD international conference on management of data, vol. 29, No. 2, pp. 452–461, 1997.

[10] Lent B., Swami A., and Widom J., *Clustering association rules*, In: Proceedings of IEEE international conference on data engineering, pp. 220-231, 2002.

[11] Ke K., Cheng J., and Ng W., *An information-theoretic approach to quantitative association rule mining*, Journal Knowledge and Information Systems, Vol. 16, No. 2, pp. 112-114, 2008.

[12] David L. O., Yanhong L., *Mining Fuzzy Weighted Association Rules.* In: 40th Annual Hawaii international conference on system, sciences HICSS, pp 53–62, 2007.

[13] Aumann Y., Lindell Y., *A statistical theory for quantitative association rules*, Journal of Intelligent Information Systems, Vol. 20, No. 3, pp.255–283, 2003.

[14] Alatas B., Erhan A., *Rough particle swarm optimization and its applications in data mining,* Soft Computing – A Fusion of Foundations, Methodologies and Applications, Vol.12, No. 12, pp. 1205-1218, 2008.

[15] Alatas B., Arslan A., *Mining of fuzzy association rules with genetic algorithms (in Turkish)*, Gazi University, J Polytech, Vol. 7, No. 4, pp. 269-276, 2004.

[16] Alatas B., Arslan A., *A novel approach based on genetic algorithm and fuzzy logic for mining of association rules (in Turkish)*, Firat University, J Sci Eng, Vol. 17, No. 1, pp. 42–51, 2005.

[17] Martens D., Baesens B., and Fawcett, T., *Editorial Survey: Swarm Intelligence for Data Mining,* Machine Learning, Vol. 82, No. 1, pp. 1-42, 2010.

[18] Dorigo M., Stutzle T., *Ant Colony Optimization,* A Bradford Book, MIT Press, Cambridge, Massachusetts, London, England, 2004.

[19] Socha K., *ACO for Continuous and Mixed-Variable Optimization,* Ant Colony Optimization and Swarm Intelligence, Computer Science, Vol. 3172, pp. 53-61, 2004.

[20] Socha K., Dorigo M., *Ant colony optimization for continuous domains*, European Journal of Operational Research, Vol. 185, No. 3, pp. 1155-1173, 2006.

[21] Socha K., *Ant Colony Optimization For Continuous and Mixed-Variable Domains*, Ph.D. Thesis, Universit´e Libre de Bruxelles, Brussels, Belgium, 2008.

[22] Tan P.-N., Steinbach M., and Kumar V., *Introduction to Data Mining*, Pearson International Edition Pearson Addison Wesley, 2006.

[23] Guvenir H.A., Uysal I., Bilkent University Function Approximation Repository. http://funapp.cs.bilkent.edu.tr, 2000.

Sine-Cosine-Taylor-Like Method for Hole-Filler ICNN Simulation

S. Senthilkumar* and Abd Rahni Mt Piah

Universiti Sains Malaysia
School of Mathematical Sciences
11800 USM Pulau Pinang, MALAYSIA

Corresponding author

Abstract

Sine-Cosine-Taylor-Like method is employed to improve the performance of image or handwritten character recognition under improved cellular non-linear network environment. The ultimate aim of this paper is focused on developing an efficient design strategy for simulating hole filler under ICNN arrays with a set of inequalities satisfying its output characteristics by considering the parameter range.

Keywords: *Improved Cellular Non-linear Network, Sine-Cosine-Taylor-Like Method, Hole-Filler Template, Edge Detection, Ordinary Differential Equations.*

1. Introduction

Cellular non-linear networks (CNNs), proposed by Chua and Yang [1, 2] are essentially non-linear analog electric circuits, locally interconnected for distributed computation. It is understood that the characteristics of cellular neural networks (CNNs) are analog, time-continuous, non-linear dynamical systems and formally belong to the class of recurrent neural networks. CNNs have been proposed by Chua and Yang [l, 2], and they have found that CNN has many important applications in signal and real-time image processing [12]. Roska et al. [3] have presented the first widely used simulation system which allows simulation of a large class of CNN and is especially suited for image processing applications including signal processing, pattern recognition and solving ordinary and partial differential equations etc.

Ahmad and Yaacob [4-6] introduced sin-cos-Taylor-like method for solving stiff ordinary differential equations and proved the results are better than other methods. Runge-Kutta (RK) methods have become very popular and efficient tools for computational purpose [19-22] and many real-time problems are solved. Particularly Runge-Kutta algorithms are used to solve differential equations efficiently that are equivalent to approximate the exact solutions by matching *'n'* terms of the Taylor series expansion. The RK-Butcher algorithm has been introduced by Bader [7, 8] for finding truncation error estimates, intrinsic accuracies and early detection of stiffness in coupled differential equations that arises in theoretical chemistry problems. Oliveira [9] introduced popular RK-Gill algorithm for evaluation of effectiveness factor of immobilized enzymes. Ponalagusamy and Senthilkumar [10] discussed about the comparison of RK-fourth orders of variety of means and embedded means on multilayer raster CNN simulation.

In this article, the hole filing scheme under ICNN paradigm with sine-cosine-Taylor-like method is carried out and compared with explicit Euler, RK-Gill, RK-classical fourth order. It is significant to note that the explicit sine-cosine-Taylor-like method for solving hole filing CNN simulation is a formulation of the combination of a polynomial and the exponential function. This method requires extra work to evaluate a number of differentiations of the function involved. The result shows smaller errors when compared to the results from the explicit classical fourth-order Runge-Kutta (RK4) is of order-6 [4-6].

Using the existing RK-Butcher fifth order method, the hole filing has been studied via CNN simulation by Murugesh and Badri [23] and the same has been studied by Murugesan and Elango [24] using RK fourth order method. Dalla Betta et al. [25] implemented CMOS implementation of an analogy programmed cellular neural network. Qiang Feng et al. [29] proposed new automatic nucleated cell method with improved cellular networks. Further, they proved that the running speed is comparatively high due to the easy hardware implementation and high speed of CNN.

2. Brief Theoretical Study: Structure and Functions of Improved Cellular Neural Network

The architecture of standard $M \times N$ CNN [1,2] is composed of cells C(i, j) 's where $1 \le i \le M$; $1 \le j \le N$. $M \times N$ can be understood as the dimension of the digital image P to be processed. The dynamics of each cell is given via the equations as below

$$c\frac{dx_{ij}(t)}{dt} = \frac{-1}{R}x_{ij} + \sum_{k,l \in S_{i,j}(r)} a_{k,l}y_{i+k,j+1} + \sum_{k,l \in S_{i,j}(r)} b_{k,l}u_{i+k,j+1} + z_{i,j},$$

$$1 \le i \le M;\ 1 \le j \le N.$$

$$= \frac{-1}{R}x_{ij} + \sum_{k=-r}^{r}\sum_{l=-r}^{r} a_{k,l}y_{i+k,j+1} + \sum_{k=-r}^{r}\sum_{l=-r}^{r} b_{k,l}u_{i+k,j+1} + z_{i,j} \quad (1)$$

where $x_{i,j}$, $y_{i,j}$, $u_{i,j}$ and $z_{i,j}$ represent state variable, output variable, input variable and threshold variable respectively; $S_{i,j}(r)$ is the sphere of influence with radius r; The $a_{k,l}$'s and $b_{k,l}$'s are said to be the elements of the A template (feedback template) and B template (feed-forward template), respectively. The output $y_{i,j}$ is the piecewise linear function. C and R_x determine the recall time.

The constraint / conditions are given by

$$|x_{i,j}(0)| \le 1, |u_{i,j}| \le 1 \quad \text{where C} > 0 \quad (2)$$

and

$R_x > 0$

Take C and R_x at 1. The output function is defined as below:

$$y_{i,j} = f(X_{i,j}) = \begin{cases} 1 & x_{i,j} \ge T^* \\ x_{i,j} & |x_{i,j}| < T^* \\ -1 & x_{i,j} \le -T^* \end{cases} \quad (3)$$

$$T^* = \sum_{i=1}^{M}\sum_{j=1}^{N} P(i, j)/(MN) \quad (4)$$

where P(i,j) denotes the unitary grayscale of the pixel (i,j) in the input image P. $M \times N$ can be understood as the dimension of the digital image P. It is obvious that the range of T* is (0,1).

3. Hole Filler with ICNN Paradigm

In a bipolar image all the holes are filled and remains unaltered outside the holes in case of hole filing ICNN simulation [26-28]. Allow $R_x = 1, C = 1$ and take +1 represent for the black pixel and –1 for the white pixel. If the bipolar image is input with $U = \{u_{ij}\}$ into CNN and the images having holes enclosed by the black pixels then initial state values are set as $x_{ij}(0) = 1$. The output values are obtained as $y_{ij}(0) = 1, 1 \le i \le M, 1 \le j \le N$ from equation (2).

Consider the templates A, B and the independent current source I are

$$A = \begin{bmatrix} 0 & a & 0 \\ a & b & a \\ 0 & a & 0 \end{bmatrix}, \qquad a > 0, b > 0$$

$$B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad I = -1 \quad (5)$$

where the template parameters a and b are to be determined. In order to make the outer edge cells become the inner ones, normally auxiliary cells are added along the outer boundary of the image, and their state values are set to be zeros by circuit realization, resulting in the zero output values. The state equation (1) can be rewritten as

$$\frac{dx_{ij}(t)}{dt} = -x_{ij}(t) + \sum_{c(k,l) \in N(i,j)} A(i, j; k, l)y_{kl}(t) + 4u_{ij}(t) - I \quad (6)$$

For instance, here the cells C(i+1,j), C(i-1,j), C(i,j+1) and C(i,j-1) are the non-diagonal cells. Designing of hole-filler template and its various sub-problems are discussed using

CNN simulations [26-28]. Consider the same two problems by Yin et al. [26].

Problem 1: The input value u_{ij} = 1 for cell C(*i*,*j*), signaling the black pixel. Because the initial state value of the cell C(*i*,*j*), has been set to 1, x_{ij} (*t*) = 1, and from equation (3) its initial output value is also y_{ij} (*t*) = 1. According to the hole-filler demands, its eventual output should be y_{ij} (∞) = 1. To attain this result, set

$$\frac{dx_{ij}(t)}{dt} \geq 0 \tag{7}$$

Substituting this input u_{ij} = 1 and equation (6) into equation (7), we obtain

$$\frac{dx_{ij}(t)}{dt} = -x_{ij}(t) + a[y_{(i-1)j}(t) + y_{(i+1)j}(t) + y_{i(j-1)}(t) + y_{i(j+1)}(t)] + by_{ij}(t) + 3 \tag{8}$$

Combining equations (7) and (8) and considering the minimum value of $x_{ij}(t)$ =1 this case yields

$$a[y_{(i-1)j}(t) + y_{(i+1)j}(t) + y_{i(j-1)}(t) + y_{i(j+1)}(t)] + by_{ij}(t) + 2 \geq 0 \tag{9}$$

Problem 2: The input value of cell C(*i*,*j*) is u_{ij} = 1, signaling the white pixel. Substituting this input value into equation (7) gives

$$\frac{dx_{ij}(t)}{dt} = -x_{ij}(t) + a[y_{(i-1)j}(t) + y_{(i+1)j}(t) + y_{i(j-1)}(t) + y_{i(j+1)}(t)] + by_{ij}(t) - 5 \tag{10}$$

4. Numerical Integration Methods

The ICNN is described by a system of nonlinear differential equations. Therefore, it is necessary to discretize the differential equation for performing behavioral simulation. For computational purposes, a normalized time differential equation describing CNN is used by Nossek et al [11]:

$$f'(x(n\tau)) = \frac{dx_{ij}(n\tau)}{dt} = -x_{ij}(n\tau) + \sum_{k=-r}^{r}\sum_{l=-r}^{r} a_{k,l}y_{i+k,j+1} + \sum_{k=-r}^{r}\sum_{l=-r}^{r} b_{k,l}u_{i+k,j+1} + z_{i,j} \tag{11}$$

$1 \leq i \leq M; 1 \leq j \leq N$

$$y_{ij}(n\tau) = \frac{1}{2}\left[\left|x_{ij}(n\tau)+1\right| - \left|x_{ij}(n\tau)-1\right|\right],$$

$1 \leq i \leq M; 1 \leq j \leq N$ (12)

where τ is the normalized time. For the purpose of solving the initial-value problem, well established single step methods of numerical integration methods are used [13].

These methods can be derived using the definition of the definite integral

$$x_{ij}((n+1)\tau) - x_{ij}(n\tau)) = \int_{\tau_n}^{\tau_{n+1}} f'(x(n\tau))d(n\tau) \tag{13}$$

Explicit Euler's, the improved Euler predictor-corrector and the fourth-order (quartic) Runge-Kutta are the mostly widely used single step algorithm in the CNN behavioral raster simulation. These methods vary in the way they evaluate the integral presented in [7].

4.1 Explicit Euler's Method

Euler's method is the simplest of all algorithms for solving ordinary differential equations. It is an explicit formula which uses the Taylor-series expansion to calculate the approximation.

$$x_{ij}((n+1)\tau) = x_{ij}(n\tau) + \tau f'(x(n\tau)) \tag{14}$$

4.2 RK-Gill Method

The RK-Gill method discussed by Oliveira [9] is an explicit method which requires the computation of four derivatives per time step. The increase of the state variable x^{ij} is stored in the constant k_1^{ij}. This result is used in the next iteration for evaluating k_2^{ij} and repeat the same process to obtain the values of k_3^{ij} and k_4^{ij}.

$$k_1^{ij} = f'(x_{ij}(n\tau)),$$

$$k_2^{ij} = f'(x_{ij}(n\tau) + \frac{1}{2}k_1^{ij}),$$

$$k_3^{ij} = f'(x_{ij}(n\tau) + (\frac{1}{\sqrt{2}} - \frac{1}{2})k_1^{ij}) + (1 - \frac{1}{\sqrt{2}})k_2^{ij}), \tag{15}$$

$$k_4^{ij} = f'(x_{ij}(n\tau) - \frac{1}{\sqrt{2}}k_2^{ij} + (1 + \frac{1}{\sqrt{2}})k_3^{ij}),$$

Therefore, the final integration is a weighted sum of the four calculated derivates per time step given by

$$x_{ij}((n+1)\tau) = x_{ij}(n\tau) + \frac{1}{6}[k_1^{ij} + (2-\sqrt{2})k_2^{ij} + (2+\sqrt{2})k_3^{ij} + k_4^{ij}] \quad (16)$$

4.3 Classical Runge-Kutta Method

The RK-classical fourth order method is an explicit method which requires the computation of four derivatives per time step. The increase of the state variable x^{ij} is stored in the constant k_1^{ij}. This result is used in the next iteration for evaluating k_2^{ij} and repeat the same process to obtain the values of k_3^{ij} and k_4^{ij}.

$$\begin{aligned} k_1^{ij} &= \tau f'(x_{ij}(n\tau)), \\ k_2^{ij} &= \tau f'(x_{ij}(n\tau) + \frac{1}{2}k_1^{ij}), \\ k_3^{ij} &= \tau f'(x_{ij}(n\tau)) + \frac{1}{2}k_2^{ij}, \\ k_4^{ij} &= \tau f'(x_{ij}(n\tau)) + k_3^{ij}, \end{aligned} \quad (17)$$

Therefore, the final integration is a weighted sum of the four calculated derivatives per time step which is given by

$$x_{ij}((n+1)\tau) = x_{ij}(n\tau) + [(k_1^{ij} + 2k_2^{ij} + 2k_3^{ij} + k_4^{ij})/6] \quad (18)$$

where $f(.)$ is computed according to (2).

4.4 Sine-Cosine-Taylor-Like Method

Ahmad and Yaacob [4-6] discussed the explicit one step method by the composition of a polynomial and exponential function.

$$y_{n+1} = y_n + h(f_n + h(\frac{f_n^1}{2} + h(\frac{f_n^2}{6} + h(\frac{f_n^3}{24} + h\frac{f_n^4}{120})))) + \frac{f_n^5(\sin(z_n h) + \cos z_n h))}{z_n^6}(\exp z_n h) - 1 - hz_n (1 + hz_n(\frac{1}{2} + hz_n(\frac{1}{6} + hz_n(\frac{1}{24} + \frac{z_n h}{120}))))) \quad (19)$$

5. Simulation Results and Comparisons

The hole-filing simulated output presented is performed using a high power workstation, and the simulation time used for comparisons is the actual CPU time used. The settling time *Ts* is the time from start of computation until the last cell leaves the interval [-1.0, 1.0] which is based on a specified limit (e.g., |dx/dt|< 0.01). The computation time *Tc* is the time taken for settling the network and adjusting the cell for proper position once the network is settled. The simulation shows the desired output for every cell. Take +1 and -1 to indicate the black and white pixels, respectively. It is marked that the selected template parameters *a* and *b*, are restricted to the shaded area as shown in figure 2 for the simulation. The speed is one of the major concerns in the simulation. Hence, determination of the maximum step-size (Δt) that still yields convergence for a template can be helpful in speeding up the system. The speed-up can be achieved by selecting an appropriate step-size (Δt) for that particular template. Figure 3 shows the image before and after hole-filling by employing sine-cosine-Taylor-like method.

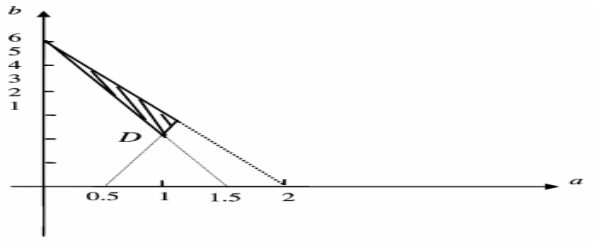


Fig. 2 Range of the template

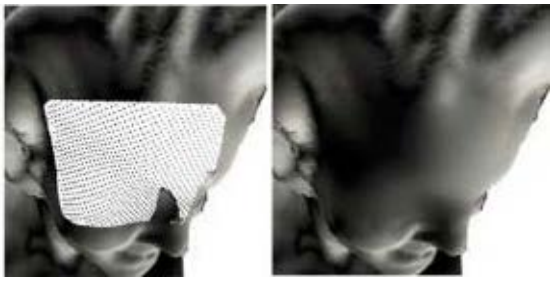


Fig.3 Image before and after hole-filling by employing sine-cosine-Taylor-like method

6. Conclusion

The importance of the hole filing CNN simulator is capable of performing hole-filing simulation for any kind as well as any size of input image or hand written character. It is a powerful tool for researchers to investigate the potential applications of CNN. It is of interest to mention that using the sine-cosine-Taylor-like

method in the hole filing simulation shows better feasible and effective output. It is observed that the hole is filled and the outside image remains unaffected that is, the edges of the images are preserved and intact. The templates of the improved cellular neural network are not unique and this is important in its implementation. In many language scripts, numerals and in images etc, there are many holes and the CNN as described above can be used in addition to the connected component detector. It is noteworthy to mention that each pixel receives the information only from its immediate neighbors, while the result $u_{xij}(\infty)$ is completely global in nature.

Acknowledgments

The first author would like to extend his sincere gratitude to Universiti Sains Malaysia for supporting this work under its post-doctoral fellowship scheme. Much of this work was carried out during his stay at Universiti Sains Malaysia in 2011. He wishes to acknowledge Universiti Sains Malaysia's financial support.

References

[1] Chua, L.O., Yang, L., "Cellular neural networks: Theory", IEEE Transactions on Circuits and Systems, Vol.35,1988,pp. 1257 – 1272.
[2] Chua, L.O., Yang, L., "Cellular neural networks: Applications", IEEE Transactions on Circuits and Systems, Vol. 35, 1988,pp. 1273 – 1290.
[3] Roska et al., "CNNM Users Guide", Version 5.3x, Budapest 1994.
[4] R. Ahmad and N. Yaacob, "Sin-Cos-Taylor-like method for solving stiff ordinary differential equations", Journal of Fundamental Sciences, 2005 pp.34-43.
[5] R. Ahmad, N. Yaacob, "Explicit Taylor-like method for solving stiff differential equations", Technical Report. LT/M Bil. 3/2004.
[6] R. Ahmad, N. Yaacob, "Explicit Sine-Taylor-like method for solving stiff differential equations", Technical Report. LT/M Bil. 8/2004.
[7] Bader, M., "A comparative study of new truncation error estimates and intrinsic accuracies of some higher order Runge-Kutta algorithms", Computers & Chemistry, Vol. 11, 1987, pp. 121-124.
[8] Bader, M., "A new technique for the early detection of stiffness in coupled differential equations and application to standard Runge-Kutta algorithms", Theoretical Chemistry Accounts, Vol. 99, 1988, pp. 215-219.
[9] Oliveira, S.C., "Evaluation of effectiveness factor of immobilized enzymes using Runge-Kutta-Gill method: How to solve mathematical undetermination at particle center point?", Bio Process Engineering, Vol.20, 1999, pp. 185-187.
[10] R. Ponalagusamy and S. Senthilkumar, "A comparison of RK-fourth orders of variety of means and embedded means on multilayer raster CNN simulation", Journal of Theoretical and Applied Information Technology, 2007, pp.7-14.
[11] Nossek, J.A., Seiler, G., Roska, T., Chua, L.O., "Cellular neural networks: Theory and circuit design", Int. J. of Circuit Theory and Applications, Vol.20, 1992, pp. 533-553.
[12] L.O. Chua and T. Roska, "The CNN universal machine part 1: The architecture", in Int. Workshop on Cellular Neural Networks and their Applications (CNNA)", 1992, pp. 1-10.
[13] W. H. Press, B. P. Flannery, S.A. Teukolsky and W.T. Vetterling, "Numerical Recipes. The Art of Scientific Computing", Cambridge University Press, New York, 1986.
[14] Gonzalez, R.C., Woods, R.E. and Eddin, S.L., "Digital Image Processing Using MATLAB", Pearson Education Asia, Upper Saddle River, N.J,2009.
[15] K.K. Lai and P.H.W. Leong,"Implementation of time-multiplexed CNN building block cell", IEEE. Proc. Of Microwave, 1996, pp. 80-85.
[16] K.K. Lai and P.H.W. Leong, "An area efficient implementation of a cellular neural network", IEEE, 1995, pp.51-54.
[17] T. Roska, "CNN software library", Hungarian Academy of Sciences, Analogical and Neural Computing Laboratory, 2000 [Online]. Available:http://lab.analogic.sztaki.hu/Candy/csl.html, 1.1
[18] Roska et al.,"CNNM Users Guide" , Version 5.3x, Budapest,1994.
[19] J.C. Butcher, "On Runge processes of higher order", Journal of Australian Mathematical Society, Vol.4, 1964, p.179.
[20] J.C. Butcher, "The Numerical Analysis of Ordinary Differential Equations: Runge-Kutta and General Linear Methods", John Wiley & Sons, U.K, 1987.
[21] J.C. Butcher,(1990), "On order reduction for Runge-Kutta methods applied to differential-algebraic systems and to stiff systems of ODEs", SIAM Journal of Numerical Analysis, Vol.27, pp. 447-456.
[22] Butcher, J.C., "The Numerical Analysis of Ordinary Differential Equations", John Wiley & Sons, U.K, 2003.
[23] V. Murugesh and Krishnan Badri, "An efficient numerical integration algorithm for cellular neural network based hole-filler template design", International Journal of Computers, Communications and Control, Vol.2, 2007, pp. 367-374.
[24] K. Murugesan and P. Elango, "CNN based hole filler template design using numerical integration technique", LNCS 4668, 2007, pp. 490-500.
[25] Dalla Betta, G.F., Graffi, S., Kovacs, M., Masetti, G, "CMOS implementation of an analogy programmed cellular neural network". IEEE Transactions on Circuits and Systems-Part–II Vol. 40(3), 1993, pp. 206–214.
[26] Chun-Li Yin, Jin-Liang Wan, Hai Lin, Wai-Kai Chen, "Brief communication, the cloning template design of a cellular neural network", Journal of the Franklin Institute Vol. 336, No. 199, pp.903-909.
[27] L. O. Chua and P. Thiran, "An analytic method for designing simple cellular neural networks", IEEE Transactions on Circuits and Systems-I, Vol. 38, No. 11, 1991, pp.1332-1341.
[28] Matsumoto, T., Chua, L.O. and Furukawa, R. "CNN cloning template: Hole filler", IEEE Transactions on Circuits and Systems, Vol. 37, Issue 5, 1990, pp. 635-638.
[29] Qiang Feng, Shenglin Yu and Huaiyin Wang, "A new automatic nucleated cell counting method with improved cellular neural networks", (ICNN), IEEE 2006 10th International

Workshop on Cellular Neural Networks and Their Applications, pp.1-4, 2006.

SUKUMAR SENTHILKUMAR was born in Neyveli Township, Cuddalore District, Tamilnadu, India on 18th July 1974. He received his B.Sc in Mathematics from Madras University in 1994, M.Sc in Mathematics from Bharathidasan University in 1996, M.Phil in Mathematics from Bharathidasan University in 1999 and M.Phil in Computer Science & Engineering from Bharathiar University in 2000. Also, he received PGDCA and PGDCH in Computer Science and Applications and Computer Hardware from Bharathidasan University in 1996 and 1997 respectively. He obtained a doctoral degree in the field of Mathematics and Computer Applications from National Institute of Technology [REC], Tiruchirappalli, Tamilnadu, India. Currently he is working as a post doctoral fellow at the School of Mathematical Sciences, Universiti Sains Malaysia in Pulau Pinang Malaysia. He was a lecturer / assistant professor in the Department of Computer Science at Asan Memorial College of Arts and Science, Chennai, Tamilnadu, India. He has published many good research papers in international conference proceedings and peer-reviewed / highly refereed international journals with high impact factor. He has made significant and outstanding contributions to various activities related to research work. He is also an associate editor, editorial board member, reviewer and referee for many scientific international journals. His current research interests include advanced cellular neural networks, advanced digital image processing, advanced numerical analysis and methods, advanced simulation and computing and other related areas.

ABD RAHNI MT PIAH was born in Baling, Kedah Malaysia on 8th May 1956. He received his B.A. (Cum Laude) in Mathematics from Knox College, Illinois, USA in 1979. He received his M.Sc in Mathematics from Universiti Sains Malaysia in 1986. He obtained his Ph.D in Approximation Theory from the University of Dundee, Scotland UK in 1993. He has been an academic staff member of the School of Mathematical Sciences, Universiti Sains Malaysia since 1981 and at present is an Associate Professor. He was a program chairman and deputy dean in the School of Mathematical Sciences, Universiti Sains Malaysia for many years. He has published various research papers in refereed national and international conference proceedings and journals. His current research areas include Computer Aided Geometric Design (CAGD), Medical Imaging, Numerical Analysis and Techniques and other related areas.

A Dynamic Load Balancing Algorithm in Computational Grid Using Fair Scheduling

U.Karthick Kumar[1]
[1]**Department of MCA & Software Systems,VLB Janki Ammal Arts and Science College, Coimbatore,TamilNadu – 641 042,India**

Abstract

Grid Computing has emerged as an important new field focusing on resource sharing. One of the most challenging issues in Grid Computing is efficient scheduling of tasks. In this paper, we propose a Load balancing algorithm for fair scheduling, and we compare it to other scheduling schemes such as the Earliest Deadline First, Simple Fair Task order, Adjusted Fair Task Order and Max Min Fair Scheduling for a computational grid. It addresses the fairness issues by using mean waiting time. It scheduled the task by using fair completion time and rescheduled by using mean waiting time of each task to obtain load balance. This algorithm scheme tries to provide optimal solution so that it reduces the execution time and expected price for the execution of all the jobs in the grid system is minimized. The performance of the proposed algorithm compared with other algorithm by using simulation.

Keywords: *Computational Grid, Scheduling, Load balancing, Fair scheduling, Mean Waiting Time, Execution Cost*

1. Introduction

Grid computing has been increasingly considered as a promising next-generation computing platform that supports wide area parallel and distributed computing since its advent in the mid-1990s [1]. It couples a wide variety of geographically distributed computational resources such as PCs, workstations, and clusters, storage systems, data sources, databases, computational kernels, and special purpose scientific instruments and presents them as a unified integrated resource [2].

In computational grids, heterogeneous resources with different systems in different places are dynamically available and distributed geographically. The user's resource requirements in the grids vary depending on their goals, time constraints, priorities and budgets. Allocating their tasks to the appropriate resources in the grids so that performance requirements are satisfied and costs are subject to an extraordinarily complicated problem. Allocating the resources to the proper users so that utilization of resources and the profits generated are maximized is also an extremely complex problem. From a computational perspective, it is impractical to build a centralized resource allocation mechanism in such a large scale distributed environment [3].

A computational grid is less expensive than purchasing more computational resources while obtaining the same amount of computational power for their computational tasks. A key characteristic of Grids is resources are shared among various applications, and therefore, the amount of resources available to any given application highly varies over time.

1.1 Dynamic Load Balancing

Load balancing is a technique to enhance resources, utilizing parallelism, exploiting throughput improvisation, and to reduce response time through an appropriate distribution of the application. Load balancing algorithms can be defined by their implementation of the following policies [15]

Information policy: It states the workload of a task information to be collected, when it is to be collected and from where.

Triggering policy: It determines the appropriate period to start a load balancing operation.

Resource type policy: It order a resource as *server* or *receiver* of tasks according to its availability status.

Location policy: It uses the results of the resource type policy to find a suitable partner for a server or receiver.

Selection policy: defines the tasks that should be migrated from overloaded resources (source) to most idle resources (receiver).

Load balancing algorithms are defined by two types such as static and dynamic [16]. Static load balancing algorithms allocate the tasks of a parallel program to workstations. Multicomputers with dynamic load balancing allocate or reallocate resources at runtime based on task information, which may determine when and whose tasks can be migrated. In this paper Dynamic Load Balancing Algorithm is implemented to multicomputers based on resource type policy.

The remaining section of this paper is organized as follows. Section 2 explains the related work. Section 3 detailed Problem formulation, Section 4 explained Fair Scheduling and Section 5 detailed the Dynamic Load Balancing Algorithm and section 6 Results and Discussion are detailed and conclusion and future work is presented in section 7.

2. Related Work

Fair Share scheduling [4] is compared with Simple Fair Task Order Scheduling (SFTO), Adjusted Fair Task Order Scheduling (AFTO) and Max-Min Fair Share Scheduling (MMFS) algorithm are developed and tested with existing scheduling algorithms. Somasundaram, S. Radhakrishnan compares Swift Scheduler with First Come First Serve (FCFS),Shortest Job First (SJF) and with Simple Fair Task Order (SFTO) based on processing time analysis, cost analysis and resource utilization[5]. For a multiprocessor system, the authors in [6] have shown that heuristic schemes that takes into account both the task deadline and EST better performs than the EDF, LLF, and MPTF algorithms. Finally, evaluation of different scheduling mechanisms for Grid computing is also presented in [7], such as the First Come First Served (FCFS), the Largest Time First (LTF), the Largest Cost First (LCF),the Largest Job First (LJF), the Largest Machine First (LMF), the Smallest Machine First (SMF), and the Minimum Effective Execution Time (MEET).

Pal Nilsson and Michal Pioro have discussed Max Min Fair Allocation for routing problem in a communication Network [8]. Hans Jorgen Bang, Torbjorn Ekman and David Gesbert has proposed proportional fair scheduling which addresses the problem of multi-user diversity scheduling together with channel prediction[9]. Daphne Lopez, S. V. Kasmir raja has described and compared Fair Scheduling algorithm with First Come First Serve (FCFS) and Round Robin(RR) schemes[10].

Load Balancing is one of the big issues in Grid Computing [11], [12]. Grosu and Chronopoulos [13], Penmatsa and Chronopoulos [14] considered static load balancing in a system with servers and computers where servers balance load among all computers in a round robin fashion. Qin Zheng, Chen-Khong Tham, Bharadwaj Veradale to address the problem of determining which group an arriving job should be allocated to and how its load can be distributed among computers in the group to optimize the performance and also proposed algorithms which guarantee finding a load distribution over computers in a group that leads to the minimum response time or computational cost [12]. Saravanakumar E. and Gomathy Prathima, discussing A novel load balancing algorithm in computational Grid [17]. M.Kamarunisha, S.Ranichandra, T.K.P.Rajagopal, dicuss about Load balancing Algorithm types and three policies are Information policy, Triggering Policy, and Selection Policy in Grid Environment[15][16].

3. Problem Formulation

Let the number of tasks be N that have to be scheduled as T_i, i=1, 2… N, is the duration of the task when executed on a processor in million instruction per second (MIPS). Let number of processors is M and its total computation capacity C is defined as

$$C = \sum_{j=1}^{M} c_j \quad (1)$$

Let M is the multiprocessor and its computation capacity of processor j is defined by c_j. The earliest time of task i started from processor j is the maximum of communication delay and completion time between i^{th} task and j^{th} processor. The completion time of task is zero, when no task allocated to processor j, otherwise it estimated the remaining time that are already allocated to processor j.

In the fair scheduling algorithm, the demanded computation rate X_i of a task T_i will play an important role. It estimated by the computation capacity that the Grid should allocate to task T_i for it to finish just before its requested deadline

4. Fair Scheduling

The scheduling algorithms do not adequately address congestion, and they do not take fairness considerations into account. Fairness [4] is most essential for scheduling of task. In Fair Scheduling, the tasks are allocated to multiple processors so that the task with unsatisfied demand get equal shares of time is as follows:

- Tasks are queued for scheduling according to their fair completion times.
- The fair completion time of a task is estimated by its fair task rates using a max-min fair sharing algorithm.
- The tasks are assigned to processor by increasing order of fair completion time.

In this algorithm, tasks with a higher order are completed first which means that tasks are taken a higher priority than the others which leads to starvation that increases the completion time of tasks and load balance is not guaranteed.

For this issue we propose a Load Balance (LB) Algorithm to give uniform load to the resources so that all task are fairly allocated to processor based on balanced fair rates. The main objective of this algorithm is to reduce the overall makespan.

5. Dynamic Load Balancing Algorithm

Dynamic load balancing algorithms make changes to the distribution of work among workstations at run-time; they

use current or recent load information when making distribution decisions. Multicomputers with dynamic load balancing allocate/reallocate resources at runtime based on a priori task information, which may determine when and whose tasks can be migrated. As a result, dynamic load balancing algorithms can provide a significant improvement in Performance over other algorithms.

Load balancing should take place when the scheduler schedules the task to all processors. There are some particular activities which change the load configuration in Grid environment. The activities can be categorized as following:

- Arrival of any new job and queuing of that job to any particular node.
- Scheduler schedules the job to particular processor.
- Reschedule the jobs if load is not balanced
- Allocate the job to processor when its free.
- Release the processor after it complete the whole job

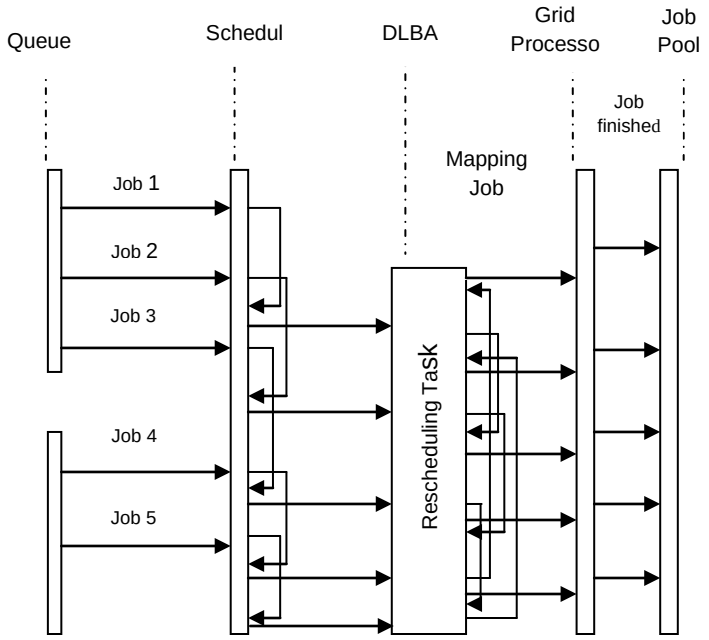


Fig.1: An Event Diagram for Dynamic Load Balancing Algorithm

Initialization of algorithm: N number of tasks that have to be scheduled and workload $w_i(x)$ of tasks are submitted to M number of processors.

Scheduling task: Scheduler allocates number of demanded tasks to M number of processors based on fair completion time of each task.

Load Balancing Algorithm: It applied when the processor task allocation is excessive than the other after scheduling the task.

Balancing criterion: Rescheduled the task for upper bound and lower bound processor based on $W_t(x)$.

Termination: This process is repeated until all the processor is balanced. Finally, obtain the optimal solution from the above process.

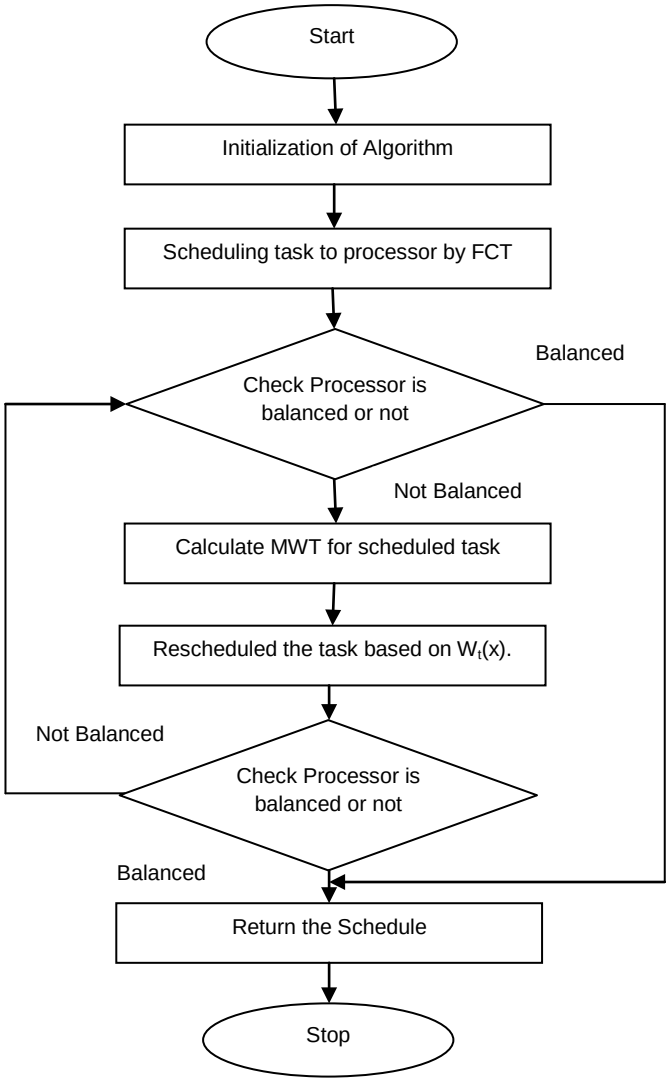


Fig.2 Flow Chart of Algorithm

5.1 Segment of code related to Algorithm

Input: A set of N task and M number of processor with computational capacity c_j.
Output: A schedule of N task
1. Create set of Queues.
2. qsize < N/M.
3. For each queue q_i in Q
4. While there are tasks in the queue do,
5. Assign demand rate of the task, X_i
6. k= C/N

7. If $X_i < k$
8. Assign X_i to i^{th} task as fair rate.
9. Else
10. Assign k to i^{th} task as fair rate.
11. Calculate fair completion time $t_i(x)$.
12. End while
13. End Loop
14. Arrange the task in increasing order based on their $t_i(x)$ and submitted to processor.
15. While (Load of any processor is greater than average load processor) do
16. Calculate mean waiting time for each scheduled task
17. If $Z_x^y > 0$
18 Migrated tasks are determined by using criteria of processor capacity.
19. Each processor which has least capacity is selected for migration.
20. End If
20. End While

5.2 Objective Evaluation

The task are scheduled by fair completion time $t_i(x)$, which is obtained by

$$t_i(x) = \int_1^M \left(d(xy)\, c(y) \Big/ c(y) \right) + w(x)/r(x)\ dx \qquad (2)$$

Here d(xy) is the earliest start time of i^{th} task to j^{th} processor i=0,1,…N and j=0,1,2,…,c(y) is the computational capacity of j^{th} processor, w(x) is workload of the task and r(x) is the fair rate of task computed by Max Min Fair Share approach.

Mean waiting time $W_t(x)$ is given by

$$W_t(x) = \sqrt{(\rho^n \lambda x^2 + x^n (1-\rho))/2(1-\rho)^2} + W(x) \qquad (3)$$

Where, W(x) is the constant delay made by the resource manager to assign to processor and arrival of all files necessary to run the task on processor.

$$W(x) = \rho x^2/2(1-\rho) \qquad (4)$$

To find the migration of processor by

$$Z_x^y = {}^1/_{W_t(x)} \int_1^M (W_t(x) - c(x))\, dx \qquad (5)$$

$$t(x_m) = max\{t[min(o, Z_x^y)], o\} \qquad (6)$$

$$t(y_n) = min\{t[max(o, Z_x^y)], o\} \qquad (7)$$

$$t(x_m) > t(y_n)\ and\ t(x_m), t(y_n) \geq 0 \qquad (8)$$

Based on mean waiting time task are rescheduled and allocated to processor. This is continued until all the processors are equally balanced to reach their minimum makespan.

5.3 Execution Cost

Our main objective is to reduce makespan and total execution cost by using load balancing algorithm. Specifically, we define the following for cost as

- $C(W(x), x)$ is cost incurred by a customer with seconds x ,if the expected constant delay is *W(x).*
- $W_t(x, l)$ is Mean Waiting time of processor with seconds x, if the rescheduling load balance algorithm is l .
- $Cost_{exe}(x, l)$ is Total execution cost of using load balance algorithm.

Cost optimization is defined by

$$Cost_{exe}(x) = \int_1^M C(W(x), x)\, dx \qquad (9)$$

The optimization problem is formulated by

$$Cost_{exe}(x, l) = \underbrace{Min}_{w_t(x,l)} \int_1^M C(W(x), x)\, W_t(x, l)\, dx \qquad (10)$$

As the primary function of a scheduler is to select a client to execute their tasks to processors when it is free. A key benefit of this algorithm is to reschedule the task by using $W_t(x)$ so that overall execution time and cost is reduced.

6. Result and Discussion

In this section proposed algorithm is simulated against

- Large set of Tasks as 256, 512, 1024, 2048 Million Instruction (MI).
- Large and varying number of processors as 8, 16, 32, 64 Million Instruction Per Second (MIPS).

Here, cost rate range is taken from 5 – 10 units is randomly chosen and assigned according to speed of the processor. Speed of the processor ranges from 0 – 1MIPS are randomly assigned to M processor. Below table shows the comparison results of proposed algorithm The work is approximately gives 45% - 25% less than EDF and 7% - 5% less than SFTO and AFTO and 5% - 2% less than MMFS for makespan. Also, LBA approximately show 30% - 25% less than EDF and 7% - 6% less than SFTO and AFTO 2% - 1% less than MMFS for Execution cost. The result shows better performance for Higher Matrix also. The following are the comparison result of existing and proposed method.

Table 1: Performance Comparison for 8 processors

Parameters	*Resource Matrix*	*EDF*	*SFTO*	*AFTO*	*MMFS*	*LBA*
Makespan	256 x 8	917.82	447.74	444.39	439.61	418.13
Cost		5506.91	4487.44	4468.54	4446.77	4023.57
Makespan	512 x 8	1121.32	1022.36	1010.09	858.54	836.72
Cost		7849.21	5111.8	5048.45	4292.71	4183.58
Makespan	1024 x 8	1825.33	1651.45	1686.17	1643.32	1599.82
Cost		10951.97	13211.63	11803.21	13180.96	12798.55
Makespan	2048 x 8	3596.42	3280.39	3247.63	3137.59	3095.82
Cost		28174.94	26243.11	25981.06	25100.75	24766.55

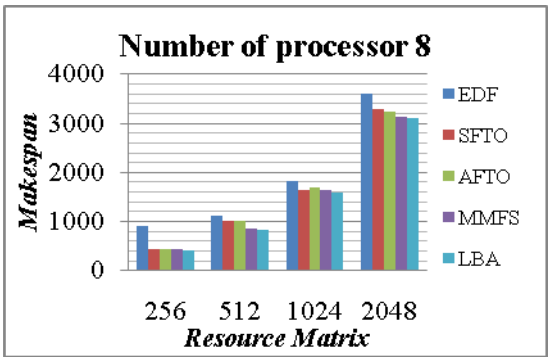


Fig 3: Performance Comparison for Makespan

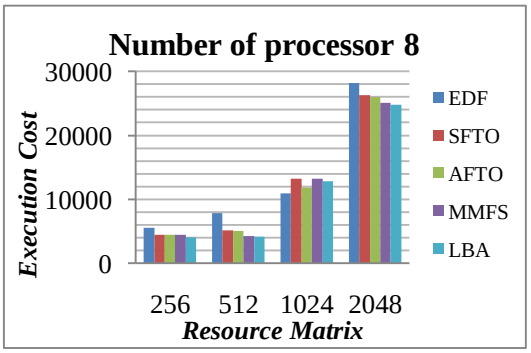


Fig 4 : Performance Comparison for Execution Cost

Table 2: Performance Comparison for 16 processors

Parameters	*Resource Matrix*	*EDF*	*SFTO*	*AFTO*	*MMFS*	*LBA*
Makespan	256 x 16	1466.72	304	300.65	295087	209
Cost		7332.11	1520	1511.0	1489.33	1045
Makespan	512 x 16	1366.48	553.89	555.37	545.76	483.57
Cost		8664.81	5868.91	5603..75	5508.24	4435.69
Makespan	1024 x 16	1540.27	1309.94	1296.35	1301.81	1231.43
Cost		9241.6	6549.72	6481.77	6519.05	6157.14
Makespan	2048 x 16	3352.67	2742.53	2761.98	2734.4	2641.04
Cost		26468.72	24682.76	27619.81	24652.09	23769.35

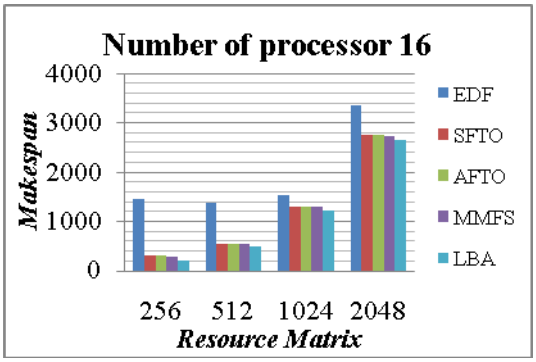


Fig 5: Performance Comparison for Makespan

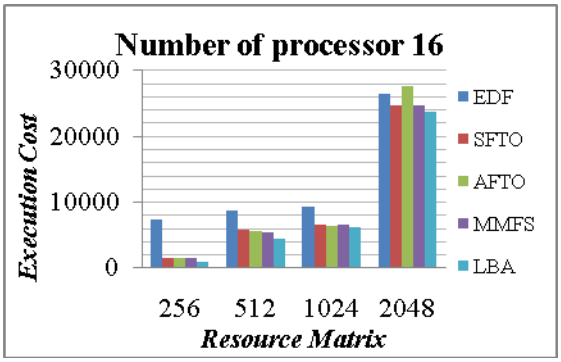


Fig 6: Performance Comparison for Execution Cost

Table 3: Performance Comparison for 32 processors

Parameters	*Resource Matrix*	*EDF*	*SFTO*	*AFTO*	*MMFS*	*LBA*
Makespan	256 x 32	206.05	183.43	180.08	175.30	114.64
Cost		1648.40	917.15	908.25	886.48	573.22
Makespan	512 x 32	744.80	580.60	577.25	574.27	464.54
Cost		5958.36	5225.43	5216.53	3445.64	2787.23
Makespan	1024 x 32	966.47	912.96	937.07	904.83	863.54
Cost		7731.78	4564.78	7512.53	4534.11	4317.72
Makespan	2048 x 32	1675.05	1427.97	1375.23	1370.11	1262
Cost		18725.38	11851.76	11377.09	11330.99	10359.85

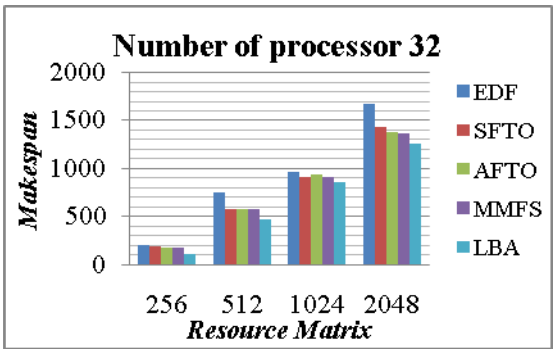


Fig 7: Performance Comparison for Makespan

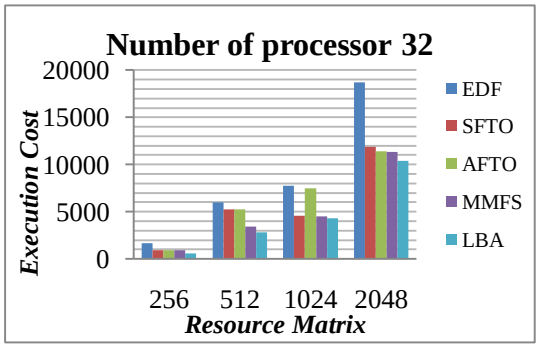


Fig 8: Performance Comparison for Execution Cost

Table 4: Performance Comparison for 64 processors

Parameters	*Resource Matrix*	*EDF*	*SFTO*	*AFTO*	*MMFS*	*LBA*
Makespan	256 x 64	305.35	281.93	278.80	273.80	211.45
Cost		2748.13	1691.60	1682.70	1660.93	1268.7
Makespan	512 x 64	966.67	600	596.65	591.87	450
Cost		5800.02	3000	2991.10	2969.33	2700
Makespan	1024 x 64	968.49	978.87	975.52	970.74	795.34
Cost		9810.95	9600.75	9265.85	8500.08	7735.36
Makespan	2048 x 64	1984.98	1630.08	1626.73	1621.95	1330.86
Cost		26879.85	26300.85	26291.95	26270.18	23308.63

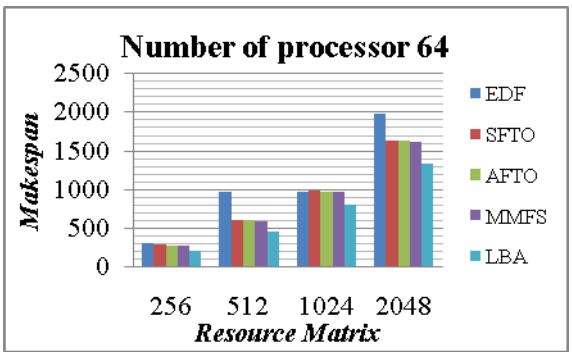


Fig 9: Performance Comparison for Makespan

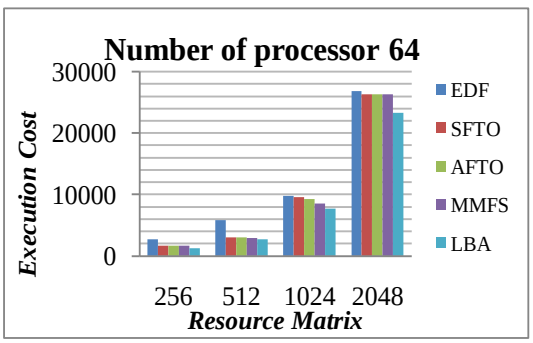


Fig 10: Performance Comparison for Execution Cost

7. Conclusion

In this paper we have proposed a Dynamic load balancing algorithm for the Grid environment that could be used to implement scheduling in a fair way. This algorithm has proved the best results in terms of makespan and Execution Cost In particular the algorithm allocates the task to the available processors so that all requesting task get equal amount of time that satisfied their demand.

Future work will focus on

- Fair scheduling can be applied to optimization techniques
- QoS Constrains such as reliability can be used as performance measure.

Reference

[1] Rajkumar Buyya, David Abramson, and Jonathan Giddy," A Case for Economy Grid Architecture for Service Oriented Grid Computing "**.**

[2] Foster, I., and Kesselman, C. (editors*)*, "The Grid:Blueprint for a New Computing Infrastructure", Morgan Kaufmann Publishers, USA, 1999.

[3] Wolski, R., Brevik, J., Plank, J., and Bryan, T., "Grid Resource Allocation and Control Using Computational Economies, In Grid Computing: Making the Global Infrastructure a Reality" Berman, F, Fox, G., and Hey, T. editors, Wiley and Sons, pp. 747--772, 2003.

[4] Doulamis, N.D.; Doulamis, A.D.; Varvarigos, E.A.; Varvarigou, T.A "Fair Scheduling Algorithms in Grids" IEEE Transactions on Parallel and Distributed Systems, Volume18, Issue 11, Nov. 2007 Page(s):1630 – 1648.

[5] K.Somasundaram, S.Radhakrishnan," Task Resource Allocation in Grid using Swift Scheduler", International Journal of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844 Vol. IV, 2009.

[6] Ramamritham, J.A. Stankovic, and P.-F. Shiah, "Efficient Scheduling Algorithms for Real-Time Multiprocessor Systems,"IEEE Trans. Parallel and Distributed Systems, vol.1, no. 2, pp. 184- 194, Apr. 1990.

[7] Ahmad, Y.-K. Kwok, M.-Y. Wu, and K. Li, "Experimental Performance Evaluation of Job Scheduling and Processor Allocation Algorithms for Grid Computing on metacomputers," Proc.IEEE 18th Int'l Parallel and Distributed Processing Symp. (IPDPS '04),pp. 170-177, 2004.

[8] Pal Nilsson1 and Michał Pi´oro," Unsplittable max-min demand allocation – a routing problem".

[9] Hans Jorgen Bang, Torbjorn Ekman and David Gesbert," A Channel Predictive Proportional Fair Scheduling Algorithm".

[10] Daphne Lopez, S. V. Kasmir raja," A Dynamic Error Based Fair Scheduling Algorithm For A Computational Grid", Journal Of Theoretical And Applied Information Technology - 2009 JATIT.

[11] Qin Zheng, Chen-Khong Tham, Bharadwaj Veeravalli," Dynamic Load Balancing and Pricing in Grid Computing with Communication Delay," Journal in Grid Computing (2008).

[12] Stefan Schamberger,"A Shape Optimizing Load Distribution Heuristic for Parallel Adaptive FEM Computations," Springer-Verlag Berlin Heidelberg 2005.

[13] Grosu, D., Chronopoulos, A.T." Noncooperative load balancing in distributed systems",Journal of Parallel Distrib. Comput. **65**(9), 1022–1034 (2005).

[14] Penmatsa, S., Chronopoulos, A.T." Job allocation schemes in computational Grids based on cost optimization",In: Proceedings of 19th IEEE International Parallel and Distributed Processing Symposium,Denver, (2005).

[15] M.Kamarunisha, S.Ranichandra, T.K.P.Rajagopal," Recitation of Load Balancing Algorithms In Grid Computing Environment Using Policies And Strategies An Approach," International Journal of Scientific & Engineering Research Volume 2, Issue 3, March-2011

[16] Prabhat Kr.Srivastava, Sonu Gupta, Dheerendra Singh Yadav," Improving Performance in Load Balancing Problem on the Grid Computing System", International Journal of Computer Applications (0975 – 8887) Volume 16– No.1, February 2011.

[17] Saravanakumar E. and Gomathy Prathima," A novel load balancing algorithm for computational grid," International Journal of Computational Intelligence Techniques, ISSN: 0976–0466 & E-ISSN: 0976–0474 Volume 1, Issue 1, 2010, PP-20-26

U.Karthick Kumar MSc., MCA., M.Phil.,He is a Post Graduate with M.Phil from Bharathiar University, Coimbatore.Now, he is working as a Assistant Professor in VLB Janaki Ammal Arts and Science College, Coimbatore. He has three years of experience in research. He presented paper in International Conference. His Interest areas are Grid Computing, Mobile Computing and Data Structures.

An Effective Intelligent Model for Medical Diagnosis

Mohamed El-Rashidy [1], Taha Taha [2], Nabil Ayad [3] and Hoda Sroor [4]

[1] Dept. of Computer Science & Eng., Faculty of Electronic Engineering, Menoufiya University, Menouf, Egypt

[2] Dept. of Electronics& Electrical Communications, Faculty of Electronic Engineering, Menoufiya University, Menouf, Egypt

[3] Nuclear Research Center, Atomic Energy Authority, Cairo, Egypt.

[4] Dept. of Computer Science & Eng., Faculty of Electronic Engineering, Menoufiya University, Menouf, Egypt

Abstract

A hybrid data mining model is proposed for finding an optimal number of different pathological types of any disease, and extracting the most significant features for each pathological type. This model is improved in order to reach the fewer subsets of features that have the most impact distinctive of each pathological type. This improvement is lead to the great importance in the decision making of the diagnosis process without confusion or ambiguity between the different variations of the diseases. This model and its optimization are based on fuzzy clustering, nearest neighbor classification, sequential backward search method, and averaging schema for features selection. Experiments have been conducted on three real medical datasets that have different diagnoses. The results show that the highest classification performance is obtained using our optimized model, and this is very promising compared to Naïvebayes, Linear and Polykernal Support Vector Machine (SVM), Artificial Neural Network (ANN), and Support Feature Machines (SFM) models.

Keywords: *Data Mining, Fuzzy Clustering, Nearest Neighbor Classification, Features Selection.*

1. Introduction

Healthcare organizations are facing a major challenge in the patient diagnosis correctly and administering treatments that are effective. This challenge is related to the multiplicity of pathological types of diseases, which makes the diagnostic process more complex, especially if the symptoms and the results of the investigations indicated to these types are several and similarities. Therefore, it is important to find out an optimal number of different pathological types for each disease, and extracts the fewer subsets of features that have high classifiability for each type. This is due to the great importance of this information in the accuracy and speed of diagnostic process without confusion or ambiguity between the different variations of the diseases, and the need to avoid poor treatments that can lead to disastrous consequences. The practice of ignoring this vital knowledge leads to unwanted biases, errors and excessive medical costs which affect the quality of medical services that are provided to patients. This practice moved us toward the developing and optimizing data mining techniques, to get the most accurate knowledge that can be extracted from the medical databases to possess the highest quality of services. Data mining techniques have been successfully applied in various biomedical domains, for example the diagnosis and prognosis of cancers, liver diseases, diabetes, heart disease and other complex diseases [1-9]. These models are omitted the multiplicity of different pathological types of diseases in diagnosis process, and they deal with the disease as its one type and have only one set of distinctive features which distinguish it.

We proposed a hybrid approach based on fuzzy clustering, max-min, and feature selection models that employ extensive advances in classification medical data. We called this approach an Optimal Clustering for Support Feature Machine (OCSFM). The goal of OCSFM is to classify the disease into optimal number of classes, and select the fewer subsets of features that have high classifiability for each class. The advantage of OCSFM is that it uses fuzzy clustering that has classes with less sensitive to noise since noise data points will have very low degrees in all classes, which yields very accurate classification upon diagnosis.

OCSFM is tested on many diseases, similarities and multiplicities of features that are extracted for each of different pathological types of disease are founded, these practices may render the convergence impossible and are leading to random classification decisions. Therefore, we worked to derive an optimization for this model. This optimization is based on a new hybrid feature selection model that used averaging schema as a filter method, and sequential backward search as a wrapper method. The goal of this optimization is to extract the fewer subsets of features that have the most impact distinctive of each pathological type, and access the highest diagnostic accuracy in less time that provide the efficiency of treatment service.

We evaluated the performance of the optimized OCSFM model on the Wisconsin breast cancer (WBCD) [10], the Cleveland heart disease [10], and surgical patient's datasets compared to NaïveBayes [11], Linear SVM [12], Polykernal SVM [13], ANN [14], and SFM [9] models. The sections organization of this paper is as follows. In section 2, classification criteria's is described. Fuzzy clustering and averaging schema for feature selection are offered in section 3, and 4 respectively. In section 5, each step of OCSFM and our optimization is detailed. In section 6, the results and the performance characteristics of the proposed approach will be discussed. The concluding remarks are offered in section 7.

2. Classification Criteria's

The performance of data classification is commonly presented in terms of sensitivity and specificity. Sensitivity measures the fraction of positive test samples that are correctly classified as positive, then we define

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (1)$$

where TP and FN denote the number of true positives and false negatives, respectively. Specificity measures the fraction of negative test samples that are correctly classified as negative. Let FP and TN denotes the number of false positives and true negatives, respectively, then we define

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (2)$$

An overall accuracy is defined as

$$\text{Accuracy} = \frac{TN + TP}{TP + FP + TN + FN} \quad (3)$$

The Matthew's correlation coefficient (MCC) is a powerful accuracy evaluation criterion of machine learning methods. Especially, when the number of negative samples and positive samples are obviously unbalanced [1].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

3. Fuzzy C-means Algorithm

Existing clustering models could be classified into three subcategories hierarchical, density based, and partition based approaches. Hierarchical algorithms organize objects into a hierarchy of nested clusters; hierarchical clustering can be divided into agglomerative and divisive methods [15-18]. Density based algorithms describe the density of data which are set by the density of its objects; the clustering involves the search for dense areas in the object space [19-21]. The idea of Partition based algorithms is to partition data directly into disjoint classes, this subcategory includes several algorithms as k-means, fuzzy c-means, P3M, SOM, graph theoretical approaches, and model based approaches [18] and [22-27]. These approaches assume a predefined number of classes. In addition, these approaches (except the fuzzy/possibilistic ones) always make brute force decisions on the class borders, for this, it may be easily biased by noisy data. This fact makes these fuzzy/possibilistic approaches less sensitive to noisy data.

Fuzzy c-means algorithm (FCM) is an iterative partitioning method [28]. It partitions data samples into c fuzzy classes, where each sample x_j belongs to a class k with a degree of believe which is specified by a membership value u_{kj} between zero and one such that the generalized least squared error function J is minimized.

$$J = \sum_{j=1}^{n} \sum_{k=1}^{c} (u_{kj})^m d(x_j, y_k) \quad (5)$$

Where m is a parameter of fuzziness, c is the number of classes, y_k is the center of class k, and $d(x_j, y_k)$ expresses the similarity between the sample x_j and the center y_k. The summation of the membership values for each sample is equal to one, and this guarantees that no class is empty.

$$0 < \sum_{k=1}^{c} u_{kj} \text{ And } \sum_{k=1}^{c} u_{kj} = 1 \quad \forall j = 1, \ldots, n \quad (6)$$

Because of calling this approach as a probabilistic clustering, since that the membership degrees for a given data point formally resemble the probabilities of its being a member of the corresponding class. This makes the possibilistic clustering less sensitive to noise since noise data points will have very low degrees in all classes. The

minimizations of J are resulted in the following membership function and class center.

$$u_{kj} = \frac{1}{\sum_{i=1}^{c}\left(\frac{d(x_j, y_k)}{d(x_j, y_i)}\right)^{\frac{2}{m-1}}} \quad (7)$$

where u_{kj} is a possibility degree that measures how much typical is data point x_j to class k. The membership degree of x_j to a cluster not only depends on the distance between x_j and that class, but also the distances between x_j and the other classes. The partitioning property of a probabilistic clustering algorithm, which distributes the weight of x_j on the different classes, is due to this equation. Although it is often desirable, the relative character of the membership degrees in a probabilistic clustering approach can lead to counterintuitive results.

$$y_k = \frac{\sum_{j=1}^{n}(u_{kj})^m x_j}{\sum_{j=1}^{n}(u_{kj})^m} \quad (8)$$

This choice makes y_k proportional to the average intra class distance of k, and is related to the overall size and shape of the class.

4. Averaging Scheme

Feature selection algorithms could be classified into two subcategories; filter methods, and wrapper methods [29]. The filter methods estimate the classification performance by some indirect assessments such as distance measures which reflect how well the classes separate from each other. The wrapper methods based on a classifier to select the best subset of features that have the highest classification accuracy, these methods have many types in the searching process as sequential backward search (SBS), and sequential forward search (SFS).

Averaging scheme is a kind of filter methods. The selection feature of averaging scheme is based on two matrices. The first is an n×m intra class distance matrix D = (d_{ij}), and the other is an n×m inter class distance matrix $\overline{D} = (\bar{d}ij)$. The entry of the intra class matrix d_{ij} is the intra class distance, and the entry of the inter class matrix $\bar{d}ij$ is the inter class distance. After the two matrices are constructed, the selection of features is derived from the sum of intra class average distances (d_{ij}) are smaller than the sum of inter class average distances ($\bar{d}ij$) in the selected features [9].

5. OCSFM Model

The model is based on fuzzy C-means, max-min, and averaging schema to classify the data points into optimal number of representative classes, this representation is not aimed only to acquire less average distance (intra class distance), and highest average distance to all different classes (inter class distance), but it takes also into consideration the access to the highest classification accuracy. This model is an integration of both characteristics of supervised and unsupervised models, that makes OCSFM has classes less sensitive to noise, since it is of lowest noise data point's degree in all classes, and maximizes classification accuracy. The flowchart of OCSFM model is shown in Figure 1, where the inputs are the data set D = { $d_0, d_1, \ldots, d_n$ }, $c_{\min}$ and $c_{\max}$ are the minimal and the maximal numbers of expected clusters respectively.

In recapitulation, OCSFM model is composed of six main steps. The first step (Clustering), clusters data points in order to form optimal partitioning representation of classes with smallest intra class distance and greatest inter class distance using Fuzzy c-means algorithm. The second step (Selected Features), finds the optimal subset of features that have high classifiability for each class in order to have the maximum number of training samples correctly classified into those partitioning classes. The third step (Classification), training samples are classified according to those selected features by using nearest neighbor classifier, and computing the performance of data classification which is presented in terms of TP, TN, FP, and FN to obtain MCC. In the fourth step (Classes representatives points), FCM is sensitive to the initial center choices especially for noisy data. We use max-min approach [30], it is desirable to select the initial centers which are well separated; these centers make FCM classes separately groups in a feature space, it chooses a median of one class from those partitioning classes as a start point to select another classes representation points as separate as possible from start point. The fifth step (Multi step max-min algorithm), finds an optimal representative partitioning for a fixed number of classes, each iteration of the optimization process is based on clustering, selected features, and classification steps which is obtained by the max-min method but it changes start point with another class median. Iteration is stopped when each of classes medians is selected as a start point, therefore number of iteration for multi step max-min algorithm are equal to classes medians (number of classes). The sixth step

(Optimal classes number), computes an optimal classes number of partitioning classes by highest classification accuracy of the representation classes. For this, multi step max-min algorithm is repeated with increasing the number of partitioning classes from $c_{\min}$ to $c_{\max}$, using MCC as a validity measure in Equ.(4) which gives a better evaluation than overall accuracy with a lot of machine learning methods, such as SVM, ANN and BNN [1].

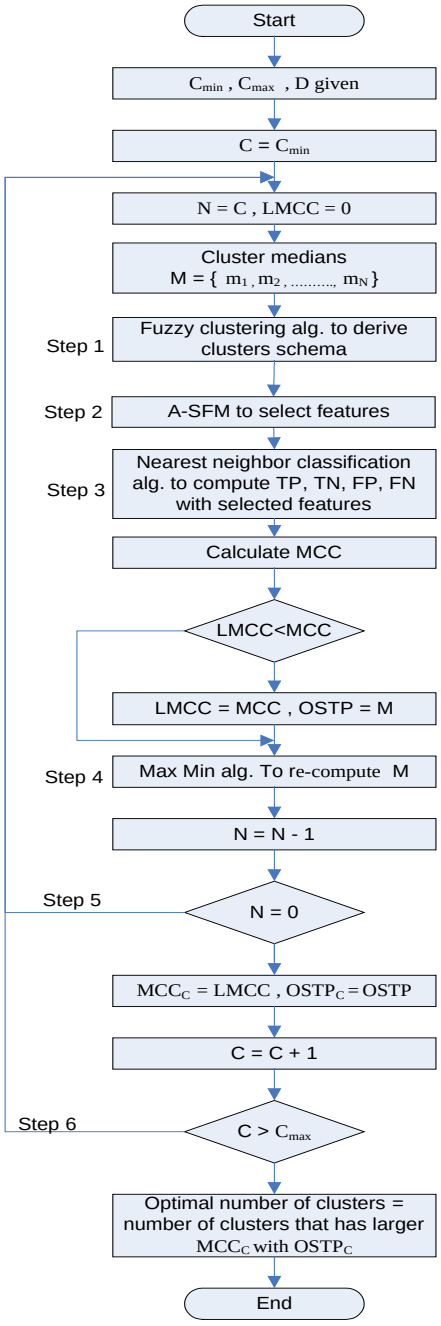


Fig. 1 Flowchart of OCSFM model.

5.1 Optimization of OCSFM Model

We improved OCSFM model in order to reach the highest classification accuracy in less time as possible, through access the fewer subsets of features that have the most impact distinctive of each class in the classification process. We proposed a new hybrid method for the feature selection instead of the averaging schema which used in the second step of the OCSFM model. This optimization is avoiding the irrelevant features that increase the computation time and may render the convergence between classes. This hybrid method consists of two steps which are explained in Figure 2. In the first step, averaging model is used as a filter method selecting the best subset of a given features set for each class, and reducing the number of features that have to be tested through the training of nearest neighbor classifier. In the second step, sequential backward search is used as a wrapper method; nearest neighbor classifier is used to select the subset of fewer features that has the highest classification accuracy for each class from the different subsets of features that are estimated from SBS method.

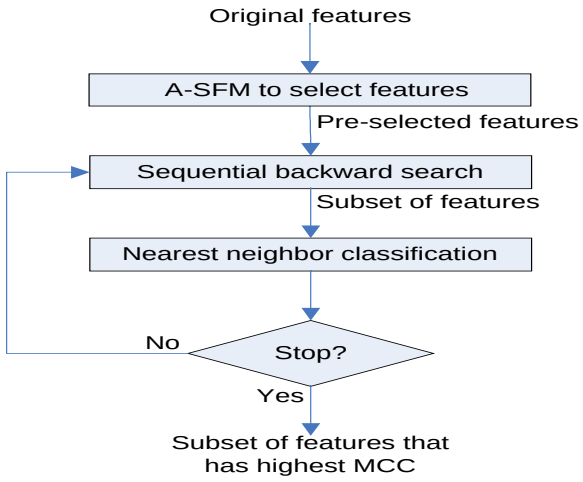


Fig. 2 The hybrid method of selected features.

6. Experimental Results and Discussion

All experiments were implemented and performed on AMD Phenom 9550 Quad Core 2.2 GHz workstation with 4 gigabytes of memory running on Windows Server 2003. The calculations and algorithms were implemented and run on ORACLE 10G. All programs were written by Java language. In the experiments, we apply our optimization model to diagnose several diseases that are related to breast cancer, heart disease, and post-operative infections. The first dataset acquired from the Breast Cancer Wisconsin Diagnostic (WDBC) database, they have been collected by Dr. William H. Wolberg at the University of

Wisconsin Madison Hospitals. There are 699 records in this database. Each record in the database has nine features which were computed from a digitized image of a fine needle aspirate of a breast mass. Those features, computed for each cell nucleus, are considered to be important characteristics for breast cancer diagnosis; those features include Clump thickness (Clump), uniformity of cell size (Ucellsize), uniformity of cell shape (Ucellshap), marginal adhesion (Mgadhesion), single epithelial cell size (Sepics), bare nuclei (Bnuclei), bland chromatin (Bchromatin), normal nucleoli (Normnuct), and Mitoses. In this database, 241 (34.5%) records are malignant and 458 (65.5%) records are benign.

The second dataset acquired from the Cleveland Heart Disease Database, they have been collected by Dr. Andras Janosi, at the Hungarian Institute of Cardiology. There are 297 records in this database; each record in the database has 13 features which are believed to be a good indicator for the angiographic disease status. Those features include chest pain type (Cp) (typical and a typical angina, non angina pain, and asymptomatic), resting blood pressure, serum cholesterol (Chol), resting electro cardio graphic results (Restecg) (normal, abnormality, probable), maximum heart rate (Thalach), indicator of exercise induced angina (Exang), Thal (normal, fixed detect, reversable detect), ST depression (Oldpeak), Slope of the peak exercise ST segment (Slope), number of major vessels colored by fluoroscopy (Ca), and the main criterion that physicians use to determine the diagnosis of heart disease is the narrowing in diameter of any major blood vessel. The diagnosis was considered to be positive (presence of heart disease) if the diameter of any major vessel was narrowed by more than 50%; and negative otherwise. In this database, 160 records (patients) have heart disease and 137 records (patients) have not heart disease.

The third dataset acquired from surgical patient's database, they have been collected from more than one server of Egyptian hospitals. There are 446 records in this database. Each record in the database has 15 features which are believed to be a good indicator for the infections. Those features include age, gender, clinical department name, operation name, operation risk index, health degree of patient (from 1 to 5), actual duration for operation, duration ideal for operation, wound class (none, mild, moderate, severe) of inflammation, length of stay sick before and after the operation, the period between first dose of anti biotic and starting operation, patient temperature during the operation, infection index (non-infected, infected), and name of organism that cause infection. In this database, 101 records (patients) have infection and 345 records (patients) have not infection.

We divided the data into training and testing phases, in test stage, 5-fold cross validation method was applied. First, we show the features that selected by OCSFM and optimization models, these features are summarized in Tables 1, 2 and 3. The tables shows the optimal number of different pathological types for each of the datasets, and the selected features for each of different pathological types of the disease by using both of averaging schema and our hybrid model, where the number of negative and positive classes reflects the number of different pathological types injured and not injured by a disease on respectively. Multiplicity of negative classes is points to existence of many different subsets of features that are caused the disease, but the diversity of suspected cases reflects the multiplicity of positive classes. The results show that similarities of the subsets of multiplicity features in the different pathological types of a disease by using OCSFM model, which leading to the confusion or the ambiguity in the diagnostic process. The distinctive subsets of the fewer features for each disease type are extracted by using our optimization model, which lead to the altitude of diagnostic accuracy in less time as possible.

Second, the altitude of diagnostic accuracy using optimization model can be appeared by sensitive rates comparing with NaïveBayes, Linear SVM, Polykernal SVM, artificial ANN, SFM, OCSFM models in Tables 4, 5 and 6. The results show that our optimization model achieves the highest classification accuracy, which leads to provide the efficiency of treatment service, helps the pathologist to better detect the type of tumor (benign or malignant), the avoidance of diseases complication, chemotherapy complication, exposure to radiation, and mastectomy. And also the improvement is considered as the important purpose that can help physicians to better detect heart disease and post-operative infections.

Table 1: Selected features for each of different pathological types of Breast Cancer Dataset used OCSFM and optimized OCSFM models.

Optimal classes number	***Negative classes number***	***Positive classes number***
14	10	4
The main features for each negative class		

Class No.	***OCSFM***	***No. of Feat.***	***Optimized OCSFM***	***No. of Feat.***
1	Clump, Ucellsize, Ucellshap, Mgadhesion, Sepics, Bnuclei, Bchromatin, and Normnuct.	8	Ucellsize, Ucellshap, Bchromatin, Bnuclei, Normnuct.	5
2	Clump, Ucellsize, Ucellshap, Mgadhesion, Sepics,	9	Clump, Ucellsize, Ucellshap, Sepics, Mitoses.	5

	Bnuclei, Bchromatin, Normnuct, and Mitoses.			
3	Clump, Ucellsize, Ucellshap, Mgadhesion, Sepics, Bnuclei, Bchromatin, Normnuct, and Mitoses.	9	Ucellsize, Ucellshap, Sepics, Bnuclei, Bchromatin, Normnuct, Mitoses.	7
4	Clump, Ucellsize, Ucellshap, Mgadhesion, Sepics, Bnuclei, Bchromatin, and Normnuct.	8	Bnuclei.	1
5	Clump, Ucellsize, Ucellshap, Mgadhesion, Sepics, Bnuclei, Bchromatin, and Normnuct.	8	Ucellsize, Ucellshap, Sepics, Bnuclei, Bchromatin, Normnuct.	6
6	Clump, Ucellsize, Ucellshap, Sepics, Bnuclei, Bchromatin, Normnuct, and Mitoses.	8	Ucellsize, Ucellshap, Sepics, Bnuclei, Bchromatin, Normnuct, Mitoses.	7
7	Clump, Ucellsize, Ucellshap, Mgadhesion, Sepics, Bnuclei, Bchromatin, and Normnuct.	8	Ucellsize, Sepics, Normnuct.	3
8	Clump, Ucellsize, Ucellshap, Bnuclei, Bchromatin, and Normnuct.	6	Bnuclei, Bchromatin, Normnuct.	3
9	Clump, Ucellsize, Ucellshap, Mgadhesion, Sepics, Bnuclei, and Bchromatin.	7	Ucellsize, Sepics, Bnuclei, Bchromatin.	4
10	Clump, Ucellsize, Ucellshap, Mgadhesion, Sepics, Bnuclei, Bchromatin, and Normnuct.	8	Ucellsize, Ucellshap, Sepics, Bchromatin, Normnuct.	5

Table 2: Selected features for each of different pathological types of Heart Disease Dataset used OCSFM and optimized OCSFM models.

Optimal classes number	***Negative classes number***	***Positive classes number***
3	1	2
The main features for each negative class		

Class No.	***OCSFM***	***No. of Feat.***	***Optimized OCSFM***	***No. of Feat.***
1	Cp, Oldpeak, Ca, Slope, Thalach, Thal, Exang, Restecg, Chol.	9	Ca, Slope, Thal, Exang.	4

Table 3: Selected features for each of different pathological types of Surgical Patient's Dataset used OCSFM and optimized OCSFM models.

Optimal classes number	***Negative classes number***	***Positive classes number***
5	2	3
The main features for each negative class		

Class No.	**OCSFM**	**No. of Feat.**	**Optimized OCSFM**	**No. of Feat.**
1	age, gender, clinical department name, operation name, operation risk index, health degree of patient, actual duration for operation, duration ideal for operation, wound class of inflammation, length of stay sick before and after the operation, the period between first dose of anti biotic and starting operation, patient temperature during the operation	13	operation risk index.	1
2	age, gender, clinical department name, operation name, operation risk index, health degree of patient, actual duration for operation, duration ideal for operation, wound class of inflammation, length of stay sick before and after the operation, the period between first dose of anti biotic and starting operation, patient temperature during the operation	13	duration ideal for operation, actual duration for operation, operation risk index	3

Table 4: Training and testing performance in % sensitivity, specificity, overall accuracy and MCC of NaïveBayes, Linear SVM, Polykernal SVM, ANN, SFM, OCSFM and Optimized OCSFM approaches for Diagnosis of Breast Cancer in WDBC Database.

Classification algorithm	Training Data				Testing Data			
	Sens.	Spec.	Accu.	MCC	Sens.	Spec.	Accu.	MCC
NaïveBayes	97.19	97.09	97.12	94.05	95.55	97.64	96.92	93.20
Linear SVM	94.54	94.60	94.56	88.17	94.73	22.22	80.85	23.91
Polykernal SVM	97.59	96.26	97.14	93.68	93.33	97.64	96.15	91.47
ANN	97.37	98.75	97.85	95.33	97.77	97.64	97.69	94.94
SFM	97.85	91.32	95.60	90.22	97.64	88.88	94.61	88.03
OCSFM	97.31	98.97	97.89	95.42	97.64	97.77	97.69	94.94
Optimized OCSFM	97.58	98.48	98.24	96.19	96.47	100	97.71	95.10

Table 5: Training and Testing Performance in % sensitivity, specificity, overall accuracy and MCC of NaïveBayes, Linear SVM, Polykernal SVM, ANN, SFM, OCSFM and Optimized OCSFM approaches for Diagnosis of Cleveland Heart Disease Database.

Classification algorithm	Training Data				Testing Data			
	Sens.	Spec.	Accu.	MCC	Sens.	Spec.	Accu.	MCC
NaïveBayes	79.56	88.12	84.17	68.14	72.72	74.28	73.68	46.12
Linear SVM	80.29	90.00	85.52	70.89	77.27	88.57	87.71	67.99
Polykernal SVM	79.56	89.37	84.84	69.53	80.36	85.71	82.96	69.07
ANN	86.86	88.75	87.87	75.61	77.27	88.57	84.21	66.45
SFM	82.60	85.60	84.16	68.26	74.28	95.45	82.45	67.99
OCSFM	86.08	88.80	87.50	74.94	82.85	90.90	85.96	72.10
Optimized OCSFM	88.69	89.60	89.16	78.29	84.53	93.71	87.72	76.35

Table 6: Training and Testing Performance in % sensitivity, specificity, overall accuracy and MCC of NaïveBayes, Linear SVM, Polykernal SVM, ANN, SFM, OCSFM and Optimized OCSFM approaches for Diagnosis of Surgical Patient's Database.

Classification algorithm	Training Data				Testing Data			
	Sens.	Spec.	Accu.	MCC	Sens.	Spec.	Accu.	MCC
NaïveBayes	92.72	72.13	89.91	60.57	92.11	55.55	85.11	49.90
Linear SVM	99.74	26.22	89.68	46.60	100	11.11	82.97	30.29
Polykernal SVM	99.74	24.59	89.46	44.95	100	11.11	82.97	30.29
ANN	100	39.34	91.70	59.91	100	22.22	85.11	43.32
SFM	90.75	39.62	83.95	30.37	94.74	11.11	78.72	9.41
OCSFM	95.08	71.69	91.97	65.74	94.74	66.66	89.36	64.29
Optimized OCSFM	98.26	89.28	97.01	87.55	97.22	93.89	95.74	86.25

7. Conclusions

In this paper, the OCSFM approach has been improved and applied to the tasks of breast cancer, heart diseases, and post-operative infections diagnosis. Results are indicated that our proposed approach found out the optimal number of different pathological types of these diseases, and extracted the fewer subsets of features that have the most impact distinctive of each pathological type without confusion or ambiguity between the different variations of these diseases. Here, after applying the proposed approach, the accuracy of the diseases diagnosis in WDBC, Cleveland Heart Disease, and surgical patient's datasets has been improved by sensitive rates.

References

[1] H. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultra sound images: Asurvey", Pattern Recognition, 43, 299-317, 2010.

[2] B. Riccardo, and Blaz Z., "Predictive data mining in clinical medicine: Current issues and guidelines", international journal of medical informatics, 77, 81-97, 2008.

[3] L. Rong-Ho, "An intelligent model for liver disease diagnosis", Artificial Intelligence in Medicine, 47, 53-62, 2009.

[4] H. Yue, M. Paul, B. Norman, and H. Roy, "Feature selection and classification model construction on type 2 diabetic patient's data", Artificial Intelligence in Medicine, 41, 251-262, 2007.

[5] M. Choua, T. Leeb, Y. Shaoc, and I. Chenb, "Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines", Expert Systems with Applications, 27, 133-142, 2004.

[6] J. Elmore, M. Wells, M. Carol, H. Lee, D. Howard, and A. Feinstein, "Variability in radiologists interpretation of mammograms", New England Journal of Medicine, 331(22), 1493-1499, 1994.

[7] A. Mehmet, "Support vector machines combined with feature selection for breast cancer diagnosis", Expert Systems with Applications, 36, 3240-3247, 2009.

[8] M. Ilias, Z. Elias, and A. Ioannis, "An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers", Appl Intell, 30, 24-36, 2009.

[9] Ya-Ju F., and Wanpracha A. Ch., "Optimizing feature selection to improve medical diagnosis", Ann Oper Res, 174, 169-183, 2010.

[10] Cleveland Heart Disease and Wisconsin Breast Cancer Datasets are originally available on UCI Machine Learning Repository website http://archive.ics.uci.edu.

[11] P. Bhargavi, and S. Jyothi, "Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils", International Journal of Computer Science and Network Security, 9(8), 117-122, 2009.

[12] L. Zhizheng, and Z. Tuo, "Feature selection for linear support vector machines", The 18th International Conference on Pattern Recognition IEEE, 2006.

[13] I. Bhattacharya, and M. P. S. Bhatia, "SVM classification to distinguish Parkinson disease patients", A2CWiC '10 Amrita ACM-W Celebration on Women in Computing in India, 2010.

[14] J. Paulo, and F. Azzam, "The use of artificial neural networks in decision support in cancer: A systematic review ", Neural Networks, 19(4), 408-415, 2006.

[15] M. Eisen, P. Spellman, P. Brown, and D. B otstein, "Cluster analysis and display of genome wide expression patterns", Natl Acad Sci USA, 95(25), 14863-14868, 1998.

[16] M. Blatt, S. Wiseman, and E. Domany, "Super-paramagnetic clustering of data", Phys Rev Lett, 76, 3251-3254, 1996.

[17] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems", IEEE, 86(11), 2210-2239, 1998.

[18] J. Herrero, A. Valencia, and J. Dopazo, "A hierarchical unsupervised growing neural network for clustering gene expression patterns", Bioinformatics, 17(2), 126-136, 2001.

[19] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: a survey", IEEE Trans Knowl Data Eng, 16(11), 1370-1386, 2004.

[20] D. Jiang, J. Pei, and A. Zhang, "DHC: a density-based hierarchical clustering method for time series gene expression data", the 3rd IEEE symp on bioinformatics and bioengineering, Maryland, USA, 393-400, 10-12 March 2003.

[21] A. Hinneburg, and D. Keim, "An efficient approach to clustering in large multimedia database with noise", the 4th int conf on knowledge discovery and data mining, NY, USA, 58–65, 27–31 August 1998.

[22] W. Au, K. Chan, A. Wong, and Y. Wang, "Attribute clustering for grouping, selection, and classification of gene expression data", IEEE/ACMTrans Comput Biol Bioinform, 2(2), 83–101, 2005.

[23] D. Bickel, "Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically", Bioinformatics, 19(7), 818–824, 2003.

[24] R. Guthke, W. Schmidt-Heck, D. Hann, and M. Pfaff, "Gene expression data mining for functional genomics", the European symp on intel techn, Aachen, Germany, 170–177, 2000.

[25] L.B. Romdhane, H. Shili, and B. Ayeb," Mining microarray gene expression data with unsupervised possibilistic clustering and proximity graphs", Appl Intell, 10.1007/s, 10489-009, 2009.

[26] R. Shamir, and R. Sharan, "CLICK: A clustering algorithm for gene expression analysis", the int conf on intelligent systems for molecular biology, CA, USA, 307–316, 19–23 August 2000.

[27] K. Yeung, C. Fraley, A. Murua, A. Raftery, and W. Ruzz, "Model-based clustering and data transformations for gene expression data", Bioinformatics, 17(10), 977–987, 2001.

[28] J. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", New York: Plenum, 1981.

[29] L. Ming-Chi, "Using support vector machine with a hybrid feature selection method to the stock trend prediction", Expert Systems with Applications, 36, 10896–10904, 2009.

[30] J. Tou, and R. Gonzalez, "Pattern recognition principles", Addison-Wesley, Reading, 1974.

First Author obtained his Master degree in computer science and engineering, 2008. Currently, he is working as a Lecturer Assistant in the Dept. of Computer Science and Engineering, Faculty of Electronic Engineering, 32952, Menouf, Menoufiya University -Egypt. Areas of interest of the author include data mining and bioinformatics.

Second Author was born in Tanta, Egypt, on October 11, 1946. He received the B.Sc. degree (with distinction) in communication engineering from Menoufiya University, Egypt, in June 1969, the M.Sc. degree in communication engineering from Helwan University, Egypt, in April 1978, and the Ph.D. degree (very honorable) in electronic engineering from the National Polytechnic Institute, Toulouse, France, in June 1985. From September 1969 to July 1978, he was a Demonstrator, in July 1978, he was an Assistant Lecturer, in November 1985, he was a Lecturer, in February 1990, he was an Assistant Professor, and in September 1995, he was named Professor, all in the Faculty of Electronic Engineering, Menoufiya University, Communication Department,. He was appointed Vice Dean from February 2002 to October 2005, and Head of the Communication Department, from November 2005 to July 2007. At present, he is an Emeritus Professor at the same department. His main research interests are surface acoustic wave devices, optical devices, superconductor devices, medical applications of ultrasound, and bioinformatics.

Third Author received Ph.D degree in CSE from Cairo University, in 1984. He is working as vice chairman for reactors division, Nuclear Research Center, Atomic Energy Authority- Egypt. He is a member of IEEE. His main research interests database and networks.

Fourth Author received Ph.D degree in CSE from Menoufiya University, in 1991. She is working as Professor in Dept. of Computer Science and Engineering, Faculty of Electronic Engineering, 32952, Menouf, Menoufiya University- Egypt, her main research interests parallel processing and database.

Selecting Features of Single Lead ECG Signal for Automatic Sleep Stages Classification using Correlation-based Feature Subset Selection

Ary Noviyanto[1], Sani M. Isa[2], Ito Wasito[3] and Aniati Murni Arymurthy[4]

[1] Computer Science, Universitas Indonesia
Depok 16424/Jawa Barat, Indonesia

[2] Computer Science, Universitas Indonesia
Depok 16424/Jawa Barat, Indonesia

[3] Computer Science, Universitas Indonesia
Depok 16424/Jawa Barat, Indonesia

[4] Computer Science, Universitas Indonesia
Depok 16424/Jawa Barat, Indonesia

Abstract

Knowing about our sleep quality will help human life to maximize our life performance. ECG signal has potency to determine the sleep stages so that sleep quality can be measured. The data that used in this research is single lead ECG signal from the MIT-BIH Polysomnographic Database. The ECG's features can be derived from RR interval, EDR information and raw ECG signal. Correlation-based Feature Subset Selection (CFS) is used to choose the features which are significant to determine the sleep stages. Those features will be evaluated using four different characteristic classifiers (Bayesian network, multilayer perceptron, IB1 and random forest). Performance evaluations by Bayesian network, IB1 and random forest show that CFS performs excellent. It can reduce the number of features significantly with small decreasing accuracy. The best classification result based on this research is a combination of the feature set derived from raw ECG signal and the random forest classifier.

Keywords: *ECG features, Correlation-based Feature Subset Selection, RR interval, EDR, Raw ECG Signal, Sleep stages.*

1. Introduction

The quality of sleep directly affects the quality of life. Using a particular measure [1], we can calculate the sleep quality of somebody by knowing the composition of his/her sleep stages. Sleep experts analyze polysomnogram data as the standard technique to determine the sleep stages. Polysomnogram is a simultaneous recording of physiological variables during sleep that include brain activity (*electroencephalogram*, EEG), eye movements (*electroocculogram*, EOG), and chin muscle activity (*electromyohram*, EMG) [2].

Based on the previous works [3, 4, 5, 6], ECG as a substitute of the standard technique to determine the sleep stages has promising results. The main reason of using ECG is the expensiveness of recording process of the polysomnogram data. The data gathering has to do in a sleep laboratory that is expensive with uncomfortable processes for patients and also require trained staff. Another issue is that the sleep study (polysomnography) is costly. It means that the manual determination of the sleep stages in a long sequence of the polysomnogram data is a work that requires endurance and high accuracy. The manual determination of the sleep stages can also trigger lack standard of the sleep stages determination (i.e. every sleep expert may have different results in the sleep stages determination). The automatic processes in the polysomnography are necessary to handle the issue in the manual determination of sleep stages.

There are two groups of sleep in sleep architecture; non-rapid eye movement (NREM) sleep and rapid eye movement (REM) sleep [7]. The NREM sleep can be divided into NREM 1, NREM 2, NREM 3 and NREM 4. The graphic that represents the sleep stages sequence, called *hypnogram*, is shown in Figure 1. ECG (*Electrocardiogram*), or sometimes called EKG, simply is

a signal from the heart during it is beating [8]. The ECG's waveform and its attributes are shown in Figure 2.

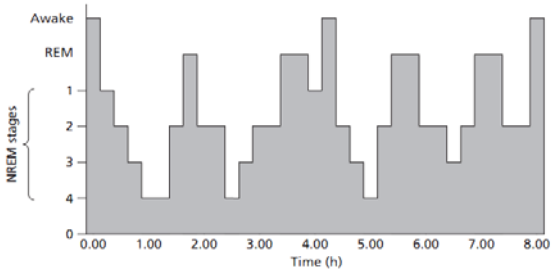


Fig. 1 Normal adult hypnogram [7].

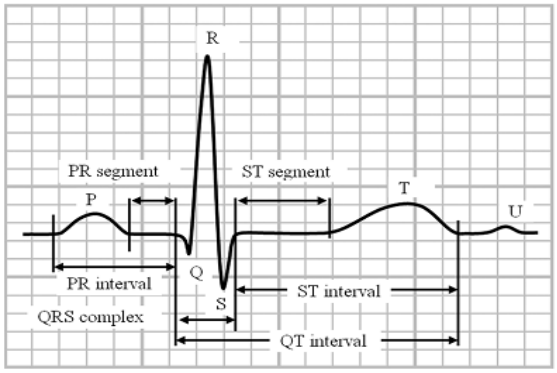


Fig. 2 ECG's waveforms and intervals identified [9].

Many features can be derived from only a single lead ECG signal. Several previous researches used various different features. We will define several feature sets that are constructed from single lead ECG's features and determine the best performance feature set for the automatic sleep stages classification.

2. Previous Works

Several previous researches have given promising result that ECG can be a substitute of the standard data to determine the sleep stages. Shinar *et al.* [3], in 2001, have used only ECG signal to detect Slow Wave Sleep (SWS) with 80% correct identification. They have used time dependent spectral component (Very Low Frequency, Low Frequency and High Frequency) that decomposed from RR interval using wavelet transform as the features.

Lewicke *et al.* [4], in 2008, have determined sleep and wake only using ECG signal of infants. Using fuzzy C-means (FCM) clustering algorithm, they determined that the best feature from standard Heart Rate Variability measure derived from RR interval is the mean. Using rejection approach, the model has achieved 85%-87% correct classification while rejecting 30% of the data with a kappa statistic of 0.65-0.68.

Yilmaz *et al.* [5], in 2010, have investigated features that derived only from single-lead ECG to classify sleep stages and obstructive apneaic epochs. Using quadratic discriminant analysis (QDA) and support vector machines (SVM) methods have 60% or 70% accuracy for specific sleep stage on healthy subject and 89% accuracy for five obstructive sleep apnea (OSA). They have used the median, inter-quartile range, and mean absolute deviation values that derived from RR interval as the features.

Bsoul *et al.* [6], in 2010, have done research in sleep quality assessment based on ECG measurements. They have used a multi-stage Support Vector Machines (SVM) classifier. Using a binary decision tree (BDT) technique for four classes, three-stage multi-class SVMs are needed and perform good result with high accuracy. This research uses a lot of features. For every ECG segment (i.e. 30 seconds), 112 feature measures can be extracted; 60 for RR time series and 52 for EDR time series.

3. Dataset and Features

In our research, we use ECG signal from the MIT-BIH Polysomnographic Database that can be downloaded from http://www.physionet.org [10]. The MIT-BIH Polysomnographic Database contains multiple physiologic signals during sleep. We have total 18 records of ECG signal with various recorded length from 16 subjects.

From 18 records of ECG signal, we extract several features of single lead ECG signal that potentially can be used to classify the sleep stages. We have three categories of ECG's features; the features that derived from RR interval, EDR information and raw ECG signal. We can get 39 features in total with composition: 12 features from RR interval, 12 features from EDR information and 15 features from raw ECG signal. Those features will be selected and evaluated in order to determine the optimal features for the sleep stages classification.

3.1 Features Derived from RR Interval

RR interval is a distance of two successive top R waves. If the R waves are in normal beats, we can call it as NN interval (Normal to Normal interval). To describe variations of RR interval, we used Heart Rate Variability (HRV) [11]. HRV is divided into time domain measures and frequency domain measures. The common time domain measures consist of [12],

1. Average of all NN intervals (AVNN);

2. Standard deviation of all NN intervals (SDNN);
3. Square root of the mean of the squares of differences between adjacent NN intervals (rMSSD);
4. Percentage of differences between adjacent NN intervals that are greater than 50 ms; a member of the larger pNNx family (pNN50);

and the common frequency domain measures consist of [12],

1. Total spectral power of all NN intervals up to 0.04 Hz (TOTPWR);
2. Total spectral power of all NN intervals between 0.003 and 0.04 Hz (VLF);
3. Total spectral power of all NN intervals between 0.04 and 0.15 Hz. (LF);
4. Total spectral power of all NN intervals between 0.15 and 0.4 Hz (HF);
5. Ratio of low to high frequency power (LF/HF).

Based on Yilmaz *et al.* [8], we can also derived three features that belong to time domain measures; they are median, Inter-quartile range (IQR) and Mean absolute deviation (MAD). All of the features that derived from RR interval are coded using prefix "RR-" (e.g. RR-AVNN, RR-SDNN, and so on).

3.2 Features Derived from EDR

ECG signal has been related to respiratory signal [5]. Respiratory information can be derived from ECG signal, called ECG-derived respiration (EDR) information. EDR information is obtained by calculating regions under the QRS segments of ECG signal. The region is a fixed window around R point (100ms centered in R point [13]). Before we can calculate EDR from ECG signal, we have to correct ECG signal by subtracting original ECG signal with the baseline of the ECG signal. The baseline can be calculated by filtering the original ECG signal using median filter of 200 ms and 600 ms respectively [13].

From EDR information, we can also extract same features as RR interval. All of the features that derived from EDR information are coded using prefix "EDR-" (e.g. EDR-AVNN, EDR-SDNN, and so on).

3.3 Features Derived from Raw ECG Signal

Raw ECG signal is the original ECG signal without transformation into other forms (e.g. RR Interval or EDR information). They are 15 features that can be derived from raw ECG signal. The features are listed in Table 1 [14, 15].

Table 1: List of Features Derived from Raw Signal ECG [14, 15]

Features	Equations
Energy	$\sum_i x_i^2$
4^{th} Power	$\sum_i x_i^4$
Curve Length	$\sum_i \|x_i - x_{i-1}\|$
Nonlinear Energy	$\sum_i -x_i \cdot x_{i-2} + x_{i-1}^2$
Peak Power (Max PSD)	$max(PSD)$
Peak Frequency	$index(max(PSD))$
Mean PSD	$mean(PSD)$
Median PSD	$median(PSD)$
Spectral Entropy	$\sum_j PSD_j \cdot \log PSD_j$
Katz Fractal Dimension	$\frac{\log N}{\log N + \log \frac{max(\sqrt{(x_i - x_0)^2 + i^2})}{\sum_i \sqrt{(x_i - x_{i+1})^2 + 1}}}$
Detrended Fluctuation Analysis (DFA)	the slope of the line relating log of root-mean-square fluctuation to log n.
Higuchi Fractal Dimension (HFD)	based on Higuchis algorithm
Hjorth Mobility	$\sqrt{\frac{\sum_i \frac{x_i - x_{i-1}}{N}}{\sum_i \frac{x_i}{N}}}$
Hjorth Complexity	$\sqrt{\frac{\sum_i \frac{(x_i - 2x_{i-1} + x_{i-2})^2}{N} \cdot \sum_i \frac{x_i}{N}}{(\sum_i \frac{x_i - x_{i-1}}{N})^2}}$
Petrosian Fractal Dimension (PFD)	$\frac{\log N}{\log N + \log(\frac{N}{N + 0.4 N_\delta})}$

3.4 Dataset Construction

We construct dataset into two groups. In the first group, all of the records (18 records) are combined into a large data. In the second group, each of the record is treated as a separate data; we can call it as *subject-based dataset*. It is intended to find out the robustness of the classifiers and the features. Every group consists of:

1. Feature set *a*: it contains features that are derived from RR interval;
2. Feature set b: it contains features that are derived from EDR information;
3. Feature set *c*: it contains features that are derived from raw ECG signal; and
4. Feature set *d*: it is a combination of the features that are derived from RR interval, EDR information and from raw ECG signal.

4. Methodology

To investigate the best feature set for the sleep stages classification, we evaluate the performance of the feature sets using four different characteristics classifier; they are Bayesian Network learning, Multilayer perceptron, IB1, and Random Forest Classifier. The feature sets can be full set features or selected features. The correlation-based feature subset selection is used to get the best selected features in every feature set.

4.1 The Classifiers

A Bayesian network is a probabilistic based classifier. A Bayesian network contains a set of variables and a directed acyclic graph (DAG) as the structure that is constructed using K2 algorithm. A Bayesian network represents a probabilities distribution as Equation 1 [16] where *U* is a set of variable and *pa(u)* is parents of $u \in U$.

$$P(U) = \prod_{u \in U} p(u \mid pa(u)) \quad (1)$$

MLP (Multilayer Perceptron) is one of the neural network approaches. MLP is a generalization of single layer perceptron that consists of an input layer, one or more hidden layer and an output layer [17]. The visualization of MLP is depicted as Figure 3. The MLP uses the back-propagation learning as the learning algorithm. For this research, the number of neuron in the hidden layer is sum of the number of input neuron and the number of output neuron divided by 2.

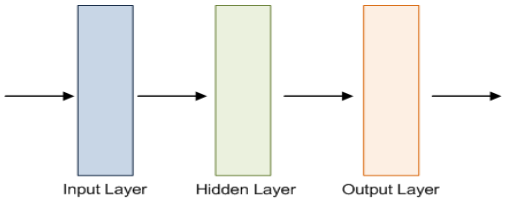


Fig. 3 The Multilayer Perceptron.

IB1 is the simplest instance based learner that known as *K*-nearest neighbors with *K*=1. In this algorithm, the similarity measure is defined as a negative value of a Euclidean distance, according to Equation 2. How IB1 works can be described in Algorithm 1 [18].

$$sim(x, y) = -\sqrt{\sum_i (x_i - y_i)^2} \quad (2)$$

Algorithm 1: Instance Based Learning [18]

```
1: CD ← ∅
2: for all x ∈ TrainningSet do
3:     y_max ← argmax_{y∈CD}(sim(x, y))
4:     if class(x) == class(y_max) then
5:         classification ← correct
6:     else
7:         classification ← incorrect
8:     end if
9:     CD ← CD ∪ {x}
10: end for
```

Random forest [19] is a decision tree based classifier that contains many classification trees (in this research, we generate 10 trees). The basic idea of random forest classifier is that every single input will be classified using each classification tree in the forest and the final classification result is done by selecting the most votes in the forest.

4.2 The Feature Selection Method

The Correlation-based Feature Subset Selection (CFS) is a feature selection method based on correlation. Correlation is a degree of dependence or predictability of one variable with another [20]. Based on this technique, the feature subsets that can be signatures have two criterions: correlated with the class and uncorrelated one feature with the other. CFS calculate an evaluation measure of feature subset, called *merit*, using an evaluation function according to Equation 3 [20], where *S* is a feature subset, M_S is the heuristic *merit* of *S*, *k* is the feature size, $\overline{r_{cf}}$ is the average of feature-class correlation, and $\overline{r_{ff}}$ is the average of feature-feature inter-correlation.

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (3)$$

To get the best feature subset, there are three possible searching techniques [20];

1. Forward selection.
   This searching technique greedily adds one feature until the feature set has no higher value of the evaluation measure.
2. Backward elimination.
   This searching technique greedily eliminates one by one feature until degrading value of the evaluation measure.
3. Best first.
   Best first searches through the search space, either forward or backward. Because of this pure technique is exhaustive; this searching technique has a stopping criterion. The stopping criterion is no improvement value of the evaluation measure over the current best subset in five consecutive fully expanded subsets. In this research, we use the Best first with forward direction as the searching technique.

5. Experimental Results

5.1 Evaluation Measures

We use two evaluation measures; there are percentage of instances correctly classified and Kappa statistic [21]. Kappa statistic gives us a numerical rating of the degree of agreement between the ground truths and the predictions. Table 2 is the interpretation of Kappa statistic. To

convince about the results of evaluation measures, we use *k*-fold validation method with $k = 10$.

Table 2: Interpretation of Kappa Statistic [21]

Kappa	***Agreement***
< 0	Less than chance agreement
0.01 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Almost perfect agreement

5.2 Feature Selection Results

Using CFS, we can take the best selected features for the feature set that derived from RR interval, EDR information and Raw ECG signal.

1. Features derived from RR interval.
   From 12 features in total, there are 6 selected features; RR-rMSSD, RR-pNN50, RR-TOTPWR, RR-VLF, RR-LF/HF and RR-MAD.
2. Features derived from EDR information.
   From 12 features in total, there are 5 selected features; EDR-TOTPWR, EDR-VLF, EDR-LF, EDR-HF and EDR-LF/HF.
3. Features derived from Raw ECG signal.
   From 15 features in total, there are 5 selected features; 4th Power, Katz Fractal Dim, Hjorth Mobility, Hjorth Complexity and PFD.
4. Features derived from RR interval, EDR information and Raw ECG signal.
   From 39 features in total, there are 14 selected features; RR-rMSSD, RR-TOTPWR, RR-LF/HF, RR-MAD, EDR-VLF, EDR-LF, EDR-HF, EDR-LF/HF, 4th Power, Katz Fractal Dim, HFD, Hjorth Mobility, Hjorth Complexity and PFD.

Based on these results, the features that are selected in the features selection process of the feature set *d* (i.e. combination of RR interval, EDR information and Raw ECG signal) are also selected in the features selection process of the feature set *a* (i.e. RR interval), *b* (i.e. EDR information) or *c* (i.e. Raw ECG signal) but not vice versa. For example, RR-rMSSD is selected in the features selection process of the feature set *d* and also selected in the features selection of the feature set *a*; RR-pNN50 is selected in the features selection process of the feature set *a* but not selected in features selection process of the feature set *d*.

5.2 Result and Discussion

The classification results of the first group dataset are presented in Table 3 and Table 4. We can depict them into line chart in Figure 4. Based on Figure 4, the graphic of dash line and continues line in Bayesian network, IB1 and random forest have shape same pattern. We can see in Figure 4b; multilayer perceptron does not have same shape pattern. Reducing features in the feature set *c* with MLP as the classifier has decreased the accuracy for 10.65%.

Table 3: The classification result using the full set features; %C is percent correct classification and *K* is the Kappa statistic

Feature Set	**BN**		**MLP**		**IB1**		**RF**	
	%C	***K***	**%C**	***K***	**%C**	***K***	**%C**	***K***
a	41.01	0.24	48.87	0.23	40.00	0.17	50.33	0.28
b	42.69	0.26	54.10	0.30	48.35	0.29	55.73	0.36
c	59.01	0.45	59.32	0.41	79.34	0.71	79.80	0.72
d	60.53	0.48	68.12	0.55	70.56	0.59	76.24	0.66

Table 4: The classification result using the best selected features; %C is percent correct classification and *K* is the Kappa statistic

Feature Set	**BN**		**MLP**		**IB1**		**RF**	
	%C	***K***	**%C**	***K***	**%C**	***K***	**%C**	***K***
a	44.60	0.24	47.55	0.19	38.75	0.15	49.10	0.26
b	41.42	0.24	52.35	0.27	46.99	0.27	52.59	0.33
c	58.52	0.43	48.66	0.24	74.87	0.65	78.85	0.71
d	60.65	0.48	62.40	0.45	67.82	0.55	76.50	0.67

Table 5: Comparison of decreasing accuracy and number of features; DA is Decreasing of Accuracy in percent and DN is Decreasing of Number of Features in percent.

Feature set	**BN**		**MLP**		**IB1**		**RF**	
	DA	**DN**	**DA**	**DN**	**DA**	**DN**	**DA**	**DN**
a	-3.59	50.00	1.32	50.00	1.25	50.00	1.22	50.00
b	1.27	58.33	1.75	58.33	1.36	58.33	3.14	58.33
c	0.49	66.67	10.65	66.67	4.47	66.67	0.95	66.67
d	-0.13	64.10	5.72	64.10	2.74	64.10	-0.26	64.10

Table 5 shows the comparison of decreasing accuracy and number of features. The negative value in Table 5 (e.g. BN with feature set *a* and feature set *d*, RF with feature set *d*) means that the accuracy is increase. For Bayesian network and random forest classifiers, selecting appropriate features will increase the accuracy. For multilayer perceptron and IB1, reducing the number of features through features selection process generally will always reduce accuracy.

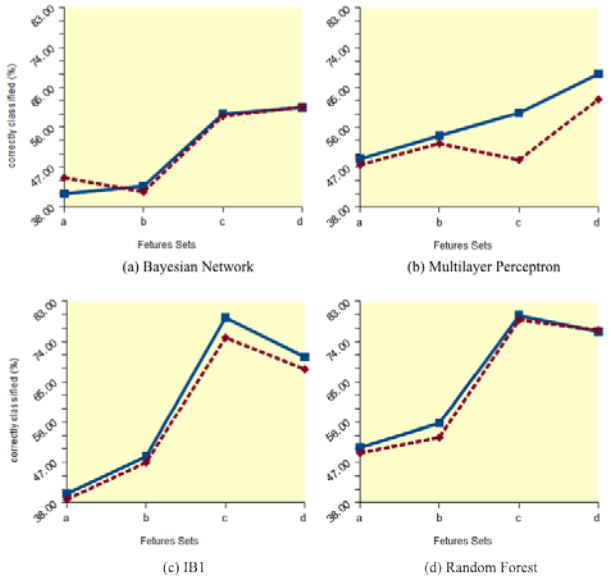


Fig. 4 The effect of feature selection in accuracy over the classifiers; y-axis is correctly classified in percent and x-axis is the feature set; dash line is using selected features and continues line is using full set features.

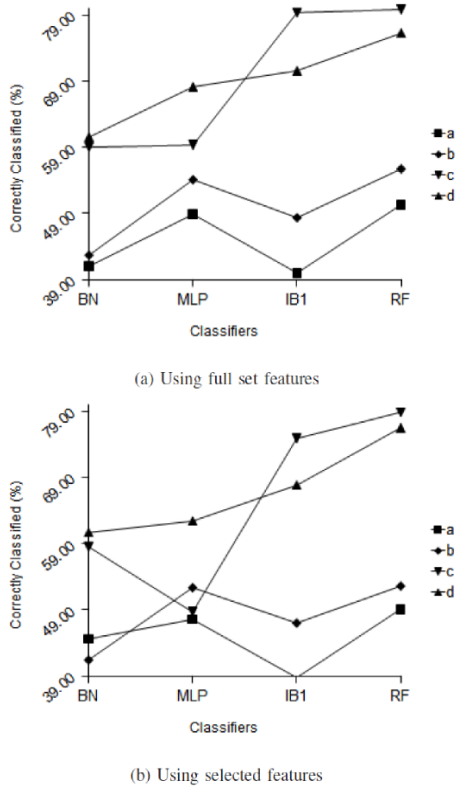


Fig. 5 Comparison of the feature sets; a, b, c, and d are the feature set; y-axis is correctly classified in percent and x-axis is the classifiers.

The line chart in Figure 5 that derived from Table 3 and Table 4 shows the performance of the feature sets. Overall, feature set *c* and *d* are better than feature set *a* and feature set *b* for all classifiers. Based on this experiment, the feature set *c* and feature set *d* which include raw ECG signal show good performance to determine the sleep stages for the all classifiers.

According to the result in Table 3 and Table 4, the best performance of Bayesian network and Multilayer perceptron are achieved when using the feature set *d* (i.e. combination of features that derived from RR interval, EDR information and raw ECG signal) in both cases; with and without features selection. Whereas the best performance of IB1 and Random forest are achieved when using the feature set *c* (i.e. the features that derived from raw signal ECG) in both cases.

In the experiment using combination all of records (i.e. first group dataset); the best accuracy is 79.80% with 0.72 kappa statistic. This performance is achieved using the random forest as the classifier and 15 features that derived from raw ECG signal (i.e. the feature set *c* without feature selection). Using CFS, the best accuracy is 78.85% with 0.71 kappa statistic. This performance is achieved using the random forest as the classifier and 5 selected features that derived from raw ECG signal. CFS can reduce 66.67% number of features in the feature set *c*, from 15 features becomes 5 features. The accuracy is reduced by 0.95% and the kappa statistic is reduced by 0.01.

The comparison results of subject based dataset (i.e. second group dataset) are show in Table 6, 8, 10, and 12 for full set features and Table 7, 9, 11, and 13 for selected features. To simplify our analysis we can resume Table 6, 8, 10, and 12 into Table 14 and Table 7, 9, 11, and 13 into Table 15. Using subject based dataset, the random forest classifier and the feature set *c* that derived from raw ECG signal do the best performance with overall average percent correct classification is 80.95% and the average of kappa statistic is 0.70 in full set features. The random forest classifier and the feature set *c* do the best performance also in the selected features using CFS; overall average percent correct classification is 79.78% and the average kappa statistic is 0.68. CFS can decrease 66.67% number of features, from 15 features becomes 5 features with only 1.17% decreasing of average percent correct classification and 0,02 decreasing of average kappa statistic. This best result is close to IB1. IB1 using the feature set *c* can achieve 80.54% average percent correct classification and 0.70 average kappa statistics in full set features; 79.29% average percent correct classification and 0.68 average kappa statistics in selected features.

Table 6: Percent correct comparison of every records using feature set *a* without feature selection; %C is percent correct classification and *K* is the Kappa statistic

Rec.	BN		MLP		IB1		RF	
	%C	*K*	%C	*K*	%C	*K*	%C	*K*
slp01a	60.40	0.42	58.10	0.38	55.21	0.35	61.09	0.42
slp01b	56.31	0.25	60.18	0.31	59.39	0.33	63.93	0.37
slp02a	64.02	0.46	77.18	0.60	73.73	0.56	76.60	0.58
slp02b	72.30	0.57	77.80	0.64	76.04	0.62	77.26	0.63
slp03	44.99	0.29	54.96	0.33	45.38	0.24	57.37	0.37
slp04	58.32	0.36	66.48	0.32	64.12	0.36	69.48	0.40
slp14	45.06	0.23	52.74	0.28	45.97	0.23	53.94	0.30
slp16	62.58	0.46	64.90	0.47	58.26	0.39	66.33	0.49
slp32	72.23	0.52	77.27	0.57	72.78	0.51	78.21	0.58
slp37	77.33	0.37	89.59	0.55	84.50	0.41	89.08	0.51
slp41	43.23	0.23	49.15	0.30	45.66	0.26	52.10	0.34
slp45	52.14	0.18	58.18	0.15	53.83	0.21	61.22	0.18
slp48	51.96	0.32	57.95	0.38	53.79	0.33	58.70	0.40
slp59	54.20	0.42	56.34	0.44	50.08	0.36	57.92	0.45
slp60	71.15	0.51	71.06	0.49	58.95	0.33	74.52	0.55
slp61	47.33	0.29	54.74	0.33	46.27	0.25	54.00	0.33
slp66	65.61	0.48	67.00	0.50	58.71	0.38	67.32	0.50
slp67x	51.77	0.25	55.04	0.30	55.64	0.32	61.08	0.38
Avg.	**58.39**	**0.37**	**63.81**	**0.41**	**58.80**	**0.36**	**65.56**	**0.43**

Table 7: Percent correct comparison of every records using feature set *a* with feature selection; %C is percent correct classification and *K* is the Kappa statistic

Rec.	BN		MLP		IB1		RF	
	%C	*K*	%C	*K*	%C	*K*	%C	*K*
slp01a	61.56	0.40	60.32	0.41	57.40	0.38	62.95	0.45
slp01b	60.96	0.28	60.87	0.30	60.64	0.35	64.62	0.39
slp02a	67.12	0.47	76.63	0.59	72.42	0.53	77.01	0.59
slp02b	72.50	0.57	78.13	0.65	75.66	0.61	77.71	0.64
slp03	48.79	0.30	53.95	0.31	44.86	0.24	56.15	0.36
slp04	61.00	0.37	67.63	0.32	65.56	0.37	70.57	0.42
slp14	48.02	0.22	52.65	0.27	45.10	0.21	53.98	0.30
slp16	60.94	0.42	63.72	0.46	58.70	0.40	65.94	0.49
slp32	73.94	0.51	78.83	0.59	71.86	0.49	78.19	0.58
slp37	86.06	0.47	90.82	0.58	84.45	0.41	88.80	0.50
slp41	45.48	0.26	48.63	0.29	47.59	0.29	50.99	0.33
slp45	55.20	0.19	59.99	0.12	51.61	0.18	60.28	0.17
slp48	51.01	0.30	56.54	0.36	51.61	0.30	55.51	0.35
slp59	54.16	0.41	58.58	0.46	45.29	0.30	58.32	0.46
slp60	72.82	0.52	74.61	0.55	58.98	0.33	74.43	0.55
slp61	49.99	0.28	55.03	0.30	47.98	0.27	54.39	0.33
slp66	64.21	0.46	69.75	0.54	64.40	0.46	66.84	0.50
slp67x	52.50	0.24	58.12	0.34	54.99	0.31	58.71	0.35
Avg.	**60.35**	**0.37**	**64.71**	**0.41**	**58.84**	**0.36**	**65.30**	**0.43**

Table 8: Percent correct comparison of every records using feature set *b* without feature selection; %C is percent correct classification and *K* is the Kappa statistic

Rec.	BN		MLP		IB1		RF	
	%C	*K*	%C	*K*	%C	*K*	%C	*K*
slp01a	52.23	0.34	51.74	0.27	45.53	0.20	50.56	0.27
slp01b	65.89	0.46	72.13	0.53	69.76	0.51	75.21	0.58
slp02a	45.20	0.27	72.99	0.53	61.32	0.35	67.18	0.39
slp02b	72.60	0.59	71.79	0.54	73.19	0.57	74.48	0.58
slp03	47.91	0.31	55.91	0.35	52.97	0.34	60.12	0.41
slp04	59.28	0.37	72.65	0.44	67.84	0.42	72.68	0.47
slp14	54.51	0.35	57.20	0.36	54.36	0.33	60.32	0.40
slp16	60.38	0.43	63.33	0.46	59.52	0.42	66.41	0.50
slp32	68.77	0.47	77.88	0.58	73.59	0.53	77.93	0.57
slp37	74.76	0.29	86.63	0.42	83.88	0.32	86.56	0.39
slp41	49.03	0.33	55.63	0.39	53.63	0.37	58.88	0.43
slp45	57.79	0.24	62.74	0.16	59.89	0.30	67.03	0.32
slp48	52.19	0.34	57.82	0.39	55.76	0.36	61.02	0.44
slp59	51.98	0.39	50.78	0.37	45.65	0.31	54.58	0.41
slp60	72.28	0.54	75.82	0.59	68.01	0.47	76.62	0.60
slp61	48.33	0.35	57.49	0.38	57.73	0.41	61.09	0.44
slp66	59.82	0.39	62.43	0.43	59.01	0.38	61.64	0.42
slp67x	66.66	0.48	62.90	0.42	62.00	0.40	66.41	0.47
Avg.	**58.87**	**0.39**	**64.88**	**0.42**	**61.32**	**0.39**	**66.59**	**0.45**

Table 9: Percent correct comparison of every records using feature set *b* with feature selection; %C is percent correct classification and *K* is the Kappa statistic

Rec.	BN		MLP		IB1		RF	
	%C	*K*	%C	*K*	%C	*K*	%C	*K*
slp01a	51.04	0.30	50.09	0.26	48.06	0.25	53.16	0.31
slp01b	63.27	0.43	71.00	0.50	75.53	0.60	77.23	0.62
slp02a	44.13	0.26	59.34	0.24	70.09	0.51	69.13	0.46
slp02b	73.49	0.58	72.36	0.55	70.67	0.54	72.73	0.56
slp03	47.78	0.30	57.33	0.36	54.84	0.37	59.86	0.42
slp04	69.70	0.42	74.57	0.46	67.95	0.43	73.43	0.49
slp14	53.08	0.31	53.99	0.26	53.57	0.33	57.25	0.36
slp16	54.46	0.35	62.09	0.42	61.41	0.44	65.03	0.48
slp32	78.78	0.58	79.21	0.59	74.75	0.55	77.39	0.57
slp37	79.23	0.28	86.85	0.30	83.54	0.36	85.55	0.37
slp41	49.87	0.33	55.66	0.39	52.49	0.36	53.94	0.37
slp45	59.98	0.23	61.61	0.05	61.79	0.33	70.64	0.44
slp48	56.81	0.39	61.14	0.44	58.37	0.39	60.72	0.43
slp59	53.66	0.41	53.14	0.39	54.67	0.42	57.70	0.45
slp60	75.33	0.59	77.32	0.61	69.59	0.49	75.71	0.59
slp61	48.91	0.34	57.50	0.36	59.29	0.43	63.12	0.47
slp66	60.58	0.40	66.31	0.49	60.81	0.41	65.23	0.47
slp67x	62.90	0.41	64.30	0.44	56.16	0.32	67.45	0.49
Avg.	**60.17**	**0.38**	**64.66**	**0.39**	**62.98**	**0.42**	**66.96**	**0.47**

Table 10: Percent correct comparison of every records using feature set *c* without feature selection; %C is percent correct classification and *K* is the Kappa statistic

Rec.	BN		MLP		IB1		RF	
	%C	K	%C	K	%C	K	%C	K
slp01a	69.64	0.58	80.84	0.71	82.14	0.74	84.37	0.77
slp01b	74.97	0.61	83.11	0.72	84.80	0.75	85.29	0.75
slp02a	66.49	0.54	89.18	0.81	86.70	0.78	87.46	0.78
slp02b	81.83	0.72	89.74	0.84	91.44	0.87	90.01	0.84
slp03	67.61	0.57	70.73	0.58	74.51	0.65	80.15	0.72
slp04	68.75	0.52	85.32	0.73	88.52	0.79	87.75	0.77
slp14	61.77	0.45	69.39	0.56	73.07	0.61	72.06	0.59
slp16	73.69	0.63	83.72	0.76	84.52	0.78	83.59	0.76
slp32	61.67	0.40	81.25	0.65	80.56	0.65	82.57	0.68
slp37	78.67	0.43	94.03	0.75	93.83	0.76	94.14	0.76
slp41	58.16	0.44	63.06	0.50	67.16	0.55	68.47	0.57
slp45	67.88	0.46	80.09	0.62	84.83	0.73	83.91	0.70
slp48	63.38	0.47	70.49	0.57	70.35	0.57	69.55	0.56
slp59	64.64	0.56	77.55	0.71	78.49	0.73	76.13	0.69
slp60	71.33	0.53	79.07	0.64	81.09	0.68	83.09	0.71
slp61	61.99	0.51	76.73	0.67	77.94	0.69	77.65	0.69
slp66	73.14	0.60	78.19	0.67	76.00	0.64	75.86	0.64
slp67x	70.61	0.55	74.70	0.60	73.76	0.59	75.07	0.62
Avg.	**68.68**	**0.53**	**79.29**	**0.67**	**80.54**	**0.70**	**80.95**	**0.70**

Table 11: Percent correct comparison of every records using feature set *c* with feature selection; %C is percent correct classification and *K* is the Kappa statistic

Rec.	BN		MLP		IB1		RF	
	%C	*K*	%C	*K*	%C	*K*	%C	*K*
slp01a	72.27	0.59	73.70	0.60	81.95	0.73	80.91	0.71
slp01b	73.66	0.58	79.03	0.63	83.63	0.73	83.71	0.73
slp02a	83.48	0.72	85.43	0.75	86.52	0.77	86.37	0.76
slp02b	81.32	0.71	87.27	0.80	91.59	0.87	89.27	0.83
slp03	68.77	0.56	62.10	0.41	73.03	0.63	77.70	0.69
slp04	73.68	0.57	81.56	0.65	87.58	0.78	87.77	0.78
slp14	61.34	0.44	64.51	0.47	69.66	0.56	70.19	0.56
slp16	73.45	0.62	78.61	0.68	83.29	0.76	82.05	0.74
slp32	75.40	0.54	80.49	0.63	80.38	0.65	81.60	0.66
slp37	88.55	0.55	91.80	0.62	93.20	0.74	92.61	0.69
slp41	57.39	0.43	60.01	0.45	68.35	0.57	66.97	0.55
slp45	77.35	0.55	75.04	0.50	84.23	0.72	82.96	0.68
slp48	68.72	0.55	69.35	0.56	66.93	0.52	68.99	0.55
slp59	65.44	0.56	70.50	0.62	75.19	0.68	75.44	0.69
slp60	72.84	0.54	70.76	0.49	77.10	0.62	82.26	0.70
slp61	61.80	0.49	65.72	0.50	75.19	0.65	75.60	0.66
slp66	71.44	0.57	76.42	0.65	74.63	0.62	74.95	0.62
slp67x	71.97	0.55	74.91	0.61	74.71	0.61	76.71	0.64
Avg.	**72.16**	**0.56**	**74.84**	**0.59**	**79.29**	**0.68**	**79.78**	**0.68**

Table 12: Percent correct comparison of every records using feature set *d* without feature selection; %C is percent correct classification and *K* is the Kappa statistic

Rec.	BN		MLP		IB1		RF	
	%C	*K*	%C	*K*	%C	*K*	%C	*K*
slp01a	66.59	0.53	70.75	0.55	70.63	0.56	79.85	0.69
slp01b	79.47	0.68	84.03	0.74	81.56	0.70	86.01	0.77
slp02a	69.52	0.58	85.43	0.76	84.66	0.75	86.90	0.78
slp02b	83.02	0.75	88.40	0.82	86.31	0.79	88.56	0.82
slp03	61.05	0.48	68.59	0.55	70.30	0.59	79.09	0.70
slp04	68.93	0.54	82.42	0.68	82.81	0.69	86.89	0.75
slp14	64.53	0.48	67.41	0.53	67.12	0.53	69.96	0.56
slp16	75.57	0.66	80.94	0.72	78.63	0.69	83.25	0.76
slp32	67.96	0.48	77.58	0.60	79.46	0.64	81.66	0.66
slp37	73.01	0.37	93.18	0.73	92.49	0.70	92.68	0.67
slp41	57.99	0.44	64.68	0.52	67.02	0.55	68.83	0.58
slp45	66.55	0.45	74.54	0.54	77.25	0.60	81.90	0.65
slp48	62.35	0.47	69.21	0.55	71.44	0.59	71.47	0.59
slp59	66.73	0.58	70.21	0.62	66.52	0.58	75.15	0.68
slp60	78.81	0.65	80.48	0.67	77.17	0.62	82.97	0.71
slp61	60.10	0.49	72.50	0.61	76.52	0.66	76.36	0.66
slp66	72.12	0.58	69.46	0.54	69.78	0.55	73.68	0.60
slp67x	74.13	0.61	75.70	0.62	75.74	0.62	74.88	0.61
Avg.	**69.36**	**0.55**	**76.42**	**0.63**	**76.41**	**0.63**	**80.00**	**0.68**

Table 13: Percent correct comparison of every records using feature set *d* with feature selection; %C is percent correct classification and *K* is the Kappa statistic

Rec.	BN		MLP		IB1		RF	
	%C	*K*	%C	*K*	%C	*K*	%C	*K*
slp01a	68.60	0.55	72.97	0.59	73.82	0.60	79.33	0.68
slp01b	78.21	0.66	82.91	0.72	82.46	0.72	83.87	0.73
slp02a	68.25	0.57	85.78	0.77	84.20	0.74	85.32	0.75
slp02b	82.80	0.74	84.66	0.76	86.66	0.79	86.39	0.79
slp03	63.02	0.49	68.71	0.55	71.68	0.61	78.43	0.69
slp04	73.98	0.59	84.49	0.71	82.98	0.69	87.19	0.76
slp14	62.92	0.46	65.33	0.50	66.17	0.52	69.40	0.55
slp16	75.72	0.65	80.57	0.72	80.66	0.72	82.55	0.75
slp32	71.16	0.51	79.41	0.63	78.53	0.62	82.10	0.67
slp37	86.24	0.55	93.27	0.72	93.80	0.74	92.93	0.70
slp41	58.52	0.45	64.19	0.51	65.21	0.53	67.39	0.56
slp45	70.70	0.50	74.96	0.53	77.41	0.61	81.41	0.65
slp48	64.64	0.49	68.73	0.55	65.67	0.50	69.88	0.56
slp59	64.60	0.55	70.81	0.63	69.03	0.61	73.97	0.67
slp60	82.33	0.70	80.89	0.68	77.56	0.63	84.09	0.73
slp61	63.10	0.51	71.26	0.58	74.87	0.64	77.09	0.67
slp66	73.88	0.61	72.87	0.59	73.17	0.60	75.14	0.63
slp67x	71.85	0.56	73.98	0.59	70.10	0.54	74.44	0.60
Avg.	**71.14**	**0.56**	**76.43**	**0.63**	**76.33**	**0.63**	**79.50**	**0.67**

Table 14: Average percent correct comparison of every feature set without feature selection; %C is percent correct classification and *K* is the Kappa statistic

Feature set	BN		MLP		IB1		RF	
	%C	*K*	%C	*K*	%C	*K*	%C	*K*
a	58.39	0.37	63.81	0.41	58.80	0.36	65.56	0.43
b	58.87	0.39	64.88	0.42	61.32	0.39	66.59	0.45
c	68.68	0.53	79.29	0.67	80.54	0.70	80.95	0.70
d	69.36	0.55	76.42	0.63	76.41	0.63	80.00	0.68

Table 15: Average percent correct comparison of every feature set with feature selection; %C is percent correct classification and *K* is the Kappa statistic

Feature set	BN		MLP		IB1		RF	
	%C	*K*	%C	*K*	%C	*K*	%C	*K*
a	60.35	0.37	64.71	0.41	58.84	0.36	65.30	0.43
b	60.17	0.38	64.66	0.39	62.98	0.42	66.96	0.47
c	72.16	0.56	74.84	0.59	79.29	0.68	79.78	0.68
d	71.14	0.56	76.43	0.63	76.33	0.63	79.50	0.67

4. Conclusions

The CFS shows good result for Bayesian network, IB1 and random forest with relatively small amount in decreasing accuracy and significantly reducing the number of the features but it is not too good for multilayer perceptron. Increasing accuracy in Table 5 indicates that not all features will be useful and adding inappropriate features may decrease the accuracy of some classifiers. For example Bayesian network classifier with selected feature set *a* and feature set *d*, random forest classifier with selected feature set *d*. Multilayer perceptron and IB1 always show better results when not reducing the number of features using the features selection method.

Overall, random forest classifier is better than Bayesian network, multilayer perceptron or IB1 for any feature sets in the first group dataset. On the other hand, based on our experiment and setting, the features that derived from raw ECG signal have more potency to be the signature of sleep stages than the features that derived from RR interval or EDR information. This feature set has higher accuracy and kappa statistic than the others. The combination of the random forest classifier and the features that derived from raw ECG signal shows the best performance. Using full set features from raw ECG signal, we can get 79.80% correctly classified instances and 0.72 kappa statistic. CFS as feature selection method shows good result. We can reduce 66.67% number of features that derived from raw ECG signal and only reduce 0.95% accuracy and 0.01 kappa statistic. The accuracy becomes 78.85% and the kappa statistic becomes 0.71.

The experiments using subject based dataset support our conclusion that the combination of the random forest as the classifier and the features that derived from raw ECG signal perform better result than the others. Using full set features from raw ECG signal, we can get 80.95% of average percent correct classification and 0.70 average kappa statistic. CFS also performs well; using selected features, we can get 79.78% average percent correct classification and 0.68 average kappa statistics; it means only reduce 1.17% in the average percent correct classification and 0.02 in the average kappa statistics. This results show almost the same with the first group dataset. It means the classifiers and the feature set are robust.

Acknowledgments

This research as a part of research grant with title Development Of Sleep-Awakening Timing Controller for Occupational Safety and Health Based-On Computational Intelligent Algorithm was supported by Universitas Indonesia.

References

[1] American Academy of Sleep Medicine, *International classification of sleep disorders, revised: Diagnostic and coding manual*, Chicago, Illinois: American Academy of Sleep Medicine, 2001.

[2] C. Pollak, M.J. Thorpy, and J. Yager. *The Encyclopedia of Sleep and Sleep Disorders. Facts on File library of health and living*. New York: Facts on File, 2009.

[3] Z. Shinar et al., "Automatic detection of slow-wave-sleep using heart rate variability", *Computers in Cardiology 2001*, 2001, pp. 593-596.

[4] A. Lewicke, et al., "Sleep versus wake classification from heart rate variability using computational intelligence: consideration of rejection in classification models", *IEEE Trans Biomed Eng*, Vol. 55, No. 1, 2008, pp. 108–18.

[5] B. Yilmaz et al., "Sleep stage and obstructive apneaic epoch classification using single-lead ecg", *BioMedical Engineering OnLine*, Vol. 9, No. 1, 2010, pp. 39.

[6] M. Bsoul et al., "Real-time sleep quality assessment using single-lead ECG and multi-stage SVM classifier", in *the International Conference of IEEE Engineering in Medicine and Biology Society*, 2010, Vol 2010, pp 1178–1181.

[7] J.M. Shneerson. *Sleep medicine: a guide to sleep and its disorders*, Oxford: Blackwell Publishing, 2005.

[8] G. D. Clifford, F. Azuaje, and P. McSharry, *Advanced Methods And Tools for ECG Data Analysis*, USA:Artech House, Inc., 2006.

[9] R. V. Andreo, B. Dorizzi, and J. Boudy, "Ecg signal analysis through hidden markov models", *IEEE Transactions on Biomedical Engineering*, Vol. 53, No. 8, 2006, pp. 1541–1549.

[10] A. L. Goldberger et al., "Physiobank, physiotoolkit, and physionet : Components of a new research resource for complex physiologic signals", *Circulation*, Vol. 101, No. 23, 2000, pp. e215–220.

[11] "Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and lectrophysiology", *Circulation*, Vol. 93, No. 5, 1996, pp. 1043–1065.
[12] J. E. Mietus. *Time domain measures:from variance to pnnx*. Internet: http://physionet.org/events/hrv-2006/mietus-1.pdf. June 2011.
[13] P. de Chazal, T. Penzel, and C. Heneghan, "Automated detection of obstructive sleep apnoea at different time scales using the electrocardiogram", *Physiological Measurement*, Vol. 25, No. 4, 2004, pp. 967.
[14] F. S. Bao, X. Liu, and C. Zhang, "PyEEG:An open source python module for EEG/MEG feature extraction. *Comp. Int. and Neurosc.*, Vol. 2011, 2011.
[15] M. Wiggins et al., Evolving a bayesian classifier for ecg-based age classification in medical applications. *Applied Soft Computing*, Vol. 8, No. 1, 2008, pp. 599 – 608.
[16] R. R. Bouckaert, "Bayesian network classifiers in weka", Technical report, University of Waikato, May 2008.
[17] S. Haykin, *Neural Networks: A Comprehensive Foundation (2nd edition)*, Upper Saddle River, NJ: Prentice Hall, 1999.
[18] D. W. Aha and D. Kibler. "Instance-based learning algorithms", In *Machine Learning*, 1991, pp 37–66.
[19] L. Breiman, "Random forests", *Mach. Learn.*, Vol. 45, 2001, pp. 5–32.
[20] M. A. Hall, "Correlation-based Feature Subset Selection for Machine Learning", PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.
[21] A. J. Viera and J. M. Garrett, "Understanding Interobserver Agreement: The Kappa Statistic", *Family Medicine*, Vol. 37, No. 5, 2005, pp. 360–363.

Ary Noviyanto holds Bachelor degree in Computer Science from Department of Computer Science, Universitas Gadjah Mada, Indonesia. He is a research assistant for image processing and pattern recognition laboratory in faculty of computer science, Universitas Indonesia.

Sani M. Isa holds Bachelor degree in Mathematics from Faculty of Natural Science, Padjadjaran University, Indonesia; and Master degree from Faculty of Computer Science, University of Indonesia.

Ito Wasito holds Ph.D in Computer Science, School of Computer Science and Information Systems, Birkbeck College, University of London, United Kingdom.

Aniati Murni Arymurthy holds Bachelor degree in Electrical Engineering from Faculty of Engineering, University of Indonesia, Master degree in Computer and Information Science, The Ohio State University, United States of America, and Doctor degree in Dept. Of Optoelectronics and Laser Application, University of Indonesia.

Specification and Verification of Uplink Framework for Application of Software Engineering using RM-ODP

Krit Salahddine[1], Laassiri Jalal[2] and El Hajji Said[3]

[1] Polydisciplinary Faculty of Ouarzazate, University Ibn Zohr, BP/638, Morocco

[2] Department of informatics, Faculty of Sciences, University Ibn Tofail, BP 33, Morocco

[3] Department of Mathematic Informatics, University Mohamed V-Agdal, BP 1040, Morocco

Abstract

This paper present a survey and discussion of the Reference Model for Open Distributed Processing (RM-ODP) viewpoints; oriented approaches to requirements engineering viewpoint and a presentation of new work in the application wireless mobile phone, this area which has been designed with practical application using the Unified Modelling Language (UML)/VHDL_AMS (VHSIC Hardware Description Language Analog and Mixed-Signal). We mainly focus on rising and fulling time, action, uplink behaviour constraints (sequentiality, non determinism and concurrency constraints).We discuss the practical problems of introducing viewpoint; oriented requirements engineering into industrial software engineering practice and why these have prevented the widespread use of existing approaches.

The goal of this article is to check the uplink path using the MIC (Microphone amplifier) with all analog inputs, and check the amplifier gain.

This paper provides an example of using the Uplink Framework to build a comprehensive, good solution for Application Wireless Mobile Phone.

Finally, we discuss how well this approach addresses some outstanding problems in requirements engineering (RE) and the practical industrial problems of introducing new requirements engineering methods.

Keywords: *RM-ODP, UML, VHDL-AMS, Uplink Behaviour, Software engineering, RE.*

1. Introduction

The rapid growth of distributed processing has led to a need for coordinating framework for the standardization of Open Distributed Processing (ODP). The Reference Model for Open Distributed Processing (RM-ODP) [1]-[4] provides a framework within which support of distribution, networking and portability can be integrated. The foundations part [2] contains the definition of the concepts and analytical framework for normalized description of (arbitrary) distributed processing systems. These concepts are grouped in several categories. The architecture part [3] contains the specifications of the required characteristics that qualify distributed processing to be open. It defines a framework comprising five viewpoints, viewpoint language, ODP functions and ODP transparencies. The five viewpoints, called enterprise, information, computational, engineering and technology provide a basis for the specification of ODP systems.

Each viewpoint language defines concepts and rules for specifying ODP systems from the corresponding viewpoint. The ODP functions are required to support ODP systems.

In this context, VHDL_AMS is used to specify the properties to be tested. The UML meta-models provide a precise core of any ODP tester. We use in this paper ModelSim under Cadence to verify process behavior based on interaction and the binding object in the ODP systems.

VHDL_AMS is an industry standard modeling language for mixed signal circuits. It provides both continuous-time and event-driven modeling semantics, and so is suitable for analog, digital, and mixed analog/digital circuits. It is particularly well suited for verification of very complex analog, mixed-signal and radio frequency integrated circuits.

This capability is used to highlight some benefits of the Architectural realized (uplink path): raising the level of abstraction at which development occurs; which, in turn, will deliver greater productivity, better quality, and insulation from underlying changes in technology.

We treated the need of formal notation for Uplink behavioral concepts in the Computational language [8]. Indeed, the viewpoint languages are abstract in the sense that they define what concepts should be supported, not how these concepts should be represented. It is important

to note that, RM-ODP uses the term language in its broadest sense: “a set of terms and rules for the construction of statements from the terms”. It does not propose any notation to support the viewpoint languages. Using the Unified Modeling Language (UML)/VHDL_AMS [9], [10] we defined a formal semantic for a fragment of ODP uplink behavior concepts defined in the RM-ODP foundations part and in the engineering language [11]. These concepts (time, action, uplink behavior constraints) are suitable for describing and constraining the uplink behavior of ODP engineering viewpoint specifications.

2. Meta-MODELING Time and Behavioral Constraints

Behavioral constraints may include sequentiality, non-determinism, concurrency, real time” (RM-ODP, part 2, clause 8.6). In this work we consider constraints of sequentiality, non-determinism and concurrency. The concept of constraints of sequentiality is related with the concept of time.

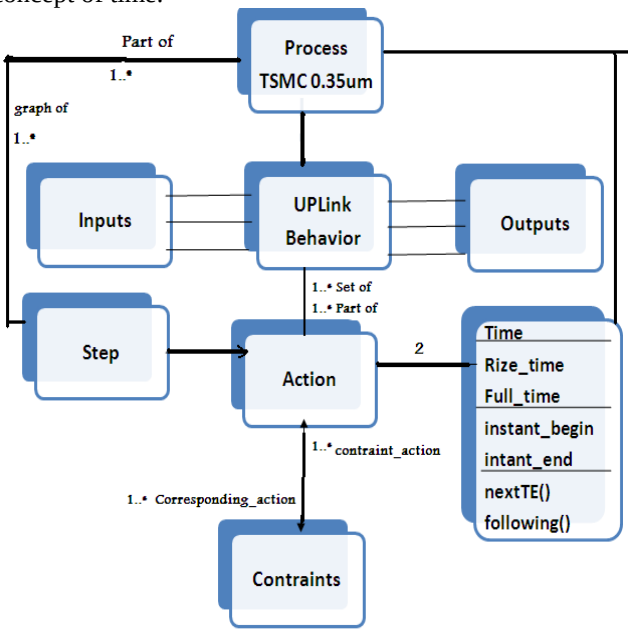


Fig.1 Meta-Modelling core Behavior concept in RM-ODP foundations part

2.1Time

Time has two following important roles in system design [26]:

•It serves for the purpose of synchronization of actions inside and between processes, the synchronization of a system with system users, the synchronization of user requirements with an actual performance of a system.

•It defines sequences of events (action sequences)

To fulfil the first goal, we have to be able to measure time intervals. However, a precise clock that can be used for time measurement does not exist in practice but only in theory [27]. So the measurement of the time is always approximate. In this case we should not choose the most precise clocks, but ones that explain the investigated phenomena in the best way. Simultaneity of two events or their sequentiality, equality of two durations should be defined in the way that the formulation of the physical laws is the easiest” [27]. For example, for the actions synchronization, internal computer clocks can be used and, for the synchronization of user requirements, common clocks can be used that measure time in seconds, minutes and hours.

We consider the second role of time. According to [27] we can build some special kind of clock that can be used for specifying sequences of actions. RM-ODP confirms this idea by saying that “a location in space or time is defined relative to some suitable coordinate system” (RM_ODP, part 2, clause 8.10). The time coordinate system defines a clock used for system modeling. We define a time coordinate system as a set of time events. Each event can be used to specify the beginning or end of an action. A time coordinate system must have the following fundamental properties [26]:

•Time is always increasing. This means that time cannot have cycles.

•Time is always relative. Any time moment is defined in relation to other time moments (next, previous or not related). This corresponds to the partial order defined for the set of time events.

We use the UML (fig1) and OCL to define time: Time is defined as a set of time events.

nextTE: defines the closest following time events for any time events [26].

We use the followingTE relation to define the set of the following time events or transitive closure for the time event t over the nextTE relation:

followingTE: defines all possible following time events Using followingTE we can define the following invariant that defines the transitive closure and guarantees that time event sequences do not have loops :

Context t : time **inv** :

Time->forAll(t:Time | (t.nextTE->isempty implies t.follwingTE->isempty)

and (t.nextTE->notempty and t.follwingTE->isempty implies t.follwingTE =t.nextTE) and (t.nextTE->notempty and t.follwingTE->notempty implies t.follwingTE->

includes(t.nextTE.follwingTE->union(t.nextTE)) and t.follwingTE->exludes(t)).

This definition of time is used in the next section to define sequential constraints.

2.2 Behavioral constraints

We define the behavior like a finite state automaton (FSA). For example, figure 2 shows a specification that has constraints of sequentiality and non determinism. The system is specified using constraints of non-determinism since state S1 has a non-deterministic choice between two actions a and b.

Based on RM-ODP, the definition of behavior must link a set of actions with the corresponding constraints. In the following we give definition of constraints of sequentiality, of concurrency and of non-determinism.

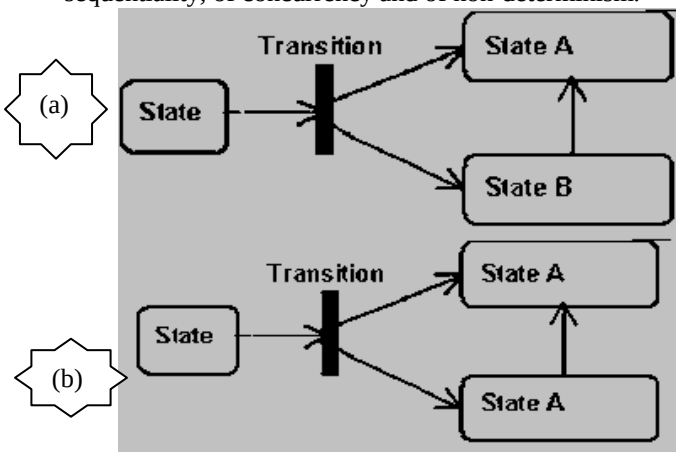


Fig. 2. a - Sequential deterministic constraints;

b - Sequential non deterministic constraints.

B.1 Constraints of sequentiality

Each constraint of sequentiality should have the following properties [26]:

•It is defined between two or more actions.

•Sequentiality has to guarantee that one action is finished before the next one starts. Since RM-ODP uses the notion of time intervals it means that we have to guarantee that one time interval follows the other one:

Context sc : constraintseq **inv** :
Behavior.actions-> forAll(a1,a2 | a1<> a2 and a1.constraints->includes(sc)
and a2.constraints->includes(sc) and
((a1.instant_end.followingTE->includes(a2.instant_begin)
or(a2.instant_end.followingTE->includes(a1.instant_begin))

For all SeqConstraints sc, there are two different actions a1, a2, sc is defined between a1 and a2 and a1 is before a2 or a2 is before a1.

B.2 Constraints of concurrency

Figure 3 shows a system specification that has constraints of concurrency since state a1 has a simultaneous choice of two actions a2 and a3.

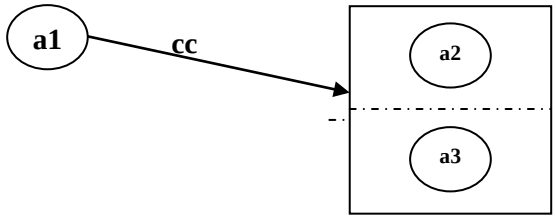


Fig. 3. RM-ODP diagram: Example constraints of concurrency

For all concuConstraints cc there is a action a1, there are two different internal actions a2, a3, cc is defined between a1 and a2 and a3, a1 is before a2 and a1 is before a3

Context cc: constraintconc inv:
Behavior.actions-> forAll(a1 :Action ,a2 ,a3 : internalaction | (a1 <> a2) and
(a2 <> a3) and (a3 <> a1) and a1.constraints->includes(cc) and a2.constraints->includes(cc) and a3.constraints->includes(cc) and
a1.instant_end.followingTE-> includes(a2.instant_begin) and
a1.instant_end.followingTE-> includes(a3.instant_begin))

B.3 Constraints of non-determinism

In order to define constraints of non-determinism we consider the following definition given in [24]: “A system is called non-deterministic if it is likely to have shown number of different behavior, where the choice of the behavior cannot be influenced by its environment”. This means that constraints of non-determinism should be defined between a minimum of three actions. The first action should precede the two following actions and these actions should be internal (see figure 4).

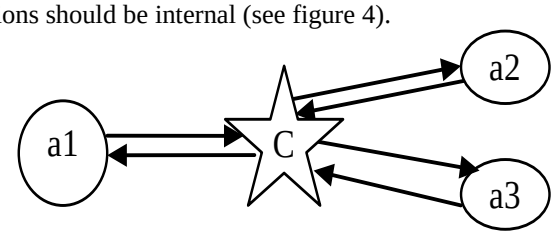


Fig. 4. Example Constraints example of non-determinism

We define this constraint as follows:
Context ndc: NonDetermConstraints **inv**:
Behavior.actions-> forAll(a1 :Action ,a2 ,a3 : internalaction | (a1 <> a2) and
(a2 <> a3) and (a3 <> a1) and a1.constraints->includes(ndc) and
a2.constraints->includes(ndc) and
a3.constraints->includes(ndc) and
a1.instant_end.followingTE-> includes(a2.instant_begin) or
a1.instant_end.followingTE-> includes(a3.instant_begin)) .

We note that, since the choice of the behavior should not be influenced by environment, actions a2 and a3 have to be internal actions (not interactions). Otherwise the choice between actions would be the choice of environment [26].

3. Simulation results and discussion

Simulations are carried on using VHDL-AMS after verification all of the connectivity between the different actions.
The goal of action 1 is to check the uplink path using the MIC amplifier with all analog inputs, and check the amplifier gain.
Sine waves (at different frequencies) are sent from all analog inputs to the VSIF through the uplink path (MIC amplifier, ADC).
First MICN/P differential input is selected Fig.5.
Then FML mono input is selected, and then HSMIC mono input is selected fig.6.
The MIC amplifier gain is set from 3dB to 33dB.
Checks are done on the MIC amp outputs and ADC outputs: signal, gain fig.7

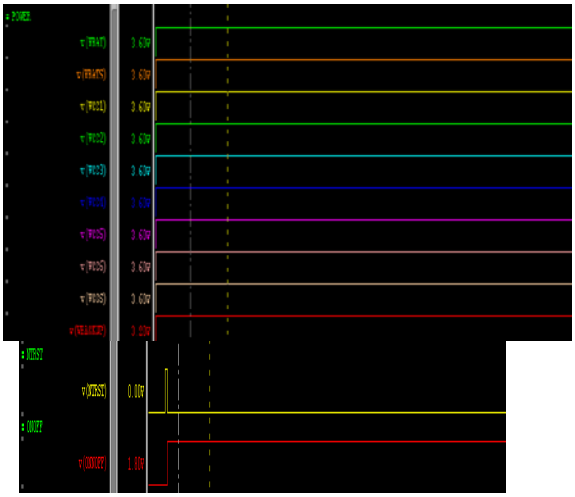


Fig.5 NPUT signals

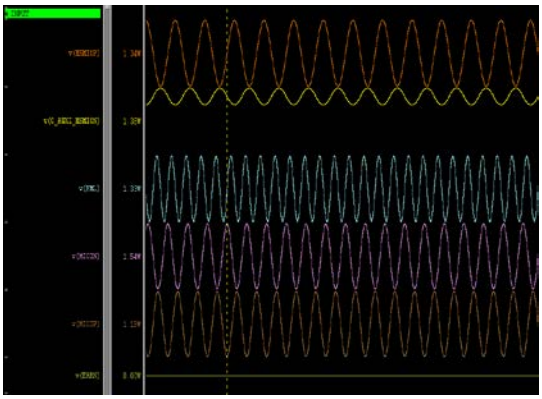


Fig.6 Input analog signals

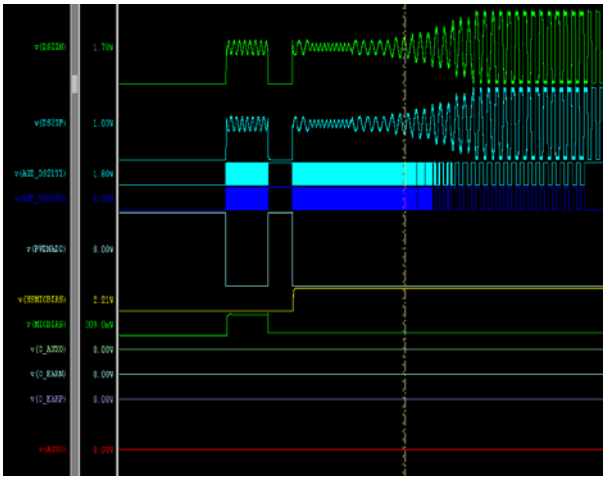


Fig.7 microphone amplifier signals

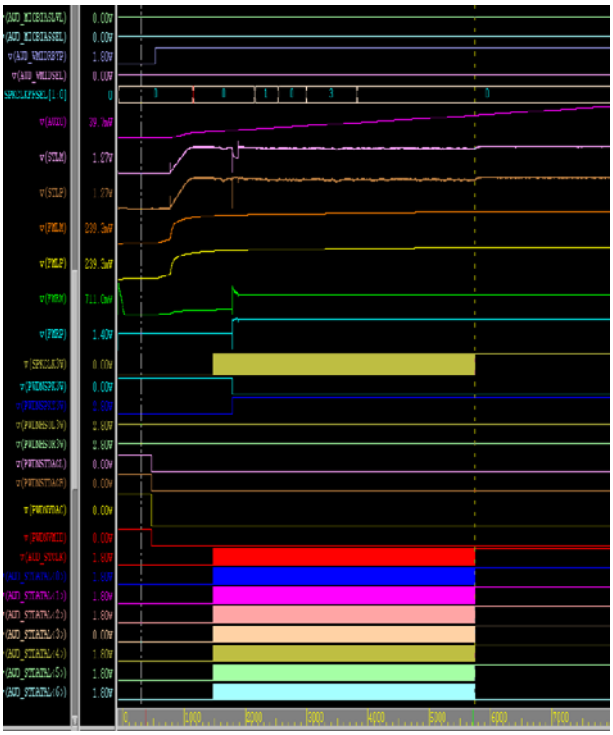


Fig.8 Speaker output

4. Conclusions

An Architectural Framework for Wireless Mobile has been proposed. This Architectural Framework will allow Uplink to be used in any wireless environment, as required, to provide any type of services demanded by the user regardless whether the uplink is a stand alone, a pure Wireless sensor network or integrated with other networks.
In this paper we have presented our contribution to the RMODP standard-related research. This contribution resolves an important problem of the RM-ODP standard: the absence of a single consistent formalization of the RM-ODP conceptual framework. A realization of such formalization was officially verified.
The goal of our work is to help promote the practical applications of RM-ODP. The formal model of RM-ODP Part 2 that we presented in this paper can indeed serve for the promotion of RM-ODP towards a wider use in the modern modeling practices. Some of the applications of our results have already justified this claim.

References

[1] ISO/IEC, ‘’Basic Reference Model of Open Distributed Processing-Part1: Overview and Guide to Use, ‘’ISO/IEC CD 10746-1, 1994
[2] ISO/IEC, ‘’RM-ODP-Part2: Descriptive Model, ‘’ ISO/IEC DIS 10746-2, 1994.
[3] ISO/IEC, ‘’RM-ODP-Part3: Prescriptive Model, ‘’ ISO/IEC DIS 10746-3, 1994.
[4] ISO/IEC, ‘’RM-ODP-Part4: Architectural Semantics, ‘’ ISO/IEC DIS 10746-4, July 1994.
[5] M. Bouhdadi et al., ‘’A UML-Based Meta-language for the QoS-aware Enterprise Specification of Open Distributed Systems’’ IFIP Series, Vol 85, Springer, (2002) 255-264.
[6] J. Rumbaugh et al., The Unified Modeling Language, Addison Wesley, 1999.
[7] B. Rumpe, ‘’A Note on Semantics with an Emphasis on UML, ‘’ Second ECOOP Workshop on Precise Behavioral Semantics, LNCS 1543, Springer, (1998) 167-188.
[8] A. Evans et al., ‘’Making UML precise, ‘’ Object Oriented Programming, Systems languages and Applications, (OOPSLA'98), Vancouver, Canada, ACM Press (1998)
[9] A. Evans et al. The UML as a Formal Modeling Notation, ‘’ UML, LNCS 1618, Springer, (1999) 349-274
[10] J. Warmer and A. Kleppe, The Object Constraint Language: Precise Modeling with UML, Addison Wesley, (1998).
[11] S. Kent, et al. ‘’A meta-model semantics for structural constraints in UML,, In H. Kilov, B. Rumpe, and I. Simmonds, editors, Behavioral specifications for businesses and systems, Kluwer , (1999). chapter 9
[12] E. Evans et al., Meta-Modeling Semantics of UML, In H. Kilov, B. Rumpe, and I. Simmonds, eds, Behavioral specifications for businesses and systems, Kluwer , (1999). ch. 4.
[13] D.A. Schmidt, ‘’Denotational semantics: A Methodology for Language Development, ‘’ Allyn and Bacon, Massachusetts, (1986)
[14] G. Myers, ‘’The art of Software Testing, ‘’, John Wiley &Sons, (1979)
[15] Binder, R. ‘’ Testing Object Oriented Systems. Models. Patterns, and Tools, ‘’ Addison-Wesley, (1999)
[16] A. Cockburn, ‘’Agile Software Development. ‘’Addison-Wesley, (2002).
[17] B. Rumpe, ‘’ Agile Modeling with UML, ‘’ LNCS vol. 2941, Springer, (2004) 297-309.
[18] Beck K. Column on Test-First Approach. IEEE Software, Vol. 18, No. 5, (2001) 87-89
[19] L. Briand , ‘’A UML-based Approach to System testing, ‘’ LNCS Vol. 2185. Springer, (2001) 194-208,
[20] B. Rumpe, ‘’ Model-Based Testing of Object-Oriented Systems; ‘’ LNCS Vol.. 2852, Springer; (2003) 380-402.
[21] B. Rumpe, Executable Modeling UML. A Vision or a Nightmare?, In: Issues and Trends of Information technology management in Contemporary Associations, Seattle, Idea Group, London, (2002) pp. 697-701.
[22] M. Bouhdadi, Y. Balouki, E. Chabbar. ‘’ Meta-Modeling Syntax and Semantics of Structural Concepts for Open Networked Enterprises”, ICCSA 2007, Kuala Lumpor, 26-29 August, LNCS 4707, Springer, (2007) 45-54
[23] Lamport, L. and N.A. Lynch, Distributed Computing: Models and Methods, in Handbook of Theoretical Computer Science, Volume B: Formal Models and Semantics. 1990, Elsevier and MIT Press.
[24] Broy, M., “Formal treatment of concurrency and time,’’ in Software Engineer's Reference Book,J. McDermid, Editor, Oxford: Butterworth-Heinemann, (1991),pp 23
[25] Wegmann, A. et al. ‘’ Conceptual Modeling of Complex Systems Using RMODP Based Ontology‘’ . in 5th IEEE International Enterprise Distributed Object Computing Conference -EDOC (2001). September 4-7 USA. IEEE Computer Society pp. 200-211
[26] P. Balabko, A. Wegmann, “From RM-ODP to the formal behavior representation” Proceedings of Tenth OOPSLA Workshop on Behavioral Semantics ¨Back to Basics¨, Tampa, Florida, USA , pp. 11-23 (2001).
[27] Henri Poincaré, The value of science, Moscow «Science», 1983
[28] Harel, D. and E. Gery, “Executable object modeling with statecharts“, IEEE Computer.30(7) pp. 31-42 (1997)
[29] Jean-Raymond Abrial: A System Development Process with Event-B and the Rodin Platform. ICFEM (2007) 1-3.

Salah-ddine Krit received the B.S. and Ph.D degrees in Microectronics Engineering from Sidi Mohammed Ben Abdellah university, Fez, Morroco. Institute in 2004 and 2009, respectively. During 2002-2008, he is also an engineer Team leader in audio and power management Integrated Circuits (ICs) Research.

Design, simulation and layout of analog and digital blocks dedicated for mobile phone and satellite communication systems using CMOS technology. He is currently a professor of informatics with Polydisciplinary Faculty of Ouarzazate, Ibn Zohr university, Agadir, Morroco. His research interests include wireless sensor Networks (Software and Hardware), computer engineering and wireless communications.

Jalal Laassiri received his Bachelor's degree (License es Sciences) in Mathematics and Informatics in 2001 and his Master's degree (DESA) in computer sciences and engineering from the faculty of sciences, university Mohammed V, Rabat, Morocco, in 2005, and he developed He received his Ph.D. degree in computer sciences and engineering from University of Mohammed V, Rabat, Morocco, in Juin, 2010. He was a visiting scientific with the Imperial College London, in London, U.K. He is Member of the International Association of Engineers (IAENG), He joined the Faculty of Sciences of Kénitra, Department of Computer Science , Ibn Tofail University, Morocco, as an Professor in October 2010, His current research interests include Software and Systems Engineering, UML-OCL, B-Method, ..

3D Graphical User Interface on personal computer using p5 Data Glove

Khyati R. Nirmal [1] , Nitin Mishra [2]

[1] M.Tech *, Departement of IT , NRI Institutions, RGPV
Bhopal, Madhya Pradesh, India

[2] Prof, Departement of IT, NRI Institutions, RGPV
Bhopal, Madhya Pradesh, India

Abstract

*T*his paper presents Essential Reality works on 3D HCI for changing 2D visual to 3D visual. The mouse is the critical interface to handle 3D graphical objects. Using data glove it's possible to put it on like a normal glove and it then acts as an input device that senses finger movements and hand position and orientation (3 coordinates) in real time. The limitation of surface do not allow large no of windows and icons to be positioned on the screen. If more no of windows are forcibly open some of them get overlie. However in 3D spaces we get one more dimension and it's better to work with 3D instead of 2D. Number of windows and icons can be situated five surface.. It looks like a tunnel in which icons and windows glide in the space as the windows are dragged to the for front they turn in to bigger and they pushed back they turn in to smaller due to perspective projections**.**

Keywords:3 Dimensional, Human Computer Interface, perspective projection,WIMP

1. Introduction

The graphical user interfaces (GUIs) that are known as WIMP (which stands for Windows, Icons, Menus and Pointing) have been dominant for more than two decades. In a time when computer tools and applications became used by millions of people, software engineers realized that the interface was one of the main factors that determined the success of an application. WIMP user interfaces provided a "de facto"standard that, thanks to the existent consistencies in the look and feel of application interfaces, gave the user ease of learning and ease of use. As a result, user interface design was introduced in the life cycle of software development, and many methodologies have been proposed which support the development of user interfaces [1].

Interface technology advances towards a fourth generation of user interfaces, a "post-WIMP" generation which is not only based on 2D widgets such as menus, forms or toolbars, but also on 3D user interfaces that enable a more natural and intuitive style of interaction Human-computer interaction currently faces the challenge of taking advantage of new technologies, which efficiently use the capabilities of the computing systems of today and more effectively match human capabilities and perception. We investigate how 3D GUI can be useful in software applications of today and identify general areas where 3D technologies can be more useful than other technologies [2]. This paper consists of a survey of 3D standards, GUIs and in particular 3D GUIs. There is no doubt that 3D is going to be the way of the future when it comes to software. The revolutionary 3D Wonder 3D Desktop improves and makes it easier to use your Windows PC. Its 3D GUI replaces the decade old 2D desktop with a more productive and enjoyable 3D interface. Lets you organize your tasks and icons in previously impossible way. Animated real-time .3D Worlds create an overwhelming entertaining and fun to use desktop.

2. The existing 3 D environment in the world of proprietary software

In the last few years some interesting software application has appeared that have brought 3 dimensionality to the world of computer graphical interface. In the literature the term "3D interface" is used to describe a wide variety of interfaces for displaying and interacting with 3D objects. True 3D interfaces i.e. interfaces with all its components in a 3D environment have not yet had any major impact outside the laboratory. The interfaces nowadays referred to as 3D GUIs are almost exclusively "hybrids" between 2D and 3D interfaces. In this thesis we are introducing complete 3D GUI having 3D Windows, Icons, Menus, and Pointer (WIMP) and more on those 3D characters.

Apple was the first mainstream vendor to choose three-dimensionality when it released MAC OS X in

2001.This operating system interface introduced some functions able to take advantages of the 3d graphics cards installed in Macintosh computers. There are semi transparency effects able to permit a better management of overlapped windows, something that recalls the 'glass metaphor 'used 6 years later in Microsoft Window Vista. [3]

Without doubt the "glass" is one of the main 'new' features of **Aero** the 3D GUI present in the Windows Vista pack. Aero's main effects are, beyond the semi-transparencies, , a 3 dimensional windows managing system called Windows Flip 3D. With this feature a new concept of Spatial organization and management based on a perspective effect that allows better management of the desktop space in the Windows graphical interface has been introduced. This makes possible more natural navigation. This new function requires a powerful graphics card creating technologies limitation because not all existing computer have similar capabilities, this trend, with the pretext of 3d GUIs, will surely continue. [3]

Now talking about open source we must begin with the Sun looking **Glass project**, supported by the Sun Microsystems software house. In this system, windows start in 2d normal mode but can be manipulated as 3d objects that can be set any angle or turned completely around the user but still in appearance of 3D not the real one. [4]

In the Linux/Unix world we focus in particular on the ***Compiz* project**, the first software to popularize the 3D interfaces on this operating system. The effects of virtual desktop and cubes are effects that make possible fast access to all six segments of off-screen space. Instead, with the concept of the virtual desktop and its 3D visualization, the off-screen space is now shown. But the limitation is still the 2D characters and 2D Input Device. [5]

3D Top uses the icons that are present on your normal desktop and represents them in 3 dimensions instead. This approach enables you to fly around your desktop, change the shape of the icons and arrange or drag them in 3D also. In addition you can create colored spotlights, paintings (bitmaps), background and floor textures, clocks, flags (shortcuts) and whatsoever the future may bring. But It is Hybrid version of 3D GUI not the real one. [6]

The Cubic Eye is the revolutionary 3D Application Platform transforming the way people view and interact with their data. The only 3D platform based on cubes, Cubic Eye is the evolution of the traditional 2D user interface into a real-time 3D environment, providing a richer, more productive, more enjoyable experience. But this application is virtual vision not the real vision.[7]

Currently available application will provide the features of rotation, scaling, tilting, Selecting, moving, windows, and icons. But all this functionality is done using mouse and keyboard. This way of handling of 3D objects is some what difficult and not that much efficient in working.

One more cons with current application available in market are that it does not provide any facility for adjusting the fonts according to the rotation of 3D objects. So its reliability is not that much powerful,

3. Visualization

In a 3D GUI the visual elements are genuinely three-dimensional: they are situated in xyz space, are defined in terms of 3D coordinates, need not be flat and may contain spatial regions (volumes). At this stage we plan on changing the structure of WIMP and characters, which are essentially 2D constructs, which gives real of 3D presentation and provides more advantageous.

To implement this 3D WIMP components our prototype use perspective projection. Here to display real 3D interface to 2D display device conversion is the necessity. Formulas for this are

$$xd1 = ((xc * z1) - (zc * x1)) / (z1 - zc)$$
$$yd1 = ((yc * z1) - (zc * y1)) / (z1 - zc)$$
$$xd2 = ((xc * z2) - (zc * x2)) / (z2 - zc)$$
$$yd2 = ((yc * z2) - (zc * y2)) / (z2 - zc)$$

(xc ,yc, zc) is the point from where user look inside the object

(x1,y1,z1) and (x2,y2,z2) two original end points of segment

(xd1,yd1) and (yd2,zd2) after perspective projection two end points

There are 5 main thrusts to our prototyping work.

3.1 The desktop manager

The desktop manager metaphor we have used for the 3D space in which we arrange windows is that of a *tunnel.* The user is positioned in the middle of the tunnel looking toward the other end. The tunnel, and windows in it, is displayed with a perspective projection. The tunnel shaped desktop provides 5 surfaces to manipulate with multiple windows.

Fig. 1 3D tunnel shaped desktop

3.2 Window Manager

Windows may be positioned at arbitrary depths in the tunnel. The normal orientation is orthogonal to the longitudinal axis of the tunnel, or more simply, ``front-on". As a user pushes a window further into the tunnel its size is diminished in the normal inverse size to distance relationship of perspective projection. In addition to the front-on window display mode there is a ``hanging" mode where the windows are hung on the left or right walls of the tunnel, as shown in figure 1. Hanging all the windows allows the user to quickly gain an overall idea of where the windows are in the tunnel.

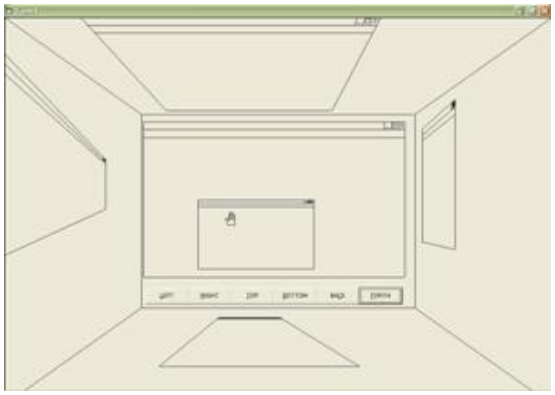


Figure 2. Multiple windows placed in 5 surface of desktop

The first is an *overview* area to the right of the tunnel which may be seen in figure 1. The overview area provides a plan or “top view of the tunnel” and is provided in addition to the main ``down the tunnel" view. Windows may be selected and then moved up and down in the overview area, corresponding to back and forward in the tunnel. The last major component of our window manager is a console, positioned at the bottom of the screen as shown in figures Controls are provided on the console for changing global settings affecting windows. There is a button to hang all the windows at once, rather than hanging individual windows.

3.3 Character Manager

To tilt and rotate the multiple windows with normal 2D characters cause the problem of sharp visibility as 2D character are not compactable with 3d window manipulation. In our prototype we have introduced 3d characters which are compactable with 3d perspective projection of individual window as shown in figure 3.

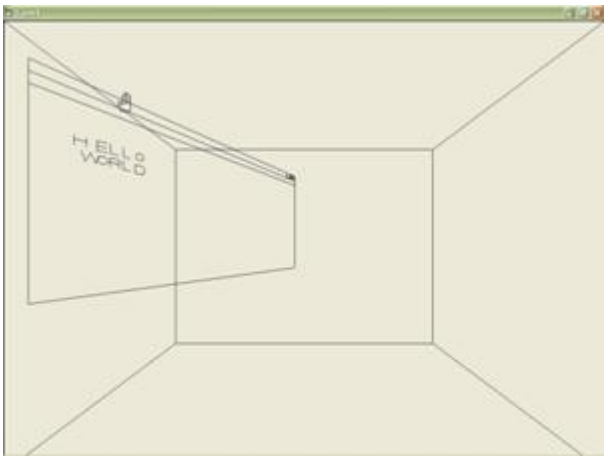


Figure 3.XY Rotation of Window & 3D Characters.

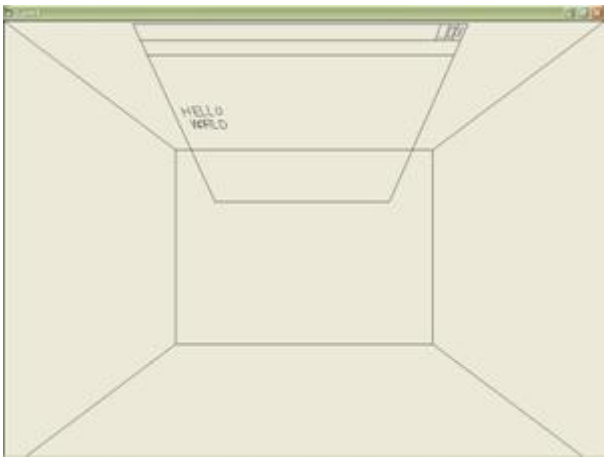


Figure 4.YZ Rotation of Window & 3D Characters

3.4 3D cursor : Virtual Hand

The fourth thrust of our 3D GUI prototyping work is exploring the design and usability of a 3D cursor controlled by a 3D input device. One of our 3D cursors is shown in figure 5 where it is being used to select a 3D window, Icon etc.

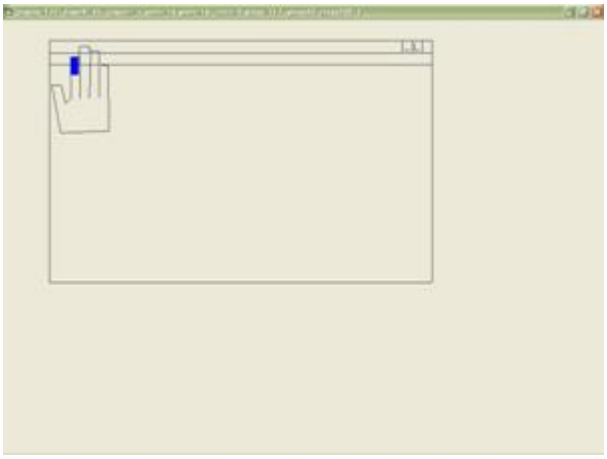


Figure 5 3D Cursor: Virtual Hand

Most interactive 3D graphics applications employ a 2D cursor and 2D pointing device for interaction. Given that the user is manipulating 3D objects or worlds in these applications, this gives rise to a fundamental mismatch in dimensionality. In the context of a 3D application the cursor ``floats" over the top of the objects rather than being part of the scene. We believe the use of a 3D cursor, introduced into the scene as an object in its own right, controlled by a 3D (six-degree-of-freedom input) device is a better approach for interactive 3D applications, including a 3D user interface, than the traditional approach. Recently, work in the area of 3D interaction has intensified. We believe this to be a significant area for research in its own right, with the potential for many important practical outcomes.

3.5 Icon, Device and Taskbar, Menu Manager

Icon manager will active the window by just selecting or clicking on particular icon. Taskbar manager will generate icons on the taskbar for active window.

3D Start menu is implemented with our own 3D characters to apply realistic view.

This four feature are visualized in the Figure 6 .

Figure 6. 3D start menu and sub menu

Device manager will start timer and captures the 8 values (ø1 to ø5 and x, y, and z) from p5 Data glove. It will maintain circular queue and generate average value, global variables are used for storing purpose.

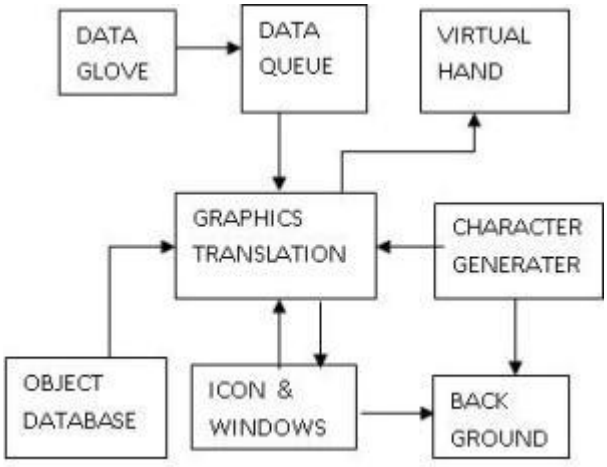


Figure 7 Actual Working and Interfacing

4. Interfacing with Interface

The most natural way for humans to manipulate their surroundings, including the windows e.g. on their desktop, is of course by using their hands. Hands are used to grab and move objects or manipulate them in other ways. Hands can be used to communicate with others and state intentions by making Postures or gestures. the most promising approach to minimize the cognitive load required for learning and using a user interface in a virtual environment is to employ a gesture recognition engine that lets the user interact with the application in a natural way by just utilizing his hands in ways he is already used to.[7]

4.1 Applied Hardware:

The glove hardware we used to realize our gesture recognition engine is a P5 Glove from Essential reality, shown in Figure. The P5 is a consumer data glove originally designed as a Game controller. It features five bend sensors to track the position of the wearer's fingers as well as an infrared-based optical tracking system, allowing computation of the glove's position and Orientation without the need for additional hardware. The P5 consists of a stationary base station housing the infrared receptors enabling the spatial tracking. The attainment of position and Orientation data is achieved with the help of reflectors mounted on prominent positions on the glove housing dependent on how many of these reflectors are visible for the base station and on which positions the visible reflectors are registered, the glove's driver is able to calculate the orientation and position of the glove.

Figure 6. Essential Reality P5 Data Glove

During our work with the P5, we learned that the calculated values for the flexion of the fingers were quite accurate, while the spatial tracking data was, as expected, much less reliable. The estimated position information was fairly dependable, whereas the values for yaw, pitch and roll of the glove were, dependent on lighting conditions, very unstable, with sudden jumps in the calculated data. Because of this, additional adequate filtering mechanisms had to be applied to ascertain sufficiently reliable values. Of special attention is the very low price of the P5. It costs about 50 €, by comparison to about 4000 € for a professional data glove, which of course provides much more accurate data but on the other side doesn't come with integrated and transportable position tracking. Indeed, the low price was one reason we chose the P5 for our gesture recognition, because it shows that serviceable interaction hardware for virtual environments can be realized at a cost that makes it an option for the normal consumer market. The other reason for our choice was to show that our recognition engine is powerful and flexible enough to enable reliable gesture recognition even when used with inexpensive gamer hardware.

4.2 Recognition Process

Our recognition engine consists of two components: the data acquisition and the gesture manager. The data acquisition runs as a separate thread and is constantly checking the received data from the glove for possible matches from the gesture manager. As mentioned before, position and especially orientation data received from the P5 can be very noisy, so they have to be appropriately filtered and smoothed out to enable a sufficiently reliable matching to the known postures.

First, the tracking data is piped through a deadband filter to reduce the chance of jumping error values in the tracked data.Alterations in the position or orientation data that exceed a given deadband limit are discarded as improbable and replacedwith their previous values to eliminate changes in position and orientation that can only be considered as erroneous calculation of the glove's position. The resulting data is then straightened out by a dynamically adjusting average filter. Depending on the variations of the acquired data, the size of the averaging values is altered within a defined range. If the data is fluctuating in a small region, the size of the filter is increased to compensate jittering data. If the values show larger changes, the filter size is reduced to reduce latency in the consequential position and orientation

Figure 8.P5 Gesture Trainer

The resulting data is reasonably correct enough to provide a good basis for the matching process of the gesture manager. To lower the possibility of misrecognition, a posture is only accredited as recognized when held for an adjustable minimum time span. During our tests it showed that values between 300 and 800 milliseconds are suitable to allow a reliable recognition without forcing the user to hold the posture for too long.[8]

5. System Evaluation

FEATURES	COMPIZ LINUX	AERO MICROSOFT VISTA	MAC OS X APPLE	GLASS PROJECT SUN MICRO	CUBIC EYE	MAW3	OUR SYSTEM
TUNNEL SHAPED DESKTOP	√	×	×	×	√	√	√
ROTATION AND TILTING OF WINDOWS	√	√	√	√	√	√	√
3D CHARACTERS	×	×	×	×	×	×	√
3D INPUT DEVICE	×	×	×	×	×	√	√
REALISTIC PERSPECTIVE PROJECTION OF 3D WIMP	×	×	×	×	×	×	√

6. Conclusions

In this paper, we suggested the 3D Desktop System which can help users of personal computers to work under more comfortable environments. Our system provides users with powerful 3D desktop which is handled by 3D input device .Our system is efficient and flexible in handling the numbers of task spaces and provides more smooth graphics by specially generated 3D characters and 3D Cursor.

References:

[1] José Pascual Molina, Pascual González, M.DoloresLozano, Francisco Montero and Víctor López-Jaquero Bridging the gap: developing 2D and 3D user interfaces with the IDEAS methodology

[2] Laura Dipietro, Angelo M. Sabatini, *Senior Member, IEEE*, and Paolo Dario, *Fellow, IEEE* A Survey of

[3] Glove-Based Systemsand Their Applications, JULY 2008

[4] giacomo Andreucci, 3D graphical User Interface on Personal Computer ,Aether:The journal of Media Geography,Spring 2008

[5] Sun Looking Glass Project home page. http://www.sun.com/software/looking_glass/

[6] All peculiarities of these interface are shot in hundred of videos on You Tube. For example see these searches: http://www.youtube.com/results?search_query=Linux+Compiz

[7] http:// BumpTop-Pro-3D-Desktop-BumpTop-Pro-Cubic-Eye_969571.html

[8] Andreas Dengel, Stefan A gne, Bertin Klein, Achim Ebert, Matthias Deller Human-Centered Interaction with Documents**,** *HCM'06*, October 27, 2006

[9] The P5 Glove.
URL:*www.essentialreality.com/VGA/video_game/P5.php*

[10] Seungpyo Hong, Donsu Lee, Sangjun Lee, Efficient and Flexible 3d virtual desktop system in Windows enviorment, Journal of Measurement Science and Instrumentation 2010

Dynamic Reputation Based Trust Management Using Neural Network Approach

Reza Azmi[1] Mahdieh Hakimi[2], and Zahra Bahmani[3]

[1] College of Engineering, Alzahra University
Tehran, Iran

[2] College of Engineering, Alzahra University
Tehran, Iran

[3] College of Engineering, Alzahra University
Tehran, Iran

Abstract

Multi-agent systems like Peer-to-Peer (P2P) Networks employ scalable mechanisms that allow anyone to offer content and services to other system users. The open accessibility of these networks makes them vulnerable to malicious users wishing to poison the system. This paper proposed a novel trust and reputation system, using RBF artificial neural network to determine trust level and mitigate the number of unreliable downloads.

Keywords: *Peer-to-Peer Network, Trust, Reputation, RBF Artificial Neural Network.*

1. Introduction

Agents in P2P Networks are anonymous and heterogeneous also there is no central authentication system. In these distributed systems flexibility and low participation cost encourages a much larger number of participants. Accessing to data and shared services in dynamic networks like P2P environments are related to trust and reputation of peers. Recently these systems are usually applied to file sharing and social networks. P2P networks tend to be more scalable, robust and adaptive than other forms of distributed systems. In this paper we proposed a novel dynamic trust model as one kind of decision support systems, using RBF artificial neural network to determine trust level and mitigate the number of unreliable downloads. Recommended trust models are applied broadly that help peers to download from reliable providers. They differ in selections of recommenders and in aggregations of recommendations.

2. Related Work

Various techniques have been proposed to secure P2P networks over the last decade. Abdul-Rahman[1] captured the most important characteristics of trust and reputation and proposed the general structure for developing trust and reputation in a distributed system. Most of the later works in this area followed their ideas, but in different application domain, such as [2, 3, 4, 5]. EigenTrust model of Kamvar[4] is built on the notion of transitive trust. A major issue of applying this model is to find pre-trusted peers that guarantee convergence of the algorithm and avoid malicious collectives. Wang[5] applied Naive Bayesian network to recommendation trust. The model can be used to solve the problem of different estimation process of the same online service. In [6, 7, 8, 9] proposed the trust and reputation system based on artificial neural network which are built on back-propagation algorithm to train the MLPs neural network. In proposed paper we used RBF neural network since these networks tend to learn much faster and require fewer training samples than MLPs.

3. Determination of Trust and Reputation Level by Neural Network

In this paper we focus on pure P2P networks for file exchange and, more precisely, on the Gnutella architecture because it is closest to the ideal structure of the P2P networks, where all participants have a uniform role.

3.1 Basic idea of proposed system

In a P2P overlay network like Gnutella, exchanging a file is containing two phases such as 1.Searching a file, 2.Downloading it. But some other researches like [2, 3, 4, 5] proposed this protocol by adding two other phases such as 3.Pooling and 4.Evaluating the votes. In this way before a requester decides about downloading the file from a provider, first it asks other peers about reputation of him/her.

In addition this paper studied about influence of layering concept on evaluating trust, for the first time and we use artificial neural network to achieve higher reliable trust and reputation about provider. In other word, as it's showed by literature review about trust and reputation systems, all the collected votes about reputation of a file provider from other peers, are in the same level. Whereas being more hops between responses to the requester causes more malicious responses and spoofing. In this paper we assume that peers in the first layer get to the requester through one hope, for the second layer there are two hopes and so on.
Since peers are heterogeneous, they may have different preferences and judge issues by different criteria. On the other hand in this paper same as [1] we consider different level of trust like Very Trustworthy, Trustworthy, Untrustworthy, Very Untrustworthy (see Table 1). We use these levels of trust to determine the output of neural network.

Table 1: Different level of Trust

Meaning	***Trust Level***
Very Trustworthy	VT
Trustworthy	T
Untrustworthy	U
Very Untrustworthy	VU

It is a difficult problem to predict the character of a client. Furthermore, because the open environments are dynamic, it is much more complicated to predict the distribution of clients with different characters on specific time. Therefore we use RBF neural network to evaluate other peer's recommendation. Artificial neural networks are robust to noise data and support incremental training. So our proposed trust model tries to overcome some disadvantages of previous trust systems by training the RBF neural network and tuning its weights by similarity matching to find the index of best center.
Before describing more details about application of neural network in our model, we need to explain about Gnutella phases as they presented in [3].

3.2 Phase 1: Resource Searching

At first peer **A** who is a peer that looking for a resource, broadcasts a Query indicating the resource it is looking for. Every other peers which receiving the query and willing to offer the requested resource for download, sends back a QueryHit message stating how it satisfies the query and providing its ID and its pair(IP, port), which peer **A** can use for downloading.

3.3 Phase 2: Polling

Upon reception of the QueryHit messages, peer **A** selects a top list of favorite peers T and polls its peers about the reputations of them. In the poll request, peer **A** includes IDs of peers in T about which it is enquiring and a public key PKpoll generated on the fly for the poll request, with which responses to the poll will need to be encrypted. The poll request is sent through the P2P network and therefore peer **A** does not need to disclose its ID or its IP to be able to receive back the response. Peers receiving the poll request and wishing to express an opinion on any of the peers in the list, send back a PollReply expressing their votes and declaring their (IP, port) pair. Peer A hash of the votes and pair (IP, port) is also added in order to allow peer **A** to check the integrity of the message. The PollReply is then encrypted with PKpoll to ensure its confidentiality (of both the vote and the voters) when in transit.

3.4 Phase 3: Vote Evaluation

As a result of the previous phase, peer **A** receives a set of votes, where, for each peer in T, some votes can express a good opinion while some others can express a bad opinion. To base its decision on the votes received, peer **A** needs to trust the reliability of the votes. Thus, peer **A** first uses the hash to detect tampered-with votes and discard them. Second, peer **A** detects votes that appear suspicious, for example since they are coming from IPs suspected of representing a clique. Third, peer **A** selects a set of voters that it directly contacts (by using the (IP, port) pair they provided) to check whether they actually expressed that vote. For each selected voter, peer **A** directly sends a TrueVote request reporting the votes it has received, and expects back a confirmation message TrueVoteReply. This forces potential malicious peers to pay the cost of using real IPs as false witnesses.

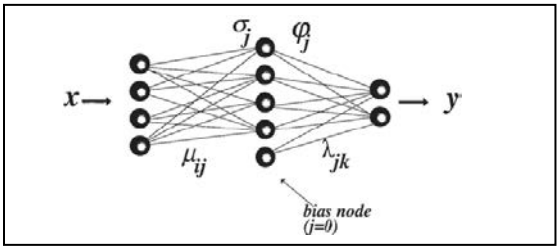
3.5 Application of RBF neural network in proposed trust model

When a peer's looking for a resource, broadcasts a query. Every other peers which willing to offer the requested resource, sends back a message. Then the requester polls about the reputations of the provider. After evaluating received recommendations or votes, finally the requester has to select the provider who seems to be the best on the list, in different aspects like download speed, file type and file quality. In this point we use Radial-Basis Function (RBF) artificial neural network to solve the problem (see Figure 1). RBF is a single-hidden-layer feed forward network with linear output transfer functions and nonlinear transfer functions, on the hidden layer nodes. RBF networks provide a powerful alternative to Multi-Layer Perception (MLPs) for function approximation or classification. They train faster and require fewer training samples than MLPs.

There are several techniques for training these networks. RBF networks are nonlinear hybrid networks typically containing a single hidden layer of processing elements. This layer uses Gaussian transfer functions, rather than the standard sigmoid functions employed by MLPs. The primary adjustable parameters in Figure 1 are the final layer weights, $\{\lambda_{jk}\}$, connecting the jth hidden node to the kth output node. There are also weights $\{\mu_{ij}\}$ connecting the ith input node with the jth hidden node [10].

The mathematical embodiment of the RBF takes the following form. The kth component of the output vector y_p corresponding to the pth input pattern x_p is expressed as:

$$[y(x_p)]_k = \sum_{j=0}^{h} \lambda_{jk}\, \phi_j\left(\|x_p - \mu_j\|; \Sigma_j\right) \quad (1)$$



Where $\phi_j(\ldots)$ denotes the nonlinear transfer function of hidden node j. In the RBF neural network there are three basic parameters like 1- centers, 2- spreads, 3- weights.

Fig. 1 The basic radial basis function structure[10].

We use K-means clustering algorithm for determining centers:

1. Initialization: random $\mu_j(t=0)$

2. Sampling: draw $x_p(t)$ from input space

3. Similarity matching: find index of best center
$k = \arg min_j \|x_p(t) - \mu_j(t)\|$ (2)

4. Updating: adjust centers
$\mu_k(t+1) = \mu_k(t) + \eta * [x_p(t) - \mu_k(t)]$ (3)

5. Continuation: increment t by 1, go to 2 and continue until no noticeable changes of centers occurred.

Next we use normalizing method to find spreads:
$$\sigma = \frac{Maximum\ \ distance\ \ between\ \ any\ \ 2\ centers}{\sqrt{number\ \ of\ \ centers}} = \frac{d_{max}}{\sqrt{m_1}} \quad (4)$$

$$\varphi_i(\|x - t_i\|^2) = exp\left(-\frac{m_1}{d_{max}^2}\|x - t_i\|^2\right)$$

$$i \in [1, m_1] \quad (5)$$

At last using LMS method for tuning weights:
they are part of a sentence, as in

$$\lambda_{jk} = \lambda_{jk} + \eta * [y(k) - y'(k)]$$

$$0 < \eta < 1, \quad (6)$$

y′ is predicted output of network and y is actual output of network.

$$y'(k) = \begin{cases} 1 & if\ confidence\ condition\ is\ true \\ 0 & otherwise \end{cases} \quad (7)$$

, y(k) = output of function φ from [0,1] (8)

Generally inputs of this RBF neural network are other peer's recommendations about a file provider based on their assigned layer. As mentioned before these recommendations are in VT, T, U, VU format. Thus total number of input neurons of this neural network will be equals to product of network TTL in these 6 trust levels. In order to normalization of inputs of trust levels, we divide number of received recommendations from each trust level into the all number of them. In RBF network hidden layer's neurons do the similarity matching and find index of best center. As previous terms express weights of hidden layer's neurons calculate by K-means clustering algorithm. Finally neuron's weights of output layer are difference between predicted output of network and actual output. The requester can finally download the resource and, depending on its satisfaction for the download, update its reputation information for the provider. Every peer keeps the last trained neural network in its memory. After each interaction, the neural network could be trained by requester's experiment and tuned network's weights. If the neural network was built beyond a certain period of time or some recommenders have changed their trust models, or the requester changes his trust estimation accuracy requirement, the requester collects up-to-date trust data and retrains the neural network. After a while the network is well trained and it would help peers to find reliable file providers.

3.6 Phase 4: Resource downloading

Peer A can finally download the resource and, depending on its satisfaction for the download, update its reputation information for peer **B**.

4. Experiments

We evaluate our system in a simulation of a peer-to-peer network with implementation of the trust computation model in RBF neural network developed with Matrics programming language on the Matlab.

4.1 Simulation setup

Our simulation involves 20 peers with 2 very trusted peers, 10 trusted peers, 6 untrusted peers and 2 very untrusted peers (malicious peers), and with a random topology of these nodes. For implementation of RBF neural network we assumed 20 nodes, each node has 4 states of pooling like VT, T, U and VU, so we had 4^{20} number of sampling spaces. First we made 1,000 samples randomly and applied K-means clustering algorithm for determining centers. Next the RBF neural network has been made up of 100 centers and it has been trained for 20,000 iterations.

4.2 Simulation results

The goal of this simulation is to see whether the trust and reputation system based on RBF neural network will help peers with different characters make accurate decisions and decrease the number of unreliable downloads. Thus we compare the percent of accurate suggestions for each training steps. In figure 2 we can see that the curve of *accurate suggestions* is increasing during the training steps. It means that output of the neural network finds the trust level of the file provider. It's been showed in figure 2 that training of the RBF neural network starts with 63 % of accurate suggestions in first step and then it's received to 89.9 % of accurate suggestions.

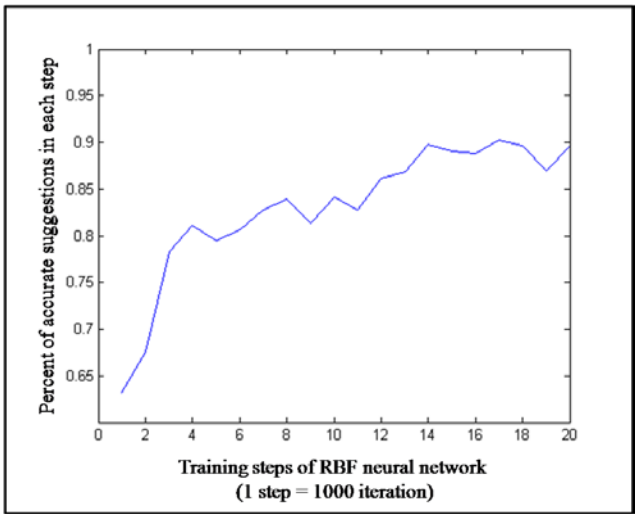


Fig. 2 Simulation result with 2 malicious in the network

4.3 Behavior analysis

In order to see the impact of increasing malicious peers in P2P network on output of reputation system that is based on RBF neural network, we have continued our simulation experiments by a P2P network with 2, 4 and 8 malicious peers in each time. Figure 2 is about the network with 2 malicious peers and in Figure 3 we can see the output of RBF trust model, while number of malicious peers has been duplicated. As shown in figure 3, training of the RBF neural network starts with 61% of accurate suggestions in first step and finally it reaches to 88%. Similarly, figure 4 with 8 malicious peers in P2P network indicates that rate of accurate suggestions from 66% reaches to 92% .

By comparing the result of simulation experiments which mentioned above, we can conclude that despite of increasing malicious peers in P2P network, there is no significant impact on training of the RBF neural network.

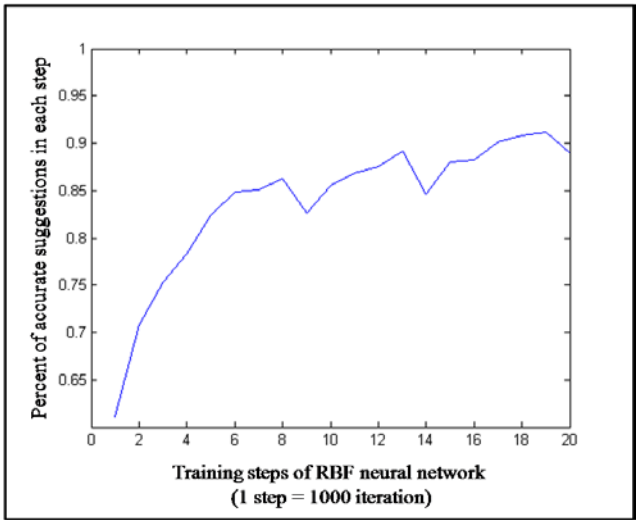


Fig. 3 Simulation result with 4 malicious in the network

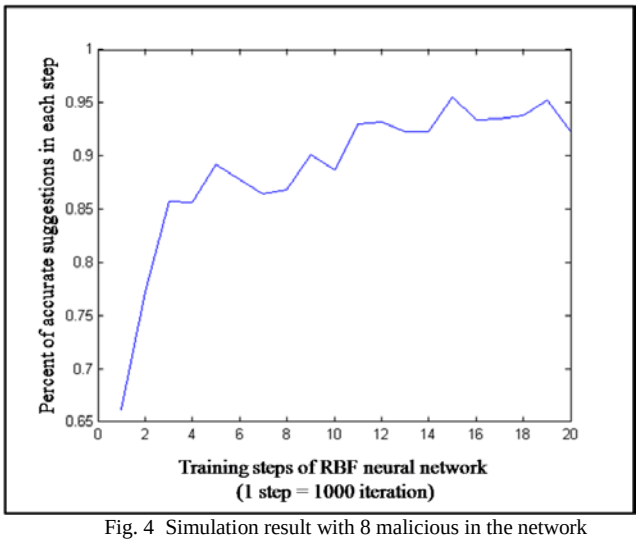


Fig. 4 Simulation result with 8 malicious in the network

5. Conclusions

In this paper we decide to use the RBF neural-network based recommendation trust model since experiments in this area have

discovered that hidden variables capture by hidden layers of the model, and these networks are robust to noise in the training data, also they have fast speed with high accuracy. In addition adaptability and non-linear aggregation of heterogeneous agent's recommendations are other properties of RBF based trust system.

References

[1] A. Abdul-Rahman, S. Hailes, "Supporting trust in virtual communities", In Proceedings of the Hawai'i International Conference on System Sciences, Maui, Hawaii, Jan 4-7 2000.

[2] F. Cornelli, E. Damiani, "Implementing a Reputation-Aware Gnutella Servent". In Proc. International Workshop on Peer-to-Peer Computing, Pisa, Italy, May 24, 2002.

[3] E. Damiani, S. D. C. di Vimercati, S. Paraboschi, P. Samarati, "Managing and sharing servents' reputations in p2p systems", IEEE Transactions on Knowledge and Data Engineering, 15(4), 840-854, 2003.

[4] S. Kamvar, M. T. Schlosser, and H. Garcia-Molina." The EigenTrust Algorithm for Reputation Management in P2P Networks". In Proc.12th International World Wide Web Conference (WWW'03), May, 2003.

[5] Y. Wang, J. Vassileva, " Trust and Reputation Model in Peer-to-Peer Networks", Proc. IEEE Conference on P2P Computing, Linkoeping, Sweden, 2003.

[6] W. Song, V. Phoha, X. Xu, "An Adaptive Recommendation Trust Model in Multiagent System", IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'04).

[7] B. Zong, F. Xu, J. Jiao, J. Lv , "A Broker-Assisting Trust and Reputation System Based on Artificial Neural Network", In IEEE 978-1-4244-2794-9/09/ 2009.

[8] Baohua H., Heping H., Zhengding L.,"Identifying Local Trust Value with Neural Network in P2P Environment". IEEE 0-7803-9179-9/05, 2005.

[9] Fuke Sh., Pan Ch., Xiaoli R.,"Research of P2P Traffic Identification Based on BP Neural Network". in Proc. 3rd Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing(IIHMSP), Taiwan, Vol.2, pp. 75-78, 2007.

[10] D. Lowe, "Radial basis function networks and statistics, in Statistics and Neural Networks: Advances at the Interface", New York: Oxford University Press, pp. 65–95,1999.

Farsi/Arabic Document Image Retrieval through Sub -Letter Shape Coding for mixed Farsi/Arabic and English text

Zahra bahmani [1], Reza Azmi[2]

[1,2] Computer department
Alzahra University
Tehran, Iran

Abstract

A retrieval method for explicit recognition free Farsi/Arabic document is proposed in this paper. The system can be used in mixed Farsi/Arabic and English text. The method consists of Preprocessing, word and sub_word extraction, detection and cancelation of sub_letter connectors, annotation sub_letters by shape coding, classifier of sub_letters by use of decision tree and using of RBF neural network for sub_letter recognition. The Proposed system retrieves document images by a new sub_letter shape coding scheme in Farsi/Arabic documents. In this method document content captures through sub_letter coding of words. The decision tree-based classifier partitions the sub_letters space into a number of sub regions by splitting the sub_letter space, using one topological shape features at a time. Topological shape Features include height, width, holes, openings, valleys, jags, sub_letter ascenders/descanters. Experimental results show advantages of this method in Farsi/Arabic Document Image Retrieval.

Keywords: shape **code, sub-word, sub-letter, RBF neural network.**

1. Introduction

By Incremental use of digital libraries and the promise of paperless offices, large amount of document images are scanned and archived. Also Modern Technology has made it possible to produce, process, transmit and store digital images efficiently.

Although, the image processing technology can be used for automatic conversion of digital document images to readable text format by computer using optical recognition of characters (OCR), but this method is not practical and optimal for using in a huge volume of documents. Moreover, OCR for Persian texts still has some drawbacks. Thus new content based image retrieval techniques are required in Farsi printed documents.

Most of retrieval and recognition methods are divided into two category[1]:

The first category methods retrieval and recognation document images based on description of global shape of words or sub-words. In this method the descriptor are directly extracted from the image of the word or sub-word[1,3,4,9,10,15].

The second category methods segment a word to its letters and then extract features from the image of each letter. These features constitute a word descriptor[2,14,17].

In addition of such problems as noise, variety of fonts and size, in the segmentation based methods (especially in Farsi and cursive writing) there is the problem of segmentation. Because of the variety of shape, size and length of letters, estimation of segmentaion points of the letters is done erroneously.

In this paper we used a new segmentation method based on detection of sub-letter connectors and cancellation of them [1]. This method is free from explicit segmentation point detection. For this purpose, the connectors of main elements of the letter which have been named sub-letters, is detected. Sub-letters are exteracted by detection and cancelation of their connectors. Then sub-letters are classifier by using of their topological shape features including their height, width, holes, openings, valleys, jags, sub_letter ascenders/descanters and position of sub-letter to the base line.RBF neural network is used for recognition of sub_letters. Finally, sub-letters are encoded according to defined dictionary.

The remainder of this paper is organized as follows: secrtin2 reviews related works, section3 describe abaout Farsi script .Section 4, presents proposed system briefly. In section 5, preprocessing phase is discussed, in section 6 processing phase illustrate. Experimental results represent in section 7. Section 8 concludes the paper.

2. Review related works

In recent years, there has been much interest in the research area of Farsi Document Image and handwriting recognition [2, 3 and 4]. Some works has been done in the field of Farsi Document Image retrieval. Ebrahimi proposed a method in Persian document images recognition and retrieval. This method uses the whole shape of sub-word and sub-word's body by local features. These features are became robust against noise. Principal

Components Analysis (PCA) has been used to dimension reduction in feature space. In this proposed system, the sub-word images are clustered with k_means algorithm by using Euclidean distance for search space reduction. Two evaluation indices are used for determining the appropriate number of clusters. Moreover for clustering evaluation, the qualitative analysis has been done using classification of a test series to the centers of the clusters.

Akbari and Azmi [5] had introduced a recognition free method for Persian document images retrieval. In this method, first, upper contours of sub-words has been extracted, then a pictorial dictionary has been developed based on this featur. Document images can be retrieved by either query keywords or a query document image based on their content similarity. In another paper, Azmi and Habibi represent content based document image retrieval with support vectors clustering. In this proposed approach, sub-words feature vectors are extracted with wavelet transform and clustered with support vector clustering (SVC) algorithm.

In [6], a classification and retrival system was introducted for document images baseed on the column structure, the size of font, the density of text of regions and statistical features of continuous components of regions. This system uses this feature for document images classification and retrieval based on the visual similarity of the layout structure.

In our earlier work [21], we have proposed a Novel method for Recognition free Farsi document retrieval. In this method, the retrieval is done through recognition of sub-letters and other elements of letters such as dots and some signs like Sarkesh. Novel algorithmic technologies are used for pure Farsi/Arabic text documents. Pure text means those documents that do not include images or formulas and those in which a single language is used.

Many works had been done in English documents images retrieval [7, 8]. One of them is presented by Jilin Li, and et al [19]. They present Document Image Retrieval system with Local Feature Sequences. This paper proposed a fast, accurate and OCR-free image retrieval algorithm using local feature sequences which can describe the intrinsic, unique and page-layout-free characteristics of document images. It well handles the challenges including low resolution, different language, rotation and incompleteness and N-up.

Shijian Lu, and et al are perposed [9] a document retrieval technique that is capable of searching document images without optical character recognition (OCR). The proposed technique retrieves document images by a new word shape coding scheme, which captures the document content through annotating each word image by a word shape code. In particular, thay annotate word images by using a set of topological shape features including character ascenders/descenders, character holes, and character water reservoirs. With the annotated word shape codes, document images can be retrieved by either query keywords or a query document image.

Meshesha and Jawahar[10] described an effective word image matching scheme in their paper that achieved high performance in the presence of script variability, printing variation, degradation and word-form variants. A novel partial matching algorithm is designed for morphological matching of word form variants in a language. They formulated feature extraction scheme that extracted local features by scanning vertical strips of the word image and combining them automatically based on their discriminatory potential. They presented detailed performance analysis of the proposed approach on English, Amharic and Hindi documents.

In [11] a method was developed for automatically selecting sentences and key phrases to create a summary from an imaged document without any need for recognition of the characters in each word. In this method built word equivalence classes by using a rank blur hit-miss transform to compare word images and using a statistical classifier to determine the likelihood of each sentence being a summary sentence. Other works that used of character shape codes to document image retrieval are [12, 13, 14 and 19].

3. Farsi/Arabic Text Properties

3.1 **Farsi and Arabic Alphabets**

The Farsi alphabet has four more letters than the Arabic Alphabet which has 32 letters; whereas, the Arabic alphabet has only 28. In both Farsi and Arabic Alphabets there are several letters that share the same basic form and differ only by a small complementary part. The complementary part could be a dot, a group of dots or a slanted bar. It can lie above, below or inside the letter. In fact, all the letters in the Farsi alphabet are derived from 18 basic shapes.

3.2 **Cursiveness of the Words and Connection of the Letters**

The Arabic scripts and all of its derived forms (including Farsi Script) are inherently cursive. The letters connected to each other form a sub word. the position of each letter in the word and its preceding or following letter in the same word (if there is any), are the factors that determine the shapes of the letter. In the Farsi alphabet, similar to the Arabic alphabet, a letter can appear in four different forms according to their positions in a sub word, beginning, middle, end and single. All Farsi letters (with seven exceptions / six in Arabic) can be connected to other letters from both the right and the left sides. The seven exceptional characters can only be connected to other letters from the right side. Therefore, if any of those seven letters appear in the middle of a word, there will be a gap in connectivity. [3, 22, 21]

In the structure of the Farsi script, letters and their sub-letters are joined together by the connectors which are being along the base line. For examples, the letter "س" is written in form of " سـ "at the beginning and in the middle of the sub-word and "س" at the end of the sub-word and single state. The sub-letter for the first case consists of three jags ( ) and in the second case, it consists of two jags and one pit under the base line ( ). The letter "ی" is written in form " یـ "at the beginning and in the middle of the sub-word and "ی" at the end of the sub-word and single state. The first case has only one sub-letter that consists one jag and second case has also one sub-letter consists a hollow and one pit under base line.

4. OVERVIEW OF PROPOSED SYSTEM

Proposed system consists of 2 main phases, preprocessing and processing. Preprocessing includes binarization, text line and word extraction and overall base line detection. In processing phase each word is divided to its sub-words. Then sub-letters connectors are detected and removed from their whole body. Thus, only sub-letters remain. In the next step, Shape features of sub_letters are extracted and sub_letter annotated by its shape codes. Sub-letters space divided into a number of sub regions by use of the decision tree-based (DT) classifier and shape features. RBF are used for final recognition of sub_letters.at last sub_words are coded by their sub_letters code and this code can be used for document image content retrieval. Over view of system is shown in fig.1.

5. Pre-processing

Pre-processing phase has three steps.

Binarization: In the first step, colored and grayscale scanned document images converted to white and black ones and become binary.

Text lines and words extracted: text lines and words are extracted using blank space between lines and words. Division point between two words computed through vertical projection.

Overall base line detection: For each line, one overall base line is detected. The overall base line is detected according to this attribute that letters and their sub-letters are connected together by use of the connectors, which are being along the base line. Consequently, number of black dots along the base line is more than that of other horizontal line. Base line position is computed through horizontal projection of each line.

Noise cancelation: noises which are in the distance between lines and words are removed.

6. Processing

In this phase, sub_words are extracted from words and sub_letters are extracted from sub_words. Extracted sub_letters are given to DT. The DT classifier partitions sub_letters space to sub region by topological shape feature. Final recognition of sub_letters is done by RBF. At last all sub_words of each word are coded by use of their sub_letters cod and whole document is coded.

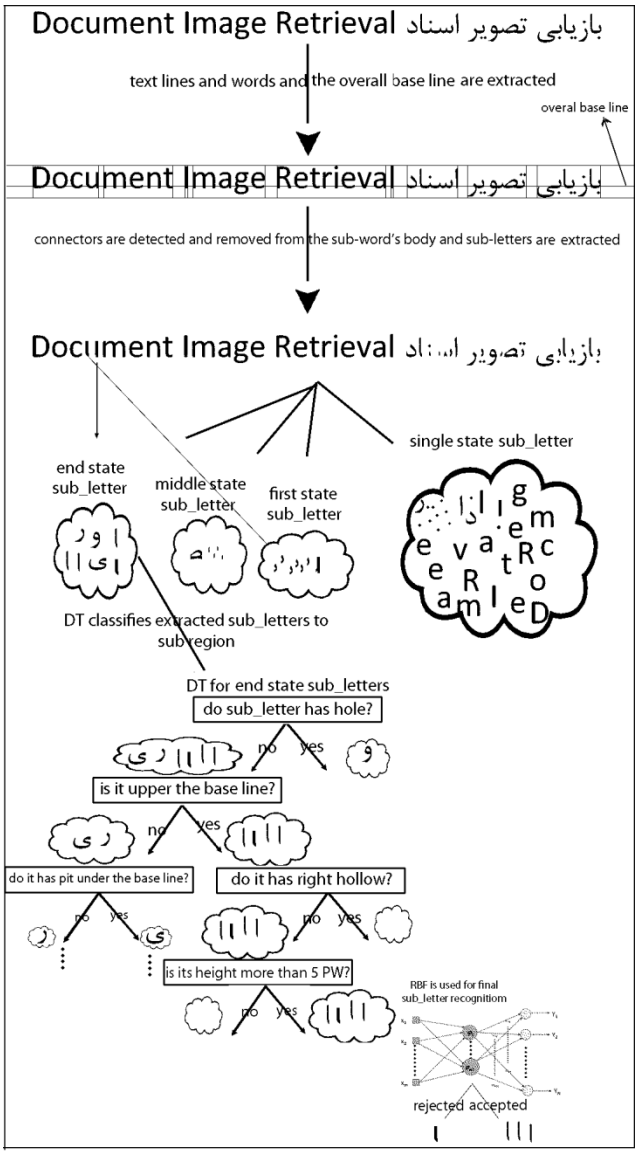


Fig. 1 over view of system

6-1. sub-letters Extraction

In considering the fact that documents will be rotated somewhat at the printing and scanning time, the overall base line usually has some error. Therefore in each word a local base line is computed through horizontal projection of word. Then, for each word, sub-words, sign and points are detected by component labeling. For sub-words processing, first connectors are detected and removed from the sub-word's body and sub-letters extracted.

6-2.topological shape feature extraction.

We used of topological shape feature for classifier of sub_letters. Features include height and width of sub_letters, holes, height and width of hole, openings, valleys, jags, sub_letter ascenders/descanters, holow and position of sub-letter to its local base line.Sub_letters are annotated by shape codes and are sended to DT classifire.

6-3.DT classifier

DT is used for classifier extracted sub_letters to sub regions. The decision tree-based classifier partitions the feature space into a number of sub regions by splitting the feature space, using one feature at a time. The regions are split until each sub region contains sub_letter that have same topological shape features. For example of single state sub_letters, letter a"ا" has only one sub_letter and its topological shape feature are height and width of sub_letters. one sub_letter is classified to sub region of "ا" if has height that is minimum six times the pen width(PW) and maximum ten and a half times PW and has width that is minimum ¾ of the PW and maximum two times the PW. According to result of experiments on mixed Farsi and English text, this error is possible occur that number "1" be classified to the class of"ا". The letter "ص" has two sub letters "ـصـ" and "ـص". The sub_letter "ـصـ" is member of middle state sub_letters and has two shape features, being upper base line (at least 2/3 of sub_letter is upper base line) and having a hole with minimum width three times of the with pen. The sub_letter "ـص" has two shape features, being under base line (at least 2/3 of sub_letter is not upper base line) and having a pit. Experimental result shows possibility of binge the sub_letter e"ی" in region of sub_lettter "ـص".

6-4. Sub-letters recognition

For final recognition of sub_letters, we used of RBF neural network and profile feature.

6.4.1 Feature extraction

Feature extraction performs in three stages

1. Image resizing: RBF recognize member of each sub region. It is fix number nodes of input layer of RBF therefore size of extracted feature vector must be fix. Then we have to do feature equalization and resize all sub_letter images of one sub region to middle size of them.

2. Profile extraction: used features are extracted profile of sub_letter. Extracted profiles for each sub_lettr image is in four directions (Up - down -left and right). These features have had reliable results in the field of printed document retrieval [23].

3. using of one dimensional Discrete Wavelet Transform (DWT): After feature extraction, the produced vector is applied to the one dimentional DWT and the output is a vector with half length of initial vector. By using DWT the system are benefited in two ways. First, it reduces the feature vector dimension. Second, it eliminates the diagram shaking caused by the jagged edge of image word during the scanning and binarization procedure.

6.4.2 NEURAL NETWORK

We used of RBF neural network in Document Image Retrieval system. The architecture of this NN is illustrated in Fig.2.

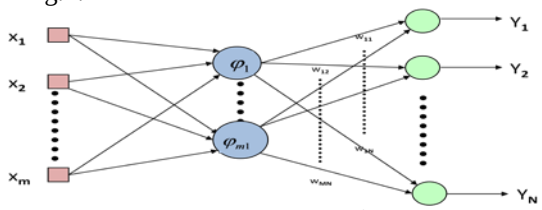


Fig. 1 Architecture of RBF NN

Architecture of RBF NEURAL NETWORK:

The RBF neural network consists of three layers, input layer, competitive layer and one output layer. Each neuron from one layer to next layer is full connected and assigned a weight to each connection.

Input layer: the extracted features are inserted into this layer.

Competitive layer: By this layer we apply a non-linear transformation from the input space to the competitive space.

Output Layer: By this layer we apply a linear transformation from the competitive space to the output space.

Training: Training is done in two phases: Unsupervised learning and supervised learning. In unsupervised learning phase the parameters of competitive layer of RBF NN are adjusted. In this stage there are 3 different strategies: first one is random selection of clusters center from data samples. Second one is use of a supervised learning method based on gradient descent that is more general case of LMS algorithm and by that minimizes total of squares of difference input and output and the third one is the using of clustering algorithms. After clustering, data are divided into special area of data space. Hence, the parameters of RBF are easily obtained from the position of centers and the distribution of clusters. Having found the parameters of RBF competitive layer, it means the number and position of centers, the weights of output layer is found by using a supervised methods like Delta rule [20].

In this research, for unsupervised learning of competitive layer, we use K_Means clustering algorithms. One supervised learning method is used to find the weights of output layer. In this, first the values of weights are set randomly and then the training samples are applied to the NN and the weights are updated according to the generated output.

Weights updating: each of the output layer units are equivalent to one of training word sets (which is one name in this case).thus, 30 training words sets produce 30 output layer units. The output value of units is computed with (1):

$$output(k) = \sum_{i=1}^{m} \varphi_{i} * w_{ik} \begin{cases} 1 & if \ \ output > 0 \\ 0 & if \ \ output \le 0 \end{cases} \quad (1)$$

Where φ_i the output of i[th] unit of competitive layer is calculated using (3) and W_{ik} is the Weight of the connection from the i[th] unit of competitive layer to the k[th] unit of output layer. Each training sample should activate its corresponding output. It means that the value of the function of the output layer unit corresponding to label of the training sample must be '1' and the values of other output layer units function must be '0'. Therefore, Weight s according to the current output and the target output will be updated by LMS rule:

$$\Delta w_i = w_i + \varepsilon \varphi_i (O_i - T_i) \quad (2).$$

Where, ε is training rate, φ_i is the output of competitive layer, O_i is the computed value of the i[th] unit, of output layer, T_i is the target output corresponding to the training sample labels. According to (2) the only weights whose corresponding outputs are different from target output are updated.

$$\varphi_i = \frac{P}{D_i + C} \quad (3).$$

Where, P and C are constant values and D_i is the distance between center of i[th] unit of competitive layer and observed sample.

Distance Function: In the proposed system, the distance between each competitive layer and training sample is computed based on a warping function.

6.5 DEFINED DICTIONARY

We defined a dictionary which includes all extracted sub-letters of sub-words. Dictionary has 4 entry classes, beginning, middle, end and single sub-letter entry. Extracted sub-letter of sub-word is shown in Tab1, 2.

7. EXPERIMENTAL RESULTS

To evaluate the proposed system we scanned 50 pages with five prevalent Farsi Font nazanin, b zar,b lotus,b mitra,b yaghot and in size 14. Percent of correct sub-letter recognition is shown in four diagrams in fig.3.

8. CONCLUSION

In this paper a retrieval method for explicit recognition free Farsi/Arabic document proposed. Proposed system retrieved the document images by a new sub_letter shape coding scheme in Farsi/Arabic documents. In the method document content captures through sub_letter coding of words. The decision tree-based classifier used for division the sub_letters space into the sub regions by splitting the sub_letter space, using topological shape features. Finally we used of RBF neural network for sub_letter recognition and coded words by their sub_word codes.

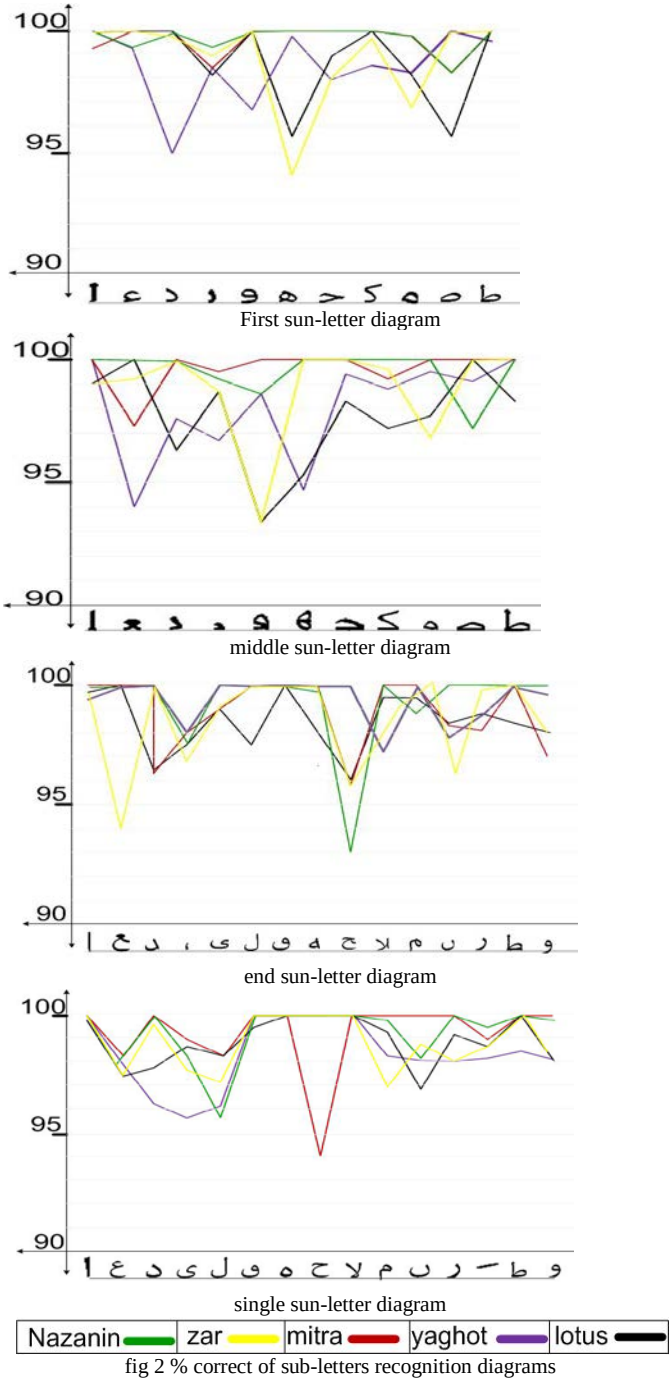


fig 2 % correct of sub-letters recognition diagrams

Table 1: letters, sub-letters and their features

num	Single	Sub Letter	Star	Sub Letter	Middle	Sub Letter	End	Sub Letter
1	آ or ا	ا	آ or ا	ا	. . . . .		ا	ا
2	ب	،،	بـ	ـ	ـبـ	ـ	ـب	،،
3	پ	،،	پـ	ـ	ـپـ	ـ	ـپ	،،
4	ت	،،	تـ	ـ	ـتـ	ـ	ـت	،،
5	ث	،،	ثـ	ـ	ـثـ	ـ	ـث	،،
6	ج	ح	جـ	ح	ـجـ	ح	ـج	ح
7	چ	ح	چـ	ح	ـچـ	ح	ـچ	ح
8	ح	ح	حـ	ح	ـحـ	ح	ـح	ح
9	خ	ح	خـ	ح	ـخـ	ح	ـخ	ح
10	د	د	.....		.....		ـد	،،
11	ذ	د	.....		.....		ـذ	،،
12	ر	ر	.....		.....		ـر	ر
13	ز	ر	.....		.....		ـز	ر
14	ژ	ر	.....		.....		ـژ	ر
15	س	ں،،	سـ	،،،	ـسـ	،،،	ـس	ں،،
16	ش	ں،،	شـ	،،،	ـشـ	،،،	ـش	ں،،

Table 2: letters, sub-letters and their features

num	Single	Sub Letter	Star	Sub Letter	Middle	Sub Letter	End	Sub Letter
17	ص	ں،	صـ	،،	ـصـ	،،	ـص	ں،
18	ض	ں،	ضـ	،،	ـضـ	،،	ـض	ں،
19	ط	ط	طـ	ط	ـطـ	ط	ـط	ط
20	ظ	ط	ظـ	ط	ـظـ	ط	ـظ	ط
21	ع	ع	عـ	ع	ـعـ	ع	ـع	ع
22	غ	ع	غـ	ع	ـغـ	ع	ـغ	ع
23	ف	و،،	فـ	و	ـفـ	ـ	ـف	،،
24	ق	ف	قـ	و	ـقـ	ـ	ـق	ف
25	ک	ک،،	کـ	ک	ـکـ	ک	ـک	ک،،
26	گ	گ،،	گـ	ک،	ـگـ	ک،	ـگ	ک،،
27	ل	ل	لـ	ا	ـلـ	ا	ـل	ل
28	م	م	مـ	ه	ـمـ	ه	ـم	م
29	ن	ں	نـ	ـ	ـنـ	ـ	ـن	ں
30	و	و					ـو	و
31	ه	ه	هـ	ه	ـهـ	ه	ـه	ه
32	ی	ی	یـ	ـ	ـیـ	ـ	ـی	ی

References

[1].Ebrahimi. A. Using printed word shape in document image retrieval and Farsi text recognition. PhD Thesis, Tarbiat Modarres University,Tehran, Iran. 2005

[2]. Hossein Khosravi ; Ehsanollah Kabir. A blackboard approach towards integrated Farsi OCR system. IJDAR, springer,2009.

[3]. Afshin Ebrahimi; Ehsanollah Kabir. A pictorial dictionary for printed Farsi sub words. Pattern Recognition Letters 29 (2008) 656–663, Elsevier, 2007.

[4]. Zahra bahmani; Reza Azmi; Fatemh Alamdar; Saman Haratizadeh. Off-Line Arabic/Farsi Handwritten Word Recognition Using RBF Neural Network and Genetic algorithm. 2th International Conference on Intelligent Computing and Intelligent Systems .ieee.2010

[5]. Reza Azmi; Mohammad Akbari; Hossain Akbari; Hossain Shirazi. LGL-DIR: Layout Graph for Layout based Document Image
Retrieval.2nd International Conference on Education Technology and Computer (ICETC). 2010, pp:262-266.

[6]. Reza Azmi; Mohammad Akbari. Document images classification and retrieval base on visual similarity. 11nd International CSI computer conference(csicc),2006.

[7] Doermann. D. The indexing and retrieval of document images: A survey. Computer Vision and Image Understanding, 1998.

[8] . Manesh B; Kokare; M.S.Shirdhonkar. Document Image Retrieval: An Overview. International Journal of Computer Applications (0975 – 8887). 2010, Vol 1, No. 7, pp: 114-119.

[9]. Shijian L; Linlin Li; and Chew Lim Tan. Document Image Retrieval through Word Shape Coding. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 30, NO. 11, NOVEMBER 2008

[10]. Million Meshesha ;C. V. Jawahar. Matching word images for content-based retrieval from printed document images. International Institute of Information Technology, Springer-Verlag, 2008.

[11]. Chen. F. R; Bloomberg. D. S.Summarization of imaged documents without OCR. Computer Vision and Image Understanding. 1998, pp: 307–319.

[12].Smeaton A. F; Spitz A. L.Using character shape coding for information Retrieval. 4th International Conference on Document Analysis and Recognition. 1997, pp 974–978.

[13] Spitz A. L; Maghbouleh A.Text categorization using character shape Codes. SPIE Symposium on Electronic Imaging Science and Technology. 1999, pp 174–181.

[14] Spitz, A. L.Using character shape codes for word spotting in document images. Shape, Structure and Pattern Recognition. 1995, pp 382–389.

[15]. Ramin Mehran ;Hamed Pirsiavash; Farbod Razzazi. A Front-end OCR for Omni-font Persian/Arabic Cursive Printed Documents, Proceedings of the Digital Imaging Computing: Techniques and Applications (DICTA).IEEE computer society. 2005.

[16]. Arundhati Tarafdar; Ranju Mondal; Srikanta Pal; Umapada Pal Fumitaka Kimura. Shape Code based Word-image Matching for Retrieval of Indian Multi-lingual Documents. International Conference on Pattern Recognition, IEEE computer society.2010

[17]. Shahab Ensafi; Mohammad Eshghi; Mahsa Naseri. Recognition of Separate and Adjoint PersianLetters Using Primitives. IEEE Symposium on Industrial Electronics and Applications (ISIEA). October 2009.
[18]. R. Azmi; E. Kabir. A new segmentation technique for omnifont Farsi text. Pattern Recognition Letters 22 (2001) 97±104. Elsevier Science,2001
[19] T. Nakayama, "Modeling Content Identification from Document Images," Proc. Fourth Conf. Applied Natural Language Processing (ANLP '94), pp. 22-27,1994.
[20]. Mohammad S. Mohammadi,Unsupervised RBF Neural Network Training Using Genetic Algorithms,9 th Iranian Electrical Stdudent Conference , 2006.(In Farsi).
[21]. Zahra bahmani; Reza Azmi. Farsi/Arabic Document Image Retrieval Through Sub –Letter Shape Coding. To be appear, International Conference on Networks and Information ICNI Chengdu, China. November 25-27, 2011
[22]. J. Sadr; S. Izadi; F. Solimanpour. STATE-OF-THE-ART IN FARSI SCRIPT RECOGNITION Invited Paper. Center for Pattern Recognition and Machine Intelligence CENPARMI. IEEE. 2007
[23]. S.Abirami, D.Manjula, "Profile Based Information Retrieval from Printed Document Images", Computer Graphics, Imaging and Visualization (CGIV 2007).

A Subnet Based Intrusion Detection Scheme for Tracking down the Origin of Man-In-The-Middle Attack

S.Vidya[1] and R.Bhaskaran[2]

[1] Department of Computer Science, Fatima College
Madurai, Tamil Nadu, India

[2] School of Mathematics, Madurai Kamaraj University
Madurai, Tamil Nadu, India

Abstract

The Address Resolution Protocol (ARP), has proved to work well under regular circumstances, but it is not equipped to cope with malicious hosts. Several methods to mitigate, detect and prevent these attacks do exist for the gateways/routers and nodes. This work is focused towards developing our own tailor made Intrusion Detection technique at the subnet level and we present an algorithm that detects the source of ARP poisoning in the Man-in-the-Middle attack. It is designed to detect both the attack and the attacker. The algorithm uses filtering rules to capture the network traffic and pass the IP packets through four phases. After the first three phases, the algorithm is made to raise an alarm on potential ARP poisoning to the user, if one exists, and the fourth phase detects the source IP that has initiated the attack and raises another alarm. This method works successfully even if there is more than one MITM attacker in the subnet. There is a proof of concept implemented for this algorithm. As a result of this experiment, it was found that the Windows 7 Operating System is also vulnerable to ARP attacks as the earlier versions of Windows.

Keywords: *ARP Poison, Intrusion Detection System (IDS), Media Access Control (MAC), Man-in-the-Middle (MITM) attack, Stateless protocol*

1. Introduction

In a Local Area Network (LAN) environment, data transfer takes place by making use of Media Access Control (MAC) addresses. MAC address operates at the Data Link layer of TCP-IP protocol stack. Communications in LAN require IP addresses to be converted into MAC addresses. Address Resolution Protocol (ARP) does this network address translations. To get good performance, ARP maintains an internal cache of IP addresses and the corresponding MAC addresses. When a node wants to send data, it searches in the ARP cache to find out the MAC address of the destination IP [1].

A MAC address consists of 48 bits (6 octets), whereas IPv4 address consists of 32 bits (4 octets). ARP deals with the two kinds of packets – ARP request and ARP reply [2].

Fig.1 shows how a typical ARP protocol communication happens. The machine with IP address 192.168.0.1 sends a broadcast request, asking for the MAC address of the machine whose IP address is 192.168.0.3. Once the request is received by 192.168.0.3, it sends a reply to 192.168.0.1 that, 00:E0:FE:09:C2:11 is the MAC address [3].

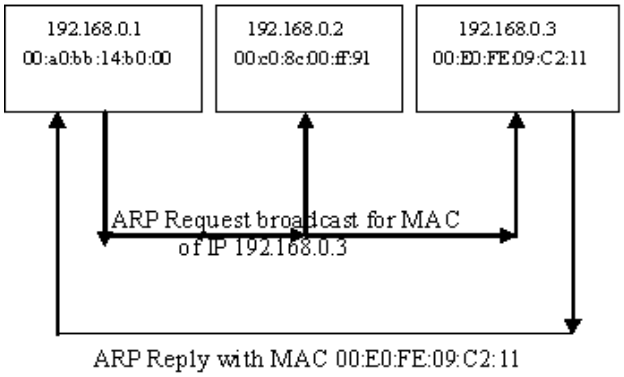


Fig.1 Communication in LAN using ARP

The potential vulnerabilities in this ARP way of communication are researched and presented here with an algorithm to detect the source of ARP poisoning in MITM attack. This paper is organized as follows : In Section 2 the problem is defined and throws light on the existing methods and attempts made by researchers, in Section 3 the proposed solution is presented and in Section 4 explains the lab environment used and the various observations that were made during the implementation of the proposed solution and finally Section 5 concludes the presentation.

2. Problem Defined

2.1 ARP Poisoning

In ARP spoofing attacks, attackers send a fake ARP response packet pretending that they own the MAC address that is wanted. This causes the requester to cache the fake data. As a result, the victim's machine, which has incorrectly cached information about the owner of the IP address, sends all traffic destined for that IP address to the attacker's node [4]. When done properly, the victim will have no idea that the information is redirected to the attacker. The process of updating a target computer's ARP cache with a forged entry is referred to as "poisoning" [5].

The trick is to associate the attacker's MAC address with the IP address of the victim instead. The network is thus made open to vulnerabilities like Man-in-the-Middle attack (MITM), Denial-of-Service (DoS) attack and such [6].

In Fig.2 the machine with IP address B does the MITM attack. So the ARP cache of machine with IP address A and that of IP address C are poisoned. All information transferred between machine A and machine C are now intercepted by machine B.

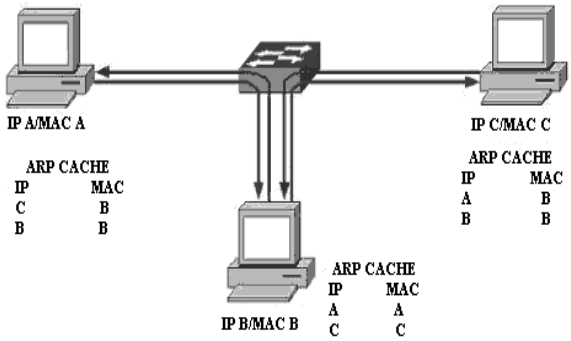


Fig.2 MITM with ARP Poisoning in a switched environment

This is extremely potent, when we consider that, not only can computers be poisoned, but routers/gateways and any device with an IP address as well.

2.2 Defense strategies against ARP poisoning

Sean Whalen [5] in his work has observed that there is no universal defense against ARP spoofing. In fact, the simple possible defense is the use of static ARP entries. Since static entries cannot be updated, spoofed ARP replies are ignored. But the overhead in deploying these tables, as well as keeping them up to date, is not practical for most LANs. Also, the other method of using Port Security does not prevent ARP spoofing.

Aside from these two methods, the only remaining defense is detection. Free tools like Arpwatch [7], Ettercap [8], are working based on defense by detection mechanism, and do not provide complete defense. Instead, at times they even seem to increase the work of the network administrator[9]. Allam Appa Rao et al [10] in their work have suggested software development based on Jpcap for IDS, involving various network vulnerabilities but not specifically meant for the vulnerabilities of ARP. The work of Wenjian Xing et al [11] is about a defense mechanism for ARP spoofing which mainly focus towards the attack of the network gateway. In it the packets are analysed against the IP of the gateway.

With the inherent vulnerabilities of the ARP protocol no complete solution is devised so far. Hence the obvious practical way to go is to devise subnet specific contextual solutions. The solution presented here applies to detecting ARP vulnerabilities due to intrusion in data link layer of the ISO-OSI stack.

3. Proposed Solution

Attacks against layer-2 of ISO-OSI stack, the data link layer, ranges from various ARP attacks like the cache poisoning for wired clients to de-authentication of wireless clients. Fairly simple to implement, these attacks can often go unnoticed by intrusion analysts since intrusion detection systems typically look at the network layer and above to detect attacks. Whatever method of attack an attacker uses to attack the data link layer, in all cases, an adversary attempts to compromise confidentiality, authentication or availability of information. The attacks succeed for the most part because of the lack of fine-grain controls for the data link layer. While layer 2 is considered a less novel platform for attacks, layer 2 attacks continue to trouble the networked systems. The implementation of each attack is unique [12]. This being the current scenario, led to the development of this new algorithm and software, proof to examine network traffic for data link layer attack and proactively respond to attacks.

The algorithm presented here detects the source of ARP poisoning in real time by monitoring traffic flow and raise alarm to the network administrators, detecting both the attack and the attacker.

3.1 Intrusion Detection Algorithm for MITM attacks

The algorithm has four phases as shown in Fig 3. The first alarm is raised when ARP poisoning is discovered. The second alarm is raised when the origin of poisoning, the IP address from which the poisoning is initiated, is discovered by the algorithm.

The first of the four phases is the process of Filtering, where the rules from the rule base are applied on the incoming IP packets and the packets are captured from the online network traffic. The entire Ethernet traffic is passed through the filtering stage and only the filtered set of packets based on the rule base is retained and passed on to the next stage for further process. In the second phase of Segregation, from the full packet with all the fields, specific fields needed for process are segregated.

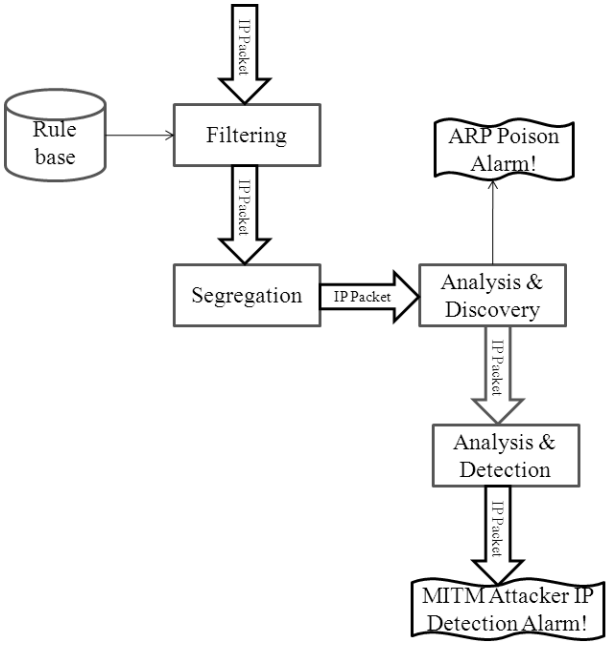


Fig.3 The Algorithm for detection of origin of MITM attack

The ARP Packet structure is as in Fig 4. All the fields in the packet are not needed for the process and hence, only the required fields, the Source IP address and the Source MAC address fields are segregated from the packet using the two functions of Jpcap.

The two functions are getSenderProtocolAddress() and getSenderHardwareAddress(). The segregated information is passed on to the third phase for analysis. As a result of analysis of the details from the captured packets, the ARP poison, if any, is discovered in the subnet and an alarm is raised.

As shown in the Fig.4, the actual ARP packet comprises of nine fields as Hardware type, Protocol type, Hardware size, Protocol size, Operation, Source and Target Hardware address and Source and Target IP address.

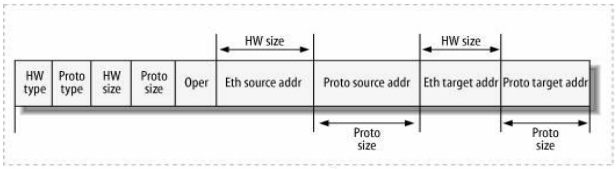


Fig.4 ARP Packet Structure

The last phase of the algorithm is to identify the culprit computer which has initiated the ARP poisoning of the victims and is performing the MITM attack. This last phase is called for, only when the earlier three phases detects an ARP poison in the subnet. In case, there is no poison detected, the last phase has no role to play and the process continues.

In the fourth phase, it captures more packets and analyses the source IP address and the MAC address after segregating the two fields from the whole ARP packet. By sufficient analysis of the information, the algorithm is able to conclude on the attacker machine's IP.

The pseudocode of the proposed Algorithm is as follows

```
Loop //Capture ARP Packets
  Split (ARP Packet ) // the IP, MAC,STATUS
    If  ARP Type != 'Unknown"  then
          Add (IP, MAC) to AddrArray
    Loop
      Split (Next ARP Packet)  //IP, MAC,STATUS
       If  NextARP Type != "Unknown" then
                  Add(IP,MAC) to NewArray
       Loop //Finding possible Poisoned IPs
         If AddrMAC=NextMAC then
           if  AddrIP != NextIP then
                    Raise Poison discovered Alarm
            Loop  //Collecting Poisoned IPs
           If ErrIP != AddrIP  and ErrMAC !=
                       NewMAC then
             Add (AddrIP,NewMAC) to ErrArray
           End if
           If ErrIP!=NewIP and ErrMAC !=NewMAC then
                    Add (NewIP,NewMAC) to ErrArray
           End if
         Else
              NOOP
         End if
    Else
      NOOP
```

```
    End if
      Loop  //Find Attacker
          If (ErrIP[Current]=ErrIP[Next]) then
              ErrIPCount++
          End if
          If  (ErrIPCount>1) then
                  Display (IP,MAC)
                  Raise Alarm attacker Found
          End if
    Else
        Fetch next ARP packet
    End if
Else
        Fetch next ARP packet
End if
```

3.2 Jpcap

The algorithm designed is implemented as a software written in JAVA using the Jpcap, a kind of wrapper class for winpcap/libpcap[13]. Jpcap is a Java library for capturing and sending network packets. It has been tested in Microsoft Windows (98/2000/XP/Vista) and Linux (Fedora, Mandriva, Ubuntu), MacOS X (Darwin), FreeBSD, and Solaris. **Now this work has shown that it works well in Windows 7 also**. It is open source and is licensed under GNU LGPL. It provides facilities to capture raw packets live from wire, save captured packets, filter the packets according to user-specified rules before dispatching them to the application. These facilities have made the use of Jpcap in this software development, trivial. Jpcap does not block, filter or manipulate the traffic generated by other programs on the same machine. Therefore, it does not provide the appropriate support for applications like traffic shapers, QoS schedulers and personal firewalls.

4. The Experiment and Results

4.1 The Experiment

The software was implemented and load tested in a simulated environment, with internet connection, from inside a University network used by the staff and students. The tool used to poison and carry out the MITM attack, was Cain and Abel [14]. The complete experiment involved Windows XP and Windows 2000 Operating Systems. The IP addresses, for the machines in the lab environment, connected to the internet, were dynamically assigned. The entire network was under a single windows domain.

If there are more than one MITM attacks in the subnet, all the attacks are detected by this algorithm based on the captured traffic flowing between poisoned machines and the attacker.

The algorithm was tested with following two experiments

1. A simple MITM attack with one attacker between two victim IPs.
2. The effect of two machines trying to attack each other at the same time.

In both the experiments the algorithm was able to detect the attacking IP/IPs creating the MITM attack.

Detecting the origin of the attack is possible only if there is active transfer of data between the victim machines. In the lab environment, minimal traffic such as a simple ping command between the victims triggers the detection of the second alarm.

The lab environment thus setup with the software developed was able to detect ARP poisoning attacks and the source of attack and raised the alarm, every time there was an attack. To cross check the authenticity of the program executed, it was run along with a free ARP attack detection tool, the XArp [15]. Every time there was a poison taking place, XArp would pop up. The program developed was also able to detect every time the attack was happening. The XArp tool does not directly point out the attacker, but shows a continous list of the changes that the MAC undergoes for each IP. Otherwise the **success rate was seen to be the same for XArp and the software developed.**

Since the detection algorithm is implemented in promiscuous mode, any computer implementing this software within the subnet can detect that there is poisoning in the subnet and identify the attacker/attackers.

4.2 Observations on Experiment 1

Contents of ARP cache of the victims and of the attacker, before and after the poison attack, in the lab is presented here. In all the screen shots below, the dots highlight the IP and MAC addresses that are involved in the MITM attack.

The Cain and Abel tool was executed from the attacker machine with IP 192.168.21.193 and the two victims involved in the attack were with the IP 192.168.21.187 and 192.168.21.178.

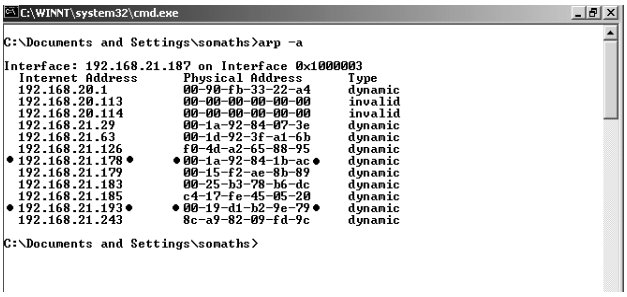


Fig.5 The ARP cache of the victim machine with IP 192.168.21.187 before MITM attack

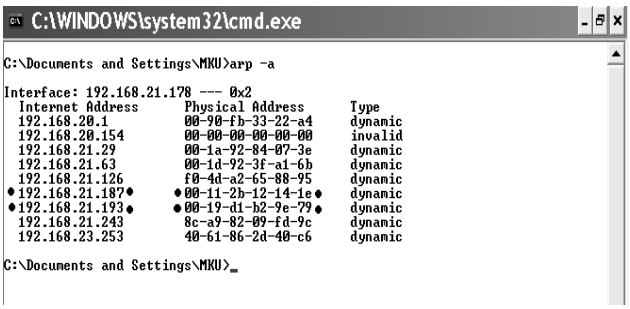


Fig.6 The ARP cache of the victim machine with IP 192.168.21.178 before MITM attack

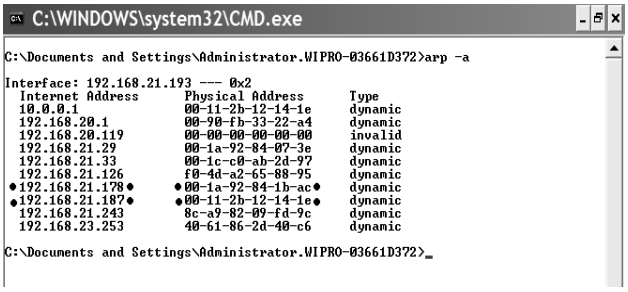


Fig.7 The ARP cache of the machine with IP 192.168.21.193 that implements the MITM attack, before the attack

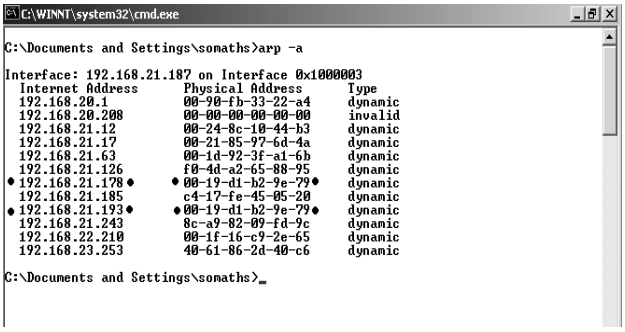


Fig.8 The Poisoned cache of the victim machine with IP 192.168.21.187

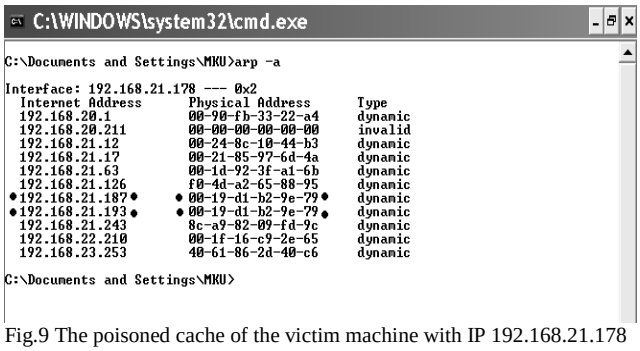


Fig.9 The poisoned cache of the victim machine with IP 192.168.21.178

4.3 Observations on Experiment 2

The IP addresses involved in the two simultaneous MITM attacks are, one set of addresses as in experiment 1 and the other set is with the attacker IP 192.168.21.187 and the two victims being 192.168.21.193 and 192.168.21.178.

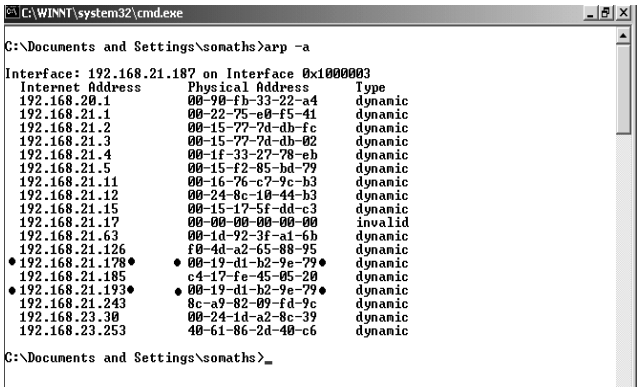


Fig.10 The ARP cache of IP 192.168.21.187 the attacker and the victim

The Fig.11 depicts the poisoned cache of the victim with both the attacker machines and their MAC address changed.

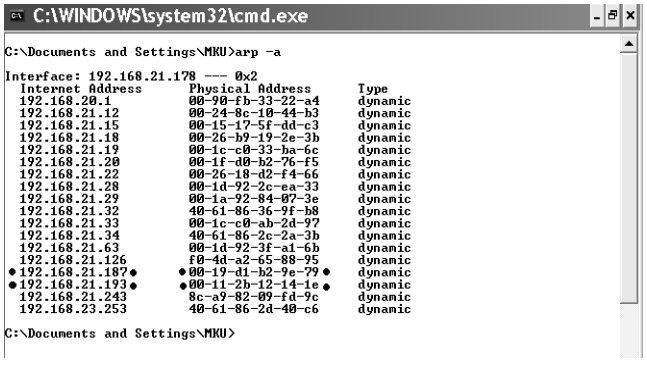


Fig.11 The ARP cache of IP 192.168.21.178 the victim of two simultaneous attacks

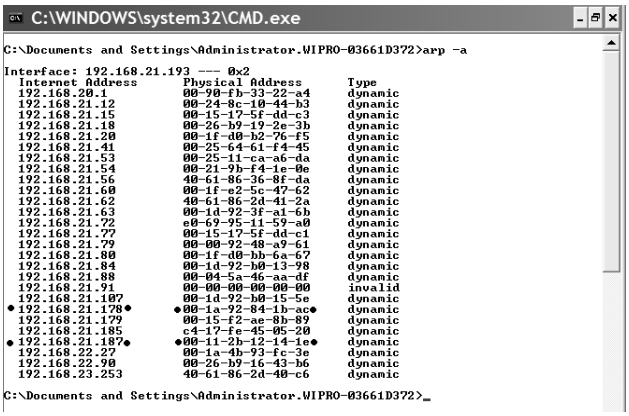


Fig.12 The ARP cache of IP 192.168.21.193, the attacker and the victim

4.4 Results

In Fig.13 and Fig.14 the result is enclosed in a box showing the discovery of poison and detection of the attacker. The dots show the packets of the victim and attacker machines captured for analysis

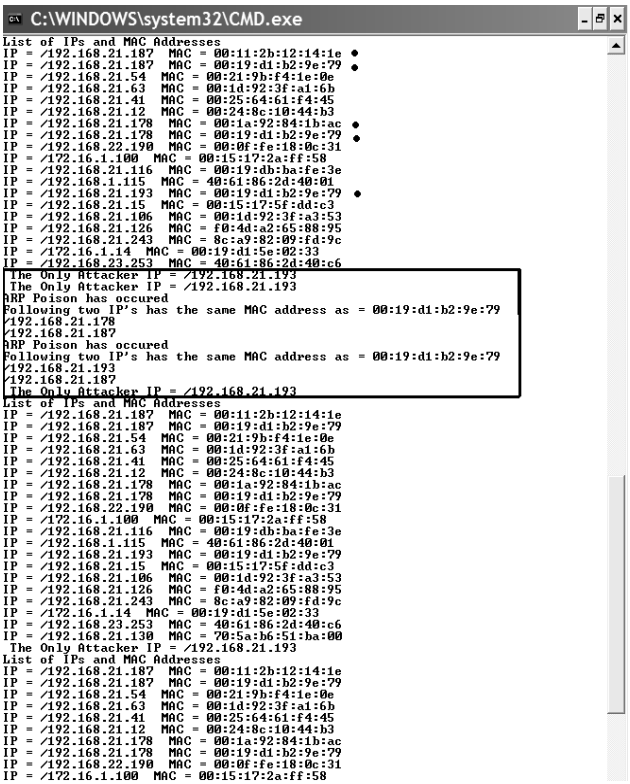


Fig.13 Screen shot of the first experiment

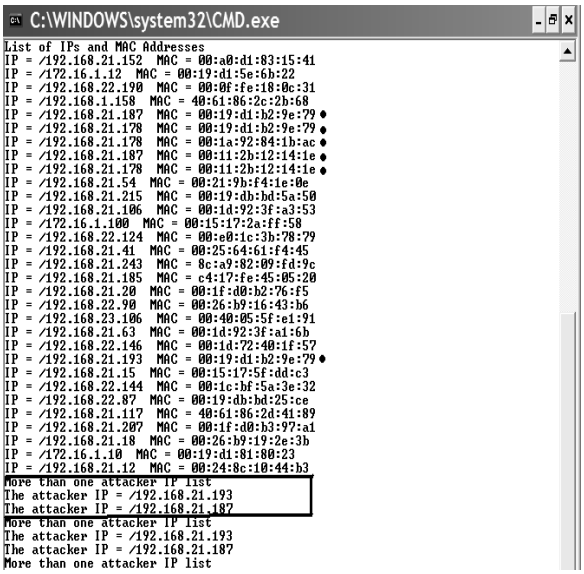


Fig.14 Screen shot of the second experiment. The results enclosed in the box shows the poison discovery and both the attackers identified

5. Finding

As a result of conducting this experiment, *it was found that Windows 7 was equally vulnerable to ARP attacks as the earlier versions of Windows.* It is documented earlier that the various Operating Systems that are vulnerable to ARP attacks are Windows 95/98/2000/NT [16], Windows XP [17]. The OS that is less vulnerable to ARP attacks is Sun Solaris System [18].

6. Conclusion

As this algorithm helps to detect attacks at an early stage, the software could be used as both continuous monitoring tool and as preventive maintenance tool. The alarms raised by the software are useful in taking necessary actions against the malicious IPs. The speed with which internetworking and malicious attacks on the internet grows, the cost of cleaning and bringing business critical systems back online is sky-rocketing. The current trend in the internetworking being so, with the automation and increase in speed and sophistication of attack tools, each year doubling of newly discovered vulnerabilities and administrators finding it difficult to keep up to date with patches, increasing permeability of firewalls with firewall friendly protocols that are designed to bypass firewall configurations, increasing threat from infrastructure attacks like Distributed DoS, Worms or attack on Domain Name Server cache and so on [19], finding generic solutions too is becoming infeasible. Hence, as a subnet based solution,

this algorithm provides a means of prevention by detection and thus provides an inexpensive method to protect systems. Also this algorithm works at data link layer while most of the known methods are based on other layers of ISO-OSI model.

References

[1] Behrouz A. Forouzan, "TCP/IP Protocol Suite", Fourth Edition, Tata McGraw Hill, pp. 220-223, 2010.
[2] D.Plummer, "An Ethernet Address Resolution Protocol", Nov. 1982, RFC 826
[3] S.Vidya, R.Bhaskaran, "ARP storm Detection and Prevention Measures", IJCSI International Journal of Computer Science Issues, Vol.8, Issue 2, March 2011, ISSN(Online): 1694-0814, pp-456-460.
[4] Gopi Nath Nayak and Shefalika Ghosh Samaddar, "Different flavors of Man-in-the-Middle attack, Consequences and feasible solutions", 978-1-4244-5540-9/10/$26.00 ©2010 IEEE.
[5] Sean Whalen, "An Introduction to ARP Spoofing", www.machacking.net/kb/files/arpspoof.pdf, April 2001.
[6] Wikipedia, The Free Encyclopedia, "ARP Spoofing", http://en.wikipedia.org/wiki/ARP_Spoofing, 2011.
[7] L. N. R. Group, "arpwatch, the Ethernet monitor program; for keeping track of Ethernet/ip address pairings," ftp://ftp.ee.lbl.gov/arpwatch.tar.gz, (Accessed on April 7,2011)
[8] Alberto Ornaghi, Marco Valleri, "Ettercap", http://ettercap.sourceforge.net/
[9] Andre P.Ortega, Xavier E. Marcos, Luis D. Chiang, Cristina L. Abad, "Preventing ARP cache poisoning Attacks: A Proof of Concept using OpenWrt", 978-1-4244-4550-9/09/$25.00 ©2009 IEEE
[10] AllamAppaRao, P.Srinivas, B.Chakravarthy, K.Marx, P.Kiran, " A Java Based Network Intrusion Detection System(IDS)", in the proceedings of the 2006 International Journal of Modern Engineering INTERTECH Conference , Session ENG 206-118
[11] Wenjian Xing, Yunlan Zhao, Tonglei Li, "Research on the defense against ARP spoofing attacks based on Winpcap", in Second International Workshop on Education Technology and Computer Science, 2010, DOI 10.1109/ETCS.2010.75, 978-0-7695-3987-4/10 $26.00 © 2010 IEEE
[12] TJ O'Connor, "Detecting and Responding to Data Link Layer Attacks", SANS Institute InfoSec Reading Room, Oct 13, 2010, http://www.sans.org/reading_room/whitepapers/detection/detecting-responding-data-link-layer-attacks_33513, 2010.
[13] Keita Fujii, "Jpcap – a Java library for capturing and sending network packets", http://netresearch.ics.uci.edu/kfujii/Jpcap/doc/, 2007.
[14] Massimilliano Montoro, "Cain and Abel", http://www.oxid.it/, 2011
[15] Chrismc de, "XArp", http://www.chrismc.de/development/xarp/,2010
[16] Roudha Khcherif, "ARP cache poisoning for the detection of sniffers in an Ethernet Network", www.docstoc.com/docs/38584195/arp-cache-poisoning-for-the-dete, 2010
[17] Vamshidhar Chillamcharla, "ARP Spoofing", www.scribd.com/doc/47803882/arp-spoofing,, 2011
[18] Cristina L.Abad, Rafael I.Bonilla, "An Analysis on the Schemes for Detecting and Preventing ARP cache poisoning attacks", 27th International Conference on Distributed Computing Systems Workshops (ICDCSW '07),0-7695-2838-4/07© 2007 IEEE
[19] CERT and CERT Coordination Center, "Overview of attack Trends",© Carnegie Mellon University, 2002, http://www.arcert.gov.ar/webs/textos/attack_trends.pdf.

S.Vidya completed M.Sc in Computer Science from SR College, Trichirappalli, TamilNadu, India in the year 1990 and M.Phil in Computer Science from Alagappa University, Karaikudi, TamilNadu, India in the year 2001. Working as Associate Professor in Computer Science in the Department of Computer Science, Fatima College, Madurai, TamilNadu, India since 1990. Areas of interest are Data Structures and Algorithms. She is currently pursuing research in Network Security. She is a Life Member of Computer Society of India and Member of ACM.

Dr.R.Bhaskaran completed M.Sc in Mathematics from IIT Chennai, TamilNadu, India in the year 1974 and got his Doctoral degree from Ramanujam Institute of Mathematical Sciences, Chennai, TamilNadu, India in the year 1979. Joined the School of Mathematics, Madurai Kamaraj University, Madurai in 1980. Currently he is the Chairperson of School of Mathematics. He has to his credit lots of publications including in IEEE conferences. His area of interest includes Linden Mayer System, Computer Applications, and developing software for learning Mathematics. He has guided students in both Mathematics and Computer Applications. He is a Life Member of the Indian Mathematical Society.

Medicinal Plants Database and Three Dimensional Structure of the Chemical Compounds from Medicinal Plants in Indonesia

Arry Yanuar[1*], Abdul Mun'im[1], Akma Bertha Aprima Lagho[1], Rezi Riadhi Syahdi[1], Marjuqi Rahmat[2], and Heru Suhartanto[2]

[1] Department of Pharmacy, Faculty of Mathematics and Natural Sciences, University of Indonesia Depok, 16424, Indonesia

[2] Faculty of Computer Sciences, University of Indonesia Depok, 16424, Indonesia

Abstract

During this era of new drug designing, medicinal plants had become a very interesting object of further research. Pharmacology screening of active compound of medicinal plants would be time consuming and costly. Molecular docking is one of the *in silico* method which is more efficient compare to *in vitro* or *in vivo* method for its capability of finding the active compound in medicinal plants. In this method, three-dimensional structure becomes very important in the molecular docking methods, so we need a database that provides information on three-dimensional structures of chemical compounds from medicinal plants in Indonesia. Therefore, this study will prepare a database which provides information of the three dimensional structures of chemical compounds of medicinal plants. The database will be prepared by using MySQL format and is designed to be placed in **http://herbaldb.farmasi.ui.ac.id** website so that eventually this database can be accessed quickly and easily by users via the Internet.

Keywords: *Indonesia's Medicinal Plants, Natural Compounds Database, Three Dimensional Structure.*

1. Introduction

Indonesia as a mega-biodiversity center has a second position in the world after Brazil. Indonesia become the first position, if marine biota is also included. In our Earth, which lived about 40,000 species of plants, of which 30,000 species live in the Indonesian archipelago. Among the 30,000 plant species that live in the Indonesian archipelago known to at least 9600 species of plants have a pharmacological activity [1]. In an era of new drug design, these plants become the basic material to study further. Screening of pharmacological activity of active ingredients in medicinal plants is an expensive process, requiring energy, qualified human resources and require a long time if done in laboratory experiments using experimental animals [2]. Attractive chance today is the use of computers as tools in drug development. Exponentially increase computing capabilities provide opportunities to develop simulations and calculations in drug design. The method used in drug design can be a structure-based drug design and ligand-based drug design [3]. In the field of structure-based drug design, molecular docking is a commonly used method. Molecular docking is a method which used to predict an intermolecular complex between the drug molecule with its target protein. When performing molecular docking, a set of data which contains information on the ligand or drug to be docked and protein targets to be used are needed. The information required in this process include three-dimensional structure of the ligand and target protein [4].

Until recently, there is no database of medicinal plants and its natural compounds in Indonesia, which presents data interactively and comes with three-dimensional structure of chemical compounds. The aim of this research is to prepare medicinal plants database and three dimensional structure of the active compounds from medicinal plants in Indonesia.

2. Design and Methods

This research is done using combination of literature study, simulation and molecular modeling. The presentation of the results as a monographs are reported descriptively through online database prototype. The experimental design is made as follows:

2.1 Screening and preparing the list of Indonesian medicinal plants

This study begins with gathering information about the collection of medicinal plants in Indonesia. Searching of medicinal plants in Indonesia obtained by literature from scientific journals, books and websites. This search is then continued with the selection of plants used as medicine in Indonesia. Various sources of classic and official books are used as a database raw data [5-12].

2.2 Searching and drawing of chemical compound

From medicinal plants found in the database table and then collected data on chemical compounds found in the medicinal plants. If the chemical compounds found contained in it, then extended the search to find a two-dimensional structure of these chemical compounds. Two-dimensional structure of the search query is performed on the chemical compound database Pubchem [13] or KNApSAcK [14]. Two-dimensional structure of the search should be available in specific file formats. In KNApSAcK metabolite database, available file format is gif file format to be converted into a file format acceptable to the program used. Conversion of file formats was done by drawing two-dimensional structures of chemical compounds of medicinal plants using Symyx Draw program. The 2D structures were then stored in a file .mol format [15].

2.3 Generating the 3D structure

Three-dimensional structure is formed using VEGAZZ [16], a compound modification program. File format .mol (storage results from Symyx Draw program) or file format .sdf which have been downloaded from PubChem (the conversion results with the program PyMOL [17]) are then included in the program VEGAZZ computer which will then be processed to form a new three-dimensional molecule. Formation of three-dimensional structure was done through several stages. The first stage is to open two-dimensional structure that is stored in a file format .mol. Two-dimensional structures that have emerged in Vega ZZ view, will then be turned into three-dimensional with the run command run scripts. Run this script and then bring up the next dialog box choose the command '2 D to 3D 'which means the two-dimensional display will be turned into a three-dimensional. Having formed three-dimensional structure, save the file with the .mol format. The .mol format was then converted into .mol2 format using the OpenBabel program [18].

2.4 Preparation of medicinal plants database.

Medicinal plant database was prepared using MySQL format. Designing databases using raw data (file. XLS) which has been established the previous step and armed with the requirements gathering process, then from there made it an entity relationship. The following entity relationship diagram illustrates the connectedness between the data to be stored in a database.

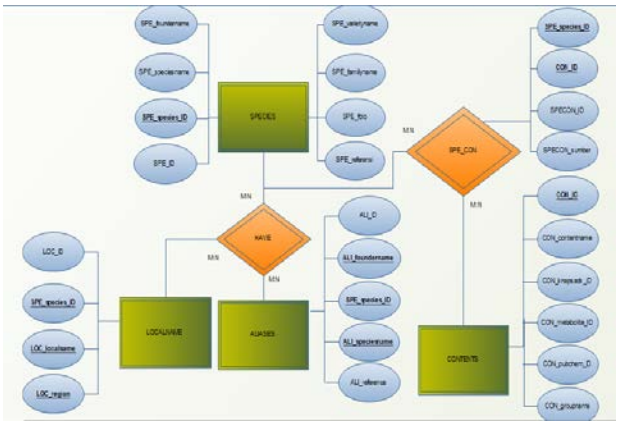


Fig 1. Entity Relationship Diagram

2.5 Creating a prototype web database

Prototype of web database of medicinal plants and three-dimensional structures of chemical compounds was created by PHP. The user interface is created are the homepage, the page data and details of species, compound data pages, profile pages, admin pages, file upload pages, user pages, search results pages, language settings and FAQs page. The authorized users are divided into the categories: administrator, expert and contributor that have a role to maintain the database via login session. The common users could also use this database on the website without login session, but they will have limited access.

3. Results and Discussion

To date data has been collected as many as 3825 species records, with various local names as many as 16,244 records, and content as many as 6776 record recorded in the (species-contents) interaction as many as 12,980 records. All of this data is collected from the various literature and noted the source. Web database prototype of medicinal plants and three-dimensional structures of chemical compounds have been created with PHP. Website address of Indonesian Medicinal Plant Database is located at URL http://herbaldb.farmasi.ui.ac.id. Initially, a total of 1412 three-dimensional structure of chemical compounds from medicinal plants of Indonesia embedded in the system. The entire database system runs well and database needs to be verified by the expert users.

The nature of open system allows wide use of plant medicine for Indonesia database by both the general public and scientists or other stakeholders from both industry sectors Herbs / Jamu, government, or university. The nature of “Open Systems”, a search algorithm and interactive data presentation allows the emergence of knowledge systems. Quality assurance and validity of the

database content can be maintained by the control from the existing users and the verification from an expert. Thus the potential for growth and wider utilization becomes very open. Global utilization is also wide open with the features of English.

Fig 2. Web database snapshot of list of species

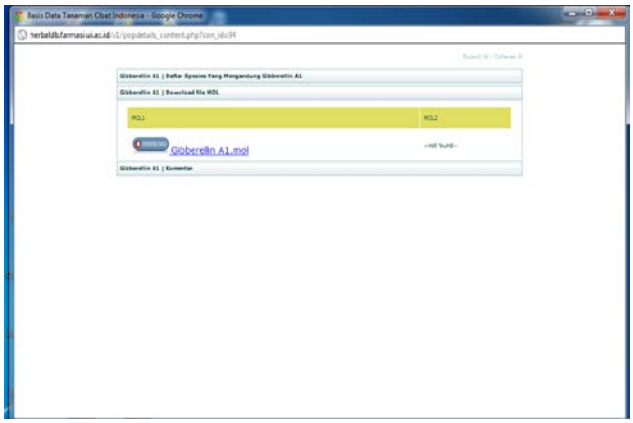


Fig. 3 Snapshot of 3D chemical structure download page.

4. Conclusions

The prototype of the web-based interactive Indonesian Medicinal Plant Database System has been created using PHP and MySQL relational database. Database contents can be accessed through a search algorithm using the species name, alias names or synonyms, local or regional names, compound name, and usage keywords. Data species are equipped with photographs that can be uploaded by the authorized user and the name of the chemical compound has three-dimensional data structure in the format .mol and .mol2 that can be uploaded by authorized users and downloaded by all users. The features of the contents are made as “open system” those can be modified, added or deleted by authorized users. The uniqueness of this system that is also seen as an open feedback system is a comment feature for species name and compound name of each entry, which allows for discussion between users and the objections of an entry. Authorized users are registered as administrator, expert or contributor, which each have a role in maintaining the content in the database. Indonesian medicinal plants database has a record with default: not verified. The task to verify the data in the database is performed by the expert users. This work is an early version of the overall road map website of Indonesian medicinal plants database and three-dimensional structure of chemical compounds. Website of Indonesian Medicinal Plant Database is located at **http://herbaldb.farmasi.ui.ac.id**.

Acknowledgments

This research is financially supported by Riset Unggulan Universitas Indonesia (RUUI), 2010. A.Y. Author thanks to Prof. Shigehiko Kanaya, PhD from Nara Institute of Science and Technology (NAIST), Japan for providing data from KNApSAcK metabolite database.

References

[1] Keputusan Menteri Kesehatan Republik Indonesia No : 381/MenKes/SK/III/2007, Menteri Kesehatan RI, Jakarta, 2007
[2] Jayaram, B., SCFBIO: What is drug design? http://www.scfbio-iitd.res.in/tutorial/drugdiscovery.htm,2011
[3] Hawkins, P., & Skillman, G., Ligand-based design workflow, http://images.apple.com/science/pdf/ligandbased_design_workflow.pdf, 2006
[4] Abraham, D.J., (ed). Burger's Medicinal Chemistry and Drug Discovery, Volume 1: Drug Discovery, 6th ed. Willey Interscience, 2003
[5] Heyne, K., De Nuttinge Planten van Indonesie, 3ed, Wageningen, H. Veenman & Zonen, 1950
[6] Departemen Kesehatan Republik Indonesia, Materia Medika Indonesia. Vol I. Jakarta: Departemen Kesehatan Republik Indonesia. 1977
[7] Departemen Kesehatan Republik Indonesia, Materia Medika Indonesia, Vol II. Jakarta: Departemen Kesehatan Republik Indonesia. 1978
[8] Departemen Kesehatan Republik Indonesia, Materia Medika Indonesia, Vol III. Jakarta : Departemen Kesehatan Republik Indonesia. 1979
[9] Departemen Kesehatan Republik Indonesia, Materia Medika Indonesia, Vol IV. Jakarta : Departemen Kesehatan Republik Indonesia. 1980
[10] Departemen Kesehatan Republik Indonesia, Materia Medika Indonesia, Vol V. Jakarta : Departemen Kesehatan Republik Indonesia. 1989
[11] Departemen Kesehatan Republik Indonesia, Materia Medika Indonesia, Vol VI. Jakarta : Departemen Kesehatan Republik Indonesia. 1995
[12] Medicinal Herb Index in Indonesia 2nd edition. Jakarta: PT. Eisai Indonesia, 1995
[13] Bolton, E., Wang, Y., Thiessen, P. A., and Bryant, S.H., PubChem: Integrated Platform of Small Molecules and

Biological Activities. Chapter 12 IN Annual Reports in Computational Chemistry, Volume 4, American Chemical Society, Washington, DC, 2008 , 217-241, DOI:10.1016/S1574-1400(08)00012-1

[14] Shinbo, Y., Nakamura, Y., Altaf-Ul-Amin, M., Asahi, H., Kurokawa, K., Arita, M., Saito, K., Ohta, D., Shibata, D. and Kanaya, S., KNApSAcK: A Comprehensive Species-Metabolite Relationship Database, Plant Metabolomics: in Biotechnology in Agriculture and Forestry, 2006, (57)165-181, DOI: 10.1007/3-540-29782-0_13

[15] Symyx Draw-An introductory guide, http://bbruner.org/obc/symyx.htm, 2011

[16] Pedretti, A., Villa. L., and Vistoli, G., Vega—an open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming. J. Comput. Aided. Mol. Des., 2004, 18(3) 167-173, DOI: 10.1023/B:JCAM.0000035186.90683.f2

[17] Delano, W., Pymol user's guide. Delano Scientific LLC.: http://pymol.sourceforge.net/newman/userman, 2004

[18] Guha, R., Howard, M. T., Hutchison, G. R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J. K., and Willighagen, E. L., "The Blue Obelisk -- Interoperability in Chemical Informatics." J. Chem. Inf. Model. (2006) 46(3) 991-998. DOI:10.1021/ci050400b

Arry Yanuar is an assistant Professor at Department of Pharmacy, Universitas Indonesia. He has been with Department of Pharmacy since 1990. He graduated from undergraduate program Department of Pharmacy, University of Indonesia in 1990. He also holds Pharmacist Profession certificate in 1991. In 1997, he finished his Master Program from Faculty of Pharmacy, Gadjah Mada University. He holds PhD in 2006 from Nara Institute of Science and Technology (NAIST), Japan, from Structural Biology/protein crystallography laboratory. In 1999-2003 he worked as pharmacy expert in ISO certification for pharmacy industries at Llyod Register Quality Assurance. In 2002, he visited National Institute of Health (NIH), Bethesda, USA. He won several research grants and published some paper at international journals and conferences.

Abdul Mun'im is an assistant Professor at Department of Pharmacy, Universitas Indonesia. He has been with Department of Pharmacy since 1990. His field of research is in natural product chemistry or phytochemistry.

Akma Bertha Aprima Lagho hold BSc from Department of Pharmacy, University of Indonesia in 2010.

Rezi Riadhi Syahdi is a Master Student from Department of Pharmacy, University of Indonesia. He hold BSc in Pharmacy in 2010.

Marjuqi Rahmat hold BSc from Faculty of Computer Sciences, University of Indonesia in 2010.

Heru Suhartanto is a Professor in Faculty of Computer Science, Universitas Indonesia (Fasilkom UI). He has been with Fasilkom UI since 1986. Previously he held some positions such as Post doctoral fellow at Advanced Computational Modelling Centre, the University of Queensland, Australia in 1998 – 2000; two periods vice Dean for General Affair at Fasilkom UI since 2000. He graduated from undergraduate study at Department of Mathematics, UI in 1986. He holds Master of Science, from Department of Computer Science, The University of Toronto, Canada since 1990. He also holds Ph.D in Parallel Computing from Department of Mathematics, The University of Queensland since 1998. His main research interests are Numerical, Parallel, Cloud and Grid computing. He is also a member of reviewer of several referred international journal such as journal of Computational and Applied Mathematics, International Journal of Computer Mathematics, and Journal of Universal Computer Science. Furthermore, he has supervised some Master and PhD students; he has won some research grants; holds several software copyrights; published a number of books in Indonesian and international papers in proceeding and journal. He is also member of IEEE and ACM.

A New Distinguisher for CubeHash-8/b and CubeHash-15/b Compression Functions

Javad Alizadeh[1] and Abdolrasoul Mirghadri[2]

[1] Faculty and Research Center of Communication and Information Technology, IHU
Tehran, Iran

[2] Faculty and Research Center of Communication and Information Technology, IHU
Tehran, Iran

Abstract

CubeHash is one of the round 2 candidates of the public SHA-3 competition hosted by NIST. It was designed by Bernstein. In this paper we find a new distinguisher to distinguish CubeHash compression function from a random function. This distinguisher principle is based on rotational analysis that formally introduced by Khovratovich and Nikolic. In order to use this technique, we need to compute the probability that four swap functions in CubeHash round function preserve the rotational property for any input pair. We compute these probabilities and find a new distinguisher that distinguish CubeHash-8/b and CubeHash-15/b compression function from a random function with probability greater than $2^{-538.5}$ and $2^{-1009.7}$, respectively. Until we know this is the first distinguisher for CubeHash compression function with more than 14 rounds.

Keywords: *SHA-3 candidate, CubeHash, rotational analysis, distinguisher.*

1. Introduction

Hash functions have a very important role in modern cryptography that used in many areas as digital signatures and various forms of authentication. A hash function is a transformation which maps a variable-length input to a fixed-size output, called message digest. After developments in the field of hash function cryptanalysis [13, 14, 15] along with new results targeted against commonly used hash functions has urged National Institute of Standards and Technology to announce a competition for the development of a new hash standard, SHA-3 [10].

The SHA-3 competition attracted a lot of attention in the cryptographic community. A total number of 64 hash function proposals was submitted, 51 of them advanced to the first round, 14 of them to the second round, and 5 of them advanced to the third, final round. One of the main requirements was the evaluation of the security of the submitted primitive, i.e. providing a detailed analysis on the resistance of the function against various attacks. Being one of the most powerful forms of attacks, the differential [4] and linear [9] analysis got most of the submitter's focus, while less attention was put on the other attacks.

Rotational analysis is a relatively new type of attack. Although this technique was mentioned and applied in previous works, such as [6], but it formally introduced by Khovratovich and Nikolic in [7]. It is also used for cryptanalysis of modified versions of BMW [5] and SIMD [8] hash functions in [11]. Unlike differential analysis, where for a pair (x,y), the attacker follows the propagation of the difference $x \oplus y$ through some transformation, in rotational analysis the adversary studies the propagation of the rotational pair $(x, x <<< r)$ through the transformation.

Khovratovich and Nikolic in [7] analyze the primitives composed of only three operations: addition, rotation, XOR (ARX). For these primitives, they prove that the probability that a rotational pair of inputs will produce a rotational pair of outputs depends only on the number of additions.

Previous Results on CubeHash: We refer to part 2. B. 5 of [2] and CubeHash profile in the SHA-3 Zoo [12] for a complete survey of cryptanalytic results on CubeHash. The currently best distinguisher attacks on CubeHash-r/b compression function for r=11 and r=14 were presented by Ashur and Dunkelman in [1]. They present a distinguisher of complexity 2^{470} for Cubehash-11/b compression function and another distinguisher of complexity 2^{812} for Cubehash-14/b compression function. In this work they use linear cryptanalysis technique [9] and linear approximations of CubeHash. Until we know no distinguisher for more than 14 rounds was presented so far.

Contribution of this Paper. The goal of this work is to extend the application of rotational analysis to CubeHash [2], one of the second round of the SHA-3 competition candidates that have four swap transformations other than additions, rotations, and XORs. In particular, we find the rotational probabilities of the four swap transformations in the CubeHash round function. This allows us that for several round parameters r and message block sizes b we present a new distinguisher for CubeHash-r/b compression function. Specially, we find a distinguisher for CubeHash-8/b and CubeHash-15/b compression functions with probability greater than $2^{-538.5}$ and $2^{-1009.7}$, respectively. In CubeHash-r/b, $b \in \{1,2,3,...,128\}$ and if in CubeHash-8/b compression function we set b=1, indeed we distinguish compression function of CubeHash-8/1, the first version of CubeHash that suggested to SHA-3 competition, from a random function. Also if in CubeHash-15/b we set b=32, we distinguish a reduced version of compression function of CubeHash-16/32 [3], only with one round less, from a random function. CubeHash-16/32 is a tweaked version of CubeHash-8/1 which is about 16 times faster than it and after presentation of cryptanalysis results on CubeHash-8/1 announced the official proposal for all digest lengths h=224, 256, 384 or 512.
Until we know this is the first distinguisher for CubeHash compression function with greater than 14 rounds.

Organization: In section 2 we review the concept of a distinguisher that use rotational analysis technique presented in [7]. In Section 3 we describe the hash function CubeHash and define its compression function. Section 4 specifies the probabilities that four swap functions in CubeHash round function preserve the rotational property for any input pair. In Section 5 we analyze CubeHash-8/b and CubeHash-15/b compression functions and present a new distinguisher for them. We conclude in Section 6 finally.

2. Distinguishers with Rotational Analysis

A rotational distinguisher explores the idea that some transforms on rotated inputs produce rotated outputs. Let $(X, X <<< r)$ be a pair of input words, it called a rotational pair, for some transform F, where $<<< r$ is a cyclic rotation to the left by r bits. If for an arbitrary input X, $F(X <<< r) = F(X) <<< r$ then it said that F preserves the rotational property, in other words, when the input composes a rotational pair, the output also composes a rotational pair. The input and the output of a transform can be a single word or a vector of words, i.e. $\tilde{X} = (X_1,...,X_n)$. Then, a rotational input/output pair is defined as $(\tilde{X}, \tilde{Y})$, where $Y_i = X_i <<< r, i = 1,...,n$. A system $\Phi(\tilde{X})$ composed of transforms $F_1,...,F_k$ preserves the rotational property, if on rotational input pair, produces a rotational output pair.
It is better that we attend two important issues. First, unlike differential analysis where usually out of the whole input pair only a few words have differences, in rotational analysis all the input pairs of words have to compose rotational pairs. Second, there are a few transforms that preserve the rotational property for any input pair. Usually, for an arbitrary X, $F(X <<< r) = F(X) <<< r$ only with some probability p_F, further called a rotational probability of F, that depends on the rotational amount r. If we assume that the outputs of the transforms are independent, then a system Φ composed of transforms $F_1,...,F_k$ preserves the rotational property with a probability $p_\Phi = p_{F_1} \cdot p_{F_2} .....p_{F_k}$ [11]. Hence, in order to find the probability that a system preserves the rotational property, one only has to find the probabilities that each instance of the underlying transforms preserves this property. For a random system with n-bit output, the probability that a rotational input will produce rotational output is 2^{-n}. Therefore, if a system Φ with n-bit output has a rotational probability $p_\Phi > 2^{-n}$, then this system can be distinguished from a random system [11].

2.1 Rotational Analysis of ARX Constructions

A thorough rotational analysis of ARX systems was given in the work of Khovratovich and Nikolic [7]. These systems are composed only of three transforms: addition, rotation and XOR. For each of them, the probabilities they preserve the rotational property were given by:

Lemma 1 (Addition): For n-bit words x, y, and a positive integer r

$$\Pr[(x+y) <<< r = x <<< r + y <<< r] = \frac{1}{4}(1 + 2^{r-n} + 2^{-r} + 2^{-n}).$$

Lemma 2 (Rotation): For n-bit word x and positive integers r, r'

$$\Pr[(x <<< r) <<< r' = (x <<< r') <<< r] = 1$$

Lemma 3 (XOR): For n-bit words x, y, and a positive integer r

$$\Pr[(x \oplus y) <<< r = x <<< r \oplus y <<< r] = 1$$

Hence, rotations and XORs preserve the rotational property with probability 1, while the probability of addition depends on the size of the words and the rotation amount. Further in our analysis, the rotation amount will be fixed to 1.

The proofs of lemma 1 and 2 are simple, and the proof of lemma 3 can be finding in [7].
In [11], the authors find the probabilities that subtractions, shifts and bitwise Boolean functions preserve the rotational property, too, and in this paper we find this probability for four swap functions used in CubeHash.

3. CubeHash Description

CubeHash [2] is Bernstein's proposal for the NIST SHA-3 competition [10]. CubeHash works with 32-bit words (n=32) and uses three simple operations of XOR, rotation and modular addition and four swap functions showed by $SWAP_1, SWAP_2, SWAP_3$, and $SWAP_4$. It has an internal state $S = (S_0, S_1, ..., S_{31})$ of 32 words and its variants, denoted by CubeHash-r/b, are identified by two parameters $r \in \{1,2,...,128\}$ and $b \in \{1,2,...,128\}$ which at each iteration process b bytes in r rounds. Selecting different values of r and b, allow the selection of a range of security/performance tradeoffs. The internal state S is set to a specified value which depends on the digest length (limited to 512 bits) and parameters r and b. The message to be hashed is appropriately padded and divided into b-byte message blocks. At each iteration one message block is processed as follows. The 32-word internal state S is considered as a 128-byte value and the message block is XORed into the first b bytes of the internal state. Then, the following fixed permutation is applied r times to the internal state to prepare it for the next iteration. This permutation called CubeHash round function and denoted by ROUND in this paper.

1. Add S_i into $S_{i\oplus 16}$, for $0 \le i \le 15$.
2. Rotate S_i to the left by seven bits, for $0 \le i \le 15$.
3. Swap S_i and $S_{i\oplus 8}$, for $0 \le i \le 7$ (We call this swap function $SWAP_1$).
4. XOR $S_{i\oplus 16}$ into S_i, for $0 \le i \le 15$.
5. Swap S_i and $S_{i\oplus 2}$ for $i \in \{16,17,20,21,24,25,28,29\}$ (We call this swap function $SWAP_2$).
6. Add S_i into $S_{i\oplus 16}$, for $0 \le i \le 15$.
7. Rotate S_i to the left by eleven bits, for $0 \le i \le 15$.
8. Swap S_i and $S_{i\oplus 4}$, for $i \in \{0,1,2,3,8,9,10,11\}$ (We call this swap function $SWAP_3$).
9. XOR $S_{i\oplus 16}$ into S_i, for $0 \le i \le 15$.
10. Swap S_i and $S_{i\oplus 1}$, for $i \in \{16,18,20,22,24,26,28,30\}$ (We call this swap function $SWAP_4$).

Having processed all message blocks, a fixed transformation is applied to the final internal state to extract the hash value as follows. First, the last state word S_{31} is XORed with integer 1 and then the above permutation is applied $10 \times r$ times to the resulting internal state. Finally, the internal state is truncated to produce the message digest of desired hash length. Refer to [2] for the full specification.

3.1 CubeHash Compression Function

Compression function of CubeHash-r/b that we denoted by COMP-r, gives a state of 1024 bits (128-byte) as input and then applies CubeHash round function, ROUND, r times to this state and output a new state of 1024 bits (128-byte).

COMP-8 and COMP-11: COMP-8 is the compression function of CubeHash-8/b and COMP-15 is the compression function of CubeHash-15/b.

4. Rotational Analysis of CubeHash Swap Functions

As mentioned, in order to extend the application of rotational analysis to CubeHash compression function we have to find the probabilities that four swap functions ($SWAP_1, SWAP_2, SWAP_3$ and $SWAP_4$) in CubeHash round function (ROUND) preserve the rotational property for any input pair. In this section we compute the probabilities in lemma 4 to lemma 7.

Lemma 4 ($SWAP_1$): Suppose $SWAP_1$ is the swap function used in step 3 of CubeHash round function. We choose a random X, rotate it to the left by 1 bit and produce a pair of rotational input, $(X, X <<< 1)$. For this pair, $SWAP_1$ preserves the rotational property with probability 1. In other words:

$$\Pr[SWAP_1(X <<< 1) = SWAP_1(X) <<< 1] = 1.$$

Lemma 5 ($SWAP_2$): Suppose $SWAP_2$ is the swap function used in step 5 of CubeHash round function. We choose a random X, rotate it to the left by 1 bit and produce a pair of rotational input, $(X, X <<< 1)$. For this pair, $SWAP_2$ preserves the rotational property with probability of 2^{-6}. In other words:

$$\Pr[SWAP_2(X <<< 1) = SWAP_2(X) <<< 1] = 2^{-6}.$$

Lemma 6 ($SWAP_3$): Suppose $SWAP_3$ is the swap function used in step 8 of CubeHash round function. We choose a random X, rotate it to the left by 1 bit and produce a pair of rotational input, $(X, X <<< 1)$. For this pair, $SWAP_3$ preserves the rotational property with probability of 2^{-2}. In other words:

$$\Pr[SWAP_3(X <<< 1) = SWAP_3(X) <<< 1] = 2^{-2}.$$

Lemma 7 ($SWAP_4$): Suppose $SWAP_4$ is the swap function used in step 10 of CubeHash round function. We choose a random X, rotate it to the left by 1 bit and produce a pair of rotational input, $(X, X <<< 1)$. For this pair, $SWAP_4$ preserves the rotational property with probability of 2^{-14}. In other words:

$$\Pr[SWAP_4(X <<< 1) = SWAP_4(X) <<< 1] = 2^{-14}.$$

Proof: The above lemmas will be proved in the Appendix A.

5. Rotational analysis of CubeHash-8/b and CubeHash-15/b compression functions

In this section we applied the rotational analysis technique to CubeHash and find a new distinguisher for CubeHash-8/b compression function (COMP-8) and CubeHash-15/b compression function (COMP-15) that distinguish these functions from a random function. For this purpose we use the results of lemma 1 to lemma 7.

Now, we consider the COMP-8 function that using three simple operations of modular addition, rotation and XOR, and four swap functions of $SWAP_1, SWAP_2, SWAP_3$, and $SWAP_4$. Suppose we want to find the probability that COMP-8 preserves the rotational property. In other words if $(X, X <<< 1)$ be a pair of rotational input, we have to compute the probability that (Comp-8(X), Comp-8(X)<<<1) is a pair of rotational output (the probability that Comp-8(X<<<1) = Comp-8(X)<<<1).

By lemma 2 and lemma 3, rotations and XORs preserve the rotational property with probability 1. So the desired probability is depending on the additions and swap functions. In the rotational analysis of COMP-8 if we consider left rotation amount fixed to 1, by lemma 1 we can find the probability that one addition (on the 32-bit words) in COMP-8 preserve the rotational property. This probability is:

$$\Pr[(x + y) <<< 1 = x <<< 1 + y <<< 1] = \frac{1}{4}(1 + 2^{1-32} + 2^{-1} + 2^{-32}) > 2^{-1.416}$$

Each round of COMP-8 has 32 additions on the 32-bit words. Hence the probability that in the one round of COMP-8 the rotational property is preserved by all addition (Pr_{ADD}) will be:

$$Pr_{ADD} = 2^{(-1.416)^{32}} > 2^{-45.312}$$

On the other hand in the each round of COMP-8 four swap functions ($SWAP_1, SWAP_2, SWAP_3$, and $SWAP_4$) is used too. For these functions we compute the probabilities that they preserve the rotational property. These probabilities are respectively:

$$Pr_{SWAP_1} = 1,$$
$$Pr_{SWAP_2} = 2^{-6},$$
$$Pr_{SWAP_3} = 2^{-2},$$
$$Pr_{SWAP_4} = 2^{-14}.$$

According to section 2 the probability (Pr_{ROUND}) that one round of COMP-8 using 32 additions on the 32-bit words, XORs, rotations, and swap functions $SWAP_1, SWAP_2$, $SWAP_3$, and $SWAP_4$ preserve the rotational property is:

$$Pr_{ROUND} = (Pr_{ADD})^{32} \times (Pr_{SWAP_1}) \times (Pr_{SWAP_2}) \times (Pr_{SWAP_3}) \times (Pr_{SWAP_4}) > 2^{-67.312}.$$

Finally the COMP-8 function has 8 rounds. Hence the probability (Pr_{COMP-8}) that this function preserve the rotational property for a pair of rotational input such as $(X, X <<< 1)$ and produce a pair of rotational output such as $(COMP-8(X), COMP-8(X) <<< 1)$ will be:

$$Pr_{COMP-8} = (Pr_{ROUND})^8 > 2^{-538.496} >>> 2^{-1024}$$

The Pr_{COMP-8} is very greater than 2^{-1024} (the probability that we can distinguish the CubeHash compression function when it is indistinguishable from a random function) and allows distinguishing 8-round CubeHash compression function using about 2^{539} call of it. The distinguisher of COMP-8 based on rotational analysis technique is working this way:

1- He chooses a random input such as X, computes $X <<< 1$ and produces a pair of the rotational input, $(X, X <<< 1)$.

2- He computes COMP-8 (X), COMP-8($X <<< 1$), and COMP-8(X)<<<1 .

3- He verifies whether

COMP-8(X<<<1)=COMP-8(X)<<<1

or not.

4- If COMP-8(X<<<1)=COMP-8(X)<<<1 , the distinguisher can produce a pair of the rotational output, $(COMP-8(X), COMP-8(X) <<< 1)$, and distinguishes the COMP-8, otherwise go to 1.

COMP-15 Distinguisher: Similar to whatever we said about distinguishing of COMP-8, we can construct a distinguisher based on rotational analysis for 15-round CubeHash compression function (COMP-15) and distinguish it from a random function.
Note that the only difference between COMP-8 and COMP-15 is the number of their rounds. COMP-8 using 8 times of the CubeHash round function (ROUND) while COMP-15 using 15 times of it. Hence the probability ($Pr_{COMP-15}$) that this function preserve the rotational property for a pair of rotational input such as $(X, X <<< 1)$ and produce a pair of rotational output such as $(COMP-15(X), COMP-15(X) <<< 1)$ will be:

$$Pr_{COMP-15} = (Pr_{ROUND})^{15} > 2^{-1009.68} >> 2^{-1024}$$

The $Pr_{COMP-15}$ is greater than 2^{-1024} (the probability that we can distinguish the CubeHash compression function when it is indistinguishable from a random function) and allows distinguishing 15-round CubeHash compression function using about 2^{1010} call of it. The distinguisher of COMP-15 based on rotational analysis technique, too and is working as the COMP-8 distinguisher.

6. Conclusion

In this paper we find a new distinguisher based on rotational analysis technique for CubeHash compression function and distinguish the compression functions of CubeHash-8/b and CubeHash-15/b with probability greater than $2^{-538.5}$ and $2^{-1009.7}$, respectively. If in the CubeHash-8/b compression function we set b=1, indeed we distinguish the compression function of CubeHash-8/1, the first version of CubeHash that suggested to SHA-3 competition, from a random function. Also if we set b=32 in CubeHash-15/b, then we distinguish a reduced version of the compression function of CubeHash-16/32 (a tweaked version of CubeHash-8/1) only with one round less, from a random function. The distinguisher of CubeHash-15/b compression function presented in this paper is the first distinguisher that distinguish the CubeHash compression function with more than 14 rounds from a random function.

References

[1] T. Ashur, O. Dunkelman, "Linear Analysis of Reduced-Round CubeHash", Cryptology ePrint Archive, Report 2010/535 (2010).
[2] D.J. Bernstein, "Cubehash", Submission to NIST, Round 2 (2009).
[3] D.J. Bernstein, "CubeHash parameter tweak: 16 times faster".
[4] E. Biham, A. Shamir, "Differential Cryptanalysis of DES-like Cryptosystems", J. Cryptology, 4(1):3-72 (1991).
[5] D. Gligoroski, V. Klima, S. J. Knapskog, M. El-Hadedy, J., S. Amundsen, F. Mj lsnes, "Cryptographic Hash Function BLUE MIDNIGHT WISH", Submission to NIST (Round 2), (2009). Available at http://people.item.ntnu.no/~danilog/Hash/BMW-SecondRound/Supporting_Documentation/BlueMidnightWishDocumentation.pdf.
[6] L. R. Knudsen, K. Matusiewicz, S. S. Thomsen, "Observations on the Shabal keyed permutation", OFFICIAL COMMENT, (2009). Available at http://www.mat.dtu.dk/people/S.Thomsen/shabal/shabal.pdf.
[7] D. Khovratovich, I. Nikolic, "Rotational Cryptanalysis of ARX", Fast Sotftware Encryption (FSE 2010), Springer (2010).
[8] G. Leurent, C. Bouillaguet, P.-A, Fouque, " SIMD Is a Message Digest", Submission to NIST (Round 2), (2009).
[9] M. Matsui, "Linear Cryptoanalysis Method for DES Cipher", T. Helleseth, editor, EUROCRYPT, volume 765 of Lecture Notes in Computer Science, pages 386-397. Springer, (1993).
[10] National Institute of Standards and Technology, Announcing Request for Candidate Algorithm Nominations for a New Cryptographic Hash Algorithm (SHA-3) Family, Federal Register Notice (November 2007), available online at: http://csrc.nist.gov
[11] I. Nikolić, J. Pieprzyk, P. Sokołowski, Ron. Steinfeld, "Rotational Cryptanalysis of (Modified) Versions of BMW and SIMD", Available online, (2010).
[12] ECRYPT II, The SHA-3 Zoo, CubeHash Profile, Available at: http://ehash.iaik.tugraz.at/wiki/CubeHash
[13] X. Wang, H. Yu, "How to Break MD5 and Other Hash Functions", Advances in Cryptology, EUROCRYPT 2005, LNCS, Springer-Verlag, (2005).
[14] X. Wang, H. Yu, Y.L. Yin, "Efficient Collision Search Attacks on SHA-0", Advances in Cryptology, Crypto 2005, LNCS 3621, Pages 1-16, Springer, (2005).
[15] X. Wang, Y.L. Yin, H. Yu, "Finding Collisions in the Full SHA-1", Advances in Cryptology, Crypto 2005, LNCS 3621, Pages 17- 36, Springer, (2005).

Javad Alizadeh received the Bachelor's degree in Applied Mathematics with the honor degree from IHU, Tehran, Iran in 2007 and Master's degree in Telecommunication in the field of

Cryptography with the honor degree from IHU, Tehran, Iran in 2010. He was chosen as a superior researcher student of IHU, in 2010. Currently, he is a researcher in the field of cryptography and teacher assistant (TA) at the faculty and research center of communication and information technology, IHU, Tehran, Iran. His research interest includes: Cryptography, Cryptanalysis, Information Systems Security, Mathematics of Cryptography. He is a member of ISC.

Abdolrasoul Mirghadri received the B.Sc., M.Sc. and PHD degrees in Mathematical Statistics, from the faculty of Science, Shiraz University in 1986, 1989 and 2001, respectively. He is an assistant professor at the faculty and research center of communication and information technology, IHU, Tehran, Iran since 1989. His research interest includes: Cryptography, Cryptanalysis, Statistics and Stochastic Processes. He is a member of ISC, ISS and IMS.

Appendix

A. Proofs of Lemma 4 to Lemma 7

Proof of lemma 4 ($SWAP_1$): We would like to prove that for $SWAP_1$ function by any input such as X:

$$\Pr[SWAP_1(X <<< 1) = SWAP_1(X) <<< 1] = 1.$$

For the proof suppose that $S = (S_0, S_1, ..., S_{31})$ be the state of CubeHash round function. By attention to definition of $SWAP_1$ in the step 3 of the round function we have

$SWAP_1$:
Swap S_i and $S_{i\oplus 8}$, for $0 \le i \le 7$.

In fact the $SWAP_1$ function operates on the left half of the S. Consider this half as

$$X = S_0 \| S_1 \| S_2 \| S_3 \| S_4 \| S_5 \| S_6 \| S_7 \| S_8 \| S_9 \| S_{10} \| S_{11} \| S_{12} \| S_{13} \| S_{14} \| S_{15}$$

where the means of the notation $\|$ is concatenating. Using the definition of X and $SWAP_1$ function we have

$$SWAP_1(X) = S_8 \| S_9 \| S_{10} \| S_{11} \| S_{12} \| S_{13} \| S_{14} \| S_{15} \| S_0 \| S_1 \| S_2 \| S_3 \| S_4 \| S_5 \| S_6 \| S_7.$$

Without loss of generality and for simplicity consider

$$X_1 = S_0 \| S_1 \| S_2 \| S_3 \| S_4 \| S_5 \| S_6 \| S_7,$$
$$X_2 = S_8 \| S_9 \| S_{10} \| S_{11} \| S_{12} \| S_{13} \| S_{14} \| S_{15},$$

where X_1 and X_2 have 512-bit length. Consider the bit representation of them as

$$X_1 = x_1^0 x_1^1 x_1^2 ..... x_1^{254} x_1^{255},$$
$$X_2 = x_2^0 x_2^1 x_2^2 ..... x_2^{254} x_2^{255}.$$

By rewriting the input of $SWAP_1$ function as $X = X_1 \| X_2$, we have $SWAP_1(X) = X_2 \| X_1$. Now we show that

$$\Pr[SWAP_1(X <<< 1) = SWAP_1(X) <<< 1] = 1.$$

In order to show it we compute

$$X <<< 1 = x_1^1 x_1^2 ..... x_1^{254} x_1^{255} x_2^0 \| x_2^1 x_2^2 ..... x_2^{254} x_2^{255} x_1^0,$$

$$SWAP_1(X <<< 1) = x_2^1 x_2^2 ..... x_2^{254} x_2^{255} x_1^0 \| x_1^1 x_1^2 ..... x_1^{254} x_1^{255} x_2^0, \quad (4.1)$$

and

$$SWAP_1(X) <<< 1 = x_2^1 x_2^2 ..... x_2^{254} x_2^{255} x_1^0 \| x_1^1 x_1^2 ..... x_1^{254} x_1^{255} x_2^0 \quad (4.2)$$

By attention to (4.1) and (4.2) we see that $SWAP_1(X <<< 1) = SWAP_1(X) <<< 1$ and for this equality no condition is need. Consequently, we showed that

$$\Pr[SWAP_1(X <<< 1) = SWAP_1(X) <<< 1] = 1.$$

■

Proof of lemma 5 ($SWAP_2$): We would like to prove that for $SWAP_2$ function by any input such as X:

$$\Pr[SWAP_2(X <<< 1) = SWAP_2(X) <<< 1] = 2^{-6}.$$

For the proof suppose that $S = (S_0, S_1, ..., S_{31})$ be the state of CubeHash round function. By attention to definition of $SWAP_2$ in the step 5 of the round function we have

$SWAP_2$:
Swap S_i and $S_{i\oplus 2}$, for $i \in \{16,17,20,21,24,25,28,29\}$.

In fact the $SWAP_2$ function operates on the Right half of the S. Consider this half as

$$X = S_{16} \square S_{17} \square S_{18} \square S_{19} \square S_{20} \square S_{21} \square S_{22} \square S_{23} \square S_{24} \square S_{25} \square S_{26} \square S_{27} \square S_{28} \square S_{29} \square S_{30} \square S_{31}$$

Using the definition of X and $SWAP_2$ function we have

$$SWAP_2(X) = S_{18} \square S_{19} \square S_{16} \square S_{17} \square S_{22} \square S_{23} \square S_{20} \square S_{21} \square S_{26} \square S_{27} \square S_{24} \square S_{25} \square S_{30} \square S_{31} \square S_{28} \square S_{29}$$

Without loss of generality and for simplicity consider

$$X_1 = S_{16} \square S_{17}, \quad X_5 = S_{24} \square S_{25},$$
$$X_2 = S_{18} \square S_{19}, \quad X_6 = S_{26} \square S_{27},$$
$$X_3 = S_{20} \square S_{21}, \quad X_7 = S_{28} \square S_{29},$$
$$X_4 = S_{22} \square S_{23}, \quad X_8 = S_{30} \square S_{31},$$

where $X_i, 1 \le i \le 8,$ have 64-bit length. Consider the bit representation of them as

$$X_i = x_i^0 x_i^1 x_i^2 .....x_i^{62} x_i^{63}$$

By rewriting the input of $SWAP_1$ function as

$$X = X_1 \square X_2 \square X_3 \square X_4 \square X_5 \square X_6 \square X_7 \square X_8,$$

we have

$$SWAP_2(X) = X_2 \square X_1 \square X_4 \square X_3 \square X_6 \square X_5 \square X_8 \square X_7.$$

Now we show that

$$\Pr[SWAP_2(X <<< 1) = SWAP_2(X) <<< 1] = 2^{-6}$$

In order to show it we compute

$$X <<< 1 = x_1^1 x_1^2 ... x_1^{63} x_2^0 \square x_2^1 x_2^2 ... x_2^{63} x_3^0 \square x_3^1 x_3^2 ... x_3^{63} x_4^0 \square x_4^1 x_4^2 ... x_4^{63} x_5^0 \square x_5^1 x_5^2 ... x_5^{63} x_6^0 \square x_6^1 x_6^2 ... x_6^{63} x_7^0 \square x_7^1 x_7^2 ... x_7^{63} x_8^0 \square x_8^1 x_8^2 ... x_8^{63} x_1^0$$

$$SWAP_2(X <<< 1) = x_2^1 x_2^2 ... x_2^{63} x_3^0 \square x_1^1 x_1^2 ... x_1^{63} x_2^0 \square x_4^1 x_4^2 ... x_4^{63} x_5^0 \square x_3^1 x_3^2 ... x_3^{63} x_4^0 \square x_6^1 x_6^2 ... x_6^{63} x_7^0 \square x_5^1 x_5^2 ... x_5^{63} x_6^0 \square x_8^1 x_8^2 ... x_8^{63} x_1^0 \square x_7^1 x_7^2 ... x_7^{63} x_8^0 \quad (5.1)$$

and

$$SWAP_2(X) <<< 1 = x_2^1 x_2^2 ... x_2^{63} x_1^0 \square x_1^1 x_1^2 ... x_1^{63} x_4^0 \square x_4^1 x_4^2 ... x_4^{63} x_3^0 \square x_3^1 x_3^2 ... x_3^{63} x_6^0 \square x_6^1 x_6^2 ... x_6^{63} x_5^0 \square x_5^1 x_5^2 ... x_5^{63} x_8^0 \square x_8^1 x_8^2 ... x_8^{63} x_7^0 \square x_7^1 x_7^2 ... x_7^{63} x_2^0 \quad (5.2)$$

By attention to (5.1) and (5.2) we see that for the equality of $SWAP_2(X <<< 1) = SWAP_2(X) <<< 1$ we need the following conditions on some bits of X:

$$x_1^0 = x_3^0 = x_5^0 = x_7^0$$

and

$$x_2^0 = x_4^0 = x_6^0 = x_8^0$$

Each of these conditions satisfies with probability of $\left(\frac{1}{2}\right)^3$. Consequently, we showed that

$$\Pr[SWAP_2(X <<< 1) = SWAP_2(X) <<< 1] = 2^{-6}.$$

■

Proof of lemma 6 ($SWAP_3$): We would like to prove that for $SWAP_3$ function by any input such as X:

$$\Pr[SWAP_3(X <<< 1) = SWAP_3(X) <<< 1] = 2^{-2}.$$

For the proof suppose that $S = (S_0, S_1, ..., S_{31})$ be the state of CubeHash round function. By attention to definition of $SWAP_3$ in the step 8 of the round function we have

$SWAP_3$:
Swap S_i and $S_{i \oplus 4}$, for $i \in \{0,1,2,3,8,9,10,11\}$

In fact the $SWAP_3$ function operates on the left half of the S. Consider this half as

$$X = S_0 \square S_1 \square S_2 \square S_3 \square S_4 \square S_5 \square S_6 \square S_7 \square S_8 \square S_9 \square S_{10} \square S_{11} \square S_{12} \square S_{13} \square S_{14} \square S_{15}$$

Using the definition of X and $SWAP_3$ function we have

$$SWAP_3(X) = S_4 \square S_5 \square S_6 \square S_7 \square S_0 \square S_1 \square S_2 \square S_3 \square S_{12} \square S_{13} \square S_{14} \square S_{15} \square S_8 \square S_9 \square S_{10} \square S_{11}$$

Without loss of generality and for simplicity consider

$$X_1 = S_0 \Box S_1 \Box S_2 \Box S_3,$$
$$X_2 = S_4 \Box S_5 \Box S_6 \Box S_7,$$
$$X_3 = S_8 \Box S_9 \Box S_{10} \Box S_{11},$$
$$X_4 = S_{12} \Box S_{13} \Box S_{14} \Box S_{15},$$

where $X_i, 1 \le i \le 4$, have 128-bit length. Consider the bit representation of them as

$$X_i = x_i^0 x_i^1 x_i^2 .....x_i^{126} x_i^{127}$$

By rewriting the input of $SWAP_3$ function as

$$X = X_1 \Box X_2 \Box X_3 \Box X_4,$$

we have

$$SWAP_3(X) = X_2 \Box X_1 \Box X_4 \Box X_3.$$

Now we show that

$$\Pr[SWAP_3(X <<< 1) = SWAP_3(X) <<< 1] = 2^{-2}$$

In order to show it we compute

$$X <<< 1 = x_1^1 x_1^2 ... x_1^{127} x_2^0 \Box x_2^1 x_2^2 ... x_2^{127} x_3^0 \Box x_3^1 x_3^2 ... x_3^{127} x_4^0 \Box x_4^1 x_4^2 ... x_4^{127} x_1^0,$$

$$SWAP_3(X <<< 1) = x_2^1 x_2^2 ... x_2^{127} x_3^0 \Box x_1^1 x_1^2 ... x_1^{127} x_2^0 \Box x_4^1 x_4^2 ... x_4^{127} x_1^0 \Box x_3^1 x_3^2 ... x_3^{127} x_4^0 \quad (6.1)$$

and

$$SWAP_3(X) <<< 1 = x_2^1 x_2^2 ... x_2^{127} x_1^0 \Box x_1^1 x_1^2 ... x_1^{127} x_4^0 \Box x_4^1 x_4^2 ... x_4^{127} x_3^0 \Box x_3^1 x_3^2 ... x_3^{127} x_2^0 \quad (6.2)$$

By attention to (6.1) and (6.2) we see that for the equality of $SWAP_3(X <<< 1) = SWAP_3(X) <<< 1$ we need the following conditions on some bits of X:

$$x_1^0 = x_3^0$$

and

$$x_2^0 = x_4^0$$

Each of these conditions satisfies with probability of $\frac{1}{2}$. Consequently, we showed that

$$\Pr[SWAP_3(X <<< 1) = SWAP_3(X) <<< 1] = 2^{-2}.$$

■

Proof of lemma 7 ($SWAP_4$): We would like to prove that for $SWAP_4$ function by any input such as X:

$$\Pr[SWAP_4(X <<< 1) = SWAP_4(X) <<< 1] = 2^{-14}$$

For the proof suppose that $S = (S_0, S_1, ..., S_{31})$ be the state of CubeHash round function. By attention to definition of $SWAP_4$ in the step 10 of the round function we have

$SWAP_4$:
Swap S_i and $S_{i \oplus 1}$, for $i \in \{16, 18, 20, 22, 24, 26, 28, 30\}$

In fact the $SWAP_4$ function operates on the right half of the S. Consider this half as

$$X = S_{16} \Box S_{17} \Box S_{18} \Box S_{19} \Box S_{20} \Box S_{21} \Box S_{22} \Box S_{23} \Box S_{24} \Box S_{25} \Box S_{26} \Box S_{27} \Box S_{28} \Box S_{29} \Box S_{30} \Box S_{31}$$

Using the definition of X and $SWAP_3$ function we have

$$SWAP_4(X) = S_{17} \Box S_{16} \Box S_{19} \Box S_{18} \Box S_{21} \Box S_{20} \Box S_{23} \Box S_{22} \Box S_{25} \Box S_{24} \Box S_{27} \Box S_{26} \Box S_{29} \Box S_{28} \Box S_{31} \Box S_{30}$$

Without loss of generality and for simplicity and also similarity to the previous proofs, consider

$$X_1 = S_{16}, \quad X_9 = S_{24},$$
$$X_2 = S_{17}, \quad X_{10} = S_{25},$$
$$X_3 = S_{18}, \quad X_{11} = S_{26},$$
$$X_4 = S_{19}, \quad X_{12} = S_{27},$$
$$X_5 = S_{20}, \quad X_{13} = S_{28},$$
$$X_6 = S_{21}, \quad X_{14} = S_{29},$$
$$X_7 = S_{22}, \quad X_{15} = S_{30},$$
$$X_8 = S_{23}, \quad X_{16} = S_{31},$$

where $X_i, 1 \le i \le 16$, have 32-bit length. Consider the bit representation of them as

$$X_i = x_i^0 x_i^1 x_i^2 .....x_i^{30} x_i^{31}$$

By rewriting the input of $SWAP_3$ function as

$$X = X_1 \| X_2 \| X_3 \| X_4 \| X_5 \| X_6 \| X_7 \| X_8 \| X_9 \| X_{10} \| X_{11} \| X_{12} \| X_{13} \| X_{14} \| X_{15} \| X_{16},$$

we have

$$SWAP_4(X) = X_2 \| X_1 \| X_4 \| X_3 \| X_6 \| X_5 \| X_8 \| X_7 \| X_{10} \| X_9 \| X_{12} \| X_{11} \| X_{14} \| X_{13} \| X_{16} \| X_{15}$$

Now we show that

$$\Pr[SWAP_4(X <<< 1) = SWAP_4(X) <<< 1] = 2^{-14}$$

In order to show it we compute

$$\begin{aligned} X <<< 1 = {} & x_1^1 x_1^2 \ldots x_1^{31} x_2^0 \| x_2^1 x_2^2 \ldots x_2^{31} x_3^0 \| x_3^1 x_3^2 \ldots x_3^{31} x_4^0 \\ & \| x_4^1 x_4^2 \ldots x_4^{31} x_5^0 \| x_5^1 x_5^2 \ldots x_5^{31} x_6^0 \| x_6^1 x_6^2 \ldots x_6^{31} x_7^0 \\ & \| x_7^1 x_7^2 \ldots x_7^{31} x_8^0 \| x_8^1 x_8^2 \ldots x_8^{31} x_9^0 \| x_9^1 x_9^2 \ldots x_9^{63} x_{10}^0 \\ & \| x_{10}^1 x_{10}^2 \ldots x_{10}^{63} x_{11}^0 \| x_{11}^1 x_{11}^2 \ldots x_{11}^{63} x_{12}^0 \| x_{12}^1 x_{12}^2 \ldots x_{12}^{63} x_{13}^0 \\ & \| x_{13}^1 x_{13}^2 \ldots x_{13}^{63} x_{14}^0 \| x_{14}^1 x_{14}^2 \ldots x_{14}^{63} x_{15}^0 \| x_{15}^1 x_{15}^2 \ldots x_{15}^{63} x_{16}^0 \\ & \| x_{16}^1 x_{16}^2 \ldots x_{16}^{63} x_1^0 \end{aligned}$$

$$\begin{aligned} SWAP_4(X <<< 1) = {} & x_2^1 x_2^2 \ldots x_2^{31} x_3^0 \| x_1^1 x_1^2 \ldots x_1^{31} x_2^0 \| x_4^1 x_4^2 \ldots x_4^{31} x_5^0 \\ & \| x_3^1 x_3^2 \ldots x_3^{31} x_4^0 \| x_6^1 x_6^2 \ldots x_6^{31} x_7^0 \| x_5^1 x_5^2 \ldots x_5^{31} x_6^0 \\ & \| x_8^1 x_8^2 \ldots x_8^{31} x_9^0 \| x_7^1 x_7^2 \ldots x_7^{31} x_8^0 \| x_{10}^1 x_{10}^2 \ldots x_{10}^{63} x_{11}^0 \\ & \| x_9^1 x_9^2 \ldots x_9^{63} x_{10}^0 \| x_{12}^1 x_{12}^2 \ldots x_{12}^{63} x_{13}^0 \| x_{11}^1 x_{11}^2 \ldots x_{11}^{63} x_{12}^0 \\ & \| x_{14}^1 x_{14}^2 \ldots x_{14}^{63} x_{15}^0 \| x_{13}^1 x_{13}^2 \ldots x_{13}^{63} x_{14}^0 \| x_{16}^1 x_{16}^2 \ldots x_{16}^{63} x_1^0 \\ & \| x_{15}^1 x_{15}^2 \ldots x_{15}^{63} x_{16}^0 \end{aligned} \quad (7.1)$$

and

$$\begin{aligned} SWAP_4(X) <<< 1 = {} & x_2^1 x_2^2 \ldots x_2^{31} x_1^0 \| x_1^1 x_1^2 \ldots x_1^{31} x_4^0 \| x_4^1 x_4^2 \ldots x_4^{31} x_3^0 \\ & \| x_3^1 x_3^2 \ldots x_3^{31} x_6^0 \| x_6^1 x_6^2 \ldots x_6^{31} x_5^0 \| x_5^1 x_5^2 \ldots x_5^{31} x_8^0 \\ & \| x_8^1 x_8^2 \ldots x_8^{31} x_7^0 \| x_7^1 x_7^2 \ldots x_7^{31} x_{10}^0 \| x_{10}^1 x_{10}^2 \ldots x_{10}^{63} x_9^0 \\ & \| x_9^1 x_9^2 \ldots x_9^{63} x_{12}^0 \| x_{12}^1 x_{12}^2 \ldots x_{12}^{63} x_{11}^0 \| x_{11}^1 x_{11}^2 \ldots x_{11}^{63} x_{14}^0 \\ & \| x_{14}^1 x_{14}^2 \ldots x_{14}^{63} x_{13}^0 \| x_{13}^1 x_{13}^2 \ldots x_{13}^{63} x_{16}^0 \| x_{16}^1 x_{16}^2 \ldots x_{16}^{63} x_{15}^0 \\ & \| x_{15}^1 x_{15}^2 \ldots x_{15}^{63} x_2^0 \end{aligned} \quad (7.2)$$

By attention to (7.1) and (7.2) we see that for the equality of $SWAP_4(X <<< 1) = SWAP_4(X) <<< 1$ we need the following conditions on some bits of X:

$$x_1^0 = x_3^0 = x_5^0 = x_7^0 = x_9^0 = x_{11}^0 = x_{13}^0 = x_{15}^0$$

and

$$x_2^0 = x_4^0 = x_6^0 = x_8^0 = x_{10}^0 = x_{12}^0 = x_{14}^0 = x_{16}^0$$

Each of these conditions satisfies with probability of $\left(\frac{1}{2}\right)^7$. Consequently, we showed that

$$\Pr[SWAP_4(X <<< 1) = SWAP_4(X) <<< 1] = 2^{-14}.$$

■

An Automated Approach to Embrace Changes During Use case Model Evolution

Dr. Amer AbuAli
Department of Software Engineering,
Faculty of Information Technology
P.O. Box 1 Philadelphia University, 19392, Amman, Jordan.
Tel: 00 962 6 4799000

Abstract:
Use case model is subject to changes throughout the software development life cycle. Impacts of these changes affect directly the requirements and consequently the resulted system. Scrapping and replacing use case is expensive; in this paper we proposed a solution that integrates changes in use case in requirement phase. This solution combines independent enhancements to some version of a use case into a new version that include the enhancements and the old use case. CASE tool implementation and experimental evaluation of the proposed approach showed promising results in terms of software development time saving and better use case models integrity.

Keywords: Requirement engineering, Functional requirements, Use case changes, Use case evolution.

1. Introduction

Understanding the requirements of a problem is among the most difficult tasks that face a software engineer. Requirement engineering (RE) helps software engineers to better understand the problem they will work to solve. It encompasses the set of tasks that lead to an understanding of what the business impact of the software will be, what the customer wants, and how end-users will interact with the software [1, 2].
Most of the changes into software can be traced back to the early requirements stage when a recovery action can still be cost-effective [3]. Such changes may become necessary because of changes in the real-world context in which the proposed system would be situated or because of changes in stakeholder perceptions of the proposed system. Requirements Evolution involves updating a description of user requirements for a target system to accommodate new requirements or to remove existing ones [4, 5].
To capture functional requirements, that are statements of the services that the system must provide or are descriptions of how some computations must be carried out [6, 7], the widespread practice is the use case model. It describes the functional requirements of a software system and is used as input to several activities in a software development project. It gives a high-level view of the requirements of a system. The quality of the use case model therefore has an important impact on the quality of software [8].
Use case model is subject to changes sometimes later in software life cycle. Changes are due to 1) market demands, such as a large customer wanting things done their way; 2) business requirement change, such as new policies or operational processes; 3) legislative and regulatory change; and 4) imaginative users. Impacts of these changes affect directly the requirements and consequently the product [3, 9].
Here, we faced two problems: (1) scrapping and replacing use cases or (2) merging changes in order to create new use case. The former is more expensive, we propose an original solution to the second problem.
Use case merging is essential to deal with parallel modifications carried out by different requirement engineers that are not necessarily aware of each other's changes. Our solution combines various independent enhancements of a given version of a use case into a new use case that includes the semantics of both the enhancements and the old use case. In this context, changes are brought in separate copies of the old use case. Copies as well as the old use case are compared and merged in order to produce a new version including all modifications. This approach provides computer aid for combining the results of several people's separate efforts. This approach is inspired from our previous researches in software merging where we have proposed a new approach for program integration [10, 11].
The outline of this paper is as follows: Section 2 presents a background about the use case concept. Section 3 discusses our approach to use case modelling. Section 4 illustrates our identification of changes by an example. Section 5 shows the manner of merging use cases. Sections 6 and 7 demonstrate the tool support and experimental use of the proposed approach. Finally, Section 8 concludes our research direction.

2. Background

Employment of use cases is now common practice in software development, and use case is now a recognized concept in development processes [12, 13].

A use case is an object-oriented modeling construct that is used to define the behavior of a system. Interactions between the user and the system are described through a prototypical course of actions along with a possible set of alternative courses of action. Primarily, use cases have been associated with requirements gathering and domain analysis. However, with the release of the Unified Modeling Language (UML) specification version 1.5 [14], the scope of use cases has broadened to include modeling constructs at all levels. Due to this expanded scope, the representation of use cases has taken on increasing importance.

A *use case* defines a goal-oriented set of interactions between external actors and the system under consideration. *Actors* are parties outside the system that interact with the system [14]. An actor may be a class of users, roles users can play, or other systems. A use case is initiated by a user with a particular goal in mind, and completes successfully when that goal is satisfied. It describes the sequence of interactions between actors and the system necessary to deliver the service that satisfies the goal. It also includes possible variants of this sequence, e.g., alternative sequences that may also satisfy the goal, as well as sequences that may lead to failure to complete the service because of exceptional behavior, error handling, etc. The system is treated as a "black box", and the interactions with system, including system responses, are as perceived from outside the system [12, 13].

According to UML version 1.5 [14] we describe, briefly, the types of relationships of use case as below:

i) Actor relationships

There is one standard relationship among actors and one between actors and use cases, called generalization and association respectively. A generalization from an actor A to an actor B indicates that an instance of A can communicate with the same kinds of use-case instances as an instance of B. Association is related to the participation of an actor in a use case, i.e. instances of the actor and instances of the use case communicate with each other.

ii) Use case relationships

In addition to the association, described previously, there are several standard relationships among use cases or between actors and use cases. A generalization from use case A to use case B indicates that A is a specialization of B. An extend relationship from use case A to use case B indicates that an instance of use case B may be augmented (subject to specific conditions specified in the extension) by the behavior specified by A. The behavior is inserted at the location defined by the extension point in B which is referenced by the extend relationship. While an include relationship from use case A to use case B indicates that an instance of the use case A will also contain the behavior as specified by B. The behavior is included at the location which defined in A.

3. Use case modelling

To permit an automatic use case analysis which is implicit in the conventional representation, an explicit representation needs an internal form.

3.1 Internal form

In the context of use case understanding and modification (evolution), a dependence relationship of a use case model is defined formally by the 5-tuples:

<As, At, Rel, Typ, Id>.

It means that target actor/use case **At** depends on actor/use **As** according to the relationship **Rel** with the type **Typ** for the relationship **Id**.

Rel is a relationship that can be a generalization between actors/use cases (**Gen**), an association between actor and use case (**Ass**), an extend (**Ext**), or an include between use cases (**Inc**).

Typ is dedicated to the type of multiplicity in a given association (n..m), it is the number of possible instances of actors associated with a single instance of use case.

Id is a unique identifier corresponding to relationship number**.**

3.2 Modeling

3.2.1 Modeling actor relationships

An association is formalized by the following 5-tuplet: <Actor, Use case, Ass, Mul, Id>.

It means that instances of **Actor** and instances of **Use case** communicate with a multiplicity **Mul** in the association **Ass** numbered **Id**.

A generalization between actors is represented by: <Actor1, Actor2, Gen, ϕ, Id>

It expresses that **Actor2** inherits (Gen) from **Actor1** in the relationship numbered **Id**.

3.2.2 Modeling use case relationships

We express a generalization between use cases by: <Use case1, Use case2, Gen, ϕ, Id>

It means that **Use case2** inherits (Gen) from **Use case1** in the relationship numbered **Id**.

An include relationship is expressed by: < Use case1, Use case2, Inc, ϕ, Id>

It indicates that an instance of the **Use case1** will also contain the behavior as specified by **Use case2** in the relationship numbered **Id**.

An extend relationship is expressed by: <Use case1, Use case2, Ext, ϕ, Id>

It means that an instance of **Use case2** may be augmented (**Ext**) by the behavior specified by **Use case1** in the relationship numbered **Id**.

In order to illustrate our approach Figure 1 presents the use case of an ordering system and the corresponding internal form (Table1).

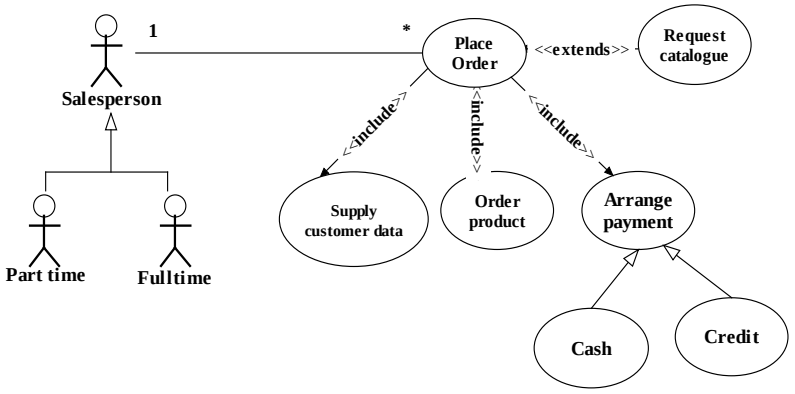


Fig. 1. Use case of an ordering system and its internal form of use case.

< Salesperson, Part time, Gen, ϕ, 1>,
< Salesperson, Full time, Gen, ϕ, 2>,
< Salesperson, Place Order, Ass, (1, *), 3>,
<Place Order, Supply Customer Data, Inc, ϕ, 4>,
<Place Order, Order Product, Inc, ϕ, 5>,
<Place Order, Arrange Payment, Inc, ϕ, 6>,
<Place Order, Request Catalogue, Ext, ϕ, 7>,
<Cash, Arrange Payment, Gen, ϕ, 8>,
<Credit, Arrange Payment, Gen, ϕ, 9>.

Table 1. Internal form of ordering system use case

4. Identification of changes

Use case changes can be syntactic or semantic. Syntactic changes concern changes of actors, use cases, and relationship names. Semantic changes concern semantic changes of actors, use cases, and relationships.

Semantic changes of actors/use cases can be adding/deleting actors/use cases. Semantic changes of relationships not only occur with previous changes but also with redirecting edges or changing type of relationships.

In order to illustrate this approach, we propose to apply it in the following example. According to use case model **Base** of figure 1, two variants are proposed. In variant A (figure 2), we add a new actor "Trainee", change the multiplicity (1..*) by (1..5), and change "Salesperson" by "Salesperson Team". In other words we make two semantic changes and one syntactic change, namely adding new actor changing the multiplicity and renaming "Salesperson" by "Salesperson Team" Table 2 gives the internal form of variant A. In variant B (figure 3), we add a new use case "Log in", redirect use case "Request catalogue" to actors and make syntactic change of "cash" and "credit" by "cash payment" and "credit payment". Table 3 gives the internal form of variant B.

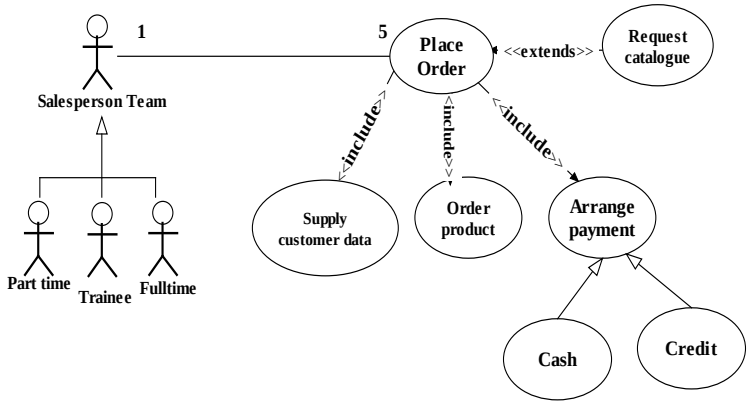


Fig. 2. Variant A of use case Base**.**

< Salesperson Team, Part time, Gen, ϕ, 1>,
< Salesperson Team, Full time, Gen, ϕ, 2>,
< Salesperson, Place Order, Ass, (1, 5), 3>,
<Place Order, Supply Customer Data, Inc, ϕ, 4>,
<Place Order, Order Product, Inc, ϕ, 5>,
<Place Order, Arrange Payment, Inc, ϕ, 6>,
<Place Order, Request Catalogue, Ext, ϕ, 7>,
<Cash, Arrange Payment, Gen, ϕ, 8>,
<Credit, Arrange Payment, Gen, ϕ, 9>
< Salesperson Team, Trainee, Gen, ϕ, 10>.

Table 2. Internal form of Variant A.

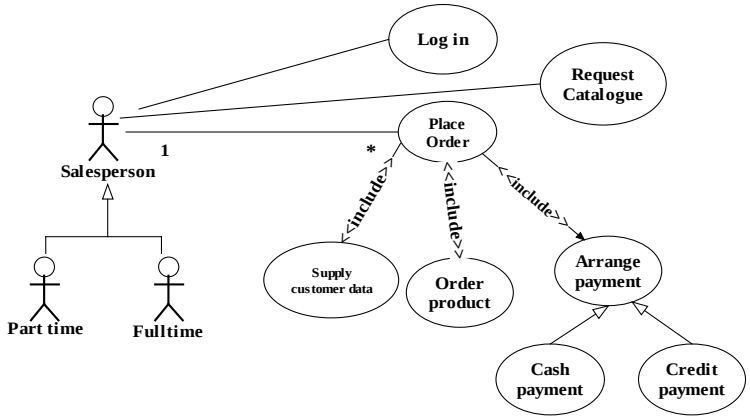


Fig. 3. Variant B of use case Base

< Salesperson, Part time, Gen, ϕ, 1>,
< Salesperson, Full time, Gen, ϕ, 2>,
< Salesperson, Place Order, Ass, (1, *), 3>,
<Place Order, Supply Customer Data, Inc, ϕ, 4>,
<Place Order, Order Product, Inc, ϕ, 5>,
<Place Order, Arrange Payment, Inc, ϕ, 6>,
<Cash Payment, Arrange Payment, Gen, ϕ, 8>,
<Credit Payment, Arrange Payment, Gen, ϕ, 9>
< Salesperson, Log in, Ass, (ϕ), 10>,
< Salesperson, Request Catalogue, Ass, (ϕ), 11>.

Table 3. Internal form of Variant B.

4.1. Actors changes

Actor changes concern the change of name (syntactic) or the behavior (semantic) of a given actor. Semantic changes can be adding, deleting actors, and/or redirecting the relationships from these actors. By comparing actors of each variant according to actors of use case Base, we can identify actor changes. Changes are grouped in four sets: UA, ACN, ACB, and ACNB.

UA set contains Unaltered Actors in all use cases. This concerns actors keeping the same name and the same behavior in all variants. Informally, it is interpreted by the same internal form (5-tuples) of actors all in variants.
Let A= (A_{11}, A_{12},…, A_{ij}, …, A_{nm}) to denote actor i in the use case model j
UA = { A_{ij} / <A_{ij}, A_{kj}, Rel_{ij}, Typ_{ij}, Id_{ij} >$_{Base}$ = <$A_{i'j'}$, $A_{k'j'}$, $Rel_{i'j'}$, $Typ_{i'j'}$, $Id_{i'j'}$>$_{variant}$ }
In our example, this set is concerned by the following actors: "Full Time" and "Part Time".

ACN is the set of Actors with Changed Names, but keeping the same behavior. Informally, it is interpreted by changing only the name of actor in the specific 5-tuples.

$ACN = \{ A_{ij} / <A_{ij}, A_{kl}, Rel_{ij}, Typ_{ij}, Id_{ij} >_{Base} = <A_{i'j'}, A_{k'l'}, Rel_{i'j'}, Typ_{i'j'}, Id_{i'j'}>_{variant} \wedge "A"_{ij} \neq "A"_{i'j'}\}$
In our example, ACN is concerned by actors: Salesperson replaced by Salesperson Team in variant A.

ACB is the set of Actors with Changed Behaviors but keeping the same names. As stated previously this concerns redirecting relationships or changing the relationship types.
$ACB = \{\forall A_{ij} / <A_{ij}, A_{kl}, Rel_{ij}, Typ_{ij}, Id_{ij} >_{Base} \neq <A_{i'j'}, A_{k'l'}, Rel_{i'j'}, Typ_{i'j'}, Id_{i'j'}>_{variant} \wedge " A"_{ij} = "A"_{i'j'}"\}$
In variant A, ACB is concerned by a new inheritance between "Salesperson" and the new actor "Trainee", while in variant B we have added two associations (with "Request catalogue" and "Log in").

ACNB (Actors with Changed Names and Behavior) set is concerned by adding/deleting actors.
$ACNB = \{\forall A_{ij} / <A_{ij}, A_{kl}, Rel_{ij}, Typ_{ij}, Id_{ij} >_{Base} \neq <A_{i'j'}, A_{k'l'}, Rel_{i'j'}, Typ_{i'j'}, Id_{i'j'}>_{variant} \}$
In our example, ACB is concerned by adding a new actor "Trainee".
We note that A_{kl} and $A_{k'j'}$ can be actor or use case.

4.2. Use cases changes
Use case changes is concerned by syntactic change or semantic of a given use case. Semantic changes can be adding, deleting use cases, and/or redirecting the relationships from these use cases. By comparing use cases of each variant according to use cases of Base, we can identify use cases changes. Changes are grouped in four sets: UUC, UUCN, UUCB, and UUCNB.
UUC set contains Unaltered Use Cases. This concerns use cases keeping the same name and the same behavior in all variants. Informally, it is interpreted by the same internal form (5-tuples) of this use case in variants.
Let $UU= (UU_{11}, UU_{12},\ldots, UU_{ij}, \ldots, UU_{nm})$ to denote use case i in the use case model j
$UUC = \{U_{ij} / <U_{ij}, U_{kj}, Rel_{ij}, Typ_{ij}, Id_{ij}>_{Base} = <U_{i'j'}, U_{k'j'}, Rel_{i'j'}, Typ_{i'j'}, Id_{i'j'}>_{variant} \}$
In our example, this set is concerned by the following use cases: "Place Order", "Supply customer data", "Order Product", and "Arrange payment".

UUCN is the set of Use Cases with Changed Names, but keeping the same behavior. Informally, it is interpreted by changing only the name of actor in the specific 5-tuples.
$UUCN = \{U_{ij} / <U_{ij}, U_{kl}, Rel_{ij}, Typ_{ij}, Id_{ij} >_{Base} = <U_{i'j'}, U_{k'l'}, Rel_{i'j'}, Typ_{i'j'}, Id_{i'j'}>_{variant} \wedge "U"_{ij} \neq "U"_{ij}\}$
In our example, UUCN is concerned by use cases Cash and Credit replaced by Cash payment and Credit payment in variant B.

UUCB is the set of use cases with Changed Behaviors but keeping the same names, this concerns redirecting relationships or changing the relationship types.

$UUCB = \{ U_{ij} / <U_{ij}, U_{kl}, Rel_{ij}, Typ_{ij}, Id_{ij}>_{Base} \neq <U_{i'j'}, U_{k'l'}, Rel_{i'j'}, Typ_{i'j'}, Id_{i'j'}>_{variant} \wedge " "U_{ij}"= "U_{ij}"\}$.

In variant B, use case "Request Catalogue" is redirected to "Salesperson" instead of "Place Order" and the type of relationship (<<include>>) is changed into a normal association.

UUCNB (Use Cases with Changed Names and Behavior) set is concerned by adding/deleting use cases.
$UUCNB = \{U_{ij} / <U_{ij}, U_{kl}, Rel_{ij}, Typ_{ij}, Id_{ij} >_{Base} \neq <U_{i'j'}, U_{k'l'}, Rel_{i'j'}, Typ_{i'j'}, Id_{i'j'}>_{variant}\}$
In our example, UUCNB is concerned by adding a new use case: "Log in" in variant B.
We note that U_{kl} and $U_{k'j'}$ can be actor or use case.

4.3 Relationships changes

Also relationship changes concern the syntactic change or semantic of a given relationship. Semantic changes can be adding, deleting, and/or redirecting relationships. By comparing relationships of each variant according to relationships of use case Base, we can identify relationship changes. Changes are grouped in four sets: UR, RCN, RCB, and RCNB.

UR set contains Unaltered Relationships in all use cases. This concerns actors keeping the same name and the same behavior in all variants. Informally, it is interpreted by the same internal form (5-tuples) of this relationship in variants.
Let $R= (R_{11}, R_{12},\ldots, R_{ij}, \ldots, R_{nm})$ to denote relationship i in the use case model j.
$UR= \{R_{ij} / <A_{ij}, A_{kl}, Rel_{ij}, Typ_{ij}, Id_{ij} >_{Base} = <A_{i'j'}, A_{k'l'}, Rel_{i'j'}, Typ_{i'j'}, Id_{i'j'}>_{variant}\}$

In the example, this set is concerned by an inheritance from "Part time" and "Full time" to "Salesperson", <<include>> associations from "Supply customer data", "Order Product", and "Arrange payment" to "Place Order", and an inheritance from "Cash" and "Credit" to "Arrange payment".

RCN is the set of Relationships with Changed Names, but keeping the same behavior. Informally, it is interpreted by changing only the name of relationship in the specific 5-tuples.
$RCN = \{A_{ij} / <A_{ij}, A_{kl}, Rel_{ij}, Typ_{ij}, Id_{ij} >_{Base} = <A_{i'j'}, A_{k'l'}, Rel_{i'j'}, Typ_{i'j'}, Id_{i'j'}>_{variant} \wedge "Rel"_{ij} \neq "Rel"_{i'j'}\}$

RCB is the set of Relationships with Changed Behaviors but keeping the same names. As stated previously this concerns redirecting relationships or changing the relationship types.
$RCB = \{A_{ij} / <A_{ij}, A_{kl}, Rel_{ij}, Typ_{ij}, Id_{ij} >_{Base} \neq <A_{i'j'}, A_{k'l'}, Rel_{i'j'}, Typ_{i'j'}, Id_{i'j'}>_{variant} \wedge "Rel"_{ij} = "Rel"_{I'j'}\}$

In variant A there is a change of multiplicity with use case "Place Order" (1..5 instead of 1..*). In varian

we have three new relationships: two associations with use cases "Log in" and "Request catalogue", and an inheritance with a new actor "Trainee".

RCNB (Relationships with Changed Names and Behavior) set is concerned by adding/deleting relationships.
RCNB = $\{A_{ij} / <A_{ij}, A_{kl}, Rel_{ij}, Typ_{ij}, Id_{ij}>_{Base} \neq <A_{i'j'}, A_{k'l'}, Rel_{i'j'}, Typ_{i'j'}, Id_{i'j'}>_{variant}\}$
In our example, RCNB is concerned by (1) adding new association from "Log in" to "Salesperson" redirecting and changing the <<include>> association between "Request Catalogue" and "Salesperson" into a normal association.

5. Generation of the new version of use case

We generate the new version of use case according changes identified previously. Unaltered actors, use cases, and relationships sets (UA, UUC, and UR) are kept. Actors, use cases, and relationships with changed names in variants replace corresponding actors, use cases, and relationships of Base (from ACN, UUCN, and RCN). Actors, use cases, and relationships with changed behavior of variants replace corresponding actors, use cases, and relationships of Base (from ACB, UUCB, and RCB). Actors, use cases, and relationships with changed names and behaviors of variants, interpreted by insertions or deletions, are inserted or deleted (from ACNB, UUCNB, and RCNB). Finally we obtain an internal form corresponding to the new use case (Table 4).
However there is a possible way in which we can fail to represent a satisfactory merged use case model. In Software merging [15, 16] these are referred as "Type I and Type II interference". Type I occurs when we make the same changes to the same actor, use case, relationship or multiplicity in different variants. In this case what is the change handled in the new version? Type II interference occurs when reconstituting the merged use case diagram from the internal form, it can be an infeasible graph.
If there are no interferences we can reconstitute the new use case diagram. Figure 4 illustrates a reconstitution of use case diagram from the internal form of Table 4.

< Salesperson, Part time, Gen, ϕ, 1>,
< Salesperson, Full time, Gen, ϕ, 2>,
< Salesperson, Place Order, Ass, (1, *), 3>,
<Place Order, Supply Customer Data, Inc, ϕ, 4>,
<Place Order, Order Product, Inc, ϕ, 5>,
<Place Order, Arrange Payment, Inc, ϕ, 6>,
<Cash Payment, Arrange Payment, Gen, ϕ, 8>,
<Credit Payment, Arrange Payment, Gen, ϕ, 9>
< Salesperson Team, Trainee, Gen, ϕ, 10>,
< Salesperson, Log in, Ass, (ϕ), 11>,
< Salesperson, Request Catalogue, Ass, (ϕ), 12>.

Table 4. Internal form of the new version of use case

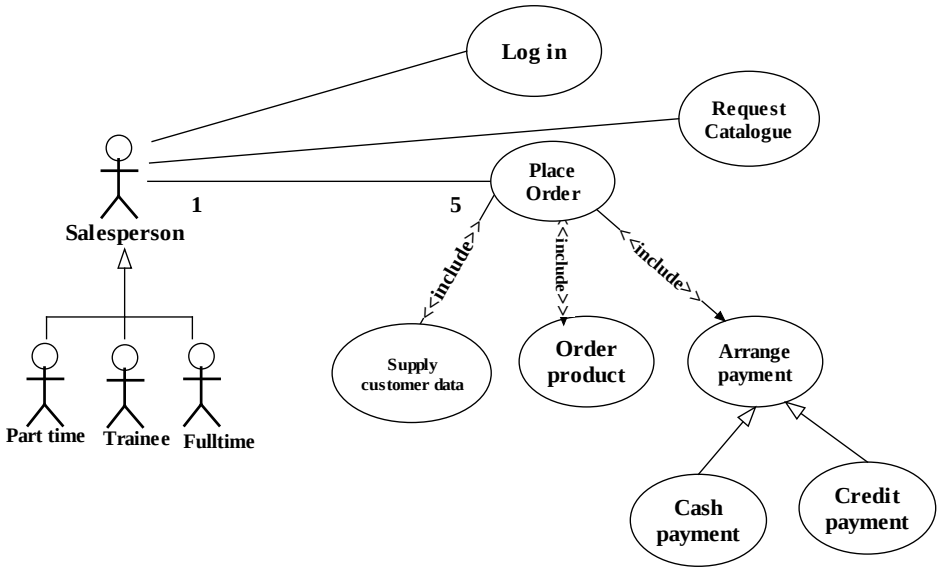


Fig. 4. The new version of use case diagram

6. Automation of Proposed Approach

The automated support of the proposed approach passed through a number of rationales. Examples include: what type of automation should be supported?, what development approach should be adopted?
The main debate that faced the research team is related to the software development approach. Two options were available: 1) develop limited capabilities stand alone CASE tool, and 2) develop an integrated shell for an already existing CASE tool. The 4+1 architectural views [17] suggest that any system has five views: design, implementation, process, deployment and use case. Activities within a view require information from other views. Elements from one view depend on or be driven by those of another. Moreover, the views may need to be ordered so that the information shared between two or more views remains consistent. An exception to this rule occurs with the use case view which is defined to drive the development of other system views. As the main output of the proposed approach is the use case model of the anticipated system and being a core model in software development, it was decided to go with the second development approach to support development of other system models using facilities of underlying CASE tool.

A survey on available CASE tools identified a number of commercial [18,19,20] and open source [21,22,23] CASE tools. Commercial tools (e.g. Rational Rose) ruled out of the candidate tools list due to expensive licensing cost which will inhibit accessibility of our approach to large number of users who are unable or unwilling to pay licensing cost. Therefore, three open source tools were short listed: StarUML [21], ArgoUML [22], and Netbeans Plug-ins [23]. Table 5 compares the features of the three tools. StarUML, as can be concluded from table 5, supersedes the other two tools in a number of factors. Hence, it was selected as a platform for the automation of the proposed approach

.

Table 5. CASE Tools Comparisons Summary**.**

Tool	UML Supported Version	Help and User Support Available	All Diagrams Supported	Portable?	Maintainable, Usable, and Extensible?	Support to Recent Trends in Software Modelling (e.g. MDA, NX)
StarUML	2.0	Yes	Yes	Yes	High	Yes
ArgoUML	1.4	No	Yes	Yes	Med.	No
Netbeans	1.4	No	No	No	Low	No

7. The Proposed Approach in Operation

TestWarehouse is a medium size software house. The main unit of software development projects is a team. Each team consists of up to 18 resources of different roles: project manager(s), IT technical support officer(s), system and business analysts, developers, and software quality engineers.

The adopted software development process in TestWarehouse projects differs from one project to another according to project context including project type, technical experience, application domain, delivery constraints, resources, and surrounding risks. However, the software development processes recently used in TestWarehouse are: eXtreme Programming, Scrum, and Rational Unified Process (RUP). These software process models are use case based and embrace frequent requirements changes which make them good test bed for the proposed use case evolution approach.

The proposed approach has been in operation for 8 months and utilized by TestWarehouse's business and system analysts in six projects. Table 6 demonstrates projects demographics in relation to project size, type, application domain, number of use cases, and number of requirements changes.

The main reported advantage of using the catalogue was the noticeable time saving in requirement engineering phase. This is attributed to reusability of use cases. Reported time saving percentages varied between 8% and 25% of the total software development project time. Analyzing the reasons behind the high fluctuation in reported time saving percentages, it was found that this is attributed to a number of factors including: (1) number of use cases in the project, and (2) use case complexity. In addition, users reported that models generated using the

proposed approach possess better completeness and comprehensiveness characteristics.

Table 6. Demographics of Experimental Projects**.**

Project	Size	Type	Application Domain	Number of Use Cases	Number of Requirements Changes
1	Small	In house development	Human Resources	20	10
2	Medium	Product	Financial	45	18
3	Medium	Custom Development	e-Commerce	42	21
4	Medium	Outsourcing	Financial	50	20
5	Large	Support Project	CRM Software	80	35
6	Large	Custom Development	Insurance	71	40

8. Conclusion and future work

Use case model is subject to changes sometimes later in software life cycle. Impacts of these changes affect directly the requirements and consequently the resulted system. Scrapping and replacing use case is expensive; in this paper we have proposed an original solution to integrate changes in old use case in requirement phase.

This solution is based on (1) an internal form to represent formally dependencies between concepts of use cases, (2) identification of changes from this internal form, and (3) merging old use cases diagrams in order to obtain a new version that takes account all modifications if there is no conflict.

The proposed approach has been implemented on top of an open source CASE tool. Actual experimental work of the automated proposed approach showed its ability to save up to 25% of software development time with better completeness and comprehensiveness characteristics.

As a future work, we plan to incorporate this technique of modification to the next diagrams of Object Oriented Analysis and Design (interaction diagrams, state diagrams, activity Diagrams, etc.). In addition, further testing using further projects and users is planned to take place for this approach.

9. References

1. Nguyen L, Swatman PA (2003) Managing Requirement Engineering Process. Requirement Engineering Journal Volume 8 No 1: pp. 55-68.
2. Lamsweerde VA (2000) Requirements engineering in the year 00: A research perspective. ICSE'2000 pp. 519 Ireland.
3. Rolland C, Salinesi C, Etien A (2004) Eliciting gaps in Requirements Change. Requirement Engineering Journal Volume 9 Number 1 pp. 1-15.
4. Anderson S, Felici M (2000) Requirements changes risk/cost analyses: An avionics case study. SRA-EUROPE Annual Conference, Volume 2 pp. 921-925 Scotland.
5. PROTEUS Project (1996) Meeting the challenge of changing requirements. Deliverable 1.3, Centre for Software Reliability, University of Newcastle.
6. Ghose A (1999) Managing Requirements Evolution: Formal Support for Functional and Non-functional Requirement. IWPSE'1999 pp. 118-124 Japan.
7. Anderson S, Felici M (2001) Requirements evolution: From process to product oriented management. PROFES'2001.
8. Lascio L (2002) Towards an inspection technique for use case models. SEKE '02 pp. 127-134.
9. Zhang L, Xie D, Zou W (2001) Viewing use cases as active objects. ACM SIGSOFT Software Engineering Notes**,** Volume 26 Issue 2.
10. Khammaci T, Bouras Z.E (2002) Versions of program integration. In World Scientific (eds). Handbook of Software Engineering and Knowledge Engineering No 2, Pittsburg (USA), pp. 495-516. 2002.
11. Bouras Z.E, Khammaci T, Ghoul S (2000) A new approach for program Integration. The South African Computer Journal, No. 25, pp.3-11.
12. Cockburn A (1997) Structuring Use Cases with Goals. Journal of Object-Oriented Programming, Sep-Oct 1997.
13. Biddle R, Noble J, Tempero E (2002) Essential use cases and responsibility in object-oriented development. CRPITS'02 Volume 24 Issue 1.
14. Unified Modeling Language, Version 1.5, http://www.rational.com/uml
15. Horwitz S, Reps T (1992) The Use of Program Dependence Graphs in Software Engineering. ICSE'92, Australia.
16. Binkley D, Horwitz S, Reps T (1995) Program integration for languages with procedure calls". ACM Trans. on Soft. Eng. and Meth. Volume 4 No 1 pp. 310-354.
17. Kruchten, P.: 'Architectural Blueprints - The 4 + 1 View Model of Software Architecture', *IEEE Software*, 1995, 12, (6), pp. 42-50.
18. Kruchten, P.: '*The Rational Unified Process: an Introduction*' (Swedish edition, Boston, Mass., London: Addison-Wesley, 2002).
19. Nicolas, J. and Toval, A.: 'On the Generation of Requirements Specifications From Software Engineering Models: A Systematic Literature', *Information and Software Technology*, 2009, 51, (9), pp. 1291-1307.
20. Jackson, M.: 'Automated Software Engineering: Supporting Understanding', *Automated Software Engineering*, 2008, 15, (3), pp. 275-281.
21. StarUML Project Team.: 'StarUML: User Guide' [online]. Available from: http://staruml.sourcefo rge.net/docs/user-guide(en)/toc.html [Accessed 15/1/2009].
22. ArgoUML Project Team: 'ArgoUML: User Manual' [online]. Available from: http://argouml-stats.tigris.org/documentation/manual-0.28/ [Accessed 15/1/2009].
23. NetBeans UML Plugin Team: 'NetBeans UML Plugin' [online]. Available from: http://netbeans.org/features/uml/index.html [Accessed 15/1/2009].

Evaluation Of Scheduling And Load Balancing Techniques In Mobile Grid Architecture

[1] Debabrata Singh, [2] Sandeep Nanda, [3]Sarbeswara Hota,[4]Manas Kumar Nanda

ITER,SOA UNIVERSITY,BBSR,ODISHA ,INDIA

Abstract

Recent advances in mobile communications and computing are strong interest of the scientific community in the *Grid* have led to research into the *Mobile Grid*. Based on a realistic Mobile Grid architecture we formulate the problem of job scheduling and load balancing techniques in a mobile environment and performance metrics. We extended the work by introducing a new scheduling policy and load balancing policy based on the notion of *installments and intra cluster load balancing algorithm* and continue the evaluation of the expanded set of scheduling and load balancing strategies in an effort to overcome the intermittent character of connectivity in a mobile environment. On real wireless traces, we demonstrated the superiority of the proposed policy, and showed the feasibility of a Mobile Grid system and design the efficient scheduling and load balancing policies subject to the underlying mobility were based a small part of the offered resources is wasted and a small part of the workload has to be processed again .The size of the installments increases as the size of the aborted fragments of the workload increases.

Keywords : **Mobile grid, replication, intermittent connectivity, wireless traces, heuristics**

1. Introduction

Grid computing emerged resource sharing and problem solving in dynamic, multi-institutional virtual organizations [1]. A grid computing system is essentially a large scaled distributed system designed to aggregate resources from multiple sites. Users of such systems have opportunity to take an advantage of enormous computational, storage or bandwidth resources that would be impossible to attain. In many cases these resources would be wasted if not aggregated inside a grid.

Grid systems [1] are very large-scale, generalized network computing systems that can scale to Internet size environments with resources distributed across multiple organizations and administrative domains. Mobile Grid Computing is making Grid Services available and accessible at anytime and anywhere from mobile devices. Advantages of mobile grid computing is mobile to mobile and mobile to desktop collaboration for resource sharing, improvement of user experience, convenience and some new application scenarios[2]. Increasing number of new autonomous , portable devices has become significant part of everyday life and work that leads to a decentralized , location independent wireless computing environment. It is natural to consider the extension of the idea of resource sharing to mobile and wireless communication environments[3] . There are different approaches on the exact character of the extension . whether mobile devices are considered as powerful enough to provide their resources or not. Here we discussed that as the number of available mobile devices is nowadays enormous and the computational power is increasingly, so then the aggregated sum of their resources can be exploited. The mobile devices are resource constrained as they can be incorporated in a grid as resource consumers. Mobile devices are increasingly becoming powerful enough to also participate in grid systems as resource providers. Mobile devices face resource limitations in comparison to their stationary counterparts , but the vast number mobile devices and their computational power constantly increases that lead us to the assessment that the aggregated sum of their resources can be exploited to overcome the limitations. So for this we intended to enrich the set of examined policy by proposing the installments policy and load balancing algorithms .

2. Scheduling in mobile grid

A. Problem Formulation :

The core functionality of an MGS in the proposed architecture is to receive a job and, in the context of *divisible load* applications, divide the submitted workload into tasks for submission to its descendant MGSs. At the lowest level, an L-MGS is responsible for distributing the received task load to the MNs currently residing in the WLAN it serves.

The number of available nodes N obviously varies in time as MNs roam in the wireless infrastructure of the campus. The whole process consists of three distinct steps: the transfer of the input workload to MN_i , i ε {0,1,…,N-1}. , the subtask execution and return of the results back to the L-MGS. In absence of disconnection events each step requires t_{in}^i, t_{exec}^i and t_{out}^i amount of time to complete. We define t_{total}^i as :

$t_{total}^i = t_{exec}^i + t_{in}^i + t_{out}^i$, i ε{0,1,…..,N-1}

The load will be distributed to available MNs :

$T_{TOTAL}=T_{IN}+T_{EXEC}+T_{OUT}$

We make the following assumptions :

- The total task load is equally distributed to all participating MNs.
- The execution of a subtask may begin only after the entirety of the input data has been received.
- The output data of a subtask can be returned only after the execution has completed.
- The output data of a subtask must be returned in whole in order to be usable.

The communication to computation ratio of a divisible load application is :

$$CCR=\frac{Communication\ \ Cost}{Computation\ \ Cost}$$

The Communication Cost factor denotes the time required by a MN to successfully receive a certain amount of workload in the absence of disconnection events and is subject to the available bandwidth[4]. The Computation Cost denotes the time required for the same amount of workload to be processed and is subject to the device characteristics and the usage of the device by its owner.

B. Performance Metrics :

In order to evaluate the performance of the scheduling strategies we need to define suitable metrics. These metrics include:

- *Response Time* (*RT*): the time required for the entire set of result data to be gathered at the scheduler.

$$RT = \max RT^i,\ i \in \{0,1,\ldots..,N-1\} \qquad (1)$$

Resource Waste (*RW*): the amount of resources wasted in the effort to compensate for intermittent connectivity. We further subdivide RW into RW_C and RW_N with respect to the type of resources wasted i.e. computational and network resources, respectively. *RW* is measured as a percentage of the actual resources required for the completion of the task.

Speedup : the comparison of the achieved *RT* with that of a single node execution. It reveals the actual degree of parallelism achieved.

3. Proposed Architecture

We have proposed a hierarchical, campus-wide computational Mobile Grid system architecture [6]. In the proposed architecture, depicted in Fig. 1, mobile nodes (MNs), willing to offer their computational resources, move between Access Points (APs) of the campus. This willingness is based on the expectation of *reciprocity* [7]. In our work we have considered *divisible load* applications [8] (e.g. query processing [9]) in which a job can be divided into tasks that can be carried out independently of each other. These tasks are distributed by the *Local-Mobile Grid Schedulers* (L-MGSs) to the collaborating MNs, which process them and return the results back. In other levels of the hierarchy *Intermediate-MGSs (I-MGSs)* may act as *meta-schedulers*. This work focuses on the last level of the hierarchy and more specifically on the load distribution performed by L-MGSs.

These traces provided us with realistic information on the mobility and connectivity characteristics of the MNs in the campus. We pointed out the important mobile networking parameters affecting the performance of a Mobile Grid system and showed that disconnection events impose a severe impact on the turn-around time of jobs executed by MNs. On an effort to smooth the effects of intermittent connectivity we examined the performance of a simple *task replication* scheme [8]. The results were very promising with respect to the achieved turn-around time. We continued our research by also investigating whether the execution of a task in a MN should be aborted upon disconnection or not, in order for the task to be rescheduled, coming up with a positive answer with respect to the resulting turn-around times [9]. In this work, we enrich the set of examined policies by proposing the *installments* policy, in which task load is further partitioned into small chunks, and demonstrate its effectiveness.

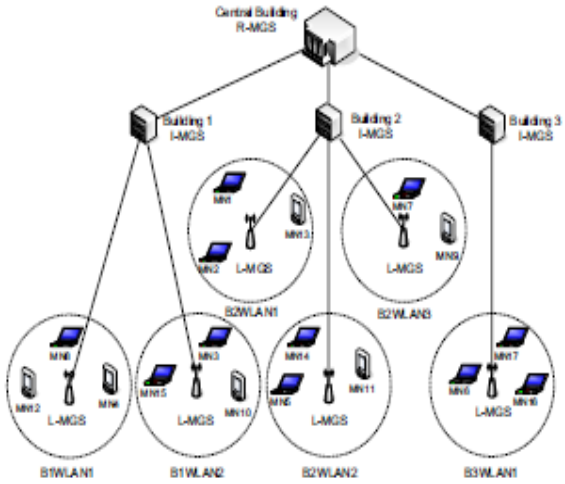


Fig. 1. Campus-wide Mobile Grid architecture.

4. Load Balancing Techniques

Load Sharing: This is the coarsest form of load distribution. Load may only be placed on idle resources, and can be viewed as a binary problem, where a resource is either idle or busy.

Load Balancing: Where load sharing is the coarsest form of load distribution, load balancing is the finest. Load balancing attempts to ensure that the workload on each resource is within a small degree, or balance criterion, of the workload present on every other resource in the system.

Load Levelling: Load levelling introduces a third category of load balancing to describe the middle ground between the two extremes of load sharing and load balancing[10][11]. Rather than trying to obtain a strictly even distribution of load across all resources, or simply utilizing idle resources, load levelling seeks to avoid congestion on any resource.

Resources are distributed in different geographic locations. Stability and performance of each resource is different. In other words, newly distributed system is dynamic and resources are composed of heterogeneous resources. Thus, an important problem is resources selection and task distribution when task are executed. This study proposed a hybrid load balancing policy, which selects effective node sets in the stage of static load balancing to lower the odds of selecting ineffective nodes and makes use of the stage of dynamic load balancing. [15]When the status of a node changes, a new substitute can be located in the shortest time to maintain the execution performance of the system.

It has four components :

1. Transfer policy determines whether a node is in a suitable state to participate in a task transfer.
2. Selection policy determines which task should be transferred.
3. Location policy determines to which node a task selected for transfer should be sent.
4. Information policy is responsible for triggering the collection of system state information.

A transfer policy requires information on the local nodes state to make decision. A location policy, is likely to require information on the state of remote nodes to make decisions.

We assume an application is composed of agents executable on any of P machines of the cluster. The structure of the application is modeled by the interdependence relationships among the agents. Specifically, we will use an undirected graph to model the application structure. An undirected graph is a generic model as a multi agent application executes perpetually and produces results continuously in response to user queries.

5. Scheduling Policies

A. Abort-Reschedule

In this policy a task executed by a mobile node must be aborted upon disconnection, in the sense that it will be rescheduled i.e. re-assigned to a new MN arriving at the WLAN served by the current L-MGS. In this policy, once a disconnection event has occurred, RTi is calculated as follows:

$$RT^{i}_{Abort} = t^{i}_{C} + t^{i}_{A} + RT^{i'}_{Abort} \quad , i \in \{0,1,\ldots\ldots,N-1\} \qquad (2)$$

In this policy, the computational and energy resources spent by a MN for the execution of a sub-task are obviously wasted if the assigned sub-task is aborted in the event of a disconnection[16]. In the following, we assume that an L-MGS is immediately aware of any disconnection event.

B. Installments

In this policy each sub-task is further partitioned into consecutive fragments (installments) of size *f*, with $f < t^{i}_{total}$ The major difference with the *Abort Reschedule* policy is that only the sub-task currently

executed by a mobile node must be aborted upon disconnection. All previously completed installments are not wasted since their results have been successfully returned to the scheduler in whole. Following the same notation, once a disconnection event has occurred, RT^i is calculated as follows:

$$RT^i_{Inst} = t^i_C + t^i_A + RT^{i'}_{ResInst} \quad , i \in \{0,1,\ldots\ldots\ldots\ldots,N-1\}$$

This policy, intends to alleviate the problem of resource wasting in the case of *Abort-Reschedule* by further fragmenting the task load. Even though the fragmentation and reassembly overhead of the scheduler increases, this policy results in a more efficient utilization of mobile resources. Upon disconnection, there is no need for the re-assignment of the completed part of the task to a new MN. Instead, only the remaining installments are submitted resulting in a decreased RT,, resource waste is inevitable in this policy, since a disconnection event may interrupt the completion of an installment.

However, the waste is limited to the size of the installments. Furthermore, it must be noted that this policy is better suited for mobile devices since it does not require large amounts of storage and memory for the computation of entire subtasks. It is also considered suitable for providing the MNs with the *flexibility* to specify the amount of resources they offer, expressed in number of installments and/or installment size (f). Moreover, *installments* can be used to implicitly inform the scheduler about the networking conditions and the processing capabilities of a MN. This can be considered as a form of quick (not waiting for the complete task) implicit feedback which can guide the scheduler to important decisions regarding scheduling and load distribution. In this case, the estimation can be used to decide whether to abort the installment or not. Finally, the proposed policy presents the advantage of quickly providing *partial results* to the scheduler. Since the results are usable only if they refer to the entirety of the input workload, it is not necessary to wait for the completion of the whole task as in the case of the *Abort-Reschedule* policy. Instead, the results from the completed installments may be returned to the Mobile Grid user. The aforementioned features of this policy are considered as especially useful in application scenarios where quick, and possibly partial, results are desirable/acceptable such as SETI like [13] scientific search applications, and in environments where the mobile devices are particularly resource constrained e.g. cellular networks.

C. Groups

In this policy tasks are replicated so that a certain task is assigned to more than one of the MNs residing in the same WLAN. This policy results in the formation of distinct *groups* of MNs processing the same sub-tasks. The reasoning behind this approach is that not all MNs present the exact same networking behavior at the same time and therefore, if the same task is submitted to multiple MNs it is highly probable that one of them will eventually return the results earlier than the others. Obviously, this policy is especially suitable for applications in which fault tolerance and reliability are on the main focus e.g. [9]. We denote R as the *replication degree* with R ε {1, ..,N}, i.e. the number of replicas produced for each task. If R = 1 we obviously have no replication. Then we have:

$$RT^i_{Groups} = \max_j RT^i_j, i\varepsilon\{0,1,\ldots\ldots\ldots\ldots..,\frac{N}{R}\} \text{ , } j\{1,\ldots,R\}$$

The *Groups* policy unavoidably results in the waste of resources. If a certain MN returns the results of a sub-task earlier than the MNs which have received the same sub-task, then the resources of the remainder of the MNs are wasted. Apart from the apparent waste of resources, excessive task replication has another important side-effect. A job of a certain workload is split to each participating MN as follows:

$$t^i_{total} = \frac{T_{TOTAL}}{\frac{N}{R}}$$

Load Balancing Strategy:

Intra-Cluster load balancing: In this first level, depending on its current workload (estimated from workloads of its worker nodes), each cluster manager decides whether to start or not a load balancing operation. If a *cluster manager* decides to start a load balancing operation, then it tries, in priority, to load balance its workload among its worker nodes. Hence, we can proceed *C* local load balancing in parallel, where *C* is the number of clusters. Load of an agent executing on machine is defined as the sum of its computational load and communication load

$U_i = H_i + G_i$ Where: H_i – Communication Load, G_i – Computational Load The load L_k of machine m_k is defined as the sum of all its local agents load. More specifically

$L_k = \sum (w_i + u_i)$ Where: w_i–Communication Load , u_i – Computational Load

Goal of a load balancing algorithm is to minimize the variance of the load among all the machines in the cluster, this will turn minimize the average response time of serving users queries.

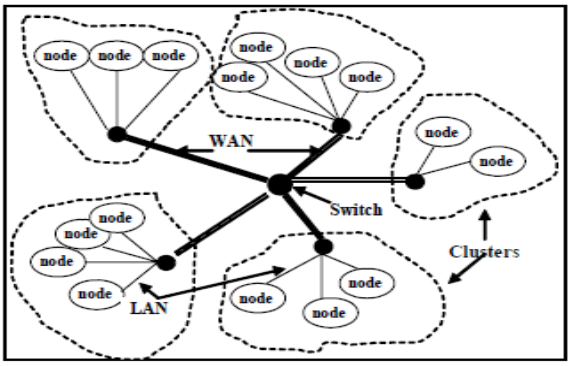


Figure 2: Example of a Grid topology

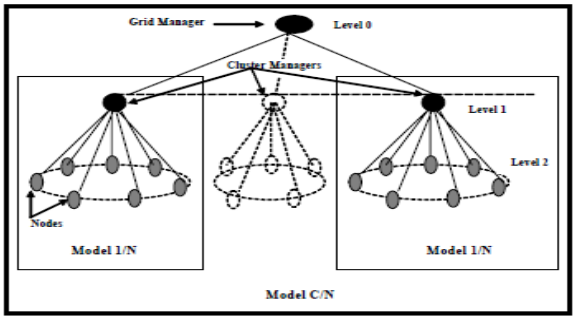


Figure 3: Tree-based representation of a Grid

6.Intra cluster load balancing algorithm

This algorithm is considered as the kernel of our load balancing strategy. The neighborhoods load balancing used by our strategy makes us think that the imbalance situations can be resolved within a cluster. It is triggered when any cluster manager finds that there is a load imbalance between the nodes which are under its control. To do this, the cluster manager receives periodically workload information from each worker node. On the basis of these information's and the estimated balance threshold ε, it analyzes the current workload of the cluster. According to the result of this analysis, it decides whether to start a local balancing in the case of imbalance state, or eventually just to inform its Grid manager about its current workload. At this level, communication costs are not taken into account in the task transfer since the worker nodes of the same cluster are interconnected by a WLAN network, of which communication cost is constant.

Step 1: Workload Estimation

1.For Every element E_i of G and according to its specific period **Do** Sends its workload LOD_i to its group manager

Endfor

2.Upon receiving all elements workloads and according to its period the group manager performs:

a- Computes speed SPD_G and capacity SAT_G of G

b- Evaluates current load LOD_G and processing time TEX_G of G

c- Computes the standard deviation σ_G over processing times

d- Sends workload information of G to its associated manager: in case where G is a cluster.

Step 2: Decision Making

3. Balance criteria

a. Cluster**: If** ($\sigma_G \leq \varepsilon$ **) Then** Cluster is balanced; Return **EndIf**

b. Grid**: If (**# (overloaded clusters)) ≤ given threshold **Then** Grid is in balance state; Return **EndIf**

4. *Saturation criteria*

If $\left(\frac{LOD_G}{SAT_G} > \delta\right)$**Then** Group G is saturated; Return **EndIf**

5. *Partitioning group G into overloaded (GES), under-loaded (GER) and balanced (GEN)*

GES←Φ; GER← Φ; GEN ← Φ

For Every element E_i of G **Do**

If (E_i is saturated) **Then** GES GES U E_i /* Saturated Overloaded */ **Else**

Switch

- $TEX_i > TEX_G + \sigma_G$: GES GES U E_i /* Source */

- $TEX_i < TEX_G - \sigma_G$: GER GER U E_i /* Receiver */

- $TEX_G - \sigma_G \leq TEX_i \leq TEX_G + \sigma_G$: GEN ← GEN U E_i/*Balanced*/

Step 3: Tasks Transfer

Test on Supply and Demand

$Supply = \sum_{E_r \, \varepsilon \, GER} \frac{LOD_G}{SPD_G} * \frac{SPD_r}{SPD_G} - LOD_r$

$Demand = \sum_{E_s \in GES} LOD_s - \frac{SPD_s}{SPD_G}\frac{LOD_G}{SPD_G}$

If$\left(\frac{Supply}{Demand} \geq \rho\right)$ **Then** local load balancing Fail; Return **EndIf**

7. Perform intra-group task transferring:

If (G = Cluster) **then** Perform Heuristic1 **else EndIf**

Heuristic 1: Intra-Cluster tasks transfer

a- Sort GES by descending order of their elements processing times.

b- Sort GER by ascending order of their elements processing times.

c- While (GES ≠ Φ .**AND.** GER ≠ Φ) **Do For** i = 1 **To** # (GER) **Do**

(i) Sort tasks of first node belonging to GES by selection criterion,

(ii) Transfer the higher priority task from first source node of GES to ith receiver node of GER

(iii) Update the current workloads of receiver and source nodes,

(iv) Update sets GES, GER and GEN,

(v) If (GES = Φ **OR** GER =Φ) **then** Return **Endif,**

(vi) Sort GES by descending order of their processing times.

Endfor

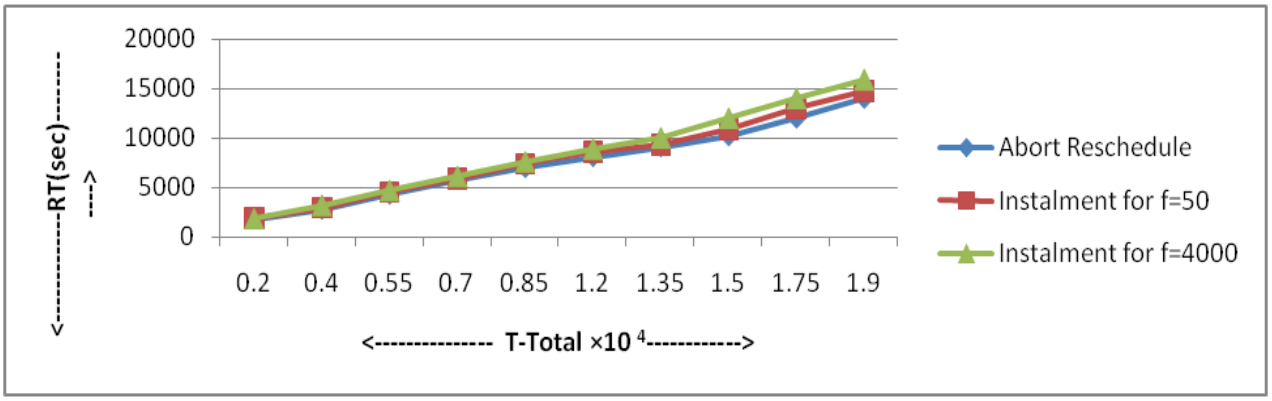
7. Experimental Study of Scheduling and Load Balancing techniques

It shows the performance of the proposed *installments* policy both it terms of RT and RW. We have examined a wide range of values for the size of the installments (f), measured in seconds. These values varied from a small portion to the entirety of the workload in order, first, to reveal the effects of the installment size on the resulting RT and RW, In both cases, the superiority of the *installments* policy is evident. As the size of the installments increases, the performance of the *installments* policy approaches that of the *Abort-Reschedule* policy .We expect that the superiority of the *installments* policy will be more evident in environments with more intermittent connectivity. Similar results were derived for the achieved Speedup. In all, for small installment sizes, any disconnection results in the abortion of a small portion of the overall subtask. Therefore, a small part of the offered resources is wasted and a small part of the workload has to be processed again. As the size of the installments increases so does the size of the aborted fragments of the workload. Hence, the increased RW and the extended re-processing leading to a higher RT.

In the first set of experimentations, we have focused on the response time, the waiting time and the processing time, according to various numbers of tasks and clusters. We have considered different numbers of clusters and we suppose that each cluster contains 50 worker nodes. For every node, we generate a random speed varying between 10 and 30 computing units per time unit. The number of instructions per task has varied between 300 and 1500 computing units. Our strategy has allowed to reduce in a very clear way the mean response time of the tasks. We obtain a gain in 100% of cases, varying between 3.09% and 24.44%. In more than 60% of cases, this gain is greater than 11%. The lower gains have been obtained when the number of clusters was fixed at 32 on the one hand and when the number of tasks was 10000 and 20000 on the other hand. We can justify this by the instability of the Grid state (either overloaded or idle).

Best improvements were obtained when the Grid were in a stable state: (for Clusters {8, 16} and for Tasks {14000, 16000}.In some infrequent cases, we have noted that the variation of the gain changes abruptly. We believe that this situation comes from the fact that the number of tasks and/or the number of clusters varies suddenly and generates instability in the Grid.

Fig .a Response Time (RT)



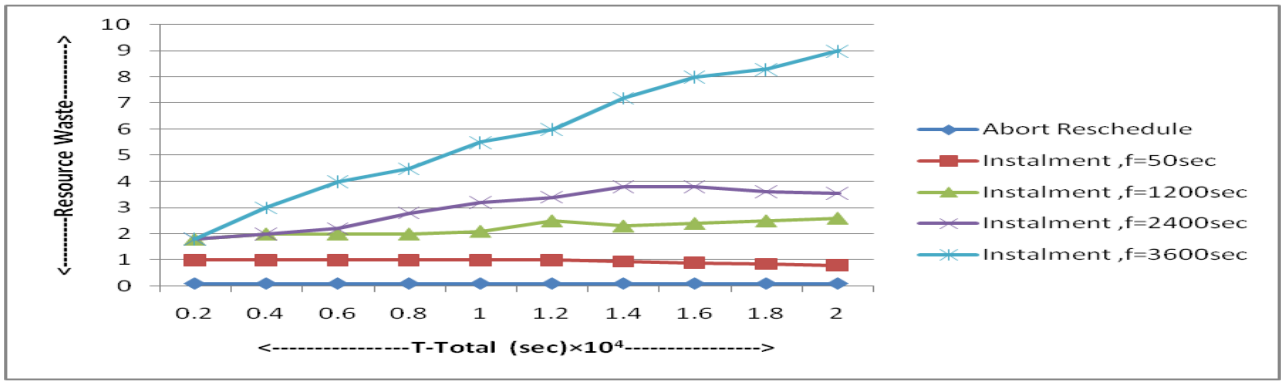


Fig b Resource Waste (RW)

Fig 4.Performance of installments policy

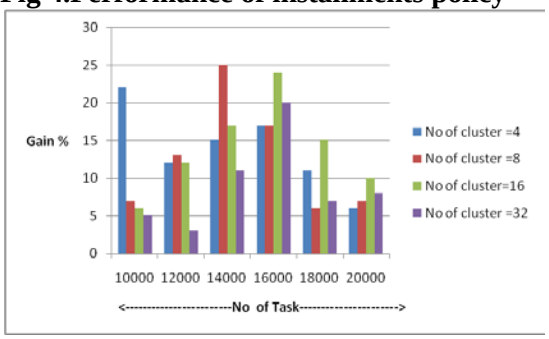


Fig 5: Gain according to various numbers of clusters by varying the number of tasks

8. Conclusion

Scalability is the precondition of algorithm to avoid poor allocation decisions. To assess stability we can measure hit-ratio, the ratio of remote execution requests concluded successfully. Another measure of stability is percentage of remote execution in the system. Activities related to remote execution should be bounded and restricted to a small proportion of the activity in the system. In both cases, the superiority of the *installments* policy is evident. As the size of the installments increases, the performance of the *installments* policy approaches that of the *Abort-Reschedule* policy. We expect that the superiority of the *installments* policy. In all, for small installment sizes, any disconnection results in the abortion of a small portion of the overall subtask. Therefore, a small part of the offered resources is wasted and a small part of the workload has to be processed again. As the size of the installments increases so does the size of the aborted fragments of the workload. Hence, the increased RW and the extended re-processing leading to a higher RT.

References

[1] K.Q. Yan1, S.C. Wang1 "The Anatomy Study of Load Balancing in Distributed System", proceeding of the seventh international conference on parallel and distributed computing, Application and Technologies (PDCAT'06)

[2] Reinhard Riedl and Lutz Richter "Classification of Load Distribution Algorithms ", CH-8057,1066-6192/96 $5.00 0 1996 IEEE Proceedings of PDP '96

[3] Ka-Po Chow and Yu-Kwong Kwok "On Load Balancing for Distributed Multiagent Computing", ieee transactions on parallel and distributed systems, vol. 13, no. 8, august 2002.

[4] G. Banavar, J. Beck, E. Gluzberg, J. Munson, J. Sussman, and D. Zukowski, "Challenges: an application model for pervasive computing," in *MobiCom '00: Proceedings of the 6th annual international conference on Mobile computing and networking*, New York, NY, USA, 2000, pp. 266–274, ACM Press

[5] A. Litke, D. Skoutas, and T. Varvarigou, "Mobile grid computing: Changes and challenges of resource management in a mobile grid environment," in *Proceedings of Practical Aspects of Knowledge Management(PAKM 2004)*, 2005.

[6] T. Phan, L. Huang, and C. Dulan, "Challenge:: integrating mobile wireless devices into the computational grid," in *MobiCom '02: Proceedings of the 8th annual international conference on Mobile computing and networking*, New York, NY, USA, 2002, pp. 271–278, ACM Press.

[7] S. Kurkovsky and Bhagyavati, "Wireless grid enables ubiquitous computing.," in *ISCA PDCS*, 2003, pp. 399–404.

[8] K. Katsaros and G.C. Polyzos, "Optimizing operation of a hierarchical campus-wide mobile grid for intermittent wireless connectivity," in *Proceedings of the 15th IEEE Workshop on Local and Metropolitan Area Networks (LANMAN)*, Princeton, NY, USA, 2007.

[9] K. Katsaros and G.C. Polyzos, "Optimizing operation of a hierarchical campus-wide mobile grid," in *Proceedings of the 18th IEEE Personal Indoor and Mobile Radio Communications conference (PIMRC)*, Athens, Greece, 2007.

[10] V. Bharadwaj, T.G. Robertazzi, and D. Ghose, *Scheduling Divisible Loads in Parallel and Distributed Systems*, IEEE Computer Society Press, Los Alamitos, CA, USA, 1996.

[11] K. Katsaros, I. Niarhos, and V. Vassalos, "DAIMON: Data Integration for a Mobile Network," in *MobiDE '05: Proceedings of the 4th ACM international workshop on Data engineering for wireless and mobile access*, New York, NY, USA, 2005, pp. 57–64, ACM.

[12] D. Kotz, T. Henderson, and I. Abyzov, "CRAWDAD trace set dartmouth/campus/movement,"*http://crawdad.cs.dartmouth.edu/dartmouth/campus/movement, Mar*.2005.

[13] "Seti@home homepage," http://setiathome.ssl.berkeley.edu.

[14] A. Kamerman and G. Aben, "Throughput performance of wireless LANs operating at 2.4 and 5 GHz," in *Personal, Indoor and Mobile Radio Communications, 2000. PIMRC 2000.*, 2000, pp. 190–195.

[15] A. Litke, D. Skoutas, K. Tserpes, and T. Varvarigou, "Efficient task replication and management for adaptive fault tolerance in mobile grid environments," *Future Generation Computer Systems*, vol. 23, no. 2, pp. 163–178, 2007.

[16] Sang-Min Park, Young-Bae Ko, and Jai-Hoon Kim, "Disconnected operation service in mobile grid computing.," in *ICSOC*, 2003, pp. 499–513.

[17] S. Kurkovsky and Bhagyavati, "Wireless grid enables ubiquitous computing.," in *ISCA PDCS*, Seong-Moo Yoo and Hee Yong Youn, Eds. 2003, pp. 399–404, ISCA.

[18] S. Kurkovsky, Bhagyavati, and A. Ray, "A collaborative problem-solving framework for mobile devices," in *ACM-SE 42: Proceedings of the 42nd annual Southeast regional conference*, New York, NY, USA, 2004, pp. 5–10, ACM Press.

Author Information

Debabrata Singh is an Assistant Professor in the Department of Information Technology, holds M.Tech in Computer Science & Engg. (BPUT,BBSR) He has nearly five years experience in teaching, software development and research. Presently, he is working as Assistant professor in ITER,SOA University, Bhubaneswar, Orissa He has published 17 papers on Multi agent technologies, Sensor Network, & Grid Computing environment in national & international journals and conferences.

Sandeep Nanda is a research scholar in the Department of Computer Applications ,holds M. Tech in CSDP branch (ITER, Bhubaneswar) He has nearly 2 years experience in teaching, and research. Presently, he is working as a lecturer in Jawahar Navodaya Vidyalaya, Sora, Odisha ,India, under central government of India. He has published 5 papers on Wireless Mesh Networks, Grid Computing environment in national & international journals and conferences.

Sarbeswara Hota is an Assistant Professor in the Department of Computer Applications, holds M. Tech in Computer Science & Engg. (ITER, Bhubaneswar) He has nearly 8 years experience in teaching, and research. Presently, he is working as Assistant professor in ITER,SOA University, Bhubaneswar ,Orissa His area of interest are Wireless Mesh Networks, MANETs, multi agent technologies & Grid Computing environment.

Manas Kumar Nanda is an Assistant Professor in the Department of Computer Applications, holds M.Tech in Computer Science & Engg. (Berhampur University, Berhampur) He has nearly 7 years experience in teaching, and research. Presently, he is working as Assistant professor in ITER,SOA University, Bhubaneswar, Orissa .He has published 5 papers on Wireless Mesh Networks, sensor network, multi agent technologies & data mining in national & international journals and conferences.

A Study of Library Databases by Translating Those SQL Queries Into Relational Algebra and Generating Query Trees

Santhi Lasya [1], Sreekar Tanuku [2]

[1] – Department of Electronics and Computer Science, Jawaharlal Nehru Technological University, SNIST, Hyderabad, India.

[2] – Amazon, Seattle, WA, USA.

Abstract

Even in this World Wide Web era where there is unrestricted access to a lot of articles and books at a mouse's click, the role of an organized library is immense. It is vital to have effective software to manage various functions in a library and the fundamental for effective software is the underlying database access and the queries used. And hence library databases become our use-case for this study.

This paper starts off with considering a basic ER model of a typical library relational database. We would also list all the basic use–cases in a library management system. The next part of the paper deals with the sql queries used for performing certain functions in a library database management system. Along with the queries, we would generate reports for some of the use cases. The final section of the paper forms the crux of this library database study, wherein we would dwell on the concepts of query processing and query optimization in the relational database domain. We would analyze the above mentioned queries, by translating the query into a relational algebra expression and generating a query tree for the same. By converting algebra, we look at optimizing the query, and by generating a query tree, we would come up a cheapest cost plan.

Keywords: *Library Databases, Query Optimization, Relational Algebra, Query Tree, SQL Server Management Studio, Microsoft Visual Studio, and Cost Analysis.*

1. Introduction

The Library Database is created to support the principal functions of a lending library's day-to-day operations. It provides access to resources across a wide spectrum of topic and subject areas. Such as: the arts, academic research, home improvement, auto repair, business and much more.

The very first step is to construct an ER-Diagram of library databases, which is just an approximate description of the information to be stored in the database. With that logical design schema, we implement our database design. The database design of the library consists of creating queries and stored procedures that satisfy some of the functionalities of library operations. Through these library operations, final reports are produced and the designed database results are given in the form of test forms.

For accessing and retrieving data from the database, we convert ER-Diagram into relational database model. Relational algebra is one of the two formal query languages associated with the relational model. As relational model supports powerful query languages, steps we take in handling the queries of library databases are:

- Initial SQL queries for library databases
- Converting these SQL queries into relational algebra based on collection of operators for manipulating relations and optimizing purpose.
- Formation of Query tree for estimating the cheapest cost plan.

2.1. LIBRARY DATABASES

A library database contains relevant and accurate information in a particular field. It is both an electronic catalog and the access point to information from published works where published information sources are of magazines, newspapers, encyclopedias, journals and other resources. Library databases are easily searchable. Database content may often be searched by: Keywords, Title, Author, or Subject. Each article or book can be given in the form of

- Full Text = entire article - Library databases sometimes omit photos, graphs, charts, and figures from articles, but most will indicate that these have been omitted.
- Abstract = summary provided by the author or database publisher.

Databases provide citation information about the items they index. A citation typically consists of information such as:

- Title
- Author
- Source (Title and type of Publication)
- Publisher
- Date of Publication

Any library visitor may access the library database collection and library card holders may access many of the library's databases from home. This database works closely with faculty administrators, computing services and the Research school to provide integrated support to researchers. There is a wide collection of electronic databases which provides full text access to eBooks, databases, and thesis.

2.2. ER- DIAGRAM

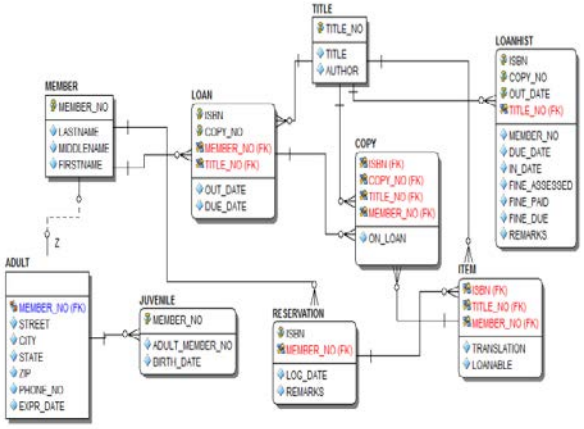


Fig.1. ER-Diagram of Library Databases

2.3. STEP BY STEP PROCEDURE

A library database contains a listing of authors that have written books on various subjects (one author per book).Here, this database has been used to-

*Create queries against the library databases that return a number of results which uses different types of joins, UNION statements, CASE statements, date manipulation, string concatenation, and aggregate functions.

*Design back-end stored procedures that satisfy some of the functionalities of library operations such as Add Adult, Add juvenile, Check in a Book, Check out a book, Add Book, Renew Membership, Change juvenile to Adult, Update Adult. The procedures incorporated input validations and provide adequate error handling using TRY/CATCH.

When we consider the E-R diagram of library databases, anyone can visit the library database collection. And for accessing many of the library databases, particular visitor should have a library card. So with the help of library card, particular member can fill up the details in 'MEMBER' entity dataset (member_no, lastname, firstname, middlename). Each member is provided with member_no in library card and whenever a library visitor wants to access particular book or any

information about the library database, using member_no we can retrieve the information.

The personal details (like address (street, city, state, and zip), phone_no, expr_date) of a visitor can be filled in the entity dataset 'ADULT' and in that, the member is given a expiry date. Once the card is expired, then the visitor has to renew the library card. If the member's age is less than 18 then that particular visitor comes under the dataset 'JUVENILE'. Juvenile entity dataset (member_no, adult_member_no, birth_date) offer online access to age-appropriate books and magazines. With the help of adult_member_no i.e, using any adult member library card number, juvenile members can get their member_no.

In the 'TITLE' dataset (title_no, title, author), the search of a specific title or any author, provides the title number (title_no). Through the dataset 'TITLE', the details of any book to be loaned are known. 'LOAN' dataset (ISBN, title_no, member_no, copy_no, out_date, due_date) contains the information about a specific book, still how many copies are there in the database if no then which member_no loans that book and the date until it is reserved by a particular member_no is given. Here in 'LOAN' dataset ISBN i.e, International Standard Book Number, is present where specific number is given to the book which is called all over world as its number. ISBN is locked in each and every dataset wherever used. 'COPY ‘dataset (ISBN copy_no, title_no, on_loan) gives the number of copies of each standard book and its details. The total data of 'LOAN' and 'COPY' dataset with excess amount of data are provided in 'LOANHIST' i.e, loan history dataset. 'LOANHIST'(ISBN, copy_no, out_data, title_no, member_no, due_date, in_date, fine_assessed, fine_paid, fine_due, remarks) contains database book details i.e, copy number ,title number, book return date, if due date is completed the fine to be paid, overall fine to be paid, fine_due and remarks given for that member_no.

The details of a book or even translation of any language can be done in the 'ITEM' dataset (ISBN, title_no, translation, loanable). Finally in 'RESERVATION' dataset (ISBN, member_no, log_date, remarks), any member can reserve a book and the details when that particular book is reserved or any remarks written on that book can be seen here. And this concludes the explanation of the overall E-R diagram of library databases.

The front-end of the database queries include many cases-

- Return member info from member and adult tables.
- Display available books.
- Search member info using member_no.
- Use UNION, list all member reserve the specific book.
- Use CASE, list all member reserve the specific book.
- Create temporary table.
- Display members have past due loan use temporary table.
- Display members who pay highest fine.
- List of members who want to reserve the specific book.

All these queries are developed on SQL Server Management Studio (SQL Server 2008). When you consider the use-case 'Display Available Books' in the library databases, the code to be written in query language on SQL Server Management Studio is shown as,

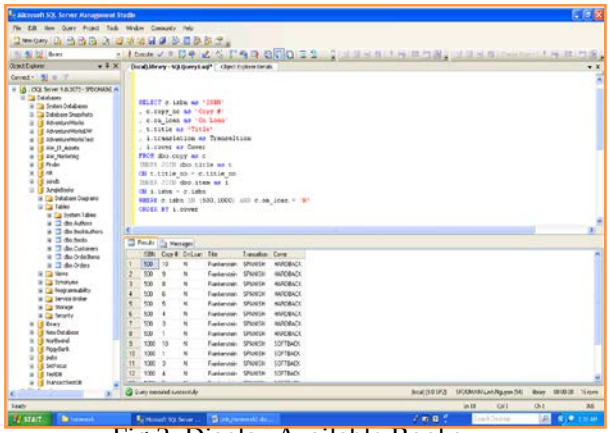


Fig.2. Display Available Books

Similarly, when we consider another case like 'Display members who pay highest fine', the output form is given as,

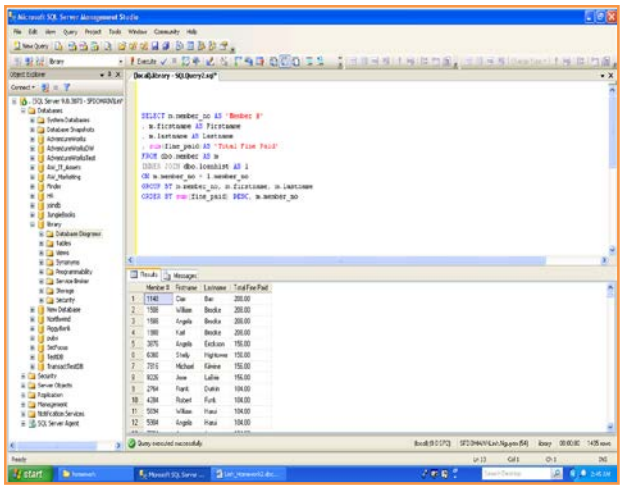


Fig.3. Display Members Who Pay Highest Fine

The back-end stored procedures support the following principal functions of a library's day-to-day operations:

Add Adult, Add Juvenile, Check In a book, Check out a book, Add Book, Renew Membership, Change Juvenile to Adult, Update Adult.

Here each and every case is designed and developed on Microsoft Visual Studio (ADO.NET). When comes to the result, Fig.4. shows 'Add Adult' test form output.

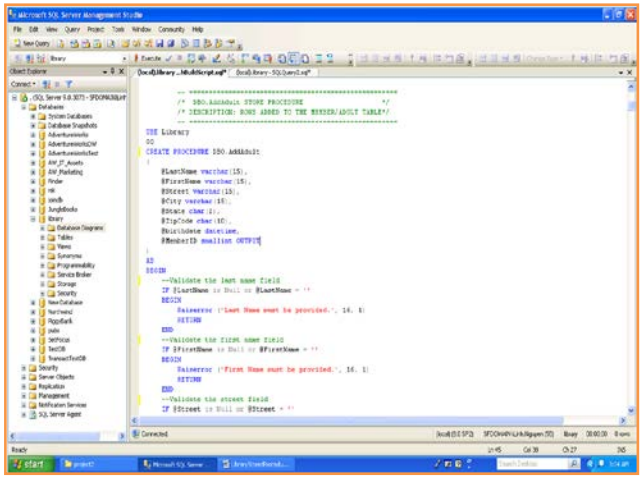


Fig.4. Add Adult

Similarly, the test form 'check out a book' output refers to be,

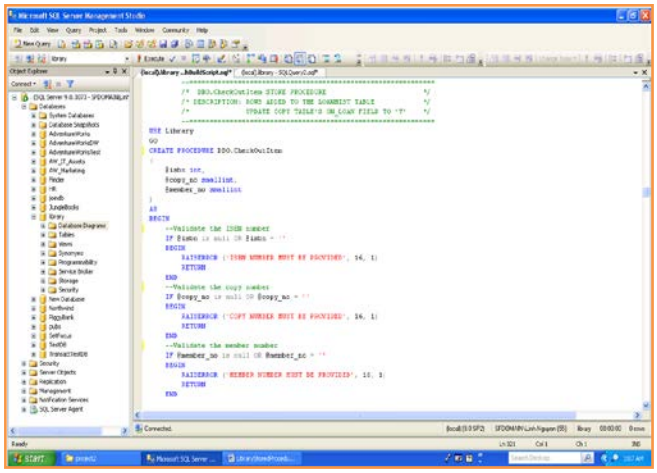


Fig.5. Check Out A Book

2.4. Reports Of Library Databases

The final reports are shown in Microsoft Visual Studios. Overall reports produced in library databases are-

- Complete list of books
- No of copies per title
- Most active members
- List of books on loan
- Adult member detail
- Dependents member detail
- Expired memberships
- Reference and special collection items
- Current fines for overdue books
- Total fines by member

The report 'List of books on loan' i.e, the list of total books under loan in the library database are shown as result in this fig.6.

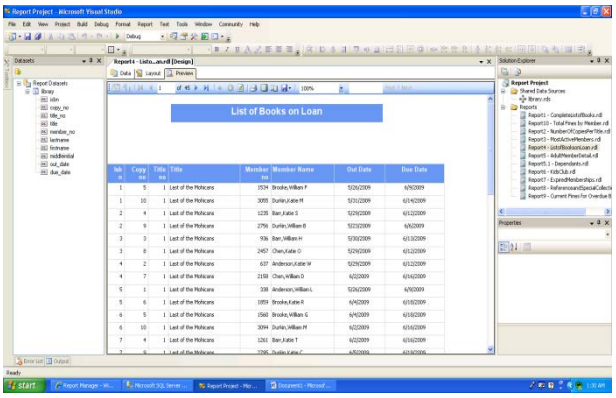


Fig.6. List Of Books On Loan

Similarly, Fig.7 shows another report 'Current fines for overdue books' where the overdue book fines are given as result in the test form.

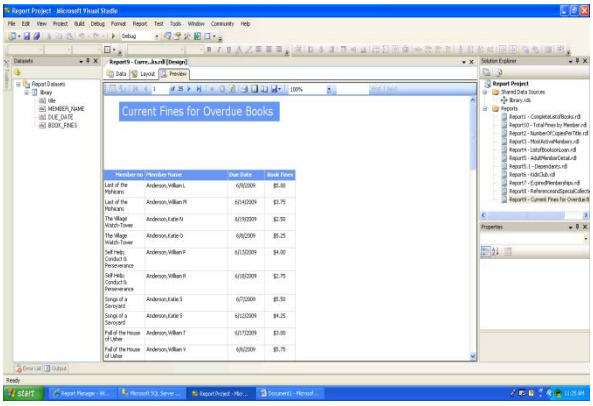


Fig.7. Current Fines For Overdue Books

We have seen two examples in each and every case of front-end queries and back-end stored procedures. In the similar way, the rest of the output test forms and reports in the library database are produced.

3.1. Translation Of SQL Queries In To Relational Algebra And Generating Query Tree

SQL queries are optimized by decomposing them into a collection of smaller units, called blocks. A typical relational query optimizer concentrates on optimizing a single block at a time. When a user submits an SQL query, the query is parsed into a collection of query blocks and then passed onto the query optimizer.

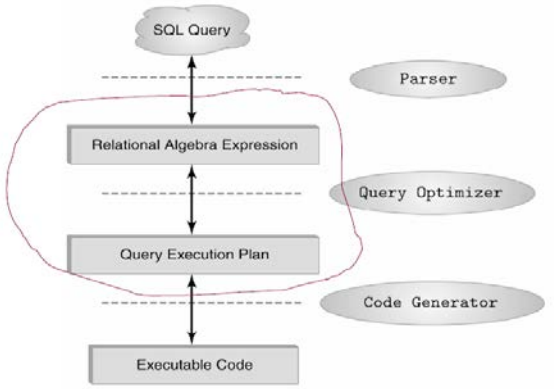


Fig.8. Query Optimizer

The optimizer examines the system catalogs to retrieve information about the types and lengths of fields, statistics about the referenced relations, and the access paths available for them. The optimizer then considers each query block and chooses a query evaluation plan for that block.

In given query, it essentially enumerates a certain set of plans and chooses the plan with the least estimated cost. The query blocks that contain two or more relations in the FROM clause require joins (or cross-products). Finding a good plan for such queries is very important because these queries can be quite expensive. Regardless of the plan chosen, the size of the final result can be estimated by taking the product of the sizes of the relations in the FROM clause and the reduction factors for the terms in the WHERE clause. But, depending on the order in which relations are joined, intermediate relations of widely varying sizes can be created, leading to plans with very different costs. In that process of enumerating multiple relation queries plan, query is taken as example for forming the cheapest plan.

3.1.1. Relational Algebra

Relational algebra is the query language associated with the relational model. Queries in algebra are composed using a collection of operators. Each relational query describes a step-by-step procedure for computing the desired output, based on the order in which operators are applied in the query.

For optimizing a query block, the SQL query must be converted into relational algebra expression.

3.1.2. Query Tree

Query tree is a tree data structure that corresponds to a relational algebra expression. It represents the input relations of the query as leaf nodes of the tree, and represents the relational algebra operations as internal nodes.

An execution of the query tree consists of executing an internal node operation whenever its operands are available and then replacing that internal node by the relation that results from executing the operation.

When we consider a case "Display available books" from front-end database queries as an example, As shown in fig.2,

CASE1: Display available books

CASE1: SQL QUERY:

```
SELECT i.ISBN, copy_no, t.title_no,
translation, title, author
FROM copy As c
INNER JOIN item As i
ON i.isbn = c.isbn
AND i.title_no = c.title_no
INNER JOIN title As t
ON t.title_no = i.title_no;
```

A query block is an SQL query with no nesting and exactly one SELECT clause and one FROM clause and at most one WHERE clause, GROUP BY clause, and HAVING clause.

So, the first step in optimizing a query block is to express it as a relational algebra expression.

Every SQL query block can be expressed as an extended algebra expression having this form. The SELECT clause corresponds to the projection operator, the WHERE clause corresponds to the selection operator, the FROM clause corresponds to the cross-product of relations, and the remaining clauses are mapped to corresponding operators in a straightforward manner.

The alternative plans examined by a typical query optimizer can be understood by recognizing that a query is essentially treated as an algebra expression. Translating that "CASE1: SQL QUERY" into relational algebra expression, we have

CASE1: RELATIONAL ALGEBRA:

$$\pi_{\text{I.ISBN, copy\_no, t.title\_no, translation, title, author}} \left(\left(\text{Copy} \bowtie_{\substack{C.ISBN = I.ISBN \\ C.title\_no = I.title\_no}} \text{Item} \right) \bowtie_{I.title\_no = T.title\_no} \text{Title} \right)$$

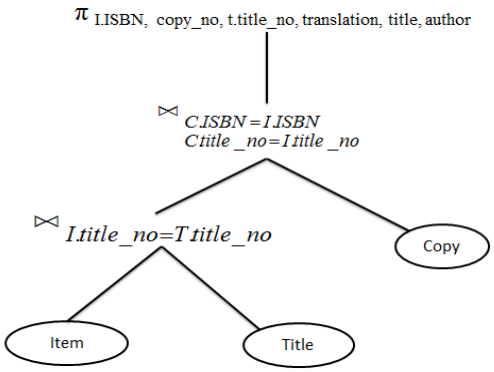
For estimating the cost of an evaluation plan for a query block, we consider query tree. For each node in the tree, we must estimate the cost of performing the corresponding operation. Costs are affected significantly by whether pipelining is used or temporary relations are created to pass the output of an operator to its parent.

Steps in converting a query tree during optimization involve:

- Initial (canonical) query tree for sql query.
- Moving SELECT operations down the query.
- Applying the more restrictive SELECT operation first.
- Replacing CARTESIAN PRODUCT and SELECT with JOIN operations.
- Moving PROJECT operations down the query tree.

So the converted query tree for "CASE1: RELATIONAL ALGEBRA" is,

CASE1: QUERY TREE:



Suppose that the following indexes are available: for Item and Title, a B+ tree index on the -(ISBN and title_no) fields and a hash tree on the title_no field(join method of item and title);similarly for Copy, a B+ tree index on the -(ISBN and title_no) fields and a hash tree on the ISBN and title_no fields.

The best plan is found for accessing each relation, regarded as the first relation in an execution. so the best plan for copy, item and title is obviously a file scan because no selections match an available index. Still now the plans generated are taken as outer relation and we consider joining another relation as the inner one. Hence, the following joins are processed: file scan of Item(outer) with Title(inner),file scan of Item(outer) with Copy(inner),file scan of Copy(outer) with Title(inner), file scan of Copy(outer) with Item(inner).

For each such pair, we consider every join method, and for each join method, we consider every available access path for the inner relation. For each pair of relations, we retain the cheapest of the plans considered for every sorted order in which the tuples are generated. Note that, since the result of the first join is produced in sorted order by title_no, whereas the second join requires its inputs to be sorted by ISBN and title_no, the result of the first join must be sorted by ISBN and title_no before being used in the second join. The tuples in the result of the second join are generated in sorted order by ISBN and title_no fields. For each plan retained, if the result is not sorted on ISBN and title_no, we add the cost of sorting on the ISBN and title_no fields.

The sample plan generated produces tuples in ISBN and title_no order, therefore, it may be the cheapest plan for the query even if a cheaper plan joins all three relations but does not produce tuples in ISBN and title_no order.

As we have seen one example of SQL query translating into relational algebra including the formation of query tree, we'll see another example for some more information.

CASE2: use UNION, list all member reserve the specific book.

CASE2: SQL QUERY:

```
SELECT i.ISBN, title, m.member_no, lastname
+','+ firstname As Name,
'Adult' As [MemberType] FROM adult As a
INNER JOIN member As m
ON a.member_no = m.member_no
INNER JOIN reservation As r
ON m.member_no = r.member_no
INNER JOIN item As i
ON r.isbn = i.isbn
INNER JOIN title As t
ON t.title_no = i.title_no WHERE i.isbn = 500;
```

UNION

```
SELECT i.ISBN, title, m.member_no, lastname
+','+ firstname As Name,
'Juvenile' As [MemberType] FROM juvenile As j
INNER JOIN member As m
ON j.member_no = m.member_no
INNER JOIN reservation As r
ON m.member_no = r.member_no
INNER JOIN item As i
ON r.isbn = i.isbn
INNER JOIN title As t
ON t.title_no = i.title_no WHERE i.isbn = 500;
```

Here we use the UNION operation to club two datasets ADULT and JUVENILE. When consider the relational algebra for optimizing that SQL query, the result can be given as,

CASE2: RELATIONAL ALGEBRA:

$$\pi_{\text{t.ISBN, title, m.member\_no, lastname + ',' + firstname As Name, a.MemberType = 'Adult'}}$$

$$((((\text{Adult} \bowtie_{Amember\_no=Mmember\_no} \text{Member}) \bowtie_{Mmember\_no=Rmember\_no} \text{Reservation}) \bowtie_{RISBN=IISBN} \text{Item}) \bowtie_{Ititle\_no=Ttitle\_no} (\sigma_{\text{t.ISBN = '500'}} \text{Title}))$$

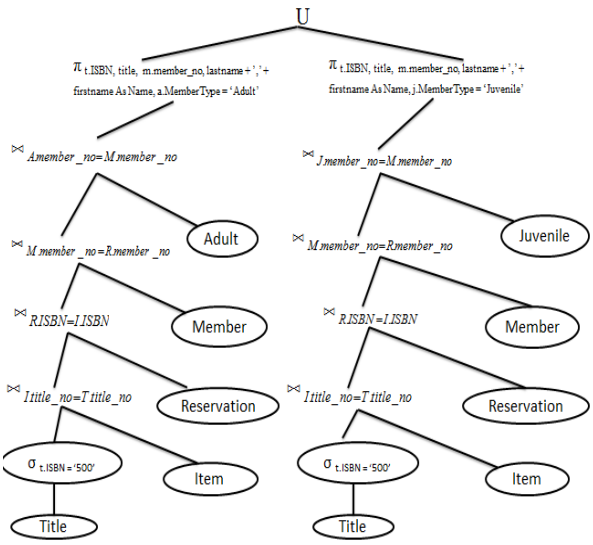
U

$$\pi_{\text{t.ISBN, title, m.member\_no, lastname + ',' + firstname As Name, j.MemberType = 'Juvenile'}}$$

$$((((\text{Juvenile} \bowtie_{Jmember\_no=Mmember\_no} \text{Member}) \bowtie_{Mmember\_no=Rmember\_no} \text{Reservation}) \bowtie_{RISBN=IISBN} \text{Item}) \bowtie_{Ititle\_no=Ttitle\_no} (\sigma_{\text{t.ISBN = '500'}} \text{Title}))$$

Many inner joins are consider for obtaining the required information taking into account the selection - t.ISBN = '500' and the optimizer uses UNION operation to gather wanted information from the datasets ADULT and JUVENILE. For estimating the cost from this query, consider query tree,

CASE2: QUERY TREE:



We have two datasets ADULT and JUVENILE combining with UNION operation. And the same information i.e, selection operator "t.ISBN = '500' " is required as a result in both the datasets. So the tree continues from UNION operation as shown in the figure, and continues with the left depth tree and the right depth tree which have the same indexes and joins to be considered.

First we'll estimate the cost for the left depth tree and the same cost is taken for the right depth tree which combines with a UNION operation.

The indexes available for this left depth tree query are: for Member, a B+ tree index on the member_no and hash tree on the member_no; for Reservation, a B+ tree index on the member_no and hash tree on the ISBN; for Item, a B+ tree index on the member_no and a clustered B+ tree on the ISBN field; for title, a B+ tree index on the ISBN(join method with title) field and a hash tree index on the ISBN(join method with title) field.

For (Adult, Member, Reservation, and Item), the best plan is obviously a file scan as seen in earlier case. The best plan for Title is to hash index on ISBN, which matches the selection t.ISBN = '500'. The B+ tree on ISBN also matches this selection and is retained even though the hash index is cheaper, because it returns tuples in stored order by ISBN field. Similarly, the joins processed in CASE2 are: all possible file scans for Adult, Member, Reservation, Item, Title as seen in earlier CASE1 and; Title accessed via B+ tree index on ISBN(outer) with Adult(inner), Title accessed via hash tree index on ISBN(outer) with Adult(inner),Title accessed via B+ tree index on ISBN(outer) with Member(inner),Title accessed via hash tree index on ISBN(outer) with Member(inner),Title accessed via B+ tree index on ISBN(outer) with Reservation(inner), Title accessed via hash tree index on ISBN(outer) with Reservation(inner), Title accessed via B+ tree index on ISBN(outer) with Item(inner), Title accessed via hash tree index on ISBN(outer) with Item(inner).

We consider here two examples for cheapest plan. First example, with Title accessed via hash index on ISBN as the outer relation, an index nested loops join accessing Item via the B+ tree index on ISBN(join method title) is likely to be a good plan; observe that there is no hash index on this field of Item. Another plan for joining Item and Title is to

access Title using the hash index on ISBN, access Item using the B+ tree on ISBN (join method title), and use a sort-merge order by ISBN (join method title).It is retained even if the previous plan is cheaper, unless an even cheaper plan produces the tuples in sorted order by ISBN (join method title). As told in the previous CASE1, considering the cost of sorting on member_no field, the cheapest plan is generated. From this, even right depth tree sorting cost is considered to be cheapest plan. And the result cost is clubbed by the UNION operation which overall is generated to be cheapest cost plan.

Similarly rest of the queries built in library databases are translated into relational algebra and the formation of query tree for estimating the least cost plan are done.

4. Conclusion

In this paper, we have described the query optimization of a single query block, which is expressed by translating SQL queries into relational algebra expression. We have implemented the design of library databases and used those SQL queries for the purpose of our optimization. Generating the cheapest cost plan is estimated in the formation of query tree. The process of finding a good plan for any SQL query gets complex when we require joins. Even for cases like this, we evaluated the cheapest query plan using left deep plans. But, the downside to this method of left deep plans is that, if the number of joins is more than 15 or so, analyzing the cost of optimization becomes complex.

5. References

[1] Raghu Ramakrishnan and Johannes Gehrke: 'Database Management Systems' - Introduction to database design and translation into relational algebra including tree structures, third edition, 2003, pages 25-110,344-385,478-507.

[2] Elmasri, Navathe: ' Fundamentals of Database systems', 2nd Edition,1994.

[3] A Swami, Optimization of Large join Queries Combining Heuristics and Combinatorial Techniques, in Proceedings of the 1989 ACM-SIGMOD Conference, Portland, OR, June 1989

[4] Harrington, Jan L.: 'Relational database design and implementation | clearly explained, third edition.

[5]Henk Ernst Blok, Djoerd Hiemstra and sunil choenni, Franciska de jong, Henk M. Blanken and peter M.G. Apers. Predicting the cost-quality tradeoff for information retrieval queries: Facilitating database and query optimization. Proceedings of the tenth international conference on information and knowledge management, October 2001, pages 207-214.

[6]Micheal L. Rupley, Jr.: 'Introduction to query processing and optimization'.

[7]Jamie MacLennan, ZhaoHui Tang, Bogdan Crivat: 'Data Mining with Microsoft SQL Server 2008'.

[8]Roger Jennings: 'Professional ADO.NET 3.5 with LINQ and the entity Framework', 2009.

[9]Carlos Coronel, Steven Morris, Peter Rob: ‘Database Systems- Design, Implementation and Management', ninth edition.

[10] 'Introduction to Databases and Programming with ADO.NET'- www.philadelphia.edu.jo/courses/ADO.NET/0782141838-1.pdf

Modify LEACH Algorithm for Wireless Sensor Network

Mortaza Fahimi Khaton Abad [1] , Mohammad Ali Jabraeil Jamali [2]

[1] Graduate Student of Islamic Azad University.
Shabestar Branch. Shabestar, Iran.

[2] Department of Computer Science.
Islamic Azad University. Shabestar Branch. Shabestar, Iran.

Abstract

Research on wireless sensor networks has recently received much attention as they offer an advantage of monitoring various kinds of environment by sensing physical phenomenon. Prolonged network lifetime, scalability, and load balancing are important requirement for many sensor network applications. Clustering sensor nodes is an effective technique for achieving these goals. In this work, we introduce an energy efficient clustering algorithm for sensor networks based on the LEACH protocol. LEACH (Low Energy Adaptive Clustering Hierarchy) is one of popular cluster-based structures, which has been widely proposed in wireless sensor networks. LEACH uses a TDMA based MAC protocol, and In order to maintain a balanced energy consumption. The proposed protocol adds feature to LEACH to reduce the consumption of the network resource in each round. The proposed protocol is simulated and the results show a significant reduction in network energy consumption compared to LEACH.

Keywords: *Wireless sensor networks, hierarchical Clustering , LEACH protocol, network lifetime*

1. INTRODUCTION

Microsensor network consist of many spatially distributed sensors, which are used to monitor various kinds of ambient conditions like temperature, humidity, etc and then transform them into electric signal. A sensor is equipped with a radio transceiver, a small microcontroller, and an energy source, usually a battery. Usually sensors are physically small and inexpensive. Small sensors are not as reliable as more expensive macrosensors, but small size and small cost of an individual sensor, allow production and deployment in large numbers. A wireless sensor network contains hundreds or thousands of these sensor devices that have ability to communicate either directly to the Base Station (BS) or among each other. The nodes in WSNs are usually battery operated sensing devices with limited energy resources and replacing or replenishing the batteries is usually not an option. Thus energy efficiency is one of the most important issues and designing power efficient protocols is critical for prolonging the lifetime. Usually, sensor nodes are scattered in the sensing field, being the area where we want to monitor some ambient conditions. Sensor nodes have to coordinate among themselves to get information about the physical environment. The information collected by sensor nodes is routed to the Base Station either directly or through other sensor nodes. The Base Station is a fixed node or mobile node, which is capable to connect the sensor network to an infrastructure networks or to the Internet where users can access and process data.

Routing in WSNs is very challenging due to the specific characteristics that distinguish WSNs from other wireless networks such as wireless ad hoc networks or cellular networks. Many new algorithms have been proposed, taking into consideration the inherent features of WSNs along with the application and architecture requirements.

Based on the network structure adopted, routing protocols for WSNs can be classified into flat network routing, hierarchical network routing, location-based network routing [3].

In flat network routing, all nodes have the same functionality and they work together to perform sensing and routing tasks.

The Sensor Protocols for Information via Negotiation (SPIN) [4] and Directed Diffusion [5] fall into this category. Hierarchical network routing divides the network into clusters to achieve energy-efficient, scalability and one of the famous hierarchical network routing protocol is low-energy adaptive clustering hierarchy (LEACH) [1]. In location-based network routing, location information of nodes is used to compute the routing path. This information can be obtained from global positioning system (GPS) devices attached to each sensor node. Examples of location-based network routing protocols include geography adaptive routing (GAF) [1] and Geographic and Energy-Aware Routing (GEAR) [6].

During the creation of network topology, the process of setting up routes in WSNs is usually influenced by energy

considerations. Because the power attenuation of a wireless link is proportional to square or even higher order of the distance between the sender and the receiver, multi-hop routing is assumed to use less energy than direct communication. However, multi-hop routing introduces significant overhead to maintain the network topology and medium access control. In the case that all the sensor nodes are close enough to the BS, direct communication could be the best choice for routing since it reduces network overhead and have a very simple nature. But in most cases, sensor nodes are randomly scattered so multi-hop routing is unquestionably defacto. Many research projects and papers have shown that the hierarchical network routing and specially the clustering mechanisms make significant improvement in WSNs in reducing energy consumption and overhead [7, 8] also have to note that most of clustering protocols proposed for WSNs assume that nodes are stationary. The reason for sensor nodes to be taken as stationary is the assumption of simple network topology. Clustering protocols can reduces signaling overhead since they do not have to manage the mobility pattern or location information of sensor nodes. As a result, it allows nodes saving more energy leading to a longer network life time. However, with some applications such as animal tracking, search and rescue activities this assumption is not very realistic; hence there are raising demands for clustering protocols to support mobile nodes.

Clustering network is efficient and scalable way to organize WSNs [1, 2]. A cluster head responsible for conveying any information gathered by the nodes in its cluster and may aggregate and compress the data before transmitting it to the sink. However, this added responsibility results in a higher rate of energy drain at the cluster heads. One of the most popular clustering mechanisms, LEACH, addresses this by probabilistically rotating the role of cluster head among all nodes. However, unless each node selects its probability of becoming a cluster head wisely, the performance of the network may be far from optimal. The main focus of this paper is modifying LEACH clustering algorithm. This algorithm fully utilizes the location information of network nodes in routing to reduce the routing cost.

2. RELATED WORK

Clustering is the method by which sensor nodes in a network organize themselves into hierarchical structures. By doing this, sensor nodes can use the scarce network resources such as radio resource, battery power more efficiently. Within a particular cluster, data aggregation and fusion are performed at cluster- head to reduce the amount of data transmitting to the base station. Cluster formation is usually based on remaining energy of sensor nodes and sensor's proximity to cluster-head [1]. Non cluster-head nodes choose their cluster-head right after deployment and transmit data to the cluster-head. The role of cluster-head is to forward these data and its own data to the base station after performing data aggregation and fusion. LEACH is one of the first hierarchical routing protocols for WSNs. The idea proposed in LEACH has inspired many other hierarchical routing protocols [9, 10].

2.1 LEACH and LEACH-C

LEACH (Low-Energy Adaptive Clustering Hierarchy), an energy-conserving routing protocol for wireless sensor network, was proposed by Heinzelman, Chandrakasan and Balakrishnan [1].

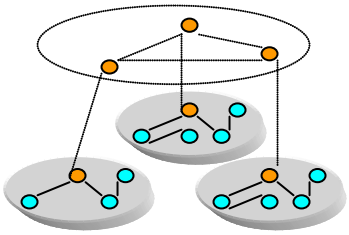


Fig.1 Structure of clustered WSNs.

The idea is to form cluster of sensor nodes based on signal strength and use the cluster-head as a router to forward data of other nodes in cluster to the base station. The data processing is performed at cluster-heads. LEACH is a dynamic clustering mechanism. Time is divided in rounds/intervals with equal length. At the beginning of the round, cluster-heads is generated randomly among the nodes which have remaining energy higher than the average remaining energy of all the nodes.

Each sensor node *n* generates a random number such that 0< random < 1 and compares it to a pre-defined threshold *T (n)*. If random < *T (n)*, the sensor node becomes cluster-head in that round, otherwise it is cluster member.

$$T(n) = \frac{P}{1 - P\left(r \bmod \frac{1}{P}\right)} \qquad \forall n \in G \tag{1}$$

In this formula, *p* is the percentage of cluster heads over all nodes in the network, i.e., the probability that a node is selected as a cluster head; r the number of rounds of selection; and G is the set of nodes that are not selected in round 1/p. As we can see here, the selection of cluster heads is totally randomly.

After becoming clusterheads, the nodes broadcast messages to all nodes to inform the status of them. Non cluster-head nodes decide which clusterhead to join based on the receiving signal strength of these messages.

The cluster-heads create schedules and send to all the nodes in the clusters. For the rest of the round, the nodes send data to their respective cluster head nodes, then the cluster-heads aggregate and send the data to the base station.

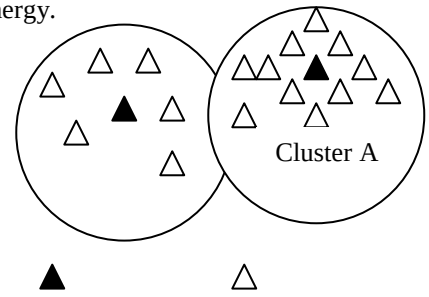
After each round, clusters-heads are re-generated to form new clusters. The cluster-head rotation allows network to spend energy equally between sensor nodes and hence it can lengthen the sensor network life time.

LEACH-centralized (LEACH-C) [1] is similar to LEACH in operation except cluster formation. In LEACH-C, the cluster head selection is carried out at BS. During the setup phase, BS receives from other nodes information about their current locations and remaining energy levels. BS uses the remaining energy level to determine the candidate set for cluster head node. The average node energy is computed and the node has remaining energy falling below this value will be removed from the candidate set.

Using the candidate set, BS finds clusters using the simulated annealing algorithm [11] to solve the NP-hard problem of finding *k* optimal clusters [12]. This algorithm attempts to minimize the total energy that noncluster head nodes use to transmit their data to cluster head nodes by minimizing the total sum of squared distance between nodes and their cluster head nodes. Once the cluster head nodes are determined, BS broadcast to all nodes the information including cluster head nodes, clusters member node and transmission schedule for each cluster. Nodes use this information to determine its TDMA slot for data transmission.

2.2 Disadvantages of LEACH

Despite the obvious advantages in using LEACH protocol for cluster organization, few features are still not supported. LEACH assumes a homogeneous distribution of sensor nodes in the given area. This scenario is not very realistic. Let us consider a scenario in which most of the sensor nodes are grouped together around one or two cluster-heads. As being shown in Figure 2, cluster-head a have more nodes close to it's than the other cluster-heads. LEACH's cluster formation algorithm will end up by assigning more cluster member nodes A. This could make cluster head nodes a quickly running out of energy.



Cluster-head Sensor nodes

Fig.2 A Sensor Network.

In addition, cluster heads are randomly selected, it is possible the scenario illustrated in Figure 3 occurs, in which two or even more cluster heads are very close to each other.

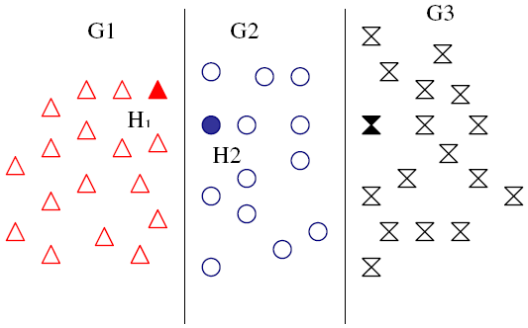


Fig.3 Multiple Cluster-head in small region.

In Figure 3, H1 and H2 are two cluster heads, nodes ▲ and ● are their cluster members, respectively. H1 and H2 are very closely located. According to data communication model, the energy that a cluster head consumes is the sum of that consumed in receiving data and that in sending data.

$$E_{ch} = L E_{bit} N_{mem} + L E (N_{mem} + 1) + L E_{bit} + L m d_{to\ Bs} \quad (1)$$

where L is the length of data, m the power consumption of transferring l bit of data, E_{bit} the power consumption of processing 1 bit of data, N_{mem} the number of members in a cluster, $d_{to\ Bs}$ the distance between the cluster head and node Sink, $LE_{elec}N_{mem}$ the power that N_{mem} cluster members consume when each of them send out length of l data to the cluster head, and LEN_{mem} the power that the cluster head consumes when it receives data of length l from its cluster members. It follows from (1) that the amount of energy that cluster heads H1 and H2 consume during data transfer is:

$$E_{h1} = LE_{bit}N_{mem1} + LE (N_{mem1} + 1) + L E_{bit} + L m d_{h1\ to\ BS} \quad (2)$$

$$E_{h2} = LE_{bit}N_{mem2} + LE (N_{mem2} + 1) + L E_{bit} + L m d_{h2\ to\ BS} \quad (3)$$

Where N_{mem1} and N_{mem2} the number of members in clusters H1 and H2, d_{h1toBs} and $d_{h2to\ Bs}$ the distance between the two cluster heads and node Sink, Therefore, the total energy consumed by the two clusters is:

$$E_{h1} + E_{h2} = L E_{bit} (N_{mem1} + N_{mem2}) + LE (N_{mem1} + N_{mem2} + 2) + 2LE_{bit} + L m (d_{h1toBS} + d_{h2toBS}) \quad (4)$$

When H1 and H2 are very close, we can have

$$d_{h1toBS} = d_{h2toBS}$$

Then (4) becomes

$$E_{h1} + E_{h2} = LE_{bit}(N_{mem1} + N_{mem2}) + LE(N_{mem1} + N_{mem2} + 2) + 2LE_{bit} + 2Lm\, d_{h1toBS} \quad (5)$$

As we can see, in this case the total energy consumption of two clusters is only $LE_{bit}+Lmd_{h1tobs}$ greater than the case that there is only one cluster head. In addition, because $LE_{bit}+Lmd_{h1tobs}$ is much greater than therefore, the total energy consumption when there are two cluster heads is approximately twice of that when there is only one cluster head.

It is clear now that when multiple cluster heads are randomly selected within a small area, a big extra energy loss occurs. The amount of lost energy is approximately proportional to the number of cluster heads in the area. Of course, there is a precondition on this conclusion, that is, cluster heads are very closely located and the distance between them becomes negligible.

3. PROTOCOL PERFORMANCE

LEACH protocol is suitable for the WSNs under the following assumptions:

1. All senor nodes are identical and charged with the same amount of initial energy. All nodes consume energy at the same rate and are able to know their residual energy and control transmission power and distance. Every node has the capability to support different MAC protocol and data processing. All communication channels are identical. The energy consumption of transferring data from node A to node B is the same as that of transferring the same amount of date from node B to node A.

2. Every node can directly communicate with every other node, including the sink node.

3. The Sink node is fixed and far away from the wireless network. Thus we can ignore the energy consumed by the sink node. We assume that it always has sufficient energy to operate.

4. Every node has data to transfer in every time frame. The data transferred by sobering nodes are related and can be fused.

5. Sensor nods are static.

WSNs are autonomous networks. Sensor nodes are independent with each other. The coordination between nodes is done through wireless communication, which costs much. This is one of the major reasons that the LEACH protocol selects cluster heads randomly. As we discussed before, this approach may cause the waste of energy because of unbalanced cluster head distribution. To solve this problem, we propose a new approach to selecting cluster heads. We assume that:

1. The network satisfies the pre-conditions of applying LEACH protocol.
2. After deployment, sensors are able to know their positions through GPS, or before deployment, their positions are accurately decided.
3. All nodes are able to adjust data transmission power. If necessary they can communicate with the base stations to acquire the initial setting information of the network.

If we modify the procedure of the calculation of $T(n)$ during the cluster head generation such that cluster heads are produced progressively, then a node could decide if it is suitable to be a new cluster head based on the locations of existing cluster heads and its own location. More specifically, if the node is very close to any existing cluster head, then this node will give up the attempt to be a cluster head.

As shown in Figure 4, the network is divided into three parts. Nodes in region G1 will compete for being a cluster head. When a node is selected as a cluster head, it will broadcast the information to nodes nearby. Nodes in region G2 will receive the message. Thus, when nodes in this region compete for being cluster head, the location information of the cluster head in region G1 will taken into consideration. If a node in G2 is close to the cluster head in G1, the node will be discarded. The cluster heads in all other regions will be generated in the same way.

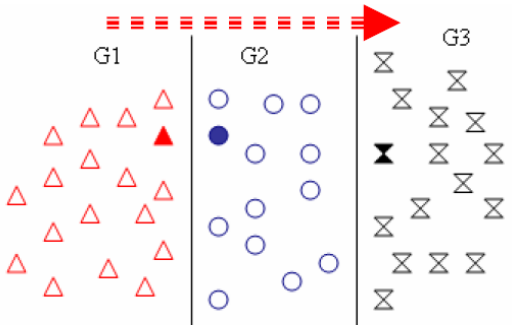


Fig.4 Selecting Cluster-heads.

The cluster heads generated with this approach will not be close to each other. However, because some nodes quit the competition for cluster head, the total number of cluster heads can be reduced, which is not good for saving the network energy. Our approach to solving this problem is when a node

is excluded in the cluster head selection, a message is broadcast to other nodes and $T(n)$ will be modified to increase the probability of others nodes being selected as cluster heads. The modified $T(n)$ is:

$$T(n) = \begin{cases} \dfrac{P}{1 - P(r \bmod (1/p)) - pk} & n \in G \\ 0 & \text{others} \end{cases}$$

K is the number of nodes that are excluded from the cluster head selection due to the location reason, with an initial value of 0. When k increases, $T(n)$ increases as well, which will ensure sufficient number of cluster heads will be generated by the progressive algorithm.

To facilitate the explanation of our improved algorithm, we introduce the following notations:

Bs The base station or node *Sink*
S_i The *i*-th sensor node
H_j The *j*-th cluster head
Mem (C_j) Members of the *j*-th cluster
Mem $(C_j)_I$ The *i*-th members of the *j*-th cluster
Loc (S_i) Location of the *i*-th sensor node
Delay (S_i) Time delay that the *i*-th sensor node start to compete for a cluster head
Num(Giveup) Number of discarded cluster heads
|| Operation of concatenation

3.1 cluster head selection

After the deployment of sensor nodes, we first acquire all nodes' location information (through GPS technology or known prior to its deployment) and report it to the base station. The base station decides Delay (S_i) for every node based on the geographic distribution of all sensor nodes.

Delay (S_i) = 0 for those in the region to start first. As illustrated in Figure 4, nodes in G1 start to compete for cluster heads at time 0, then nodes in G2 start with a delay, and then nodes in G3 start with a delay after nodes in G2 are finished, and so on. During the process, nodes need to send their location information to the base station:

S_i → BS: Loc (S_i)

The base station needs to send the delay information to each node:

Bs → S_i: Delay (S_i)

Set Num (Giveup) to 0. Start with the nodes in G1. If a cluster head is generated from G1, broadcast a *Hello* package and Num (Giveup).

H_j → broadcast: Hello, Num (Giveup)

When nodes in G1 are finished, consider nodes in G2. Now the cluster heads generated in G1 are reference points. The distance between a node in G2 and any cluster head in G1 is a factor in selecting the node as a cluster head, as well as the random value of T (n). If all conditions are satisfied, then broadcast the Hello message and Num (Giveup).

Hj → broadcast: Giveup, Num (Giveup)

Otherwise, only broadcast Num (Giveup). When nodes in other region receives this message, they will increment Num (Giveup) by 1, and then modify T (n) to increase the probability of being selected as cluster head.
Repeat the above process until all nodes in the network are considered.

4. SIMULATIONS

In this section, we evaluate the performance of My LEACH protocol implemented with NS2.

100 sensor nodes are randomly distributed in an area of 100 m x 100 m. BS is put at the location with x = 175, y = 50. The bandwidth of data channel is set to 1 Mbps, the length of data messages is 500 bytes and packet header for each type of packet was 25 bytes. The number round is set to 500s. When a node uses energy down to its energy threshold, it can no longer send data and is considered as a dead node.

E_{bit} = 50nJ/bit
E = 5nJ/bit/report
E_0 = 0.5 J

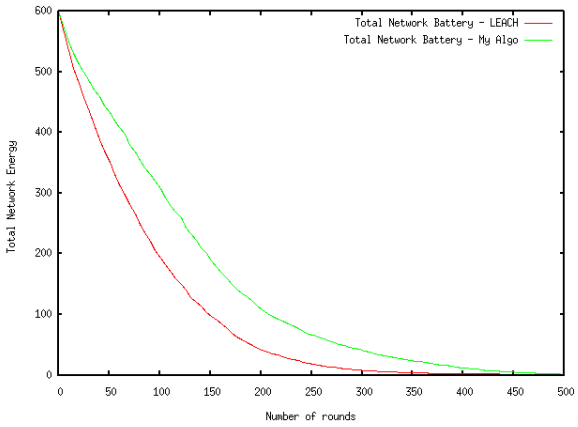


Fig.5 Compare Total Network Energy LEACH and My Algo.

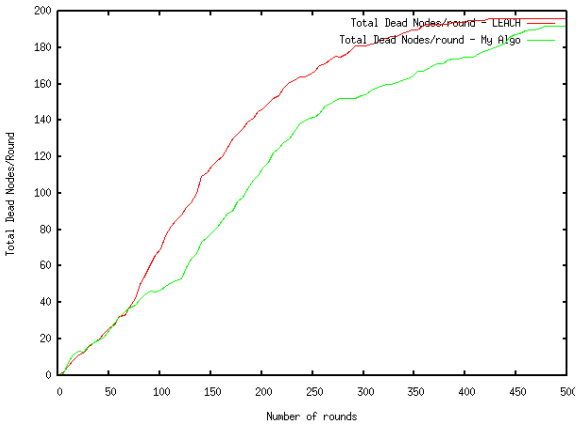


Fig.6 Compare Total Dead Nodes LEACH and My Algo

References

[1] W. Heinzelman, A. Chandrakasan and H. Balakrishnan, ”Energy-efficient communication protocols for wireless microsensor networks” Proceedings of the Hawaii International Conference on Systems Sciences, Jan. 2000.

[2] W. Heinzelman, A. Chandrakasan and H. Balakrishnan, ”An application-specific protocol architecture for wireless microsensor networks”, IEEE Transactions on Wireless Communication, vol. 1, no. 4, pp. 660-670, 2002.

[3] Lu, Ye Ming and Vincent W. S. Wong, “An energy-efficient multipath routing protocol for wireless sensor networks” research articles. Int. J.Commun. Syst., 20(7):747--766, 2007.

[4] W.Heinzelman , J.Kulik, H.Balakrishnan, “Adaptive protocols for information dissemination in wireless sensor networks” Proceedings of ACM/IEEE MobiCom’99, Seattle, WA, U.S.A., August 1999.

[5] C.Intanagonwiwat, R.Govindan, D.Estrin, “Directed diffusion: a scalable and robust communication paradigm for sensor networks” Proceedings of ACM MobiCom’00, Boston, MA, U.S.A., Aug 2000.

[6] Y.Xu, J.Heidemann, D.Estrin “Geography-informed energy conservation for ad-hoc routing” Proceedings of ACM/IEEE MobiCom’01, Rome, Italy, July 2001.

[7] T. C. Hou, T. J.Tsai, “An access-based clustering protocol for multihop wireless ad hoc networks ”IEEE Journal on Selected Areas in Communications, July 2001.

[8] M.Joa-Ng, I.T.Lu, “A Peer-to-peer Zone-based Two-level link state routing for mobile Ad Hoc Networks”, IEEE Journal on Selected Areas in Communications, Special Issue on Ad-hoc Networks, Aug1999.

[9] A. Manjeshwar, D.P. Agrawal, TEEN,”A protocol for enhanced efficiency in wireless sensor networks” Proceedings of the 1st International Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing, San Francisco, CA, April 2001.

[10] A. Manjeshwar, D.P. Agrawal, APTEEN, “A hybrid protocol for efficient routing and comprehensive information retrieval in wireless sensor networks” Proceedings of the 2nd International Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile computing, Ft. Lauderdale, FL, April 2002.

[11] T. Murata and H. Ishibuchi, “Performance evaluation of genetic algorithms for flowshop scheduling problems” Proc. 1st IEEE Conf. Evolutionary Computation, June 1994.

[12] P. Agarwal and C. Procopiuc, “Exact and approximation algorithms for clustering,” in Proc. 9th Annual. ACM-SIAM Symp. Discrete Algorithms, Baltimore, MD, Jan. 1999.

First Author Biographies : Mohammad Ali Jabraeil Jamali. received his B.Sc. in Electrical Engineering and M.Sc. in Computer Hardware Engineering from Oromie University in 1994 and Tehran Science and Research Branch of Islamic Azad University in 2003, respectively. Currently, he is a Ph.D. student and Faculty Member at the Tehran Science and Research Branch of Islamic Azad University under supervision of Dr. Ahmad khademzadeh and Islamic Azad University, Shabestar Branch, respectively. His research interests include VLSI Design, Interconnection Network, Fault Tolerant, Ad hoc Network, Sensor Network and Computer Architectures.

Education
B.S., University of Oromie
M.S., University of Oromie
Ph.D., Science and Research Branch,Islamic Azad University of Tehran

Journal Papers
1-The Optimum Location of Delay Latches Between Dynamic Pipeline Stages
2- Improved Circles Intersection Algorithm for Localization in Wireless Sensor Networks
3- An energy-efficient algorithm for connected target coverage problem in wireless sensor network
4-DAMQ-Based Schemes for chemes Efficiently Using the Buffer Spaces of a NoC Router

Second Author Biographies : Mortaza Fahimi Khaton Abad Graduate Student of Islamic Azad University, Shabestar Branch. Shabestar, Iran.

The Descriptive Study of Knowledge Discovery from Web Usage Mining

Yogish H K[1] Dr. G T Raju[2] Manjunath T N[3]

[1]Department of Computer Science and Engineering
REVA Institute of Technology and Management,
Yelahanka. Bangalore-560064, Karnataka, India.
(Research Scholar - Bharathiar University Coimbatore-641046)

[2]Department of Computer Science and Engineering
RNS Institute of Technology, Bangalore -560061,
Karnataka, India.

[3]Wipro Technologies
(Research Scholar - Bharathiar University Coimbatore-641046)

Abstract

The World Wide Web serves as huge, widely distributed, global information service centre for news, advertisements, consumer information, financial management, education, government, e-commerce and many other information services. The web also contains a rich and dynamic collection of hyperlink information and web page access and usage information, providing rich sources of data for data mining. The Web usage mining is the area of data mining which deals with the discovery and analysis of usage patterns from web logs, in order to improve web based applications. Web usage mining consists of three phases, preprocessing, pattern discovery, and pattern analysis. After the completion of these three phases the user can find the required usage patterns and use this information for the specific needs.

Keywords*: Data mining, Knowledge Discovery, bot, Preprocessing, associations, clustering, web data.*

1. Introduction

Web Usage Mining is a part of Web Mining, which, in turn, is a part of Data Mining. As Data Mining involves the concept of extraction meaningful and valuable information from large volume of web data.

The web data is:

1. Content: The visible data in the Web pages or the information which was meant to be imparted to the users. A major part of it includes text and graphics (images).
2. Structure: Data which describes the organization of the website. It is divided into two types. Intra-page structure information includes the arrangement of various HTML or XML tags within a given page. The principal kind of inter-page structure information is the hyper-links used for site navigation.
3. Usage: Data that describes the usage patterns of Web pages, such as IP addresses, page references, and the date and time of accesses and various other information depending on the log format.

In Web Usage Mining[3,5,7], data can be collected in server logs, application server logs, browsers logs, user's profiles, user's queries, book marked data, mouse clicks and scrolls, registration data, cookies, user sessions or transactions data. These data collections differ in terms of the location of the data source, the kinds of data available, the segment of population from which the data was collected, and methods of implementation.

The logs can be examined from either server or client perspective. When evaluated from a server perspective, mining uncovers informs about a sites where the service resides, it can be used to improved the design of the sites. By evaluating client's sequence of clicks information about the users or group of users or detected. This could be used to perform pre-fetching and caching of pages.

For Example: The web master at ABC [3] corporation learns that a high percentage of users have the following patterns of reference to pages :(A, B, A, C). This means that user access page A, then page B, then back to page A and finally to page C. Based on this observation he determines that a link is needed directly to page C from B. He then adds this link.

Web Usage mining involves mining the usage characteristics of the users of Web Applications. This extracted information can then be used in a variety of ways such as, improvement of the application, checking of fraudulent elements etc.

Web Usage Mining is often regarded as a part of the Business Intelligence in an organization rather than the technical aspect. It is used for deciding business strategies through the efficient use of Web Applications. It is also crucial for the Customer Relationship Management (CRM) as it can ensure customer satisfaction as far as the interaction between the customer and the organization is concerned.

The major problem with Web Mining in general and Web Usage Mining in particular is the nature of the data they deal with. With the upsurge of Internet in this millennium, the Web Data has become huge in nature and a lot of transactions and usages are taking place by the seconds. Apart from the volume of the data, the data is not completely structured. It is in a semi-structured format so that it needs a lot of preprocessing and parsing before the actual extraction of the required information.

The Web Usage Mining process consists three phases, namely preprocessing, pattern, and discovery and pattern analysis. This paper describes each of these phases in detail.

1.1 Motivation

In the current era, we are witnessing a surge of Web Usage around the globe. A large volume of data is constantly being accessed and shared among a varied type of users; both humans and intelligent machines. Thus, taking up a structured approach to control this information exchange, has what made Web Mining one of the hot topics in the field of Information Technology.

2 Data Sources [1]

The data sources used in Web Usage Mining may include web data repositories like:

Web Server Logs [7] - These are logs which maintain a history of page requests. The W3C maintains a standard format for web server log files, but other proprietary formats exist. More recent entries are typically appended to the end of the file. Information about the request, including client IP address, request date/time, page requested, HTTP code, bytes served, user agent, and referrer are typically added.

These data can be combined into a single file, or separated into distinct logs, such as an access log, error log, or referrer log. However, server logs typically do not collect user-specific information. These files are usually not accessible to general Internet users, only to the webmaster or other administrative person. A statistical analysis of the server log may be used to examine traffic patterns by time of day, day of week, referrer, or user agent. Efficient web site administration, adequate hosting resources and the fine tuning of sales efforts can be aided by analysis of the web server logs. Marketing departments of any organization that owns a website should be trained to understand these powerful tools.

Proxy Server Logs - A Web proxy is a caching mechanism which lies between client browsers and Web servers. It helps to reduce the load time of Web pages as well as the network traffic load at the server and client side. Proxy server logs contain the HTTP requests from multiple clients to multiple Web servers. This may serve as a data source to discover the usage pattern of a group of anonymous users, sharing a common proxy server.

Browser Logs - Various browsers like Mozilla, Internet Explorer Opera etc. can be modified or various JavaScript and Java applets can be used to collect client side data. This implementation of client-side data collection requires user cooperation, either in enabling the functionality of the JavaScript and Java applets, or to voluntarily use the modified browser. Client-side collection scores over server-side collection because it reduces both the bot and session identification problems.

3. Information Obtained [3]

i. ***Number of Hits*****:** This number usually signifies the number of times any resource is accessed in a Website. A hit is a request to a web server for a file (web page, image, JavaScript, etc.). When a web page is uploaded from a server the number of "hits" or "page hits" is equal to the number of files requested. Therefore, one page load does not always equal one hit because often pages are made up of other images and other files which stack up the number of hits counted.

ii. ***Number of Visitors****:* A "visitor" is exactly what it sounds like. It's a human who navigates to your website and browses one or more pages on your site.

iii. ***Visitor Referring Website****:* The referring website gives the information or URL of the website which referred the particular website in consideration.

iv. ***Visitor Referral Website:*** The referral website gives the information or URL of the website which is being referred to by the particular website in consideration.

v. Time and Duration: This information in the server logs give the time and duration for how long the Website was accessed by a particular user.

vi. Path Analysis: Path analysis gives the analysis of the path a particular user has followed in accessing contents of a Website.

vii. Visitor IP address: This information gives the Internet Protocol (I.P.) address of the visitors who visited the Website in consideration.

viii. Browser Type: This information gives the information of the type of browser that was used for accessing the Website.

ix. Cookies: A message given to a Web browser by a Web server. The browser stores the message in a text file called cookie. The message is then sent back to the server each time the browser requests a page from the server. The main purpose of cookies is to identify users and possibly prepare customized Web pages for them. When you enter a Web site using cookies, you may be asked to fill out a form providing such information as your name and interests. This information is packaged into a cookie and sent to your Web browser which stores it for later use. The next time you go to the same Web site, your browser will send the cookie to the Web server. The server can use this information to present you with custom Web pages. So, for example, instead of seeing just a generic welcome page you might see a welcome page with your name on it.

x. ***Platform:*** This information gives the type of Operating System etc. that was used to access the Website.

4. Possible Actions [9]

i. Shortening Paths of High visit Pages: The pages which are frequently accessed by the users can be seen as to follow a particular path. These pages can be included in an easily accessible part of the Website thus resulting in the decrease in the navigation path length.

ii. Eliminating or Combining Low Visit Pages: The pages which are not frequently accessed by users can be either removed or their content can be merged with pages with frequent access.

iii. Redesigning Pages to help User Navigation: To help the user to navigate through the website in the best possible manner, the information obtained can be used to redesign the structure of the Website. Redesigning Pages For Search Engine Optimization: The content as well as other information in the website can be improved from analyzing user patterns and this information can be used to redesign pages for Search Engine Optimization so that the search engines index the website at a proper rank.

iv. Help Evaluating Effectiveness of Advertising Campaigns: Important and business critical advertisements can be put up on pages that are frequently accessed.

5. Web Usage Mining Process [1]:

The main processes in Web Usage Mining are:

Preprocessing: Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

Pattern Discovery: Web Usage mining can be used to uncover patterns in server logs but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data. The following are the pattern discovery methods.

1. Statistical Analysis
2. Association Rules
3. Clustering
4. Classification
5. Sequential Patterns

Pattern Analysis [11]: This is the final step in the Web Usage Mining process. After the preprocessing and pattern discovery, the obtained usage patterns are analyzed to filter uninteresting information and extract the useful information. The methods like SQL
(Structured Query Language) processing and OLAP (Online Analytical Processing) can be used.

5.1 Web Usage Mining Areas [15]

1. Personalization
2. System Improvement
3. Site Modification
4. Business Intelligence
5. Usage Characterization

5.2 Web Usage Mining Applications [3]

i. Letizia [4]: Letizia is an application that assists a user browsing the Internet. As the user operates a conventional Web browser such as Mozilla, the application tracks usage patterns and attempts to predict items of interest by performing concurrent and autonomous exploration of links from the user's current position. The application

uses a best-first search augmented by heuristics inferring user interest from browsing behavior.

ii. WebSift[1]: The WebSIFT (Web Site Information Filter) system is another application which performs Web Usage Mining from server logs recorded in the extended NSCA format (includes referrer and agent fields. The preprocessing algorithms include identifying users, server sessions, and identifying cached page references through the use of the referrer field. It identifies interesting information and frequent item sets from mining usage data.

iii. Adaptive Websites: An adaptive website adjusts the structure, content, or presentation of information in response to measured user interaction with the site, with the objective of optimizing future user interactions. Adaptive websites are web sites that automatically improve their organization and presentation by learning from their user access patterns. User interaction patterns may be collected directly on the website or may be mined from Web server logs. A model or models are created of user interaction using artificial intelligence and statistical methods. The models are used as the basis for tailoring the website for known and specific patterns of user interaction.

6. Analysis of Web Server Logs

We used different web server log analyzers like Web Expert Lite 6.1 and Analog6.0 to analyze various sample web server logs obtained. The key information obtained was:
Total Hits, Visitor Hits, Average Hits per Day, Average Hits per Visitor, Failed Requests, Page Views Total Page Views, Average Page Views per Day , Average Page Views per Visitor, Visitors Total Visitors Average Visitors per Day, Total Unique IPs , Bandwidth, Total Bandwidth , Visitor Bandwidth , Average Bandwidth per Day, Average Bandwidth per Hit, Average Bandwidth per Visitor. Access Data like files, images etc., Referrers, User Agents etc. Analysis of above obtained information proved Web Usage Mining as a powerful technique in Web Site Management and improvement.

Fig: shows the generalized model of web usage mining [8]

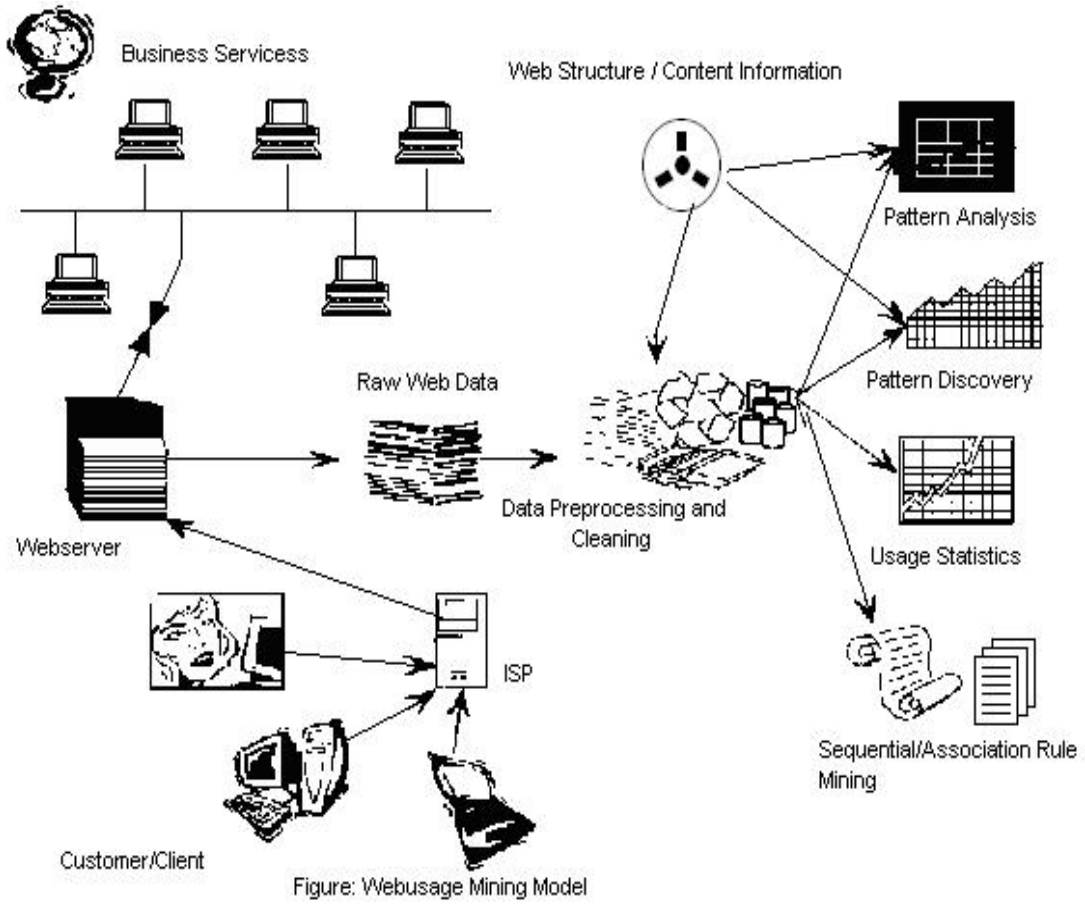


Figure: Webusage Mining Model

7. Data Preprocessing for Mining [2] [7]

Web data is collected in various ways [7], each mechanism collecting attributes relevant for its purpose. There is a need to preprocess the data to make it easier to mine for knowledge, specifically; we believe that issues such as instrumentation and data collection, data integration and transaction identification need to be addressed.

Clearly improved data quality can improve the quality of any analysis on it. A problem in the Web domain is the inherent conflict between the analysis needs of the analysts, who want more detailed usage data collected, and the privacy needs of users, who want as little data collected as possible. This has lead to the development of cookie files on one side and cache busting on the other, the emerging OPS standard on collecting profile data may be a compromise on what can and will be collected. However, it is not clear how much compliance to this can be expected. Hence, there will be a continual need to develop better instrumentation and data collection techniques, based on whatever is possible and allowable at any point in time

Web usage data collected in various logs is at a very fine granularity, Therefore, while it has the advantage of being extremely general and fairly detailed, it also has the corresponding drawback that it cannot be analyzed directly, since the analysis may start focusing on micro trends rather than on the macro trends, On the other hand, the issue of whether a trend is micro or macro depends on the purpose of a specific analysis.

Hence, we believe there is a need to group individual data collection events into groups, called Web transactions, before feeding it to the mining system.

8. The Mining Process [5][8][9]

The key component of Web mining is the mining process itself. The Web mining has adapted techniques from the field of data mining, databases, and information retrieval, as well as developing some techniques of its own.

Web mining studies reported to data have mined for association rules, temporal sequences, clusters, and path expressions. As the manner in which the Web is used continues to expand, there is a continual need to figure out new kinds of knowledge about user behavior that needs to be mined.

The quality of a mining algorithm can be measured both in terms of how effective it is in mining for knowledge and how efficient it is in computational terms. There will always be a need to improve the performance of mining algorithms along both these dimensions.

The data collection on the Web is incremental in nature. Hence, there is a need to develop mining algorithms that take as input the existing data mined knowledge, and the new data, and develop a new model in an efficient manner.

The data collection on the Web is also distributed by its very nature. If all the data were to be integrated before mining, a lot of valuable information could be extracted.

9. Analysis of Mined Knowledge [3]

The output of knowledge mining algorithms is often not in a form suitable for direct human consumption, and hence there is a need to develop techniques and tools for helping an analyst better assimilate it. Issues that need to be addressed in this area include usage analysis tools and interpretation of mined knowledge.

There is a need to develop tools which incorporate statistical methods, visualization, and human factors to help better understand the mined knowledge.

In general one of the open issues in Web mining in particular, is the creation of intelligent tools that can assist in the interpretation of mined knowledge. Clearly, these tools need to have specific knowledge about the particular problem domain to do any more than filtering based on statistical attributes of the discovered rules or patterns. In Web mining, for example, intelligent agents could be developed that based on discovered access patterns, the topology of the Web locality, and certain heuristics derived from user behavior models, could give recommendations about changing the physical link structure of a particular site.

10. Conclusion

The term Web mining has been used to refer to techniques that encompass a broad range of issues However, while meaningful and attractive. This very broadness has caused Web mining to mean different things to different people and there is a need to develop a common vocabulary. Towards this goal we proposed a definition of Web mining and developed taxonomy of the various ongoing efforts related to it. Next, we presented a survey of the research in this area. We provided a detailed

survey of the efforts in this area, even though the survey is short because of the area's newness. This paper is useful for researcher exclusively for doing research on web mining.

11. References

[1] Arun k Pujari, "Data Mining Techniques", University press, edition 2001.

[2] Jaiwei Han, Michelinne Kamber, "Data Mining: Concepts and Techniques ".

[3] Margaret H. Dunham,"Data Mining: Introductory and advanced topics".

[4] Srivastava, J., Cooley, R., Deshpande, M., and Tan, P, Web usage mining: Discovery and applications of usage patterns from Web data, SIGKDD Explorations, Vol. 1(2), 12-23, 2000.

[5] Knowledge Discovery from Web Usage Data: Extraction of Sequential Patterns through ART1 Neural Network based Clustering Algorithm, International Conference on Computational Intelligence and Multimedia Applications 2007

[6] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar .,Web Mining – Accomplishments & Future Directions

[7] Mining Web logs for Prediction in Prefetching and Caching - Third 2008 International Conferences on Convergence and Hybrid Information Technology.

[8] Ajith Abraham – Business Intelligence from Web Usage Mining - Journal of Information & Knowledge Management, Vol. 2, No. 4 (2003) 375{390c iKMS & World Scientific Publishing Co.

[9] Adel T. Rahmani and B. Hoda Helmi, EIN-WUM an AIS-based lgorithm for Web Usage Mining, Proceedings of GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA, ACM978-1-60558-130-9/08/07 (Pages 291-292)

[10] Dr. G. K. Gupta, Introduction to Data Mining with Case Studies, PHI Publication.

[11] http://web.media.mit.edu/~lieber/Lieberary/Letizia/Letizia-Intro.html

[12] http://en.wikipedia.org/wiki/Web_crawler.

[13] Prefetching based on Web Usage Mining - Daby M. Sow, David P. Olshefski, Mandis Beigi, and Guruduth Banavar, IBM T. J. Watson Research Center, Hawthorne NY, 10532, USA.

AUTHORS' BIBLIOGRAPHY

YOGISH H. K received his Bachelor's Degree in Computer Science and Engineering from PES College of Engineering, Mandya, Mysore University, Karnataka, India during the year 1998 and M. Tech in Computer Engineering from Sri Jaya Chama Rajendra College of Engineering Mysore, Karnataka, India during the year 2004. Currently pursuing PhD degree in Bharathiar University, Coimbatore. He has total 13 years of Industry and Teaching experience. His areas of interests are Data Warehouse, multimedia, Databases and Operating Systems. He has published and presented various papers in Journals, National Conferences and an author of two text books.

Dr. G T Raju received his Bachelor's Degree in Computer Science and Engineering from Kalpataru Institute Of technology, Tiptur, Karnataka, India, during the year 1992 and M. E in Computer Science and Engineering from B.M.S College of Engineering, Bangalore, Karnataka, India during the year 1995 and Doctorate of Philosophy Ph.D. in the year 2008 in Computer Science and Engineering from Visveswaraya Technological University, Belgaum, Karnataka; He has 18 years of Experience. He has visited overseas to various Universities. His area of interests is Data Mining, Data Warehousing, Image Processing, Databases, Artificial Intelligence and Computer Graphics. He has published and presented papers in journals, international and national level Conferences and published a text book.

Manjunath T N. received his Bachelor's Degree in Computer Science and Engg from Bangalore, University, Karnataka, India during The year 2001 and M. Tech in Computer Science and Engineering VTU, Belgaum, Karnataka, India during the year 2004. Currently pursuing PhD degree in Bharathiar University, Coimbatore. He has having total 10 years of Industry and teaching experience. His areas of interests are Data Warehouse & Business Intelligence, multimedia and Databases. He has published and Presented several papers.

Improved Free-Form Database Query Language for Mobile Phones

Shiramshetty Gouthami[1], Pulluri Srinivas Rao[2] and Jayadev Gyani[3]

Computer Science Engineering,JNTUH, Jayamukhi Institute of Technological Sciences
Narsampet,Warangal, Andhrapradesh-506332, India

Abstract

This paper explains that it is possible to develop a database query formulation system for mobile phones which accepts unplanned queries, by allowing imprecise inputs. Since cell phones are poor in terms of resources as compared to other devices, the success of implementing such a method on them would mean it is applicable to the other devices. Imprecise query method in the form of free-form language can provide a much simpler interface for users to formulate queries. The method can also help in reducing the number of query inputs, especially in cases where joins of relations are needed. Since the majority of queries which might be issued are of this type, providing such a method would benefit users of resource-poor devices. The language which is used as the query formulation method in a database query system prototype for WAP- enabled mobile phones has been found to be effective based on results from usability tests.

Keywords—*Mobile phone, database query language, free form queries, unplanned queries.*

1. Introduction

The objective of this paper is to explain the method of generating the query at the client side. The client should be able to generate his own query according to his own request. There will be no predefined queries behind the screen. The system does not impulse any limitations. The end user who is using this mobile application will be able to retrieve and work with data in any sort. Hence the project produces the application which is to generate all sorts of queries at the client end. The Wireless Application Protocol (WAP) is the mechanism used here in case of integration of client (Mobile) and Server (Web Server). Apache Tomcat is the web server used for the interaction with Database Server. The Wireless Application Protocol (WAP) is the mechanism used here in case of integration of client (Mobile) and Server (Web Server). Apache Tomcat is the web server used for the interaction with Database Server. Here we have tested with MySQL.

1.1 Interface Design

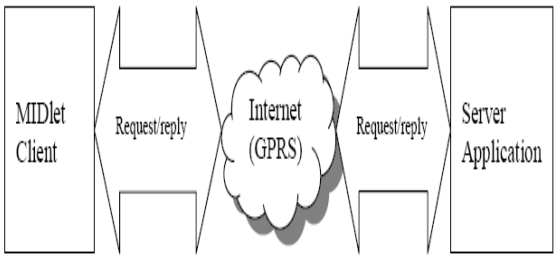


Fig.1 Communication between server and MIDlet

The structure of the system divided into two components:
The client-side MIDlet application which resides on the mobile device, e.g., mobile phone. The server-side JSP/MySQL based application.

2. Previous Work

The main objective of the system is to support unplanned queries. The end user should be able to communicate with the database in whatever option he wishes. The Programmer is not giving any limitations to the user. Since the query is generated in the client side itself, he is able to request any sort of data. The User is independent in accessing the data. When creating application upon Query Generating Systems, the main intention of the system should be to convince all the five types of classifications of queries which are Projection, Selection, Set Difference, Join and Union. In this system pull-down list contain all query terms, fields, tables, and other schema information.

3. Present Work

It is possible to develop a generic database query system for mobile phones which accepts different types of queries as well as unplanned queries, by allowing imprecise inputs. Since mobile phones are poor in terms of resources as compared to other mobile devices, the success of

implementing such a method on them would mean it is applicable to the other devices. Free-form language can provide a much simpler interface for users to formulate queries. The language helps in reducing the number of query inputs especially in cases where joins of relations are needed. Since the majority of queries which might be issued are of this type, providing such a method would benefit users of resource-poor devices. Usability tests on the prototype have also shown that the language is effective even when used by novice users. We plan to integrate recommender systems into our work so that lesser and only relevant query terms will be presented to users for selection on the pull-down list.

Modules:

i. Server Module
ii. Connection Module.
iii. Design of the Application
iv. Query Generation Module

Modules Description:

i. Server Module:

In our project, we are using MySQL as the data server and Apache Tomcat 5.0 as the web server. These two are the main core of the server side programming. Our application has been deployed in the Apache Tomcat so that all kind of Http Requests and response can be handled easily. All the parameters passed from the request can be retrieved using the getParameter() method of HttpRequest Object.

ii. Connection Module:

The Connection between the Client (i.e. J2ME application in mobile) and the Web Server is maintained by the object HttpConnectionin javax. Microedition.io. HttpConnection. Using this connection module we could retrieve the database information by passing the Http Request. The requesting attributes is sent as the parameters of the url built for HttpConnection.

iii. Design of the Application:

According to the request sent by the client, the server processes the data and it is responded to the client, which is further received by the help of openDataInputStream() in HttpConnection Object. So the client application is designed according to the Field information of the tables retrieved from the server. We would be using ChoiceGroup from the above figure. We can choose any of the radio button for composing a query. For example, get radio button for Select, remove for deleting, new for inserting and edit for updating the table.

Object for designing Radio Buttons, Check Boxes and Dropdown list boxes.

iv. Query Generation Module:

The design of the application is done by the data types which we used in the database. The choice which we are about to use, according the query will be generated behind the screen. After the complete operation of selecting the choices we are supposed to execute the query by passing it as the parameter to the server through Http Request.

4. System Prototype for Mobile Phones

In order to show that our language can support its intended capabilities, a prototype was developed. This prototype consists of a J2ME midlet for the interface on the Java phone emulator, and several Java servlets for the execution of queries. The interface developed follows the general guidelines given for mobile interfaces. J2ME such as those of Mahmoud [15] suggested that an interface for small devices should be simple and use as many as possible high-level APIs. Figure 2 below shows a screen-shot of our prototype's interface for composing free-form queries. A pull-down list is used to present all tables. These tables are presented as a linear list.

Fig 2. Interface for Composing a Query

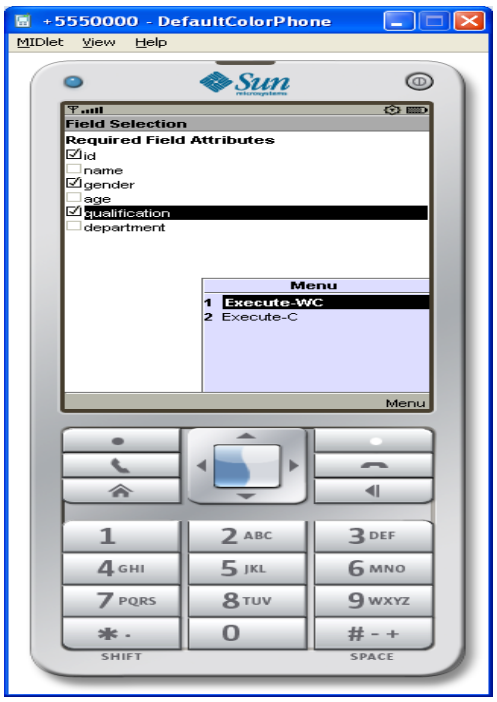


Fig 3.Interface for selecting the fields of the Contacts table

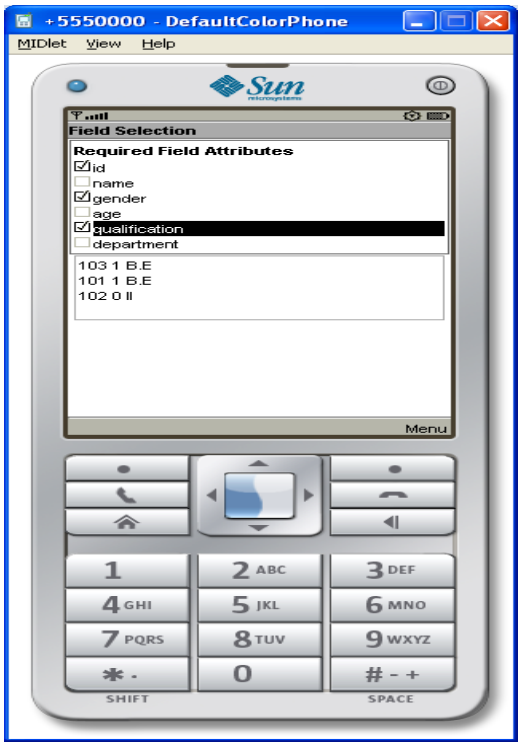


Fig 4.Interface for displaying the selected fields of the Contacts table

5. Usability Test

In order to ensure that the language is effective in supporting different query types as well as unplanned queries, usability tests were conducted on the system prototype.

Observation setup: A notebook was used as both the application and database servers for the usability tests. A Java phone emulator and a real phone, Nokia 6681, which were loaded with the interface of the prototype, were used as possible accessing devices. Besides the prototype, a test database on a university domain was used for the first two groups, and a database on asporting event was used for the third. The server was connected to the Internet for networking environment. The correctness of the query outputs expected by the observer, and the rating on the scale of 1 to 3 which was given by test subject depicting his/her acceptance of the query outputs. Allocation of credits were then determined as in Table 1 below.

Table 1. Credit allocation for measuring prototype's effectiveness

Observer's Evaluation	Participant's Ratings	Credit
Correct	3	1
	2	0.75
	1	0.5
Incorrect	3	0.70
	2	0.30
	1	0

6. Conclusions and Future Work

We conclude that it is possible to develop a database query formulation system for mobile phones which accepts unplanned queries, by allowing free-form inputs. As cell phones are poor in terms of resources as compared to other mobile devices, the successfulness of implementing such a method on them would mean it is applicable to the other devices. Indefinite query method in the form of free-form language can provide a much simpler interface for users to formulate queries. The method can also helpful in reducing the number of query inputs, particularly in cases where joins of relations are needed. Since most of queries which might be issued are of this type (as seen by the queries given by respondents), providing such a method would benefit users of resource-poor devices. In the future work, we will be trying to implement more complex queries like

aggregate functions, grouping etc.for creating and accessing database tables through the mobile phones.

Acknowledgments

We extremely thank our Principal and the management for their continuous support in Research and Development. We are also very grateful to our faculty members for their valuable suggestions and their ever ending support. Especially, we thank our college Principal and management for their financial support for receiving the sponsorship.

References

[1] R. Alonso, and H. F. Korth, "Database system issues in nomadic computing", in *Proc. 1993 SIGMOD Conference*, Washington D.C., 1993, pp. 388-392.

[2] K.Hung, and Y-T. Zhang, "Implementation of a WAP-Based telemedicine system for patient monitoring," *IEEE Transactions on Information Technology in Biomedicine*, Vol. 7, No. 2, June 2003, pp. 101-107.

[3] A. Koyama, N. Takayama, L. Barolli, Z. Cheng, and N. Kamibayashi, "An agent based campus information providing system for cellular phone," in *Proc. 1st International Symposium on Cyber Worlds*, Tokyo,2002, pp. 339-345.

[4] P. Boonsrimuang, H. Kobayashi, and T. Paungma, "Mobile Internet navigation system," in *Proc. 5th IEEE International Conference on High Speed Networks and Multimedia Communications*, Jeju Island, 2002, pp.325-328.

[5] A. Bergstrom, P. Jaksetic, and P. Nordin, "Enhancing information retrieval by automatic acquisition of textual relations using Genetic programming," in Proc IUI 2000, New Orlean, 2000, pp. 29-32.

[6] H-M. Lee, S-K. Lin, and C-W. Huang, "Interactive query expansion based on fuzzy association thesaurus for web information retrieval, " in Proc. IEEE International Fuzzy Systems Conference, Melbourne, 2001,pp. 724-727.

[7] S. Agrawal, S. Chaudhuri, and G. Das, "DBXplorer: A system for keyword-based search over relational databases," in Proc. IEEE 18th International Conference on Data Engineering (ICDE'02), San Jose, 2002, pp. 5-16.

[8] P. Calado, A.S. da Silva, A.H.F. Laender, B.A. Ribeiro-Neto, and R.C.Viera, "A Bayesian network approach to searching web databases through keyword-based queries, " Information Processing and Management, Vol. 40, No. 5, September 2004, pp. 773-790.

[9] R. Ramakrishnan, J. Gehrke, Database Management Systems. McGraw-Hill, New York, 2000.

[10] A. Chandra, "Theory of database queries," in Proc. 7th ACM Symposium on Principles of Database Systems, Texas, USA, 1988, pp. 1-9.

[11] S. Polyviou, G. Samaras, and P. Evripidou, "A relationally complete visual query language for heterogeneous data sources and pervasive querying," in Proc. 21st International Conference on Data Engineering (ICDE 2005), Tokyo, 2005, pp. 471-482.

[12] E. Chang, F. Seide, H.M. Meng, Z. Chen, Y. Shi, and Y.C. Li, "A system for spoken query information retrieval on mobile devices," IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 8,November 2002, pp. 531-541.

[13] B. R. Bai, C.L. Chen, L.F. Chien, and L.S. Lee, "Intelligent retrieval of dynamic networked information from mobile terminals using spoken natural language queries," IEEE Transactions on Consumer Electronics,Vol. 44, No. 1, February 1998, pp. 62-72.

[14] R. Ahmad, S. Abdul-Kareem, "A free-form database query language for mobile phones," in Proc. International Conference on Communications and Mobile Computing, Kunming, 2009, pp. 279-284.

[15] Q. Mahmoud, *Wireless Java*, O'Reilly, 2002. [E -book]. Available: O'Reilly e-book

First Author *Shiramshetty Gouthami* is presently working as Asst.Prof in Computer Science Engineering Department at Jayamukhi Institute of Technological Sciences,Warangal for the past 5 years.She received her M.Tech Degree from JNTU Hyderabad,(A.P, India) in 2011.

Second Author *Pulluri Srinivas Rao* is presently working as Assoc. Professor and Head for the Department of Information Technology at Jayamukhi Institute of Technological Sciences, Warangal.A.P (INDIA).He received his M.Tech Degree in 2006 from University of Allahabad (U.P, India).Presently he is a Ph.D scholar at R.U (A.P, India).

Third Author *Dr.Jayadev Gyani* is presently working as Head and Professor of Computer Science Engineering Department at Jayamukhi Institute of Technological Sciences, Warangal.A.P (INDIA).He received Ph.D. in Computer Science, from Hyderabad Central University, Hyderabad (A.P, India) in 2009.He had 18 years of Teaching Experience in the field of computer science.

FinFET Architecture Analysis and Fabrication Mechanism

Sarman K Hadia [1]. , Rohit R. Patel[2] Dr. Yogesh P. Kosta[3]

[1] Associate Professor, C.S. Patel Institute Of Technology, Charotar University of Science And Technology, Changa, Gujarat, India

[2] PG Student, C.S. Patel Institute Of Technology, Charotar University of Science And Technology, Changa, Gujarat, India

[3] Director, Marwadi Group of Institutions, Rajkot, Gujarat, India

Abstract

In view of difficulties of the planar MOSFET technology to get the acceptable gate control over the channel FinFET technology based on multiple gate devices is better technology option for further shrinking the size of the planar MOSFET [1]. For double gate SOI- MOSFET the gates control the channel created between source and drain terminal effectively. So the several short channel effects like DIBL, subthreshold swing, gate leakage current etc. without increasing the carrier concentration into the channel.

This paper mainly deals with detail description about the DG MOSFET structure and its particular type named as FinFET technology and its fabrication mechanism is also described. Below the 50nm technology FinFET has better controlling over the several short channel effects.

In section one the introduction is given, section two describe the Evaluation from previous technology, section three describe the DG MOSFET structure and its type, section four describe the FinFET technology, section five describe the fabrication mechanism of the FinFET technology and finally conclusions given in section six.

Keywords: *CMOS scaling, DG MOSFET, FinFET, Short Channel Effect, SOI Technology*

1. Introduction

When we shrinking further the size of the planar MOSFET technology several short channel effects are produced. So instead of planar MOSFET technology DG-MOSFETs technology based on multiple gates device have better controlling over the SCEs. Particularly the FinFET technology provides superior scalability of the DG-MOSFETs compare to the planar MOSFET. It provides better performance compare to the bulk Si-CMOS technology. Because of its compatibility with the recent CMOS technology FinFETs are seen to be strong candidate for replacing the bulk or planar Si-CMOS technology from 22nm node onwards. Many different ICs like digital logic, SRAM, DRAM, flash memory etc. have already been demonstrated. Due to their better controlling over subthreshold leakage current and current saturation FinFETs are advantages for the high gain analog applications and get better result in the RF applications [2].

Scaling planar CMOS to 10nm and below would be exceptionally difficult but not completely impossible, due to electrostatics, excessive leakages, mobility degradation, and many realistic fabrication issues. Particularly, control of leakage in a nano scale transistor would be critical to high performance chips such as microprocessors. Non-planar MOSFETs provide potential advantages in packing density, carrier transport, and device scalability [3].

2. Evaluation & Comparison of FinFET Technology

As devices shrinking further, the problems with the planar or bulk Si-CMOS technology are increasing. Several short channel effects like V_T rolloff, drain induced barrier lowering (DIBL), increasing leakage currents such as subthreshold S/D leakage, gate induced drain leakage (GIDL), gate direct tunneling leakage, and hot carrier effects produced in the devices which degrading the use in industry. When power supply voltage V_{dd} is reduced which helps to reduced power and hot carrier effects but the performance improvement is not good. Performance can be improved by lowering the V_T. Researcher are on search to find high-k gate dielectric so that a thicker physical oxide can be used help to reduce gate leakage and yet have adequate channel control, but this is not successful at the point of being usable. There are other problems with Si are band alignment, thermal instability problem etc. The thermal instability problem has led researchers to search for metal gate electrodes instead of polysilicon. But metal gates with suitable work functions have not been found to the point of being usable. In the absence of this, polysilicon continues to be used, whose work function

require that V_T be set by high channel doping concentration which in turn leads to random dopant fluctuations (at small gate lengths) as well as increased impurity scattering and therefore reduced mobility [4]. The off-state leakage current and standby power are increasing with shorter channel-lengths since it is becoming more difficult to keep the electrostatic integrity of devices – doping concentration into channel needs to be increased and the source and drain junctions required to become more shallower, but these trends are offset by the increased junction leakage and higher series resistances. Fully-depleted devices, double-gate devices in particular, offer significantly better electrostatic integrity and hence, better short-channel immunity [2]. In addition to excellent channel control, the FinFET transistors also offer approximately twice the on-current compare to the planar MOSFETs because of the dual gates, even without increasing channel doping. This is beneficial for the carrier mobility and results in a low gate leakage at the same time [3]. Based on discussion planar MOSFETs can be replaced by double gate MOSFETs devices at gate lengths below 50nm in order to be able to continue forth on the shrinking path [4].

3. DG – MOSFET Structure

Currently standard CMOS technology can be replaced by DG MOSFETs technology to increase the integration capacity of silicon technology in the near future. A DGSOI Structure consists, basically, of a silicon slab sandwiched between two oxide layers as illustrated in Fig.1.

The salient features of the DG MOSFETs are control of short-channel effects by device geometry, as compared to bulk FETs, where the short-channel effects are controlled by doping concentration; and a thin silicon channel leading to tight coupling of the gate potential with the channel potential. These features provide potential DG MOSFET advantages are reduced 2D short channel effects leading to a shorter allowable channel length compared to bulk FET, a sharper subthreshold slope is 60 mV/dec for FinFET as compared to 80 mV/dec for bulk FET as shown in Fig.3 which allows for a larger gate overdrive for the same power supply and the same off-current and better carrier transport as the channel doping is reduced [6].

Basically there are 2 kinds of DG-FETs: (1) Symmetric: - In Symmetric DG-FETs have identical gate electrode materials for the front and back gates means gate electrode material is same for both gate. When symmetrically driven, the channel is formed at both the surfaces. (2) Asymmetric: - In an asymmetric DG-FET, the top and bottom gate electrode materials can different. Channel is formed only in one surface [4].

In the fig.2 it is shown that there are three ways to fabricate the DG-FET. Types 1 and 2 suffer most from fabrication problems, viz. it is hard to fabricate both gates of the same size and that too exactly aligned to each other. Also, it is hard to align the source/drain regions exactly to the gate edges. Further, in Type 1 DG-FETs, it is hard to provide a low-resistance, area-efficient contact the bottom gate, since it is buried. The FinFET is the easiest one to fabricate as shown in fig. 4.

4. FinFET Structure Analysis

In Fig.2 it is shown that type 3 is called as a FinFET. This is called as FinFET because the silicon resembles the dorsal fin of a fish. It is referred to as a quasi-planar device. In the FinFET the silicon body has been rotated on its edge into a vertical orientation so only source and drain regions are placed horizontally about the body, as in a conventional planar FET. The separate biasing in DG device easily provides multiple threshold voltages [8].

A gate can also be fabricated at the top of the fin, in which case it is a triple gate FET. The width of a FinFET is quantized due to the vertical gate structure. The fin height determines the minimum transistor width (*Wmin*). With the two gates of a single-fin FET tied together, W_{min} is

$$W_{min} = 2 \times H_{fin} + T_{fin} \quad (1)$$

Where H_{fin} is the height of the fin and T_{fin} is the thickness of the silicon body as shown in Fig. 1. H_{fin} is the dominant component of the transistor width since T_{fin} is typically much smaller than H_{fin} Since H_{fin} is fixed in a FinFET technology, multiple parallel fins are utilized to increase the width of a FinFET as shown in fig.5. The total physical transistor width (W_{total}) of a tied-gate FinFET with *n* parallel fins is:

$$W_{total} = n \times W_{min} = n \times (2 \times H_{fin} + T_{fin}). \quad (2)$$

FinFETs are designed to use multiple fins to achieve larger channel widths. Source/Drain pads connect the fins in parallel. As the number of fins is increased, the current through the device increases [9].
Main features of FinFET are (1) Ultra thin Si fin for suppression of short channel effects (2) Raised source/drain to reduce parasitic resistance and improve current drive (3) Gate last process with low V_T, high k gate

dielectrics (4) Symmetric gates yield great performance, but can built asymmetric gates that target V_T [7].

The two vertical gates of a FinFET can be separated by depositing oxide on top of the silicon fin, thereby forming an independent-gate FinFET as shown in Fig. 5(b).

An independent-gate FinFET (IG-FinFET) provides two different active modes of operation with significantly different current characteristics determined by the bias conditions. Alternatively, in the Single-Gate-Mode, one gate is biased with the input signal while the other gate is disabled (disabled gate: biased with V_{GND} in an N-type FinFET and with V_{DD} in a P-type FinFET). The two gates are strongly coupled in the Dual of the two independent gates as shown in Fig. 7. In the Dual-Gate-Mode, the two gates are biased with the same signal -Gate-Mode, thereby lowering the threshold voltage V_{th} as compared to the Single-Gate-Mode. The maximum drain current produced in the Dual-Gate-Mode is therefore 2.6 times higher as compared to the Single-Gate-Mode as shown in Fig. 7. The switched gate capacitance of the FinFET is also halved in the Single-Gate-Mode due to the disabled back gate [9].
The drain current normalized by the channel width W at the same V_{gs} is almost independent of H_{fin} while fixing T_{fin}. The small differences in the normalized drain current for devices with the same H_{fin} and different T_{fin} come from the threshold voltage roll-off due to the increase in T_{fin} [10]. The dependences of V_{th} roll-off and subthreshold swing S on H_{fin} and T_{fin} are shown in Figs. 8 and 9.

5. Fabrication Mechanism of FinFET Technology

Fig. 10 shows the FinFET fabrication process flow. As the starting material SOI wafer is used with a 400-nm thick buried oxide layer and 50-nm thick silicon film. The measured standard deviation of the silicon film thickness is around 20 A. Although the silicon film thickness determines the channel width, the variation is acceptable for the device uniformity. The larger source of process variation is the variation in gate length. As the gate length will vary process variation also vary [11].

The CVD SI_3N_4 and SiO_2 stack layer is deposited on top of the silicon film to make a hard mask or cover layer. The fine Si-fin is patterned by electron beam (EB) lithography with 100 keV acceleration energy. The resist pattern is slightly ashed at 5 W and 30 sec for the reduction of the Si-fin width. Then, using top SiO_2 layer as a hard etching mask, the SOI layer is etched. The Si is exposed only at the sides of the Si-fin as shown in Fig. 10(1). Fig. 11 shows the fabricated Si-fin width versus the design size with the EB dose as a parameter. Fine Si-fins down to 20 nm are obtained. Using EB lithography, the S/D pads with a narrow gap in between them are delineated. The SiO_2 and amorphous Si layers are etched and the gap between the S/D pads is formed as shown in Fig. 10(3). While the cover layer protects the Si-fin, the amorphous Si is completely removed from the side of the Si-fin. Fig. 7 shows the simulated current density distribution in the Si-fin and pad region of FinFET. The current density contour shows that the current quickly spreads into the pads. This suggests that the parasitic resistance is reduced as shown in Fig.12.
CVD SiO_2 is deposited to make spacers around the S/D pads. The height of the Si fin is 50 nm, and the total S/D pads thickness is 400 nm. Making use of the difference in the heights, the SiO_2 spacer on the sides of the Si-fin is completely removed by sufficient over etching of SiO_2 while the cover layer protects the Si-fin. The Si surface is exposed on the sides of the Si-fin again as shown in Fig. 10(4). During this over etching, SiO_2 on the S/D pads and the buried oxide are etched.

Notice that the channel width of the devices is twice the height of the Si-fins or approximately 100 nm. By oxidizing the Si surface, gate oxide as thin as 2.5 nm is grown. Because the area of Si-fin side surface is too small, we use dummy wafers to measure the oxide thickness. During gate oxidation, the amorphous Si of the S/D pads is crystallized. Also, phosphorus diffuses from the S/D pads into the Si-fin and forms the S/D extensions under the oxide spacers. Then, boron-doped $SI_{0.4}Ge_{0.6}$ is deposited as the gate material. Because the source and drain extension is already formed and covered by thick SiO_2 layer, no high temperature steps are required after gate deposition. Therefore, the structure is suitable to use with new high gate dielectric and metal gates that are not compatible with each other under high temperature. After delineating the gate electrode as shown in Fig. 10(5), the probing windows are etched through the oxide. We directly probe on the poly-Si and poly-SiGe pads, with no metallization used in this experiment. The total parasitic resistance due to probing is about 3000 ohms [11].

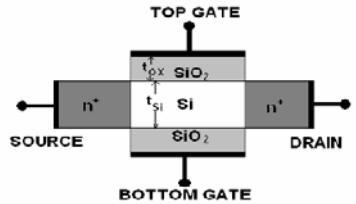
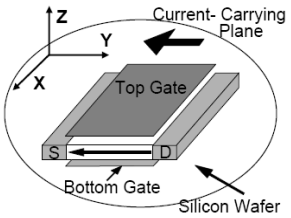
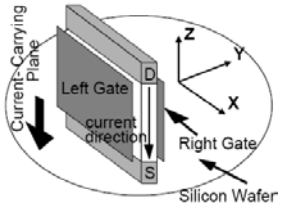


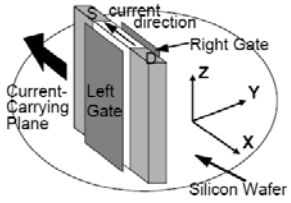
Fig. 1: Cross section of a generic planar DGFET [5]



(a)Type -1 Planar DG-FET



(b) Type-2 Vertical DG-FET



(c) Type-3 FinFET

Fig-2: Types of DG-FET (a) Planar DG-FET (b) Vertical DG-FET (c) FinFET [10]

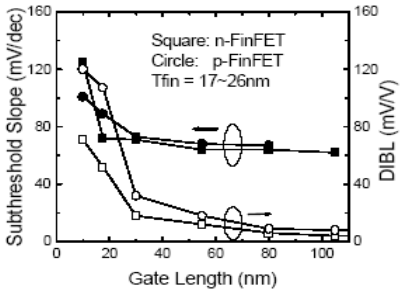


Fig.3 Short-channel effects of CMOS FinFET [3]

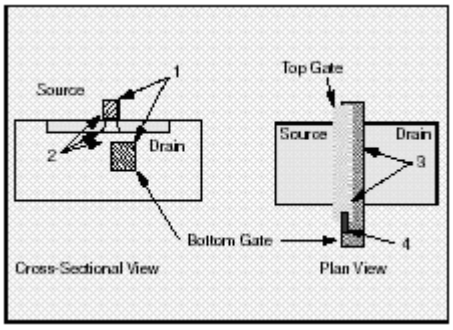
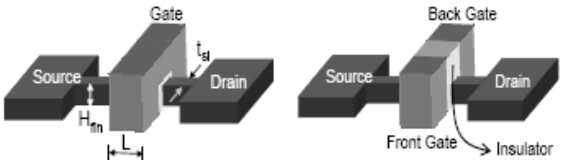
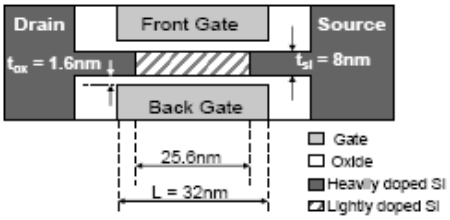


Fig.4 Schematic of the four major obstacles to DGCMOS [4]



(a) SDDG-FinFET (b) IDDG-FinFET



(c) Cross sectional top view of an independent-gate FinFET

Fig. 5 FinFET structure. (a) 3D structure of a one-fin tied-gate FinFET. (b) 3D structure of a one-fin independent-gate FinFET. (c) Cross sectional top view of an independent-gate FinFET with a drawn channel length of 32nm [9].

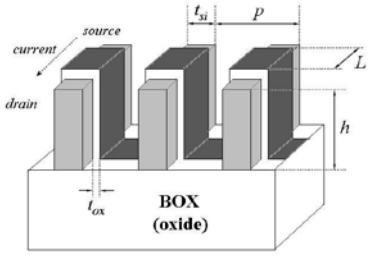


Fig.6. Multi-fin FinFET structure [12]

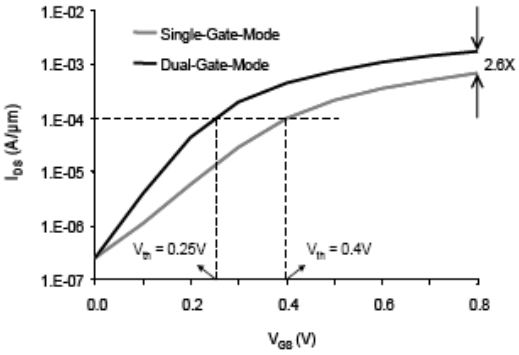


Fig. 7 Drain current characteristics of an N-type IG-FinFET. The drain-to source voltage is 0.8V. T = 70'C [9]

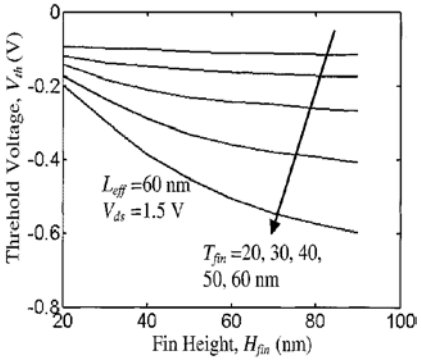


Fig. 8 Dependence of threshold voltage roll-off on H_{fin} and T_{fin} [10]

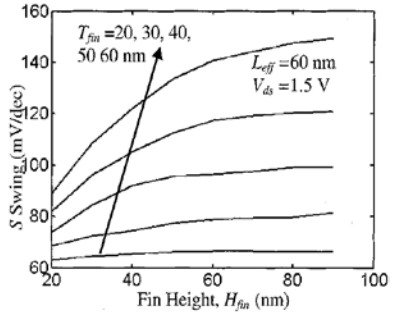


Fig. 9 Dependence of subthreshold swing on H_{fin} and T_{fin} [10]

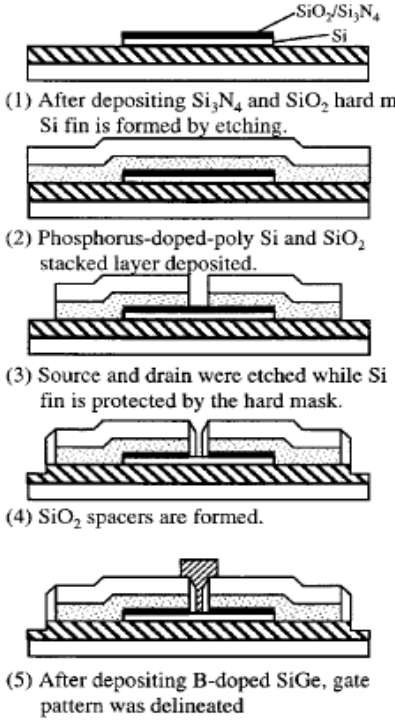


Fig.10 FinFET fabrication process flow [11].

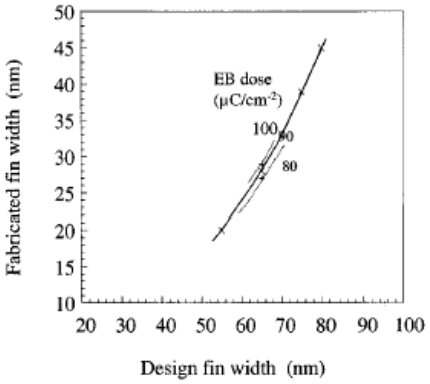


Fig. 11 Relationship between designs Si-fin width practical size with EB dose as a parameter [11].

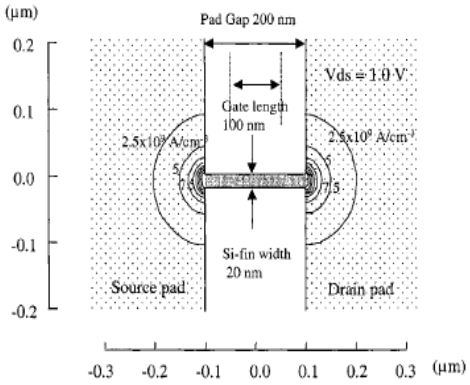


Fig. 12 Simulated current density contours in the poly-Si S/D pads. The fast spread of the current indicates effective reduction of the series resistance [11]

6. Conclusions

To summarize, in FinFET due to dual gate structure it has better controlling over several short channel effect such as V_T rolloff, DIBL, subthreshold swing, gate direct tunneling leakage and hot carrier effects compare to the planner MOSFET FinFET has higher integration density compare to the planner MOSFET. Also fabrication of the FinFET is easiest compare to the other two type of DG MOSFET. So particularly in nanometer regime the FinFET gives better performance compare to the planner MOSFET.

References

[1] Yang kyu-Choi, Leland Chang, Pushkar Ranade, Jeong-Soo Lee, Daewon Ha, Sriram Balasubramanian,

Aditya Agarwal, Mike Ameen, Tsu-Jae King and Jeffrey Bokor. "FinFET Process Refinements for Improved for Mobility and Gate Work Function Engineering," pp. 259-262, in IEDM Tech., 2002.

[2] Jovanović, T. Suligoj, P. Biljanović, L.K. Nanver, **"**FinFET technology for wide-channel devices with ultra-thin silicon body".

[3] Bin Yu, Leland Chang*, Shibly Ahmed, Haihong Wang, Scott Bell, Chih-Yuh Yang, Cyrus Tabery,Chau Ho, Qi Xiang, Tsu-Jae King*, Jeffrey Bokor*, Chenming Hu*, Ming-Ren Lin, and David Kyser, "FinFET Scaling to 10nm Gate Length," IEEE-2002.

[4] Venkatnarayan Hariharan, 2005, "EEs801 Seminar report FinFETs," http://www.ee.iitb.ac.in

[5] Asif I. Khan and Muhammad K. Ashraf, "Study of Electron Distribution of Undoped Ultra Thin Body Symmetric Double Gate SOI MOSFET in Gate Confinement Direction," pp. 1-6.

[6] Vishwas Jaju, "Silicon-on-Insulator Technology," EE 530, Advances in MOSFETs, spring 2004 pp. 1-12.

[7] http://www. techalone.com, Electronic seminar topic

[8] Nirmal, Vijaya Kumar and Sam Jabaraj, "Nand Gate Using FinFET for Nanoscale Technology," International Journal of Engineering Science and Technology, Vol. 2(5), 2010, pp. 1351-1358.

[9] Sherif A. Tawfik, Zhiyu Liu, and Volkan Kursun, "Independent-Gate and Tied-Gate FinFET SRAM Circuits: Design Guidelines for Reduced Area and Enhanced Stability," IEEE ICM, 2007.

[10] Gen Pei, Jakub Kedzierski, Phil Oldiges, Meikei Ieong, and Edwin Chih-Chuan Kan, "FinFET Design Considerations Based on 3-D Simulation and Analytical Modeling," IEEE TRANSACTIONS ON ELECTRON DEVICES, VOL. 49, NO. 8, AUGUST 2002, pp. 1411-1419.

[11] Digh Hisamoto, Wen-Chin Lee, Jakub Kedzierski, Hideki Takeuchi, Kazuya Asano, Charles Kuo, Erik Anderson, Tsu-Jae King, Jeffrey Bokor and Chenming Hu, "FinFET—A Self-Aligned Double-Gate MOSFET Scalable to 20 nm, " IEEE TRANSACTIONS ON ELECTRON DEVICES, VOL. 47, NO. 12, DECEMBER 2000, pp.2320-2325.

[12] Hari Ananthan, Aditya Bansal and Kaushik Roy, "FinFET SRAM – Device and Circuit Design Considerations," in IEEE computer society, 2004.

[13] M.Jurczak, N.Collaert, A.Veloso, T.Hoffmann, S.Biesemans, "Review of FinFET Technology," IEEE 2009

Sarman K. Hadia is Associate Professor in Electronics and Communication Department at Charotar University of Science and Technology, Changa He is presently Ph.D. pursuing in VLSI Technology. His research interests are Microelectronics and optimization techniques.

Rohit R. Patel is M.Tech. Student of C.S. Patel Institute of Technology, CHARUSAT, Changa in Communication system Engineering. He had completed his Bachelor of Engineering Degree from Vishwakarma Government Engineering Collage, Chandkheda, Gandhinagar, India.

Yogesh P. Kosta is an SCPM from Stanford University, California, USA. He did his M. Tech. in Microwave Electronics from Delhi University Delhi, and his Ph.D. in Electronics and Telecommunication. He is a member of IETE and IEEE. He worked as a scientist and designer at the Space Application Canter – ISRO Ahmedabad, and as a Sr. Designer at Teledyne USA. Presently he is Director of Marwadi Group of Institutions. His research areas are RF, Wireless Satellite Systems and Information Communications. He has guided several M. Tech students. At present, six research scholars are currently pursuing their Ph.D under his guidance. He has published many research papers and articles in referred journals and international conference proceedings.

IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011
ISSN (Online): 1694-0814
www.IJCSI.org

Indication of Efficient Technique for Detection of Check Bits in Hamming Code

Rahat Ullah, Jahangir Khan, Shahid Latif, Inayat Ullah

Department of Computer Science and IT

Sarhad University of Science and Information Technology (SUIT), Peshawar 25000 Pakistan

Abstract

When the information is transmitted through wireless or through wired channel, error may occur in data. In some techniques error may be only detected while there are such techniques which can detect as well as correct that error in the data. Hamming code is such a code which can detect a single bit error position and also correct that, this coding scheme is used now days in Wireless communication, in deep space communication etc, e.g. Fixed wireless broadband with high error rate where there is need of correction not only detection. In this coding scheme the detection of data depend on the check/Parity bits, these bits play a vital role in hamming code because they locate the position where the error has been occurred. These check bits are inserted into the data at specified position at transmitter and also at receiver side to sort out error. And to find these bits we need to design a matrix and by putting bit's information and than the value of parity bit can be determine. Normally these bits can be found by using Ex-OR logic as well as using another-logic that is AND-OR logic. In this Paper I have sketched a general Idea about these two logics that which logics performance is good and efficient. In future I will prove that by performing simulations through PSpice and will find the delays of these mentioned logics.

Keywords: *AND-OR Logic, Coding, Error Correction/Detection, Check Bits, Efficient Logic.*

1. Introduction

Error correcting codes are used in a wide range of communication systems from deep space communication, to quality of sound in compact disks and wireless phones. If we generally discuss the main method used to recover messages that might be distorted during transmission over a noisy channel is to inhabit redundancy. We make use of redundancy present in human languages to detect and than correct errors. This happens in both oral and written communication. For example, if you read the sentence "union is strenght " you can tell that something is wrong. So we can detect an error. Moreover we can even correct it. We are doing two things here: error detection and error correction. What are the principles that we are using to achieve these goals? First, because the string "strangth" is not a valid word in English, we know that there is an error. Here, the redundancy manifests itself in the form of the fact that not every possible string is a valid word in the language. Secondly, the word "strangth" is closest to the valid word "mistake" in the language, so we conclude that it is the most likely word intended. Of course we can also use the context and meaning to detect and/or correct errors but that is an additional feature, not available to computers. When I enter the string "strangthy" to Merriam-Webster online dictionary (http://www.m-w.com), no entry can be found for it, however, the computer comes up with a list of suggested words, the first of which is "strength". So, the computer is telling me that "strength" is the most likely word to be intended, because it is closest to the given string. This is called the maximum likelihood principle. As I typed this article on my computer I witness many instances of this principle used by my word processor. For instance, when I mistakenly typed "thre" it automatically corrected it to "three". There are also other ways redundancy is used in natural languages. As by now pointed out, Redundancy in context often enables us to detect and correct errors.. When humans communicate, redundancy, either explicitly introduced by the speaker/author or built into the language, comes into play to help the audience understand the message better by overcoming such obstacles as hearing, difficulties, noise, accent, etc.

The basics of wireless communication were properly recognized by Claude E. Shannon in 1948 [1]. In his attractive paper, Shannon assert that by proper encoding of information, error induced by a noisy channel or storage medium can be reduced to any desired level when the data rate is bounded by channel capacity. Since then, many coding techniques have been developed for wireless communication systems to achieve channel capacity. In 1950, Richard W. Hamming discovered the first class of linear block codes for error correction, only two years after Shannon's theory. Hamming codes are 1-error-correcting codes and are capable of correcting any single error over the span of the code block length [2]. Hamming codes are perfect codes and can be decoded easily with a look-up-table [3].

Actually Coding theory is the branch of mathematics concerned with accurate and efficient transfer of data across noisy channels as the theory of message sent. A transmission channel is the physical medium through which the information is transmitted, such as telephone lines, or atmosphere in the case of wireless communication. Undesirable disturbance (noise) can occur across the communication channel, causing the received information to be different from the original information sent [4]. Coding theory deals with detection and correction of transmission errors caused by noise in the channel. The primary goal of coding theory is competent encoding of information, easy transmission of encoded messages, fast decoding of received information and

correction of errors introduced in the channel [5]. Coding theory is used all the time: in reading CDs, receiving transmissions from satellites, or in cell phones[6]. So as my paper is about hamming code, so will discuss that this technique is used to detect single bit error and will also indicate the efficient logic. And will use check bits for the detection of error, on receiver as well as on transmitter side. I have found these check bits by using Ex-OR logic as well as AND-OR logic.

This paper is organized in such away that the first portion explains the brief introduction about occurring of errors in the information and that how the concept of information theory and than coding techniques came into being, the second portion contains information about types of errors, third portion is about error exposure, fourth portion is about Hamming Code that how the error's position will be detected and corrected using the two logics, and the last section I have discuss the propagation delay of logic gates and explained that which logic should be used to find the parity/check bits and that my future goal is to do these simulations in PSpice.

2. Types of Errors

When binary data is transmitted and processed from source to destination, which can be altered due to channel's noise. Two types of errors could be possible. Single bit error and a burst error [7]. Single bit will be inverted in a single bit error and more than one bit inversion indicated burst errors[7].

3. Error Exposure

When we be on familiar terms with what type of error can situate in the data, will we be able to recognize one when we see it? If we have a copy of the projected message for transmission, than of course we will be able to identify. But if we have no copy than we will be incapable to find the error, until we decipher the data and find that it is of no significance, by making no sagacity. If we use this concept in the engine than it will be so costly, slow of dubious value. We don't need a system than decode whatever comes in, then sit around trying to make a decision if the engine really intended to use the word chshr in the middle of an array of whether information. What we need is a means that is simple and utterly objective.

2.1 Redundancy

In this technique extra bit is added with the Information.. Figure.1 shows that how the extra bits are added to the information to check the accuracy in the information at the receiving side. Once the data stream has been generated, it passes through a device that analyses it and the redundant bit are appropriately generates. The data is know enlarged by adding redundant check, and travels to the receiver. The receiver puts the entire stream to the checking function. If the received bit stream passes the checking criteria, the data portion of the data unit is accepted and the redundant bits are unnecessary and than discarded [7].

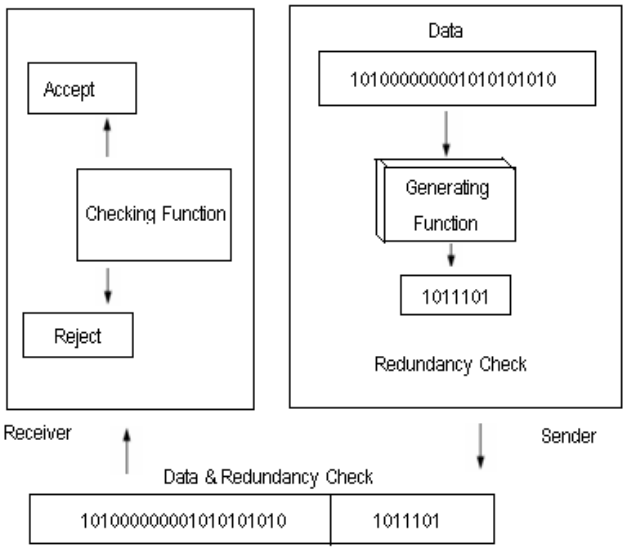


Fig. 1 Process of Using Redundant Bits

Four Types of Redundancy checks are used in data communication; Vertical Redundancy Check(VRC), Longitudinal Redundancy Check(LRC), Cyclic Redundancy Check(CRC), and Checksum.

4. Hamming Code

Data that is transmitted over a communication channel can be scratched; their bits can be masked or inverted by noise. To get rid of retransmitting the data we need to detect and correction of error in the data. Some simple codes can detect but not correct errors [7,8]; Others can detect and correct one or more errors[2]. This paper addresses one Hamming code that can correct a single-bit error which is detected with the help of parity/check bits using two logics and than indicated the efficient one. The Hamming single-bit correction code is implemented by adding check bits also called parity bits to the output message according to the following pattern:

i) The message bits are numbered from left to right, starting at 1.
ii) Every bit whose number is a power of 2 (bits 1, 2, 4, 8,…) is a check bit[8].
iii) The other are the output message bits (bits 3, 5, 6, 7, 9,…) contain the data bits, in order.
Each check bit establishes even parity over itself and a group of data bits and can be fine with the help of Ex-OR gate[9] or also by using AND-OR logic because both the circuits have same performance[10], figure.4. A data bit is in a check bit's group if the binary representation of the data bit's number contains a 1 in the position of the check bit's weight. For instance, the data bits associated with check bit 2 are all those with a 1 in the 2's position of their binary bit number—bits 2, 3, 6, 7, and so forth. Fig.2 shows an 8-bit data message. It shows the assignment of data bits and check bits to the bits of a12-bit output message. The horizontal lines below the output message box have dots to indicate the assignment of output

bits to check groups. Each check group has a ones count box, with a parity bit that is inserted into the output message..

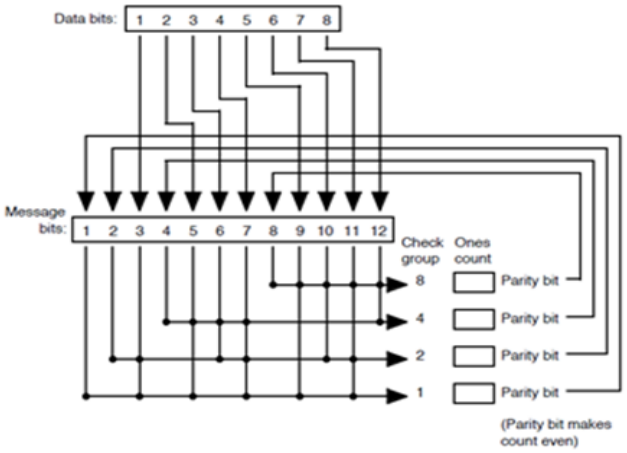


Fig.2 Indicating Positions for Parity bits

To find the value of check/parity bit, that which parity should be zero and which should be one, we will design a matrix having columns equal to the bits contained in the whole stream of data including parity bits according to the order given in the given figure. And will have rows equal to the number of parity bits used in the data, for example if we have used four parity bits, than it will have four rows, for five parity bits it will have five rows and so on.Fig.3 shows.

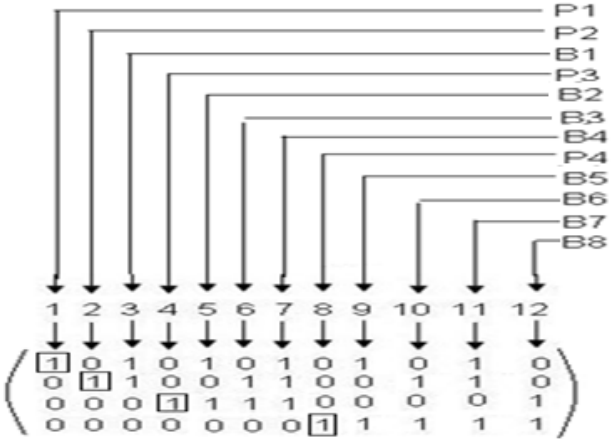


Fig.3 Designing Matrix for Parity bits

Inbox bits in Fig.5 are the starting bits for P1, P2, P3, P4. For P1 we have used 1010101….., for P2 , 1100110011….., for P3, 11110000111100., for P4, 11111111000000001111111100, and so on[16].

Now to calculate the values of P1, P2, P3, P4 from the matrix by Ex-OR or by using another logic AND-OR. Because as the figure.4 shows the outputs of both the circuits are same.

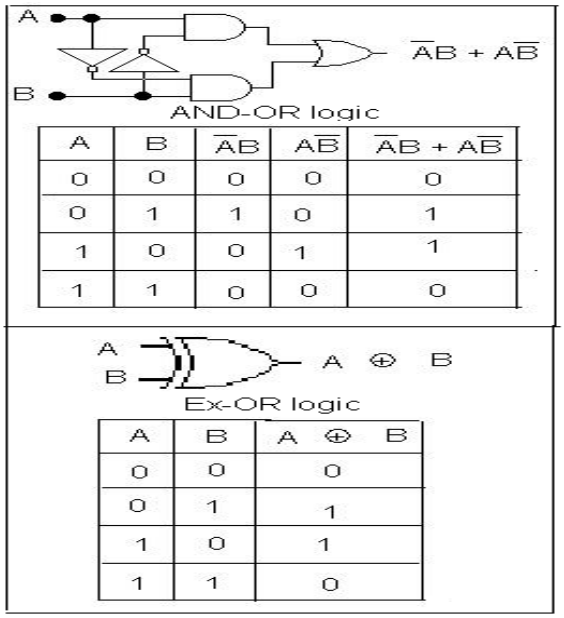
For first parity bit, from the matrix, we get the following information about the bits

$P1=3 \oplus 5 \oplus 7 \oplus 9 \oplus 11$

$=B1 \oplus B2 \oplus B4 \oplus B5 \oplus B7$

$=1 \oplus 0 \oplus 0 \oplus 1 \oplus 0$

After performing the two logics as fig.4 shows we got the parity bit “0”, as figure. 5 shows.



A	B	$\overline{A}B$	$A\overline{B}$	$\overline{A}B + A\overline{B}$
0	0	0	0	0
0	1	1	0	1
1	0	0	1	1
1	1	0	0	0

A	B	A ⊕ B
0	0	0
0	1	1
1	0	1
1	1	0

Fig.4 Showing the Two equivalent logics

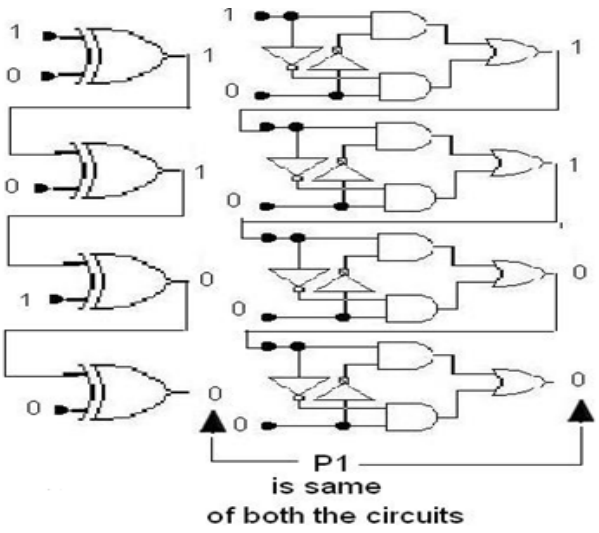


Fig. 5 Detection of First Parity bit

For second parity, from the matrix, we get the following information about the bits.

$P2=3 \oplus 6 \oplus 7 \oplus 10 \oplus 11$

$=B1 \oplus B3 \oplus B4 \oplus B6 \oplus B7$

$=1 \oplus 1 \oplus 0 \oplus 1 \oplus 0$

After performing the same logics we will get the parity bit "1".

For third parity bit, from the matrix, we get the following information.

$P3=5 \oplus 6 \oplus 7 \oplus 12$

$=B2 \oplus B3 \oplus B4 \oplus B8$

$=0 \oplus 1 \oplus 0 \oplus 1$

Here the resultant parity bit will be "0".

For the 4th parity/check bit we get the following information.

$P4=5 \oplus 6 \oplus 7 \oplus 8$

$=B5 \oplus B6 \oplus B7 \oplus B8$

$=1 \oplus 1 \oplus 0 \oplus 1$

It will give the parity bit "1", by using the two logics.

Now we are ready to send our data, but first we need to insert the parity bits in the data at specified positions, as (bits 1, 2, 4, 8), fig.6. The data becomes

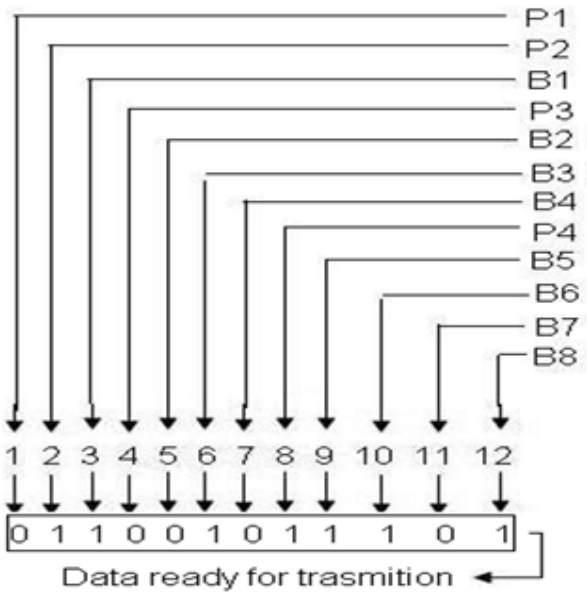


Fig. 6 Parity bits are inserted in the Data

When the message reaches the destination, we know the parity bits positions, so the data is collected from positions 3, 5, 6 ,7 ,9 ,10 ,11 ,12. Suppose the received message is 0 1 1 0 0 0 1 1 1 0 1, as the parity bits are at positions 1,2,4, and at 8 from left to right, so collect the data bits for decoding process, which is 10001101. The same process will be repeated, the matrix will be design, and than will find a parity bits at receiving side. As we have already did these things at transmitter side, we will find the parity bits by putting the received data bits. Which given below, in fig.7.

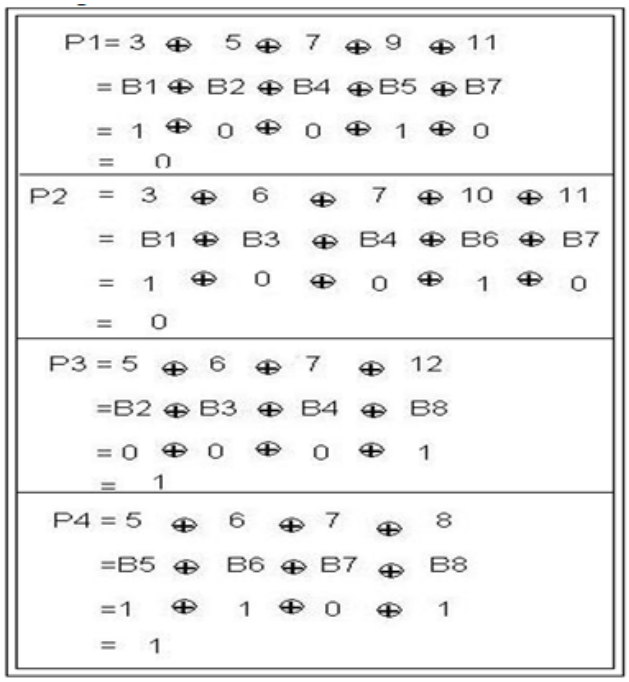


Fig. 7 Detection of Parity bits at receiving side

To check error in the data through the channel, we will compare the parity bits at receiver and transmitter ends that we have at the source and destination, if it gives the answer zero, that received data is error free, otherwise it will locate the poison.

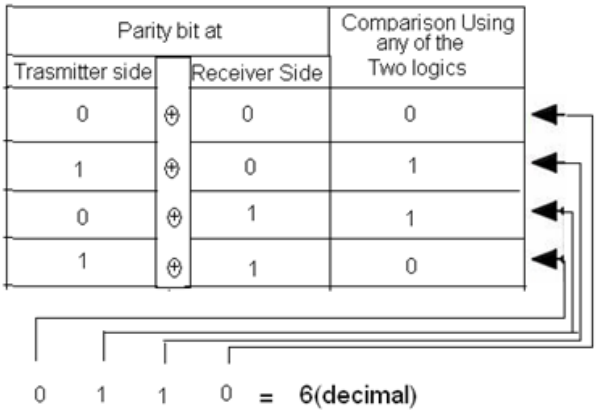


Fig. 8 Detecting the location of error

Fig.8 shows that the error was occurred at position 6, and was corrected. As shown in figure. 9.

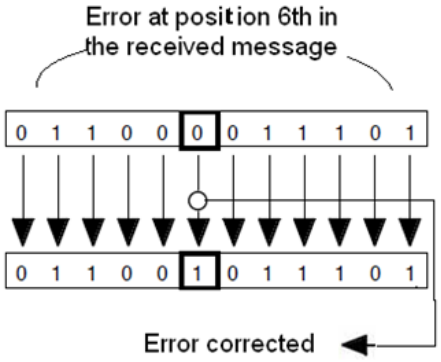


Fig. 9 Received message is corrected

5. Indication of Efficient Logic

Propagation Delay is the length of time taken for the quantity of interest to reach its destination. In computer networks, and Electronics, propagation delay is the amount of time it takes for the head of the signal to travel from the sender to the receiver over a medium. It can be computed as the ratio between the link length and the propagation speed over the specific medium [11]. Propagation delay = d / s where d is the distance and s is the wave propagation speed. In wireless communication, s=c, i.e. the speed of light. In copper wire, the speed s generally ranges from .59c to .77c [12, 13]. This delay is the major obstacle in the development of high-speed computers and is called the interconnect bottleneck in IC systems.

In electronics, digital circuits and digital electronics, the propagation delay, or gate delay, is the length of time starting from when the input to a logic gate becomes stable and valid, to the time that the output of that logic gate is stable and valid. Often this refers to the time required for the output to reach from 10% to 90% of its final output level when the input changes. Reducing gate delays in digital circuits allows them to process data at a faster rate and improve overall performance.

The difference in propagation delays of logic elements is the major contributor to glitches in asynchronous circuits as a result of race conditions.

The principle of logical effort utilizes propagation delays to compare designs implementing the same logical statement.

Propagation delay increases with operating temperature, marginal supply voltage as well as an increased output load capacitance. The latter is the largest contributor to the increase of propagation delay. If the output of a logic gate is connected to a long trace or used to drive many other gates (high fanout) the propagation delay increases substantially.

Wires have an approximate propagation delay of 1 ns for every 6 in of length [14]. Logic gates can have propagation delays ranging from more than 10 ns down to the picoseconds range, depending on the technology being used[14].

As we have used two logics for the detection of Parity Bits, & because of the number of increasing gates in any circuit the propagation delay will be increases, so in these two techniques the use of Exclusive Or gate is batter than AND-OR logic. I will prove it through simulations in PSpice[15], which is my future's goal.

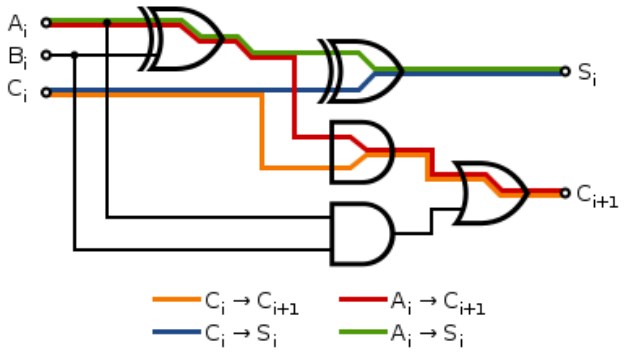


Fig.10A full adder has an overall gate delay of 3 logic gates from the inputs *A* and *B*to the carry output C_{out} shown in red

For a reference consider this full adder circuit[], it is producing delay as the figure shows, when the number of gates increases & the length of circuit increases the propagation delay increases, & slow down the functionality of a circuit and vice versa.

6. Conclusion

The coding concept was introduced for the detection of error in data. At opening level single parity bits were used for the detection of error in the information, than multiple bits were used. But the problem was the retransmission of information in case of any error occurrence in the data. Hamming introduced a coding technique for the error detection as well as correction of that error at the receiver side. In this technique hamming added some extra bits called parity/check at some specific position, but for the detection of those parity bits Ex-OR and AND-OR logic was primarily used. I have indicate the efficient logic among the two, & in future will try to prove it through simulation that which logic producing the parity/check bits efficiently.

References

[1] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, pp. 379-423, 623-656, July, August 1948.
[2] R. Hamming, "Error Detecting and Error Correcting Codes," *Bell System Technical Journal*, pp. 147-160, April 1950.
[3] S. Lin and D. J. Costello, *Error Control Coding: Fundamentals and Application.* Prentice Hall, April 2004.
[4] Shannon C. E., 1948, A mathematical theory of communication. *Bell Syst. T.J.* 27, 379-423.
[5] Hill R., 1986, A first course in coding theory, Clarendon Press, Oxford.
[6] Huffman W. C. and Pless V., 2003, Fundamentals of Error Correcting Codes, Cambridge University Press.
[7] Behrouz Forouzan, *Data Communication and Networking*,Second Edition.
[8] Theodore F. Bogart. Jr, *Introduction to Digital Circuits*, International Edition 1992.
[9] Richard B.Wells, *Applied Coding and Information Theory for Engineers*, Published by Dorling Kindersley(India) Pvt. Ltd.
[10] Floyd & Jain, *Digital Fundamentals*, Seventh/Eighth Edition.
[11] http://en.wikipedia.org/wiki/Propagation_delay
[12] "What is propagation delay? (Ethernet Physical Layer)". *Ethernet FAQ*. 2010-10-21. Retrieved 2010-11-09.
[13] "Propagation Delay and Its Relationship to Maximum Cable Length". *Networking Glossary*. Retrieved 2010-11-09.
[14] Balch, Mark (2003). *Mcgraw Hill - Complete Digital Design A Comprehensive Guide To Digital Electronics And Computer System Architecture*. McGraw-Hill Professional. pp. 430. ISBN 9780071409278.
[15] http://edmondsonengineering.com/newbie.aspx
[16] Floyd, *Digital fundamentals, (seventh/eighth edition).*

Corpus Based Context Free Grammar Design for Natural Language Interface to Database

Avinash J. Agrawal and Dr. O. G. Kakde

**Shri Ramdeobaba Kamla Nehru Engineering College,
Katol Road, Nagpur, India**

**Vishvesvarya National Institute of Technology
Bajaj Nagar, Nagpur, India**

Abstract

For practical implementation of Natural Language Interface to Database, deep semantic analysis needs to be used as it increases success rate and portability. Deep analysis uses more language knowledge rather than the domain knowledge. A primary step in deep semantic analysis is formalizing syntax of possible input questions. This paper describes a context free grammar designed for a domain specific natural language interface to database. The grammar is designed for the domain Railway Inquiry for which corpus of question is collected and analyzed.

Keywords: *Natural Language Interface to Database, Corpus of Questions, Context Free Grammar, Deep Semantic Analysis.*

1. Introduction

Natural language Interface to Database (NLIDB) can acts as an alternative interface for finding structured information from database particularly on a small handheld devices because writing questions in natural language is much easier for a casual user than the complicated and time consuming navigation required in the traditional database interfaces. NLIDB shifts a user's burden of learning use of interface to describe his or her need for information to the system. NLIDB thus demands less input-output and processing facilities which make it more useful for mobile devices [1].

Natural language interface to database is not a new area, a lot of research is going on since a long time [4]. Most of the NLIDB systems developed so far used shallow analysis [11],[12,[13]. Shallow analysis uses domain knowledge rather than linguistic knowledge to interpret the meaning of input question, which results in a low success rate. This is a main difficulty in implementing NLIDB for any practical use. To increase the success rate deep analysis of input question can be used. Deep analysis uses linguistic knowledge for detail understanding. Deep analysis has an advantage of portability that is not possible in shallow analysis due to too much dependency on database.

To explore deep analysis for NLIDB a railway inquiry is selected as a domain. For railway inquiry domain a corpus of question is collected by conducting a survey with different group of railway inquiry users. This corpus is then analyzed to find the pattern of questions used in the domain. Based on these patterns a context free grammar is designed to represent syntax of input questions. Representing syntax is an important step as in deep analysis method the subsequent steps are dependent on it.

2. Corpus Design

2.1 Domain Selection

Different restricted domains benefit from different Question Answering techniques. Some domains are particularly appropriate for the development of question answering systems. Not all domains are appropriate for natural language interface to database. For a domain specific question answering restricted domain must be circumscribed, complex and practical [12].

Circumscribed means the domain where user knows what to expect from the system and knows what questions are appropriate to the domain. A more important motivation for a circumscribed domain is the need for clearly defined knowledge sources.

A domain should be complex enough to warrant the use of a QA system. There is no need for a QA system in a

domain where a simple list of facts or a FAQ would be sufficient to satisfy the user's need for information.

Practicality is an important condition to consider when developing a QA system. The domain should be of use to a relatively large group of people. Otherwise one risks wasting effort on a system that nobody would use.

The selected domain railway inquiry system satisfies all the three requirements. It is circumscribed in a sense that user very well know what types of questions are to be asked in railway inquiry system. Authoritative and comprehensive resources containing required information are already maintained by the railway in the form of database. Thus in railway domain knowledge sources is clearly defined. To answer the question of typical railways inquiry use of extensive knowledge from outside the domain is not required.

Although the railways inquiry domain is circumscribed a QA system will prove to be very useful for the users. It is complex enough domain as simply list of facts or frequently asked questions could not satisfy the wide variety questions with different argument for different groups of people. Especially in countries like India where the network of railways is huge, a railways inquiry may include question related to availability of seats, trains, information regarding stoppage, fare and many information related to thousands of routes and trains. This domain is certainly very interesting for researchers and helpful for users.

It is very much practical domain as there is no risk of wasting of efforts. This type of system is very much required and in demand by people due to difficulty in getting information from the conventional ways. Specially in countries like India, China where railways is a part of life and is a very important medium of transportation and millions of people travel daily by rails and needs some or other kind of information related to the this domain.

2.2 Corpus Collection

More than 150 questions of railway inquiry domain are collected. This question set contains queries related to only general inquiry. Query related to reservation is not included in it. Discussing various people of different age, gender, profession and background does this question collection. The questions set contains queries related to different attribute of railways like arrival departure time, availability of trains between stations, fare, status, concessions etc. Same information may be asked in more than one way in any natural language and example of this is available in the question set. Duplicate questions are removed from the corpus.

From the collected corpus a set of questions is separated for testing purpose. The test corpus contains 24 questions. The test corpus is generated on the basis of the structure of parse tree. Each question of the test corpus is having a distinguished parse tree structure. Thus the test corpus represents the different syntactic structure of questions related to the domain.

Sample questions from the question set are listed below:
What is the position of the gitanjali Express.
What is the fare from Nagpur to gondia.
When howrah mail is coming.
How much late mangla express is running.
Whether gitanjali express having stoppage at Akola.
By what time gitanjali express reaches wardha.
In which platform Vidarbha Express will arrive.
How many trains are available from nagpur to raipur.
What is the fair for A/C two tier for the train vidarbha Express.
What is the difference in fair for ac class and sleeper class.
How much amount will be deducted if we cancel the ticket before 24 hours.
How many trains are available from nagpur to delhi.
Is any direct train to goa from nagpur.
What is the route to jaipur from nagpur.
Is any concession in the fare for student's educational tour.
What is the fare for the sport team from akola- nagpur.
What percent of concession is given to the handicappedperson.
How many stoppage does rajdhani express has.
How many trains are available for Mumbai from nagpur on Wednesday.
How many superfast trains are available for Mumbai from nagpur on Wednesday.
Whether charges of Doronto is more than superfast.
List all trains from nagpur to raipur.

The corpus is then analyzed thoroughly on different aspects that will be useful in designing the context free grammar for syntax analysis purpose and the NLIDB as a whole. Some graphs are also plotted based on the observations like length of questions in words, starting word of question and addition in question corpus with respect to number of person surveyed. These graphs are shown in (Fig.1, 2 & 3).

The graph plotted for percentage addition in corpus with respect to number of person surveyed is decreasing about exponentially with increase in number of person surveyed as shown in (Fig.1). This graph shows that for collecting sufficient number of questions in corpus it is not necessary

to survey large number of people. For a well-defined and useful domain a sufficiently large corpus can be obtained by surveying a small number of people. As we increase the number of people for surveying chances of repetitive questions increases which do not contribute in addition in corpus.

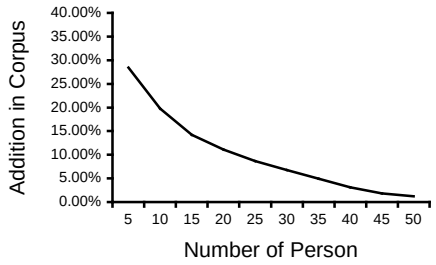


Fig.1 Corpus Collection: Addition in corpus with respect to number of person surveyed

The graph for question length in the collected corpus shows that minimum length is five words and maximum is eighteen as shown in (Fig.2). Maximum number of questions (more than three forth) in the corpus is having length between seven and eleven. It is an important observation that questions in such domains are small in length. This observation affects semantic analysis as many times question may not contains sufficient information for its interpretation. In such a situation discourse and pragmatics plays role to decide meaning of question. This observation also favors deep semantic analysis against shallow one that normally requires complete and semantically tractable questions [9].

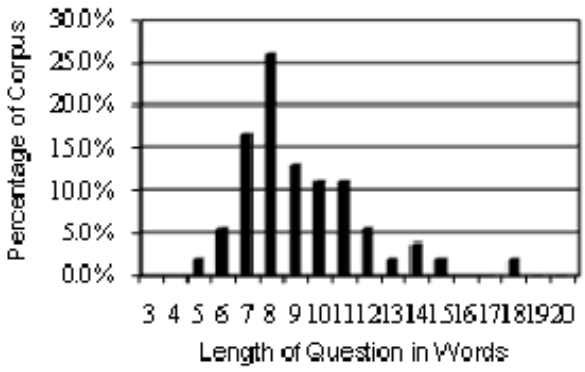


Fig.2 Question Length in Corpus

Graph showing possible starting word of question shown in (Fig.3) is simple one and is used to design context free grammar rules for representing syntax of questions. In our domain maximum questions (more than 50%) are starting from what and how but actually it depends on domain and people surveyed.

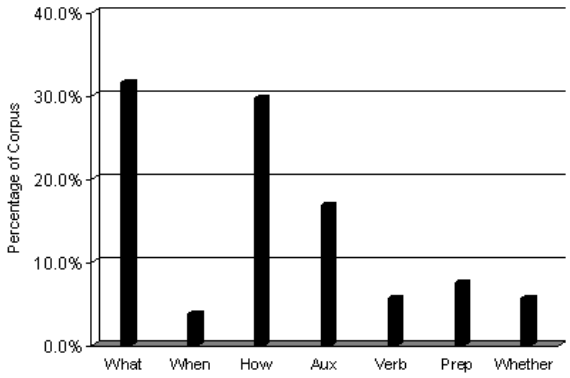


Fig.3 Starting word percentage in corpus

3. DESIGN OF GRAMMAR BASED ON THE CORPUS

The question set is analyzed thoroughly to understand the patterns of question asked for the domain. Each question is a unique in nature but still there are many common things that can be extracted like start of question. Question may start with Wh-word, Auxiliary verb, main verb or prepositions [6]. Depending on type of question information to be extracted found in different constituents of the input question. Based on such observations a grammar for the corpus needs to be designed. While designing a grammar it is very important to note that what kind of information would be required to extract by parsing the given question. After observing the questions it is found that every question contains three elements, which will be very useful in semantic analysis [7]. These elements are:

- The standard question items (“What is the ….?”, “How many …?”, “Is there any …?”)
- The goal of the query (i.e. which values have to be retrieved and reported to the user?)
- The restrictions, enabling the system to extract the relevant items

Based on the presence of these three elements a grammar is to be designed for the natural language questions. For example if question is starting from “how many” then the noun phrase immediately following it will be the goal of query and the phrases following to verb describes the restriction part. However all questions are not that much simple to interpret, it requires a detail analysis to find relationship among constituents of a sentence [10]. So basically in syntax analysis different constituents of input question needs to be identified and later on these can be related to each other to interpret meaning. Here complete grammar of English is not required as in domain specific

questions only limited variations in syntax is used while grammar of complete language is very complicated and large in size. Moreover users of such system may not be caring exact and all rules of language. So in such a case simple and tolerate grammar is required. These simple syntactic rules can be modeled by context free grammar (CFG). CFG are powerful enough to express sophisticated relations among the constituents in a sentence, yet computationally tractable enough that efficient algorithms exist for parsing sentences.

3.1 Context Free Grammar

Context free grammar is defined by four tuples [5],
G = (V, T, P, S) where
V is set of Non terminal Symbols (It can be replaced by any other string of symbol)
T is set of Terminal Symbols (It cannot be replaced by any other string)
P is set of Production and (Defines replacement rules)
S is starting symbol of the grammar. (Every valid string can be generated from it)

In the designed grammar terminal symbols are the part of speech assigned to individual words of the question like noun, preposition, verb, adjectives etc. Morphological analyzer does assignment of this part of speech. Non-terminal symbols are the syntactic categories of English language grammar, which represents a set of strings. Each string of the set is a meaningful constituent of an english sentence. Set of Non-terminal symbol includes noun phrase, verb phrase, adjectival phrase, prepositional phrase etc. The non-terminal S that also is the starting symbol of the grammar represents an input question. Production Set defines the rules for replacement used for replacing a non-terminal by any string of terminal or non-terminal symbols. The production rules with example are given in (Fig.4).

First production rule is for an input question represented by symbol S. A question can start with a verb followed by a noun phrase and then verb phrase or only noun phrase. It includes all imperative and yes/no question structure. Most commonly used questions starts with Wh-words. To represent such questions a non-terminal Wh-NP is included which can be replaced by simply a Wh-word or Wh-word followed by Nominal. In question this Wh-NP can be followed by either verb pharse only or Auxiliry verb followed by NP and then followed by VP. In last production of S question starts with a preposition followed by Wh-NP then NP and VP. Similarly productions of other non-terminals like Noun Phrase, Verb Phrase, Prepositional Phrase, Adjectival Phrase and Nominal defines rules by which that can be generated. Each rule followed by an example for easy understanding.

The grammar described in previous section is successfully tested on test corpus by using YACC [8]. YACC generates code for LALR parser [2] which uses bottom up technique. For using YACC, grammar was provided in a file called .y file [8]. It uses lexical specification defined in .l file. For testing purpose, only simple morphological rules are used in .l file [8] to define valid tokens and also it assigns part of speech to each token. The parser generated through YACC takes a question from test corpus and makes parse tree for it. An example parse tree is given in (Fig. 5).

Terminals:

N – Noun
P – Preposition
Aux – Auxiliary Verb
Verb – General Verb
Adj – Adjective
Wh-Words – Wh-Words (what, when,..)
Det – Determiner (a, an, the)
Card – cardinal number (one, two, three)
Ord – Ordinal Numbers (first, second)
Quant – Quantifiers (many, more, any)
Conj – Conjunctions (and, or)

Non Terminals

S → Query
NP → Noun Phrase
VP → Verb Phrase
NOM → Nominal
V → Verb
PP→ Preposional Phrase
Wh-NP → Wh Noun Phrase
AP → Adjectival Phrase

4. CONCLUSIONS AND FUTURE SCOPE

For the practical use of Natural language interface to database, deep analysis is to be implemented. To implement deep analysis the detail analysis of the input question is required. This method makes use of more linguistic knowledge than the domain knowledge. The first step for the detail analysis is grammar design. This paper described the context free grammar design for the railway inquiry domain. The grammar was based on the corpus of questions collected for the domain. The paper also describes results of different analysis done on the corpus that are used to design the grammar. The grammar is successfully tested on test corpus using YACC generated parser.

Context Free Grammar for the Question Set

Production Rules	Example
S ? V NP VP	Does + garibrath + has a stoppage at Mumbai
| V NP	Is + any direct train to goa from nagpur
| Wh-NP VP	What + is the departure time of Mangla express
| Wh-NP Aux NP VP	How many stoppage + does + rajdhani express + has from Mumbai to Delhi
| P Wh-NP NP VP	By + what time + vidarbha Express + reaches Madras
NP ? NP Conj NP	departure + and + arrival
| (Det) (Card) (Ord) (Quant) (AP) NOM	the + next + more + fast + train
	(Det) (Ord) (Quant) (AP) NOM
NOM ? N NOM	departure + time
| N	concession
| NOM (PP)*	departure time + of the train
VP ? V	is coming
| V NP	reaches + wardha
| V NP PP	having + stoppage + at akola
| V PP	going + to mumbai
PP ? P NP	to + Mumbai
V ? Verb	coming
| Aux	is
| Aux V	is + coming
Wh-NP ? Wh-Word	what
| Wh-Word NOM	when + howrah mail
| Wh-Word Quant	how + much
AP ? Adj AP	summer + special
| Adj	special

Fig.4 Context free grammar with example

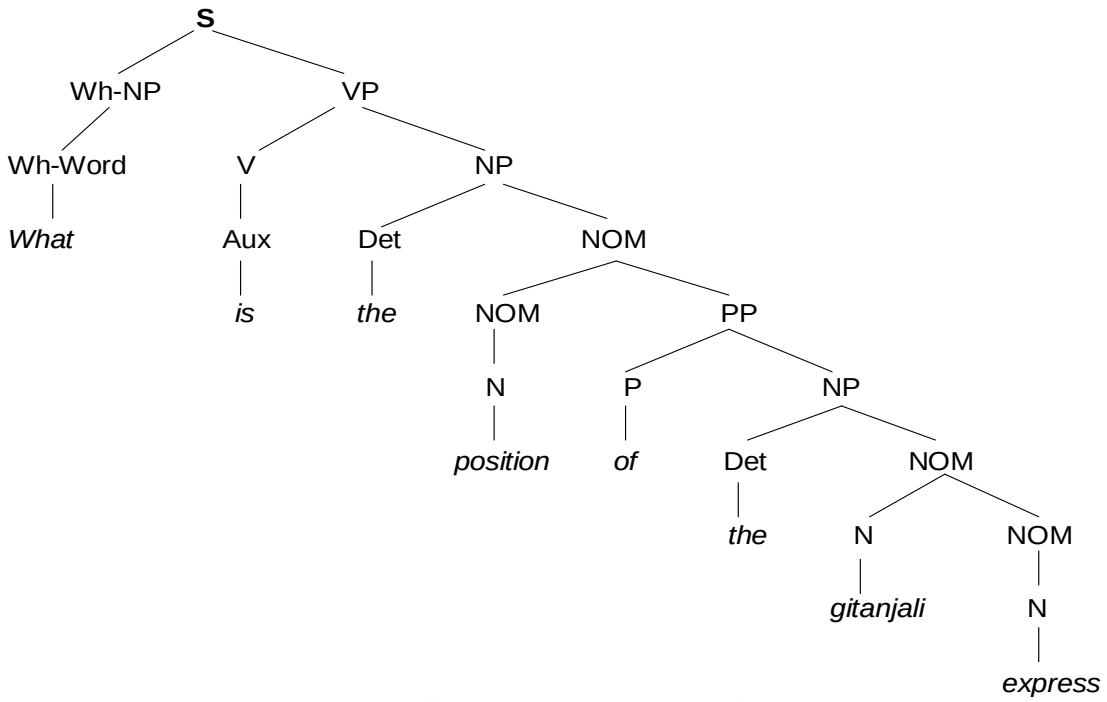


Fig.5 A sample Parse tree for question "What is the position of the gitanjali express

After deciding the syntax of input questions some semantic structure can be used to represent the meaning of the given natural language query. Now a days, NLP research has progressed tremendously and many techniques are available at lexical and syntax level. The basic problem of NLIDB lies in semantic analysis. Deep semantic analysis increases the success rate and portability by using semantic structures defined previously. This semantic structure may be either First order predicate calculus (FOPC), Head driven phrase structured grammar (HPSG), Frame structure, Semantic network etc [3]. An appropriate semantic structure needs to be selected and designed for the domain to represent meaning of given question.

References

[1] Agrawal Avinash J., Using Domain Specific Question Answering Techniques for Automatic Railways Inquiry on Mobile Phone, 5th International Conference on Information Technology: New Generations (ITNG 2008), Las Vegas, Nevada, USA, April 7-9, 2008,1111-1116.

[2] Aho, Ullman, Principle of Compiler Design, A book Published by Narosa Publications, 2002.

[3] Allen James, Natural Language Understanding, A Book Published by Addison Wesley Publications, 1994.

[4] Androutsopoulos I., Ritchie G.D., and Thanisch P., Natural Language Interfaces to Databases – An Introduction, Journal of Natural Language Engineering Part 1, 1995, 29–81.

[5] Chomsky, Noam, Three models for the description of language, Information Theory, IEEE Transaction 2, Sept. 1956,56.

[6] Jurafsky D., Martin J., Speech and Language Processing, A Book Published by Prentice Hall Publications, 2008.

[7] Lesmo Leonardo, Natural Language Query interpretation in restricted domains, 6th International Conference on NLP ICON-08, Pune, India, 2008.

[8] Levine John, Lex & Yacc, A book published by O'Reilly Publication, 1992.

[9] Minock, Michel, Where are the killer application of restricted domain question answering, In proceeding of the IJCAI Workshop on Knowledge Reasoning in Question Answering, Edinbergh, Scotland, 2005,page 4.

[10] Minock Michael, Olofsson Peter, Naslund Alexander, Towards Building Robust Natural Language Interface to Database, 13th International Conference on Applications of Natural Language to Information Systems, NLDB 2008 London, UK, 2008,187-198.

[11] Popescu Ana-Maria, Etzioni Oren, and Kautz Henry, Towards a Theory of Natural Language Interfaces to Databases, IUI, 2003,149–157.

[12] Popescu Ana-Maria, Armanasu Alex, Etzioni Oren, Ko David, and Yates Alexander, Modern Natural Language Interfaces to Databases : Composing Statistical Parsing with Semantic Tractability, 20th international conference on Computational Linguistics COLING-2004,Geneva, Switzerland, 2004,141.

[13] Wong Yuk Wah, Learning for Semantic Parsing Using Statistical Machine Translation Techniques, Technical Report UT-AI-05-323, University of Texas at Austin, Artificial Intelligence Lab, 2005.

Avinash J. Agrawal received Bachelor of Engineering Degree in Computer Technology from Nagpur University, India and Master of Technology degree in Computer Technology from National Institute of Technology, Raipur, India in 1998 and 2005 respectively. He is currently pursuing Ph.D. from Visvesvaraya National Institute of Technology, Nagpur. His research area is Natural Language Processing and Databases. He is having 12 years of teaching experience. Presently he is Assistant Professor in Shri Ramdeobaba Kamla Nehru Engineering College, Nagpur. He is the author of seven research papers in International and National Journal, Conferences.

Dr. O. G. Kakde received Bachelor of Engineering degree in Electronics and Power Engineering from Visvesvaraya National Institute of Technology (formerly Visvesvaraya Regional College of Engineering), Nagpur, India and Master of Technology degree in Computer Science and Engineering from Indian Institute of Technology, Mumbai, India in 1986 and 1989 respectively. He received Ph.D. from Visvesvaraya National Institute of Technology, Nagpur, India in 2004. His research interest includes theory of computer science, language processor, image processing, and genetic algorithms. He is having over 22 years of teaching and research experience. Presently he is Professor and Dean, Research and Development at Visvesvaraya National Institute of Technology, Nagpur, India. He is the author or co-author of more than thirty scientific publications in international journals, international conferences, and national conferences. He also authored five books on data structures, theory of computer science, and compilers. He is the life member of Institution of Engineers, India. He also worked as the reviewer for international and national journals, international conferences, and national conferences and seminars.

Rise of Data Mining: Current and Future Application Areas

Dharminder Kumar
[1]Professor and Dean, Faculty of Engineering & Technology,
Guru Jambheshwar University of Science & Technology, Hisar

Deepak Bhardwaj
Research Scholar, Department of Computer Science & Engineering
Guru Jambheshwar University of Science & Technology, Hisar

Abstract

Knowledge has played a significant role on human activities since his development. Data mining is the process of knowledge discovery where knowledge is gained by analyzing the data store in very large repositories, which are analyzed from various perspectives and the result is summarized it into useful information. Due to the importance of extracting knowledge/information from the large data repositories, data mining has become a very important and guaranteed branch of engineering affecting human life in various spheres directly or indirectly. Advancements in Statistics, Machine Learning, Artificial Intelligence, Pattern recognition and Computation capabilities have given present day's data mining functionality a new height. Data mining have various applications and these applications have enriched the various fields of human life including business, education, medical, scientific etc. Objective of this paper is to discuss various improvements and breakthroughs in the field of data mining from past to the present and also to explores the future trends.

Keywords— Current and Future of Data Mining, Data Mining, Data Mining Trends, Heterogeneous Data, KDD

I. INTRODUCTION

THE advent of information technology have affected various aspects of human life, may it be in the form of modernization of banking, land records, libraries, or data regarding population. This advent in various fields of human life has led to the very large volumes of data stored in various formats like documents, records, images, sound recordings, videos, scientific data, and many new data formats. An important new trend in information technology is to identify meaningful data collected in information systems [18]. The fact lies in that data is growing at a very rapid rate, but most of data has once been stored and have never been used. This data collected from different sources if processed properly, can provide immense hidden knowledge, which can be used further for development. As this knowledge is captured, it can serve as a key to gaining competitive advantage over competitors in industry [18]. So, there is an eminent need for developing proper mechanisms of processing these large volumes of data and extracting useful knowledge from large repositories for better decision making. Data Mining (as called as Knowledge discovery in databases (KDD), aims at the discovery of useful information from large collections of data [1] but large scale automated search and interpretation of discovered regularities belongs to KDD, but are typically not considered as part of data mining. KDD is concerned with knowledge discovery process applied to databases. KDD refers to overall process of discovering useful knowledge from data, while data mining refers to application of algorithms for extracting patterns from data [18]. The core functionalities of data mining includes applying various methods and algorithms in order to preprocess, classify, cluster and associate the data in order to discover useful patterns of stored data [2]. Data mining is best described as the union of historical and recent developments in statistics, AI, machine learning and Database technologies. These techniques are then used together to study data and find previously-hidden trends or patterns within. Data mining is finding increasing acceptance in science and business areas which need to analyze large amounts of data to discover trends which they could not otherwise find [19]. Data mining can be seen as the confluence of multiple fields including statistics, machine learning, databases, pattern discovery and visualization etc. [17]. The various application areas of data mining are Life Sciences (LS), Customer Relationship Management (CRM), Web Applications, Manufacturing, Competitive advantage, Intelligence, Retail, Finance, Banking, computer, Network, Security, Monitoring, Surveillance, Teaching Support, Climate modeling, Astronomy, and Behavioral Ecology etc. Hence, the objective of this paper is to reviews various trends of data mining and its relative areas from past to present and explores the future areas of it. This paper is organized as follows section 2 presents ground roots of data mining section 3 presents current trends in data mining section 4 presents future trends of data mining and finally conclusion follows.

II. ROOTS OF DATA MINING

Roots of Data Mining can be traced back along three lines[19]. *Final Stage*

A. Statistics

The most important lines is statistics. Without statistics, there would be no data mining, as statistics are the foundation of most technologies on which data mining is built. Statistics embrace concepts such as regression analysis, standard distribution, standard deviation, standard variance, discriminant analysis, cluster analysis, and confidence intervals, all of which are used to study data and data relationships. These are the very building blocks with which more advanced statistical analyses are underpinned. Certainly, within the heart of today's data mining tools and techniques, classical statistical analysis plays a significant role.

B. Artificial Intelligence & Machine Learning

Data mining's second longest family line is artificial intelligence and machine learning. AI is built upon heuristics as opposed to statistics, and attempts to apply human-thought-like processing to statistical problems. Because this approach requires vast computer processing power, it was not practical until the early 1980s, when computers began to offer useful power at reasonable prices. AI found a few applications at the very high end scientific/government markets, but the required supercomputers of the era priced AI out of the reach of virtually everyone else. Machine Learning could be considered as an evolution of AI, because it blends AI heuristics with advanced statistical methods. It let computer programs learn about the data they study and then apply learned knowledge to data.

C. Databases

Third family is Databases. Huge amount of data needs to be stored in a repository, and that too needs to be managed. So, comes in light the databases. Earlier data was managed in records and fields, then in various models like hierarchical, network etc. Relational model served the needs of data storage for long while. Other advanced system that emerged are object relational databases. But in data mining, volume of data is too high, so we need specialized servers for it. We call the term as Data Warehousing. Data warehousing also supports OLAP operations to be applied on it, to support decision making [20].

D. Other Technologies

Apart from these, data mining inculcates various other areas, e.g. pattern discovery, visualization, business intelligence etc. The table summarizes the evolution data mining on the grounds of development in databases.

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing &Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

Fig. 1. Evolution of Data Mining, Source [21]

III. CURRENT TRENDS AND APPLICATIONS

Data mining is formally defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [2]. The field of data mining has been growing rapidly due to its broad applicability, achievements and scientific progress, understanding. A number of data mining applications have been successfully implemented in various domains like fraud detection, retail, health care, finance, telecommunication, and risk analysis...etc. are few to name.

The ever increasing complexities in various fields and improvements in technology have posed new challenges to data mining; the various challenges include different data formats, data from disparate locations, advances in computation and networking resources, research and scientific fields, ever growing business challenges etc. [22]. Advancements in data mining with various integrations and implications of methods and techniques have shaped the present data mining applications to handle the various challenges, the current trends of data mining applications are:

A. Fight against Terrorism [23]

After 9-11 attacks, many countries imposed new laws against fighting terrorism. These laws allow intelligence agencies to effectively fight against terrorist organizations. USA launched Total Information Awareness program with the goal of creating a huge database of that consolidate all the information on population. Similar projects were also launched in European countries and rest of the world. This program faced several problems,

a. The heterogeneity of database, the target database had to deal with text, audio, image and multimedia data.

b. Second problem was scalability of algorithms. The execution time increases as size of data (which is huge). For example, 230 cameras were placed in London, to read number plates of vehicles. An estimated 40,000 vehicles pass camera every hour, in this way the camera must recognize 10 vehicles per second, which poses heavy loads on both hardware and

software.(Transport for London 2004).

B. Bio-informatics and Cure for Diseases

The second most important application trend, deals with mining and interpretation of biological sequences and structures. Data mining tools are rapidly being used in finding genes regarding cure of diseases like Cancer and AIDS [16].

C. Web and Semantic Web

Web is the hottest trend now, but it is unstructured. Data mining is helping web to be organized, which is called semantic web. The underlying technology is Resource Description Framework (RDF) which is used to describe resources. FOAF is also a supporting technology, heavily used in Facebook and Orkut for tagging. But still there are issues like combining all RDF statements and dealing with erroneous RDF statements. Data mining technologies are serving a lot to make the web, a semantic web.

D. Business Trends

Today's business environment is more dynamic, so businesses must be able to react quicker, must be more profitable, and offer high quality services that ever before. Here, data mining serves as a fundamental technology in enabling customer's transactions more accurately, faster and meaningfully. Data mining techniques of classification, regression, and cluster analysis are used for in current business trends [17]. Almost all of the current business data mining applications are based on the classification and prediction techniques for supporting business decisions, thus creating strong Business Intelligence (BI) system.

IV. DATA MINING – THE NEXT WAVES

Data mining is a promising area of engineering and it does have wide applicability. It can be applied in various domains. Data mining, as the confluence of multiple intertwined disciplines, including statistics, Machine learning, pattern recognition, database systems, information retrieval, World Wide Web, visualization, and many application domains, has made great progress in the past decade [24], [25]. Further Han in research challenges in data mining in science and engineering [26] presents major research challenges in the area of science and engineering.

A. Data Mining in Security and Privacy Preserving

Security and privacy are not very new concepts in data mining, but there is too much that can be done in this area with data mining. [50] gives a thorough analysis of impact of social networks and group dynamics. Specifying the need to understand cognitive networks, he also models knowledge network using the Enron E-mail corpus. Recording of electronic communication like email logs, and web logs have captured human process. Analysis of this can present an opportunity to understand sociological and psychological process. [27] provides various types of privacy breach and present an analysis using k-candidate anonymity [28], [29], k-degree anonymity [30] and k-neighborhood anonymity[31]. Various solutions are emerging like privacy preserving link analysis [32] which needs consideration in future. Secure Multiparty Computation (SMC) [33], [34] can be used where multiple parties, each having private input, want to communicate.

B. Challenges in Mining Financial Data

There are many motivating factors for the study of this area. Biggest is profit[35]. everyone wants profit may it be investor, speculator or operator in trading. He presents models of assets prices, and presents the modeling of relative changes of stock prices. Eraker [36] discuss the issues in modeling stochastic volatility better. [37] present a global solution for Distributed Recommendations in an adaptive decentralized network.

C. Detecting Eco-System Disturbances

This is another promising area. It comprises of many areas such as remote sensing, earth-science, biosphere, oceans and predicts the ecosystem. [38] tries to explain what are the problems in the area and what is the importance. There are also issues in mining the earth science like high dimensionality because long time series data are common in data mining. Study of this area is important due to radical changes in ecosystem has led to floods, drought, ice-storms, hurricanes, tsunami and other disasters [39]. Land Cover Change detection is also one of the areas. In a press release by NASA [40] shows the history of natural disasters.

D. Distributed Data Mining

Conventional data mining is thought to be as containing a large repository, and then mine knowledge. But there is an eminent need for mining knowledge from distributed resources. Typical algorithms which are available to us are based on assumption that the data is memory resident, which makes them unable to cope with the increasing complexity of distributed algorithms [41]. Similar issues also rise while mining data in sensor network, and grid data mining. We need distribution classification algorithms. A technique called partition tree construction approach [42] can be used for parallel decision tree construction. We also need distributed algorithms for association analysis. Distributed ARM algorithms needs to be developed as the sequential algorithms like Apriori, DIC, DHP and FP Growth [43], [44], [45], [46], [47] do not scale well in distributed environment. In his research paper the author presents a Distributed Apriori algorithm [48]. The FMGFI algorithm [49] presents a distributed FP Growth algorithm

V. CONCLUSION

In this paper we briefly reviewed the various data mining trends and applications from its inception to the future. This review puts focus on the hot and promising areas of data mining. Though very few areas are named here in this paper, yet they are those which are commonly forgotten. This paper provides a new perspective of a researcher regarding applications of data mining in social welfare.

REFERENCES

[1] Heikki, Mannila, "Data mining: machine learning, statistics, and databases", *Statistics and Scientific Data Management*, pp. 2-9. 1996.
[2] Fayadd, U., Piatesky -Shapiro, G., and Smyth, P. *From Data Mining To Knowledge Discovery in Databases*, AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–26256097–6. MIT 1996.
[3] Piatetsky-Shapiro, Gregory, "The Data-Mining Industry Coming of Age", *in IEEE Intelligent System*s, vol. 14, issue 6, Nov 1999. Doi. 10.1109/5254.809566
[4] Salmin, Sultana et al., "Ubiquitous Secretary: A Ubiquitous Computing Application Based on Web Services Architecture", *International Journal of Multimedia and Ubiquitous Engineering* Vol. 4, No. 4, October, 2009
[5] Hsu J., "Data Mining Trends and Developments: The Key Data Mining Technologies and Applications for the 21st Century", in *The Proceedings of the 19th Annual Conference for Information Systems Educators (ISECON 2002)*, ISSN: 1542-7382. Available Online: http://colton.byuh.edu/isecon/2002/224b/Hsu.pdf
[6] Shonali Krishnaswamy, "Towards Situationawareness and Ubiquitous Data Mining for Road Safety: Rationale and Architecture for a Compelling Application", *Proceedings of Conference on Intelligent Vehicles and Road Infrastructure 2005*, pages-16, 17. Available at : http://www.csse.monash.edu.au/~mgaber/CameraReadyI
[7] Kotsiantis, S., Kanellopoulos, D., Pintelas, P., "Multimedia mining", *WSEAS Transactions on Systems*, No 3, s. 3263-3268, 2005
[8] Abdulvahit, Torun. , Ebnem, Düzgün, "Using spatial data mining techniques to reveal vulnerability of people and places due to oil transportation and accidents: A case study of Istanbul strait", *ISPRS Technical Commission II Symposium*, Vienna. Addison Wesley, 1st edition. 2006
[9] T. M. Mitchell, "Generalization as Search", in *Artificial Intelligence* vol. 18 no. 2, pp.203-226. 1982
[10] R. Michalski., I. Mozetic., J. Hong., and N. Lavrac, "The AQ15 Inductive Leaning System: An Overview and Experiments", *Reports of Machine Leaning and Inference Laboratory*, MLI-86-6, George Maseon University. 1986
[11] J. R. Quinlan. *C4.5: Programs for Machine Learning*, San Francisco: Morgan Kaufmann Publishers, 1993
[12] Z. K. Baker and V. K.Prasanna. "Efficient Parallel Data Mining with the Apriori Algorithm on FPGAs" *IEEE International Parallel and Distributed Processing Symposium* (IPDPS '05), 2005.
[13] Jing He, "Advances in Data Mining: History and Future", *Third international Symposium on Information Technology Application*, 978-0-7695-3859-4/09 IEEE 2009 DOI 10.1109/IITA.2009.204
[14] Ali Meligy, "A Grid-Based Distributed SVM Data Mining Algorithm", *European Journal of Scientific Research* Vol.27 No.3. Pp.313-321 © Euro Journals Publishing, Inc. Available at :http://www.eurojournals.com/ejsr.htm
[15] S. Mitra, S. K. Pal, and P. Mitra. "Data mining in soft computing framework: A survey", *IEEE Trans. Neural Networks*, vol. 13, pp. 3 - 14., 2006
[16] Mark, J., Embrechts, "Introduction to Scientific Data Mining: Direct Kernel Methods & Applications", *Computationally Intelligent Hybrid Systems: The Fusion of Soft Computing and Hard Computing*, Wiley , New York, pp. 317-365, 2005
[17] Han, J., & Kamber, M. 2001. Data mining: Concepts and techniques .Morgan-Kaufman Series of Data Management Systems. San Diego: Academic Press.
[18] Walter Alberto Aldana, "Data Mining Industry, Emerging Trends and New Opportunities", Master's Thesis, Massachusetts Institute of Technology, 2000.
[19] Data Mining Software http://www.dataminingsoftware.com
[20] M.S. Chen, J. Han, and P.S. Yu. "Data mining: An overview from database perspective", *IEEE transactions on Knowledge and Data Eng*., 8(6):866-883, December 1999
[21] Pilot Software, "An Introduction to Data Mining", Whitepaper. Pilot Software. 1998.
[22] Venkatadari M., Dr. Lokanataha C. Reddy, "A Review on Data Mining From Past to Future", *International Journal of Computer Applications*, pp. 19-22, vol. 15, No. 7, Feb 2011.
[23] Huysmans, Baesens, Martens, Denys and Vanthienen, "New Trends in Data Mining", *Tijdschrift voor Economie en Management*, vol. L, 4, 2005.
[24] Kargupta, Han, Yu, Motwani, Vipin Kumar, "Next Generation of Data Mining", Chapman & Hall /CRC Data Mining and Knowledge Discovery Series, Taylor and Francis Group LLC, 2008.
[25] J. Han and M. Kamber. Data Mining: Concepts and Techniques (2nd ed.). Morgan Kaufmann, San Francisco, CA, 2006.
[26] J. Han, J. Gao, "Research Challenges for Data Mining in Science and Engineering", in Next Generation of Data Mining, Taylor and Francics Group LLC 2008.
[27] Liu, K. Das, T. Grandison & H. Kargupta, "Privacy Preserving Data Analysis on Graph and Social Networks", in Next Generation of Data Mining, Taylor and Francis Group LLC, 2008
[28] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava. Anonymizing social networks. Technical Report, University of Massachusetts, Amherst,MA, 2007.
[29] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In *Proceedings of the 17th ACM SIGACT-SIGMODSIGART "Symposium on Principles of Database Systems (PODS'98)*, p. 188, Seattle, WA, 1998.
[30] K. Liu and E. Terzi, "Towards identity anonymization on graphs", *In Proceedings of ACM SIGMOD*, pp. 93–106, Vancouver, Canada, June 2008.
[31] B. Zhou and J. Pei. "Preserving privacy in social networks against neighborhood attacks" In *Proceedings of the 24th International Conference on Data Engineering (ICDE'08)*, pp. 506–515, Cancun, Mexico, April 2008.
[32] Y. Duan, J. Wang, M. Kam, and J. Canny, "Privacy preserving link analysis on dynamic weighted graph" *Computational and Mathematical Organization Theory*, 11:141–159, 2005.
[33] A. C. Yao " How to generate and exchange secrets", In *Proceedings of 27th IEEE Symposium on Foundations of Computer Science*, pp. 162–167, Toronto,Canada, 1986.
[34] O. Goldreich. The Foundations of Cryptography, Vol. 2, Chapter 7. Cambridge University Press, Cambridge, UK, 2004.
[35] James E. Gentle, "Challenges in Financial Data Mining", In Next Generation of Data Mining, Taylor and Francis Group, LLC 2008.
[36] Eraker, B., 2004, "Do stock prices and volatility jump? Reconciling evidence from spot and option prices", *The Journal of Finance* 59, 1367–1403.
[37] Olfa Nasraoui and Maha Soliman, "Market-Based Profile Infrastructure: Giving Back to the User", Next Generation of Data Mining, Taylor and Francis, 2008.
[38] Shayam Boriah, Vipin Kumar, Michael Steinbach, Pang-Ning Tan, Christopher Potter, and Steven Klooster, "Detecting Ecosystem Disturbances and Land Cover Change using Data Mining", In NGDM, 2008
[39] C. Potter, P.-N. Tan, V. Kumar, C. Kucharik, S. Klooster, V. Genovese, W. Cohen, and S. Healey. "Recent history of large-scale ecosystem disturbances in North America derived from the AVHRR satellite record". *Ecosystems*, 8(7):808–824, 2005.
[40] Press Release: "Data Mining Reveals a New History of Natural Disasters",NASA. http://www.nasa.gov/centers/ames/news/releases/2003/03 51AR.html.
[41] Chris Clifton, Wei Jiang, M. Murugesan, and M.E. Nergiz, "Is Privacy Still and Issue for Data Mining", In NGDM, Taylor and Francis, 2008.
[42] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE 2006)*, pp. 23–34, Atlanta, GA, April 2006.
[43] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms", In *Proceedings of the 20th ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems*, pp. 247–255, Santa Barbara, CA, May 21–23, 2001. ACM.
[44] W. Kim, R. Agrawal, C. Faloutsos, U. Fayyad, J. Han, G. Piatetsky-Shaprio, D. Pregibon, and R. Uthurasamy. "Data mining" is not against civil liberties. Open Letter from the Directors of the Executive Committee of ACMSIGKDD, July 28, 2003.
[45] A. Kobsa. "Technical solutions for Privacy-enhanced personalization", *Communications of the ACM*, 50: 24–33, August 2007.

[46] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal of the European Communities, I (281): 31–50, October 24, 1995.
[47] L. Sweeney. "k-anonymity: A model for protecting privacy", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5): 557–570, 2002.
[48] C. Clifton. "Using sample size to limit exposure to data mining", *Journal of Computer Security*, 8(4): 281–307, November 2000.
[49] R. Feingold, J. Corzine, R. Wyden, and B. Nelson. Data Mining Moratorium Act of 2003. U.S. Senate Bill (proposed), January 16, 2003.
[50] N. Pathak, S. Mane, J. Srivastava, N. Contractor, S. Poole & D. Williams, "Analysis of Social Networks and Group Dynamics from Electronic Communication", In Next Generation of Data Mining, Taylor and Francics, 2008.

Effect of different defuzzification methods in a fuzzy based load balancing application

Sameena Naaz[1], Afshar Alam[2] and Ranjit Biswas[3]

[1]Department of Computer Science
Jamia Hamdard, New Delhi, India

[2]Department of Computer Science
Jamia Hamdard, New Delhi, India

[3]Manav Rachna International University
Faridabad, Haryana, India

Abstract

In a distributed environment the workload on the network has to be managed in such a way that the total throughput of the system can be maximized. For this to happen some of the jobs have to be migrated from one node to another. When, how and where a job has to be migrated depends upon the load balancing algorithm being used. But it is very difficult to precisely describe the behavior of a complex system as there are many factors which influence it. One way to deal with the uncertainty in the behavior of the system is to use fuzzy logic. Fuzzy logic uses the reasoning of the human mind which is not always in the form of a yes or no. The concept of linguistic variables is used to model the state of the system which is imprecise and uncertain. In this work, we have implemented the fuzzy load balancing algorithm and compared the effect of using different defuzzification methods, reported in the literature.

Keywords: Distributed Systems, Load Balancing, Fuzzy Logic, Defuzzification.

1. Introduction

Over the years the hardware technology has grown on a massive pace with the result of increase in the use of distributed systems. These systems have the advantage of sharing of resources as well as processing power. The processes arrive in the system in a random manner on different nodes. When the jobs are being executed in parallel on different systems a decision has to be made on to which system a newly arrived job has to be send. Load balancing is the technique which helps in even distribution of the jobs among the available nodes so that the throughput can be increased.

The load balancing algorithms can be categorized as static or dynamic in nature. Static algorithms collect no information and make probabilistic balancing decisions, while dynamic algorithms collect varying amounts of state information to make their decisions. Previous research on static and dynamic load balancing can be found in [l]-[5], [6, 7], respectively. It has been established from the previous studies that dynamic algorithms give better performance improvement as compared to static algorithms.

Different load balancing algorithms have different complexity which depends upon the amount of communication needed to approximate the least loaded node. In order to make a decision the information about the state of the different nodes has to be collected. However, since messages containing state information for individual nodes can only be exchanged at discrete intervals and are subject to variable latencies before reaching their destinations, the information used by nodes to estimate global system state is inevitably out of date. This uncertainty in global state has been a primary issue in the design of efficient distributed computing systems. Increasing the frequency of information exchange between nodes is not necessarily a practical solution since message overheads caused by the frequent exchange of state information may adversely affect the efficiency of the system. Moreover, the overheads of load balancing mechanisms can be highly detrimental to the performance of the system under heavy system load conditions.

When we are talking about large distributed systems there is huge amount of global state uncertainty present in it. Fuzzy logic based distributed load balancing algorithms reflect the effect of uncertainty in decision making process.

This approach has been discussed in [8]. The fuzzy logic approach for Distributed Object Computing Network has been studied in [9, 10]. Parallel and distributed computing environment is inherently best choice for solving/running distributed and parallel program applications. In such type of applications, a large process/task is divided and then distributed among multiple hosts for parallel computation. In [10] it has been pointed out that in a system of multiple hosts the probability of one of the hosts being idle while other host having multiple jobs queued up can be very high. In [11] the performance of a new Fuzzy Load balancing algorithm is compared with the existing algorithms.

In a distributed environment the processors are categorized according to workload in their CPU queues as heavily loaded (more tasks are waiting to be executed), lightly loaded (less tasks are waiting to be executed in CPU queue) and idle processors/hosts (having no pending work for execution). Here CPU queue length is used as an indicator of workload at a particular processor. The algorithms used for load balancing may require no information, or only information about individual jobs (static algorithm) or may make decisions based on the current load situation (dynamic algorithm).

In general, load balancing algorithm can be analyzed in a framework with four dimensions: selection policy, transfer policy, information policy, and location policy. Specifically, information and location policies have the most important roles.

Transfer policy: First of all the state of the different machines is determined by calculating it's workload. A transfer policy determines whether a machine is in a suitable state to participate in a task transfer, either as a sender or a receiver. For example, a heavily loaded machine could try to start process migration when its load index exceeds a certain threshold.

Selection policy: This policy determines which task should be transferred. Once the transfer policy decides that a machine is in a heavily-loaded state, the selection policy selects a task for transferring. Selection policies can be categorized into two policies: preemptive and non-preemptive. A preemptive policy selects a partially executed task. As such, a preemptive policy should also transfer the task state which can be very large or complex. Thus, transferring operation is expensive. A non-preemptive policy selects only tasks that have not begun execution and, hence, it does not require transferring the state of task.

Location policy: The objective of this policy is to find a suitable transfer partner for a machine, once the transfer policy has decided that the machine is a heavily-loaded state or lightly-loaded one. Common location policies include: random selection, dynamic selection, and state polling.

Information policy: This policy determines when the information about the state of other machines should be collected, from where it has to be collected, and what information is to be collected.

2. Fuzzy Logic Concept

In narrow sense, fuzzy logic is a logical system, which is the extension of multivalued logic. In a wider sense fuzzy logic is almost synonymous with the theory of fuzzy sets, a theory which relates to classes of object with unsharp boundaries in which membership is a matter of degree. Fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic. The mapping then provides a basis from which decisions can be made.

Basically a fuzzy logic system consists of the following 5 steps:

Fuzzification: Converting the crisp inputs to membership functions which comply with intuitive perception of system status.

Rules Processing: Calculating the response from system status inputs according to the pre-defined rules matrix (control algorithm implementation)**.**

Inference: Evaluating each case for all fuzzy rules

Composition: Combining information from rules

De-Fuzzification: Converting the result to crisp values.

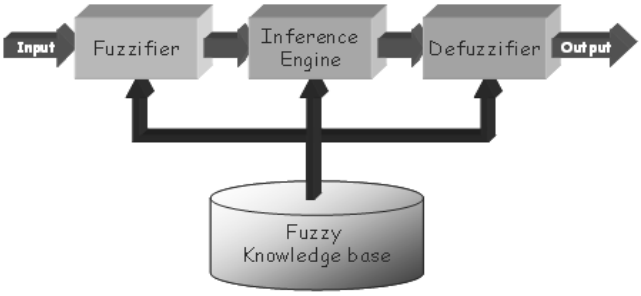


Figure 1: Fuzzy Inference System

In this paper we have compared the following five defuzzification methods

Centroid of area Z_{COG}
Bisector of area Z_{BOA}
Mean of maximum Z_{MOM}
Smallest of maximum Z_{SOM}

Largest of maximum Z_{LOM}

Centroid principle or Center of Gravity
This method is also known as center of gravity or center of area defuzzification. This technique was developed by Sugeno in 1985. This is the most commonly used technique. The only disadvantage of this method is that it is computationally difficult for complex membership functions. The centroid defuzzification technique can be expressed as

$$z_{COG} = \frac{\int_{Z} \mu_A(z) z dz}{\int_{Z} \mu_A(z) dz}$$

where z_{COG} is the crisp output, $\mu_A(z)$ is the aggregated membership function and z is the output variable

Bisector Method
The bisector is the vertical line that divides the region into two sub-regions of equal area. It is sometimes, but not always coincident with the centroid line.

$$\int_{\alpha}^{z_{BOA}} \mu_A(z) dz = \int_{z_{BOA}}^{\beta} \mu_A(z) dz$$

Largest of Maximum
Largest of maximum takes the largest amongst all z that belong to [z1, z2] as the crisp value called $Z_{LOM.}$

Smallest of Maximum
It selects the smallest output with the maximum membership function as the crisp value Z_{SOM}. In other words in Smallest of Maximum choose smallest among all z that belong to [z1, z2]

Mean of Maximum
In this method for defuzzification only active rules with the highest degree of fulfillment are taken into account. The output is computed as:

$$z^* = (a+b)/2$$

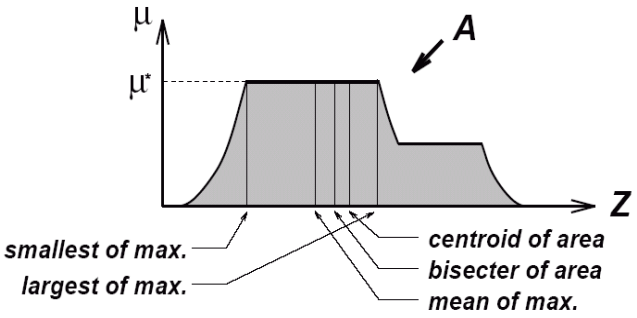


Figure 2: Results using different defuzzification methods for a particular function.

3. Distributed System Model

A simple model of a distributed system is presented here. This model consists of a decentralized decision making approach with cooperation from all the nodes. So the performance can be improved here purely by intelligent decision making and proper coordination. The various nodes of the system here are the resources and each of these resources can be in different states. A vector is used to give the state of a node which describes many characteristics of the node. The elements of this state vector are measures which imply a cost or penalty for using the resource.
The set of states of all the resources in the distributed system is known as the global system state. In distributed load balancing also the decisions are not always necessarily made using the complete global state information. In fact for each node under consideration only a subset of neighboring nodes may be needed to take a decision. Another important aspect is that a node can change state faster than the time taken to transmit state information from one state to another. Therefore there is always some amount of uncertainty in the state information used for making a decision. Hence it is necessary that the decision making process deals with these uncertainties. Fuzzy logic is one of the methods of dealing with this uncertain information and has been used in the work presented in this paper.
The Scheduler in this algorithm has to perform the following tasks.

Threshold Estimation

Decision Making

Scheduler has two functions, threshold estimation and decision making. When a scheduler is invoked, it estimates two numerical thresholds from the current states of uncertainty sources based on a fuzzy control base, and

making scheduling and state update decision using fuzzy consistency model.

We need to define fuzzy sets for the input parameters, 'load', and 'number of heavy load node' levels, and the output, 'status of load balancing node'. For this we define five membership functions for first input parameter i.e. 'load' and two membership functions for second input parameter i.e. 'number of heavy load node' and two membership functions for output parameter 'status of load balance node.

3.1 Threshold Estimation

The Threshold Estimation determines the limiting value for each membership function. Beyond this limiting value the membership function will change.

First Input parameter: Load (0-10)
Member Function 1: Very lightly (0-2)
Member Function 2: lightly (1-5)
Member Function 3: moderate (4-6)
Member Function 4: heavy (5-9)
Member Function 5: very heavy (8 -10)

Second Input Parameter: No. of heavy load node (0-5)
Member Function 1: more (0-2.5)
Member Function 2: less (2.5 – 5)

Output Parameter: Status of load balance node (0-10)
Member Function 1: receiver (0-5)
Member Function 2: sender (6-10)

In our work here we have taken the Gaussian distribution function for all the different linguistic variables for the input "load". This is shown in figure 3.

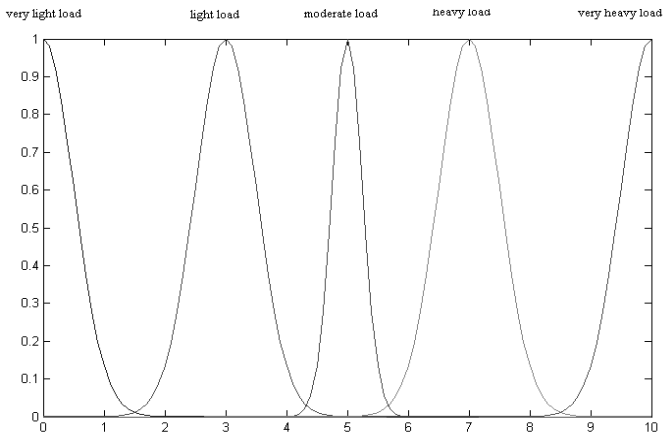


Figure 3: Input variable load of the node under consideration and it's membership function

The membership function used for the number of heavy load nodes is shown in figure 4.

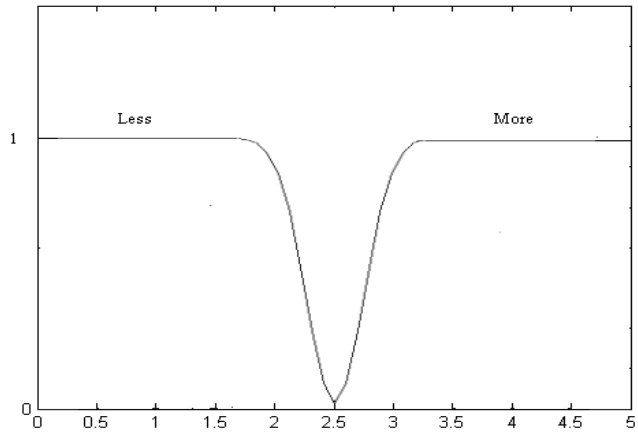


Figure 4: Input variable No. of Heavy Load Node and it's membership function.

The membership function for the output variable status of load balance node is shown in figure 5. From this figure we can see that there are two linguistic variables sender and receiver here and the load on a node determines it's value based upon the membership function.

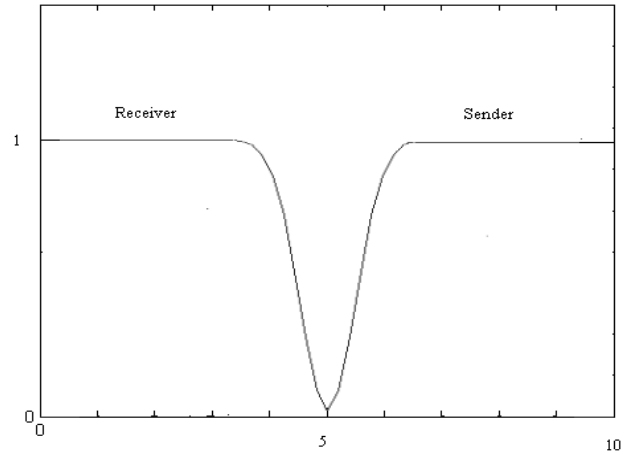


Figure 5: Membership function for the output variable Status of Load Balance Node

3.2 Decision Making

The Fuzzy rules that have been used in this work are given below:

Rule [1]. If (load is very light) then (node is receiver)

Rule [2]. If (load is very heavy) then (node is sender)

Rule [3]. If (load is heavy) and (no. of heavy load nodes is less) then (node is sender)

Rule [4]. If (load is heavy) and (no. of heavy load nodes is more) then (node is receiver)

Rule[5]. If (load is light) and (no. of heavy load nodes is more) then (node is receiver)

Rule[6]. If (load is light) and (no. of heavy load nodes is less) then (node is sender)

Rule [7]. If (load is moderate) and (no. of heavy load nodes is more) then (node is receiver)

Rule [8]. If (load is moderate) and (no. of heavy load nodes is less) then (node is sender)

This rule base is used to find out the value of the output variable using the fuzzy inference method.

4. Interpretation of Results

We have done the implementation of scheduler on MATLAB. We have taken two input parameters and one output parameter for fuzzy implementation of our logic. The first input parameter is 'load' and the second one is 'Number of heavy Load Node' and one output i.e. 'status of load balance node'. We measure the input parameters load and Number of heavy load node on a scale of 0 to 10 and 0 to 5 respectively and the output parameter status of load balancing node on a scale of 0 to 10

Based upon the crisp values that are obtained the nodes are categorized either as sender or as receiver. We have calculated this crisp value using the five defuzzification methods described above.

The surface plots that we obtain for the results are shown in Figures 6 to 10 and the input and output values obtained for 20 sets of data is shown in Table 1.

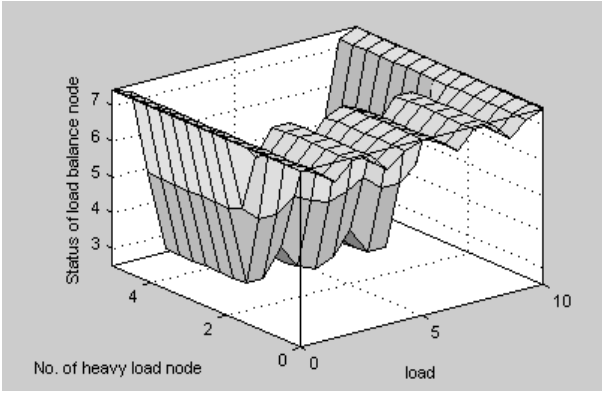


Figure 6 : Centroid Method

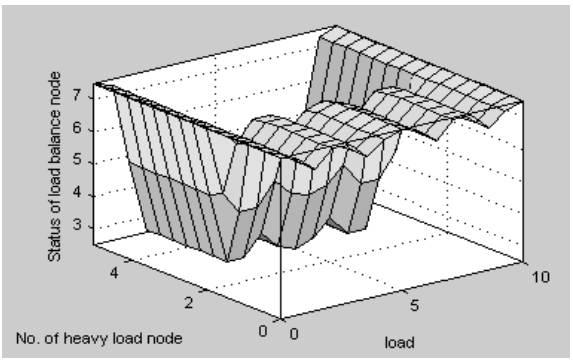


Figure 7: Bisector method

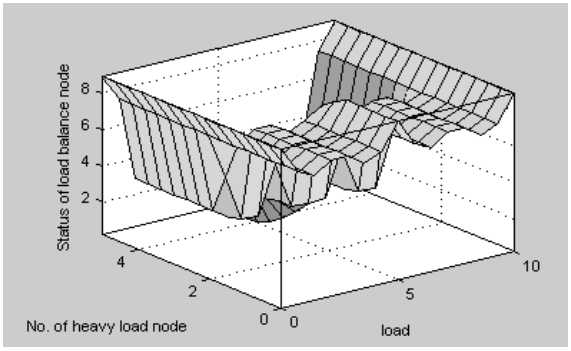


Figure 8: Mean of Maximum

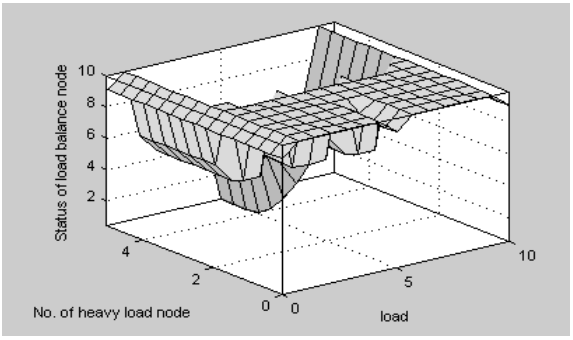


Figure 9: Largest of Maximum

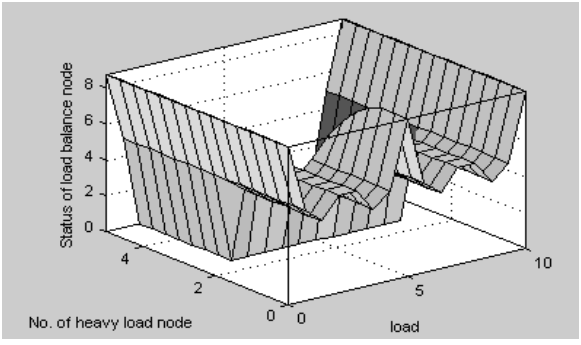


Figure 10: Smallest of Maximum

Table 1: Output values obtained for different defuzzification methods.

S.No	INPUTS		OUTPUT				
	Load	No. of Heavy Load Node	Centroid	Bisector	MOM	LOM	SOM
1	7	2	7.2647	7.3000	7.5500	10.0000	5.1000
2	4	5	2.8181	2.8000	2.6000	5.2000	0
3	9	3	7.0142	7.1000	7.3500	10.0000	4.7000
4	6	1	7.1095	7.1000	7.3500	10.0000	4.7000
5	4	5	2.8181	2.8000	2.6000	5.2000	0
6	10	5	7.4348	7.5000	8.9500	9.1000	8.8000
7	9	3	7.0142	7.1000	7.3500	10.0000	4.7000
8	6	1	7.1095	7.1000	7.3500	10.0000	4.7000
9	7	2	7.2647	7.3000	7.5500	10.0000	5.1000
10	6	3	2.8415	2.8000	2.6000	5.2000	0
11	3	3	2.6763	2.7000	2.4000	4.8000	0
12	4	2	7.0955	7.1000	7.3500	10.0000	4.7000
13	5	4	2.5057	2.5000	0.2000	0.4000	0
14	3	4	2.6635	2.6000	2.4000	4.8000	0
15	9	2	7.0955	7.1000	7.3500	10.0000	4.7000
16	2	1	7.2724	7.3000	7.5500	10.0000	5.1000
17	3	2	7.2647	7.3000	7.5500	10.0000	5.1000
18	2	2	7.2647	7.3000	7.5500	10.0000	5.1000
19	3	3	2.6763	2.7000	2.4000	4.8000	0
20	5	3	2.5143	2.5000	1.4000	2.8000	0

5. Conclusion and Future Work

The results obtained using the five defuzzification methods have been shown in table1. From this table we find that centroid method, bisector method and mean of maximum method are giving us approximately the same results in the load balancing application that we have taken. Where as for the smallest of maximum and largest of maximum approaches there is wide variations in the results that are obtained. The reason for this is that these two methods use the two extremes i.e smallest or largest values for calculation of the crisp value.

The results obtained in the tables above are graphically shown in figures 6 to 10 and from these figures also we infer the same results. Hence we conclude that centroid, bisector and MOM methods are better as compared to the LOM, SOM, as there is more consistency in the results.

In future this work has to be extended by using these methods in actual simulation for load balancing to find out the effect on response time.

References

[1]. P. V. McGregor and R. R. Boorstyn, “Optimal load balancing in a computer network,” in Proc. 1975 Int. Conf: on Commun., vol. 3, pp. 41.14-41.19.

[2]. E. D. S. Silva and M. Gerla, “Load balancing in distributed systems with multiple classes and site constraints,” in Proc. Performance’84, pp. 17-33.

[3]. A. N. Tantawi and D. Towsley, “A general model for optimal static load balancing in star network configurations, ” in Proc. Performance’84, pp. 277-291.

[4]. A. N. Tantawi and D. Towsley, “Optimal static load balancing in distributed computer systems,” J. ACM, vol. 32, no. 2, pp. 445465, Apr. 1985.

[5]. J. F. Kurose and S. Singh, “A distributed algorithm for optimum static load balancing in distributed computer systems,” in Proc. IEEE INFOCOM’86, pp. 458-467.
[6]. F. Bonomi and A. Kumar, “Adaptive optimal load balancing in a heterogeneous multiserver system with a central job scheduler,” IEEE Trans. Computer, vol. 39, pp. 1232-1250, Oct. 1990.
{7]. T. C. K. Chow and J. A. Abraham, “Load balancing in distributed systems,” IEEE Trans. Software Eng., vol. SE-8, pp. 401-412, July 1982.
[8]. Chulhye Park and Jon G.Kuhl, “A Fuzzy based distributed load balancing algorithm”, Proceedings of the Second International Symposium on Autonomous Decentralized Systems (ISADS'95) IEEE, 1995.
[9]. Lap-Sun Cheung, “A Fuzzy Approach to Load Balancing in a Distributed Object Computing Network”, First IEEE International Symposium on ClusterComputing and the Grid (CCGrid'01) risbane, Australia May 15-May 18, 2001.
[10]. Yu-Kwong Kwok And Lap-Sun Cheung, “A New Fuzzy-Decision Based Load Balancing System For Distributed Object Computing”, Journal Of Parallel And Distributed Computing, Volume 64 Issue 2, February 2004
[11]. Abbas Karimi, Faraneh Zarafshan, Adznan b. Jantan, A.R. Ramli and M. Iqbal b. Saripan, “A New Fuzzy Approach for Dynamic Load Balancing Algorithm” International Journal of Computer Science and Information Security(IJCSIS).

AUTHORS PROFILE

Sameena Naaz received the degree of B.Sc Engg. in computers from Aligarh Muslim University, in 1998 and the M..Tech Degree in Electronics from Aligarh Muslim University, in 2000. She is pursuing her Ph. D from Hamdard University. Sameena Naaz has worked as a lecturer at Amity College of Engg. And Tech. Delhi, Inti College Malaysia and is currently working as an Assistant Professor at Jamia Hamdard University, New Delhi India in the Department of Computer Science. She is a member of International Association of Computer Science and Information Technology (IACSIT). Her research interests include soft computing and load balancing and scheduling in distributed systems.

Professor Afshar Alam has an MCA and Ph. D degree and is working as a Professor at Jamia Hamdard University in the Department of Computer Science

Professor Ranjit Biswas has an M.tech , Ph. D and is currently working with Manav Rachna International University Faridabad, Haryana, India.

Visualising Pipeline Sensor Datasets with Modified Incremental Orthogonal Centroid Algorithm

Olufemi Ayinde Folorunso[1] and Shahrizal Sunar Mohd[2]

[1]UTMViCubeLab,
Department of Computer Graphics & Multimedia,
Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia, 81310, Skudai, Johor

[2] UTMViCubeLab,
Department of Computer Graphics & Multimedia,
Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia, 81310, Skudai, Johor

Abstract

Each year, millions of people suffer from after-effects of pipeline leakages, spills, and eruptions. Leakages Detection Systems (LDS) are often used to understand and analyse these phenomena but unfortunately could not offer complete solution to reducing the scale of the problem. One recent approach was to collect datasets from these pipeline sensors and analyse offline, the approach yielded questionable results due to vast nature of the datasets. These datasets together with the necessity for powerful exploration tools made most pipelines operating companies "data rich but information poor". Researchers have therefore identified problem of dimensional reduction for pipeline sensor datasets as a major research issue. Hence, systematic gap filling data mining development approaches are required to transform data "tombs" into "golden nuggets" of knowledge. This paper proposes an algorithm for this purpose based on the Incremental Orthogonal Centroid (IOC). Search time for specific data patterns may be enhanced using this algorithm

Keywords: *Piggin, Heuristics, Incremental, Centroid.*

1. Introduction

Pipelines are essential components of the energy supply chain and the monitoring of their integrities have become major tasks for the pipeline management and control systems. Nowadays pipelines are being laid over very long distances in remote areas affected by landslides and harsh environmental conditions where soil texture that changes between different weathers increase the probability of hazards not to mention the possibility of third party intrusion such as vandalism and deliberate attempt of diversions of pipeline products. It is widely accepted that leakages from pipelines have huge environmental, cost and image impacts.

Conventional monitoring techniques such as the LDSs could neither offer continuous pipeline monitoring over the whole pipeline distance nor present the required sensitivity for pipeline leakages or ground movement detection. Leakages can have various causes, including excessive deformations caused by earthquakes, landslides, corrosion, fatigue, material flaws or even intentional or malicious damaging.

Pipeline sensors datasets are structurally different and fundamentally unique for so many reasons. In the first place, these data are generated asynchronously, that is, each data attribute gets its instantaneous copy of datum at the same time at any particular time. With this, it is expected that the captured data will represent the in-state situation of the pipeline at any given time. In most cases, this is not true because some attributes "don't just get any data at all". Example of sensor datasets obtained from the velocity-vane anemometer is shown in Table 1. Secondly, pipeline sensor datasets are filled with noises. Noises are instances of repeatedly unwanted datasets clustered around a particular time. When datasets are captured this way, the noise level significantly determines the visualisation results. This is because, noises creates what is known as outliers which are exceptional data behaviour that are ordinarily are not suppose to be. They are outrageous exceptions falling into unacceptable limits for the specified data acceptable boundaries. By manual inspection, noises are difficult to detect and removed, this has greatly mediated the efficiency of the eventual visualisation results. Thirdly, pipeline sensors datasets comes in unrelated units and formats, making comparison very difficult. Example, the temperature is measured in degree Celsius while the Velocity is measured in m/s^2.

Table 1: Data Attributes and Variables from the Velocity-Vane Anemometer

Pressure (N/m^2)	*Temp. (^0C)*	*Vol. (M^3/H) x E-03*	*Flow Velocity (m/s)*	*External Body Force EBF (N)*
-	-	-	-	-
1.002312	19.302978	0.0055546	12.002302	-
1.002202	19.302990	0.0055544	12.002302	0.000344
-	19.302990	-	-	0.002765
0.903421	-	-	12.003421	-
1.002212	19.302978	0.0055546	12.004523	-
-	18.999996	0.0055544	12.005620	0.003452
0.960620	18.999996	-	-	-
1.002801	-	-	12.002302	0.003564
1.002376	19.302978	-	12.002302	0.005423
-	18.999996	-	-	0.005642
.	.	.	.	.
.	.	.	.	.

Even if every pixel on a standard display device is used to represent each datum, display device with the best resolution cannot display all the data generated by theses sensors in 1 minute at the same time, not even the 53 million pixel power wall presently being used at the University of Leeds [1]. The same goes for the memory size that is required for such computation as well as the computational time. Definitely, this will of course require greater computational effort and some compromises of visualisation results. Hence, users of visualisation applications tend to rely heavily on heuristics to arriving at decisions from their applications. Dimensionality reduction is therefore an alternative technique to explaining and understanding these vast pipeline sensors datasets more intuitively. Presence or absence of leakages or abnormal situations is gradually becoming the object of research in the recent time. In Nigeria, the 1992 pipeline explosions that claimed thousands of lives in Ejigbo is one good example of such underground oil installations that resulted to explosions and allied problems as a result of undetected leakage and bad response time to leakages due to imperfection and the errors in the visualisation and LDS systems. The mayhem was traced to inability to properly analyse and visualise leakage points about the pipeline. Although most of these pipeline failures are blamed on the activities of the vandals especially in developing nations, yet, the basic truth is that the visualisations of the various leakage detection systems are error-full. When leakages are quickly detected and fixed, it invariably reduces the vandals' activities as well as saving lives and reducing the overhead installation and administrative costs associated with pipeline installation and pigging operations.

A central problem in scientific visualisation is to develop an acceptable and resources efficient representation for such complex datasets [2, 3]. The challenges of high dimensional datasets vary significantly across many factors and fields. Some researchers including [4] and [5] viewed these challenges as scientifically significant for positive theoretical developments. There are so many problems of high dimensional datasets ranging from attributes relevance and presence to variable importance. In practical sense, not all the dimensions or attributes and not all variables or instances- presence or absence in high dimensional datasets are relevant for every specific user defined interests in understanding certain underlying phenomena represented by the datasets.

More recently, [5, 6, and 7] asserted that principal among the problems of dimensionality reduction is the issue of accuracy compromise. They all submitted that almost all data reduction algorithms and methods employ one or more procedures that lead to significant compromise of accuracy. Without any loss of generality, the problem under investigation has to do with trying to find the extent of allowable and reasonable reduction in data dimensions that could be carried out on high pipeline sensor datasets without a compromise of the desired visualisation quality obtainable from such datasets under specific or desired boundary conditions. Mathematically, given an n-dimensional random variable $x = (x_1, \ldots . x_n)^T$ a lower dimensional random variable $s = (s_1, \ldots . s_m)^T$ with n >>> m such that the entire member data of x are fully represented by s with n , m Є R (R is the set of real numbers) is required. The overall goal of reducing the data dimension is to enable a lower dimensional space reveal to us "as much as possible" details about a high dimensional data space with minimal loss of data integrity and compromise.

Often in computer graphics this is very necessary because the available devices (such as monitors) cannot display all the intrinsic elements of the voluminous datasets generated by modern day sensors and remote sensing devices. If the dimensionality of datasets could be reduced, the resulting data could be used more effectively in visualisation, verification, classification, and exploration. There are many dimensionality reduction algorithms and approaches. These are discussed in Section 2 of this paper.

2. Literature Review

Reducing dimensionality has been described as an essential task for many large-scale information processing problems involving document classification, searching over Web data sets [5]. Because of the exponential growth of the Web information and other remote sensing devices, many traditional classification techniques now require a very huge amount of memory and CPU resource if dimensionality reductions are not performed on the datasets as required. Sometimes, dimensionality reduction is a pre-processing step in data mining but may also be some steps towards data exploration and analysis such as in data clustering, visualisation etc. Historically, the Principal Components Analysis (PCA) originally credited to Pearson (1901) whose first appearance in modern literatures dates back to the work by Hotelling (1933) was a popular approach to reducing dimensionality. It was formerly called the Karhunen-Loeve procedure, eigenvector analysis and empirical orthogonal functions. The PCA is a linear technique that regards a component as linear combinations of the original variables . The goal of PCA is to find a subspace whose basis vectors correspond to the directions with maximal variances.

Let X be an dxp matrix obtained from sensor datasets for example, where d represents the individual data attributes (columns) and p the observations (or variables) that is being measured. Let us further denote the covariance matrix C that defined X explicitly as:

$$C = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T \quad (1.0)$$

Where $x_i \in X$ and $\bar{x}$ is the mean of x_i , T is the positional order of $x_i \in X$, and X is the covariance matrix of the sampled data. We can thus define an objective function as:

$$G(w) = W^T C W \quad (2.0)$$

The PCA's aims is to maximise this stated objective function G(W) in a solution space defined by:

$$H^{dxp} = \{W \in R^{dxp}, W^T W = I\} \quad (3.0)$$

It has been proved that the column vectors of W are the p higher or maxima eigenvectors of covariance matrix C defined above [see 8]). However, for very large and massive datasets like the pipeline sensors datasets, an enhancement of the PCA called the Incremental PCA developed by [9,10] could be a useful approach. The IPCA is an incremental learning algorithm with many variations. The variations differ by their ways of incrementing the internal representations of the covariance matrix. Although both the PCA and the IPCAs are very effective for most data mining applications, but, because they ignore the valuable class label information in the entire data space, they are inapplicable for sensor datasets.

The Linear Discriminant Analysis (LDA) emerged as another approach commonly used to carry out dimensionality reduction. Its background could be traced to the PCA and it works by discriminating samples in their different classes. Its goal is to maximize the Fisher criterion specified by the objective function:

$$G(w) = \frac{|W^T s_b W|}{|W^T s_w W|} \quad (4.0)$$

Where $s_b = \sum_{i=1}^{c} p_i (m_i - \bar{x})(m_i - \bar{x})^T$ and $s_w = \sum_{i=1}^{c} p_i E((x - m_i)(x - m_i)^T\}$ with $x \in ci$ are called the Inter class scatter matrix and Intra class scatter matrix respectively. E denotes the expectation and $p_i(x) = \frac{n_i}{n}$ is the prior probability of a variable (x) belonging to attribute (i).

W can therefore be computed by solving $w^* = \arg\max G(w)$ in the solution space $H^{dxp} = \{W \in R^{dxp}, W^T W = I\}$, in most reports; this is always accomplished by providing solution to the generalized eigenvalue decomposition problem represented by the equation:

$$S_b w = \lambda S_w w \quad (5.0)$$

When the captured data is very large like in the case of sensors datasets considered in this research, LDA becomes inapplicable because it is harder and computationally expensive to determine the Singular Value Decomposition (SVD) of the covariance matrix more efficiently. LDA uses attribute label information of the samples, which has been found unsuitable by many researchers including [5] for numerical datasets. [11] had developed a variant of the LDA called the Incremental LDA (ILDA) to solve the problem of inability to handle massive datasets, but, its stability for this kind of application remains an issue till present date.

The Orthogonal Centroid (OC) algorithm by [12 and13] is another acceptable algorithm that uses orthogonal transformation on centroid of the covariance matrix. It has been proved to be very effective for classification problems by [14] and it is based on the vector space computation in linear algebra by using the QR matrix decomposition where Q is an orthogonal matrix and R is an upper triangular matrix (Right Triangular Matrix) of the covariance matrix. The Orthogonal Centroid algorithm for dimensionality reduction has been successfully applied on text data (see [12]). But, the time and space cost of QR

decomposition are too expensive for large-scale data such as Web documents. Further, its application to numerical data or multivariate and multidimensional datasets of this sort remains a research challenge till date. However, its basic assumptions are extremely acceptable for development of such better algorithms.

In 2006, a highly scalable incremental algorithm based on the OC algorithm called the Incremental OC (IOC) was proposed by [5]. Because OC largely depends on the PCA, it is therefore not out of focus to state that the IOC is also a relaxed version of the conventional PCA. IOC is a one-pass algorithm. As dimensionality increases and defiles batch algorithms, IOC becomes an immediate alternative. The increase in data dimensionality could now be treated as a continuous stream of datasets similar to those obtainable from the velocity vane thermo-anemometer (VVTA) sensors and other data capturing devices, and then we can compute the low dimensional representation from the samples given, one at a time with user defined selection criterion Area of Interest (AOI) (iteratively). This reassures that the IOC is able to handle extremely large datasets. However, because of its neglect of the variables with extremely low eigenvalues, it is poised to be insensitive to outliers. Unfortunately, this is the case with the kind of data used in this research. There is therefore a necessity to improve the IOC algorithm to accommodate the insurgencies and the peculiarity presented by pipeline sensor datasets. The derivation of the IOC algorithm as well as the improvement proposed to the algorithm is discussed in detail in the following subsections.

3. IOC Derivation and the Proposed (HPDR) Improvement

Basic Assumption 1: The IOC optimization problem could be restated as

$$\max \quad \sum_{i=1}^{p} W^T S_b W \qquad (6.0)$$

The aim of this is to optimise equation 6.0 with $W \in X^{dxp}$, where the parameters have their usual meanings. However, this is conditional upon $w_i w_i^T=1$ with i=1,2,3,….p. Now, p belongs to the infinitely defined subspace of X, but, since it is not possible to select the entire variables for a particular data attribute at a time, we introduced a bias called Area of Interest (AOI) to limit each selection from the entire data space.

A Lagrange function L is then introduced such that:

$$L(w_k, \lambda_k) = \sum_{i=1}^{p} w_k S_b w_k^T - \lambda_k (w_k w_k^T - 1)$$

Or

$$L(w_k, \lambda_k) = \sum_{i=1}^{p} w_k S_b w_k^T - \lambda_k w_k w_k^T + \lambda_k) \qquad (7.0)$$

(Observe that if $w_k w_k^T=1$, then equation (7.0) is identically (6.0))

With λ_k being the Lagrange multipliers, at the saddle point, L must = 0. Therefore, it means $S_b w_k^T = \lambda_k w_k^T$ necessarily. Since obviously p >>>AOI at any point in time, this means that, w, the columns or attributes of W are p leading vectors of S_b. S_b (n) Can be computed therefore by using:

$$S_b(n) = \sum_{j=1}^{AOI} p_j(n) (m_j(n) - m(n))(m_j(n) - m(n))^T \qquad (8.0)$$

Where mj(n) is the mean of data attribute j at step i and m(i) is the mean of variables at step i. T is the order of the variable in the covariance matrix defined by data space X. To dance around this problem, the Eigen Value Decomposition (EVD) is the approach that is commonly used although it has been reported to have high computation complexity problems.

The EVD is computed by following the following procedure:
Given any finite data samples $X=\{x_1,x_2,x_3,...,x_n\}$ we first compute the mean of x_i by using the conventional formula:

$$\mu = \frac{1}{n}\sum_1^n x_i \qquad (9.0)$$

This is followed by the computation of the covariance C defined as:

$$C = \frac{1}{n}\sum_1^n (x_i - \bar{x})(x_i - \bar{x})^T \qquad (10.0)$$

Next, we compute the eigenvalue λ(s) and eigenvectors e(s) of the matrix C and iteratively solve:

$$Ce=\lambda e \qquad (11.0)$$

PCA then orders λ by their magnitudes such that λ1 >λ2>λ3>... >λn , and reduces the dimensionality by keeping direction e such that λ <<< T. In other words, the PCA works by ignoring data values whose eigenvalue(s) seems very insignificant. To apply this or make it usable for pipeline sensor datasets, we need a more adaptive incremental algorithm, to find the p leading eigenvectors of S_b in an iterative way. For sensor datasets, we present each sample of the selected AOI as: $(x\{n\}, l_n)$ where x{n} is the nth training data, l_n is its corresponding attribute label and n = 1, 2,3, ... AOI.

Basic Assumption 2: if given $\lim_{n\to\infty} a(n) = a$, then $\lim_{n\to\infty}(\frac{1}{n}\sum_{i=1}^{n} a(i)) = a$ by induction, therefore, it means that $\lim_{n\to\infty} s_b(n) = s_b$, using Assumption 1.0: which means that:

$$\lim_{n\to\infty}(\frac{1}{n}\sum_{i=1}^{n} s_b(i)) = s_b \qquad (12.0)$$

However, the general eigenvector form is Au =λu, where u is the eigenvector of A corresponding to the eigenvalue-λ. By replacing the matrix A with $s_b(n)$, we can obtain an approximate iterative eigenvector computation formulation with v = Au =λu or u= v/λ:

$$v(n) = \frac{1}{n}\sum_{i=1}^{n} s_b(i)\, u(i) \qquad (13.0)$$

Injecting equation 8.0 into equation 13.0 implies:

$$v(n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{AOI} p_j(n)\,(m_j(n) - m(n))(m_j(n) - m(n))^T\, u(i)$$

Assuming that $\Phi_j(i) = m_j(n) - m(n)$; it means

$$v(n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{AOI} p_j(n)\,\Phi_j(i)\Phi_j(i)^T\, u(i) \qquad (14.0)$$

Therefore, since u= v/λ: the eigenvector $\vec{u}$ can be computed using

$$\vec{u} = \frac{v}{\|v\|} \qquad (15.0)$$

But, vector $\vec{u}(i)$ could be explicitly defined as $\vec{u}(i) = \frac{v(i-1)}{\|v(i-1)\|}$, with i=1,2,3,…n. Therefore,

$$v(n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{AOI}\left(p_j(n)\Phi_j(i)\Phi_j(i)^T\right)\frac{v(i-1)}{\|v(i-1)\|} \qquad (16.0)$$

Hence;

$$v(n) = \frac{n-1}{n}v(n-1) + \frac{1}{n}\sum_{j=1}^{AOI}\left(p_j(n)\Phi_j(n)\Phi_j(n)^T\right)\frac{v(n-1)}{\|v(n-1)\|} \qquad (17.0)$$

If we substitute $\xi_j(n) = \Phi_j(n)^T\frac{v(n-1)}{\|v(n-1)\|}$, j=1,2,3….AOI, and if we set v(0)=x(1) as a starting point, then it is comfortable to write v(n) as:

$$v(n) = \frac{v(n-1)^2}{n} + \frac{1}{n}\sum_{j=1}^{AOI}\left(p_j(n)\Phi_j(n)\,\xi_j(n)\right) \qquad (18.0)$$

Since the eigenvectors must be orthogonal to each other by definition. Therefore, we could span variables in a complementary space for computation of the higher order eigenvectors of the underlying covariance matrix. To compute the (j+α)th eigenvector, where α=1,2,3…AOI, we then subtract its projection on the estimated jth eigenvector from the data.

$$x^{j+\alpha}(n) = x^j(n) - \frac{(x^j(n)^T v^j(n))}{\|v^j(n)\|^2} \qquad (19.0)$$

(Note that j+α = AOI for any particular selection)

Where $x_1(n) = x(n)$. Using this approach, we have been able to address the problem of high time consumption. This is because the orthogonality could now only be enforced when there is convergence which may not be at the beginning but may occur at any point at the extreme end of the selected and repeated AOIs. Through the projection procedure at each step, we can then get the eigenvectors of S_b one by one (i.e for each set of the predetermined AOI). The IOC algorithm summary as presented by [5] is shown in Algorithm 1 and improved IOC called the HPDR algorithm is presented in Algorithm 2.0, the solution of step n is given as:

$$v^j(n) = \frac{v^j(n)}{\|v^j(n)\|} \quad \text{with j=1,2,3…p} \qquad (20.0)$$

3.1The IOC Algorithm and the HPDR

By going through the algorithm an example could be used to illustrate how HPDR solves the leading eigenvectors of S_b incrementally and sequentially. Let us assume that input sensor datasets obtained from the two sources (manually and experimentally) are represented by $\{a_i\}$, i= 1,2,3,… and $\{b_i\}$, i= 1,2,3… When there is no data input, the means m(0), $m_1(0)$, $m_2(0)$, are all zero. If we let the initial eigenvector $v^1(1) = a_1$ for a start, then HPDR algorithm can be used to compute the initial values or the leading samples of the datasets $a_i(s)$ and $b_i(s)$ of the entire data space X. These initial values are given as: a_1, a_2, and b_1, b_2, and they can then be computed using equation 20.0.

3.2 The Expected Likelihoods (EL)

Because of IOC's insensitivity to outliers and less significant variables in the dataset, the expected likelihood is then computed. The computation is achieved by undergoing the following processes:

Given an arbitrary unordered set of data X defined by $X=\{x_1,x_2,x_3,\ldots x_n\}^k$ along with a set of unordered attributes $Z=\{X_1,X_2,X_3,\ldots X_N\}^{k-n}$ such that the attitudinal vector Zψ depends on the covariance matrix or X . The rowsum (RS), columsum (CS) and Grandtotal (GT) of the covariance matrix $X \mid X\psi$ are defined as:

$$RS = \sum_{i=1}^{k}\{X_i\}^N \quad (21.0)$$

$$CS = \sum_{i=1}^{k-n}\{Z_i\}^N \quad (22.0)$$

And

$$GT = \sum_{i=1}^{k-n}\{Z_i\}^N + \sum_{i=1}^{k}\{X_i\}^N \quad (23.0)$$

Using the product of the respective Row Sum (RS) the Column Sum (CS) divided by the Ground Total (GT), the expected for each of the covariance matrix elements could be estimated. The computation begins with the initialisation of counters for the row, the column and the Area of Interest (AOI) selected as i , j, and N respectively. The datum in the first data value in the first row and the first column is read and the expected value for this position is computed. The jth column positional value is advanced until all the five dimensions (J=5) are all traversed. The system then increment i and moves to the (i+1)th row positional value and the process continues until the entire value of the AOI=N is completely traversed. The WAEL is thus computed by finding the weighted average value of the data attributes as shown in Algorithm 2.0. Thus, the expected variable xi of {X} belonging to position {xi,yi} of the covariance matrix X is computed using the expected likelihood function:

$$E_k(x_i, y_i) = \frac{RS+CS}{GT} \quad (24.0)$$

The Averaged Expected Likelihood Al for $E_k(x_i, y_i)$ is defined further by

$$A_l = \sum_{k=1}^{k-n} E_k \begin{cases} E_{k-n} & \rightarrow \text{on major axis} \\ & \cdot \\ & \cdot \\ 0 & \text{elsewhere} \end{cases} \quad (25.0)$$

This gives a unit dimensional matrix A representing the original data X.

3.3 Weighted Average Expected Likelihoods (WAEL)

The WAEL is the weighted mean of the expected likelihoods and it is comparable but not the same as the arithmetic mean. It is based on the assumption that although each data value is important, they do not contribute equally to the flow dynamics and the selected datasets. It is determined by computing the average for the reduced expected likelihoods with a weight factor for each data entity. The weight factor (or the Information Gain (IG)) is the degree of sensitivity of the attribute to the entire data space. This idea plays a role in descriptive statistics and it also occurs in more general forms other areas of statistics and mathematics. This position is justified because judging from the IG computation for each of the attributes; we could see that each of the sensor data attributes contributes differently to the entire flow process. It will therefore be very illogical to use the simple average for the computation of the likelihoods.

Although WAEL will behave similar to the normal statistical means, if all the sensor datasets are equally weighted, then what is computed is just the arithmetic mean which is considered unsuitable for sensor datasets due to its variability. Example of such effects is found in what the statisticians know as the Simpson's Paradox. This paradox illustrates how correlation in different groups of data is completely reversed by just combining the two data groups. This is always the case when frequency of data is given causal interpretations hastily. However, Simpson's Paradox will disappear if causal relations (in terms of frequencies) are brought into consideration. The computation follows the conventional weighted average formula for the reduced dimension. For example in Table 2. we expanded IG computation for datasets represented in Table 1. to reflect the percentage contributions of each attributes. The percentage contribution is then calculated by the formula %Contribution = (Gain/Total Gain)*100.

Table 2: Percentage Contributions of Attributes

Data Attribute	***Information Gain***	***Percentage Contribution (%)***
Pressure (p)	0.898	26.5
Temperature (t)	0.673	19.86
Volume (v)	0.944	27.85
Flow Velocity (f)	0.445	13.13
Ext.Body force (e)	0.429	12.66

3.4 Data Attributes Selection

Based on the generated metadata, the data attributes selection could be performed by using the modified back propagation algorithm. Without modification, back propagation algorithm lacks robustness this is because errors grow exponentially while the attribute weight diminishes. It is observed that as the bias increases, there is heavy tendency for the error inherited into the visualisation to also rise. In the conventional back propagation algorithm, each attribute is given a weight which equals the sum of the errors inherited multiplied by the mean data entity. This condition of the back propagation algorithm has greatly mediated its use for modern applications of this sort.

With this modification however, it is possible to reduce the errors inherited by inversing the error threshold as shown in the modified version in Section 4. There are alternative methods for carrying out data classification, however, due to its robustness and wider acceptability, the decision tree algorithm by [15] is employed to carry out data classification. This algorithm works by computing the Information Gain (I.G) for each data attribute and promoting the one with the highest gain as the root for the tree as the test or lead attribute. This method forces the lead attribute to "inherit" transferable qualities of the other attributes which in turn provided a basis for quicker visualisation. The computation of the IG is achieved by using the conventional information gain formula:

$$I(s_1, s_2, \ldots s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (26.0)$$

Where pi = si/s is the probability that an arbitrary sensor data belong to a class Ci . Log base 2 has been used because the data are encoded in bits and si is the number of sample S in class Ci. m is the number of case attributes.

3.5 High Performance Dimensionality Reduction Algorithm (HPDR)

To achieve high performance in dimensionality reduction, this paper is structured as a form of combinational framework (like a bridge) between the Feature Extraction based method -IOC and the Feature Selection based method- the EL. The strength is derived by the introduction of a mechanism for users' choice of Areas of Interest (AOI). This is made possible by effectively determining the IG by each of the attributes and determining the lead attribute. Fixing the expected likelihood for the cases of emptiness completely remove the shortfalls insensitivity to outliers and less significant variables in the dataset [16]. Using this approach is not completely new; it has been found extremely advantageous in statistical and mathematical applications see examples in [17, 18 19]. It is often used for the computation of the popular Chi-square in non-parametric statistics for example. The normalised data is simply subjected to the HPDR (Algorithm 2).

Algorithm 1: Conventional IOC Dimensionality Reduction Algorithm

for $n = 1, 2, \ldots$, do the following steps,

$$m(n) = ((n-1)m(n-1) + x(n))/n$$
$$N_{l_n}(n) = N_{l_n}(n-1) + 1$$
$$m_{l_n}(n) = (N_{l_n}(n-1)m_{l_n}(n-1) + x(n))/N_{l_n}(n)$$
$$\Phi_i^1(n) = m_i(n) - m(n), i = 1, 2, \cdots, c$$

for $j = 1, 2, \ldots, \min\{p, n\}$

if $j = n$ then

$$v^j(n) = x(n)$$

else

$$\alpha_i^j(n) = \Phi_i^j(n)^T v^j(n-1) \Big/ \left\| v^j(n-1) \right\|$$
$$v^j(n) = \frac{n-1}{n} v^j(n-1) + \frac{1}{n} \sum_{i=1}^{c} \alpha_i^j(n) p_i(n) \Phi_i^j(n)$$
$$\Phi_i^{j+1}(n) = \Phi_i^j(n) - \Phi_i^j(n)^T v^j(n) v^j(n) / \left\| v^j(n) \right\| \left\| v^j(n) \right\|$$

end if

end for

end for

*Yan et al.(2006)

Dimensionality reduction algorithms are extremely useful in improving the efficiency and the effectiveness of datasets classifiers [5]. Reducing dimensionality this way is of great importance to ensure quality and efficiency of data classifiers for large scale and continuous data streams like sensor's datasets, this is because of the poor classification efficiency of earlier approach such as the IOC powered by the high dimension of the data space. It has been viewed and described as an essential data mining and data pre-processing approach for large scale and streaming datasets classification tasks.

3.6 Analysing the HPDR Algorithm

This algorithm must be repeated p number of time (iterations) and for each iteration there is the need to predetermine AOI set of variables {j}. This is free for any user to determine the area where specific data intuition is needed such that:

$\alpha_i^j(n)$, with i = 1,2,3, ... AOI
(α (n) has its usual meaning)

Algorithm 2. High Performance Dimensionality Reduction Algorithm

for n=1,2,3,...AOI do the following steps:

$M(n)=((n-1)m(n-1)+x(n))/n$

$N_{ln}(n)= N_{ln}(n-1)+1$

$M_{ln}(n)=(N_{ln}(n-1)m_{ln}(n-1)+x(n))/N_{ln}(n)$

$\Phi_i^1(n)=m_i(n)-m(n)$, i=1,2,...5

for i=1,2,...5; j=1,2,3,... AOI (max i=5, because we have just 5 dimensions)

If j=n then $V^j(n)=x(n)$

else

$$\alpha_i^j(n) = \Phi_i^j(n)^T \frac{v^j(n-1)}{\|v^j(n-1)\|}$$

$$v^j(n) = \frac{v^j(n-1)^2}{n} + \frac{1}{n}\sum_{j=1}^{AOI}\left(p_j(n)\Phi_j(n)\,\alpha_i^j(n)\right)$$

$$x^{j+\alpha}(n) = \Phi_i^j(n) - \frac{(\Phi_i^j(n)^T v^j(n))}{\|v^j(n)\|^2}$$

10Compute the expected E(i)for each j of the AOI Є C

$$E_x = \frac{RSi * CSi}{GT}$$

E_x *into position* P_i;

n--

if n>1, then i++;

If i>5, j++; go to Step 10 otherwise;

Compute Weighted Averaged Expected Likelihood (WAEL)-A_l

$$A_l = \sum_{x=1}^{n} \lambda * E_x/5$$

end if

end if

Return A_l into position pi

end if

end for

end for

When computational complexities are out of it, HPDR offers a faster approach to reducing the dimensionality of the datasets based on the predefined criteria. The strength of this algorithm lies in the interaction with subtlety of the intrinsic data interdependencies. When users are empowered to make their choice of the area to visualise or explore, better results are obtained. Because the computation is done one after the other in an iterative manner, HPDR offers the advantage of improved memory usage, this is a good and better promise that the earlier approaches in terms of the storage requirements. Viewing from another angle, considering the volume and nature of the pipeline sensor datasets, it is practically impossible to render the whole data, even after the dimensionality has been reduced. The HPDR offers the benefit of AOI selection; this enables step by step and continuous processing of the data in a manner that supersedes the conventional batch processing technique.

4. Procedures

Given D = n x m data space and two disjointed datasets {X, Sk Є D} Assuming that dataset (X)= {xi; $1 \le i \le \xi$ Є N+} and dataset (Sk) = { sj; $1 \le j \le \lambda$ Є N+} Є D such that X∩ Sk = ϕ, then X and Sk are independent variables (vectors) of the set D it follows that:

$$\text{Centroid (cXi)} = \overline{X} + \overline{Sk} = \left(\frac{\frac{1}{\lambda}\sum_{j=1}^{\lambda} s_j + \frac{1}{\xi}\sum_{i=1}^{I} x_i}{2}\right) \qquad (27.0)$$

or

$$2cXi = \frac{1}{\lambda}\sum_{j=1}^{\lambda} s_j + \frac{1}{\xi}\sum_{i=1}^{\xi} x_i \qquad (28.0)$$

$\overline{X}$ and $\overline{Sk}$ denotes the means of X and Sk respectively, λ and ξ are arbitrary constants. If all missing λsand ξs can be computed and inserted by "any means" into D such that nλ = nξ, it follows that:

$$cXi = \frac{1}{2\lambda}\left(\sum_{j=1}^{\lambda} s_j + \sum_{i=1}^{\lambda} x_i\right) \qquad (29.0)$$

If Sk represents a specific scenario Ap Є D. Therefore with the new centres for each classes or attributes, dataset D can be regrouped more effectively.

5. Results and Evaluation

Generally, there is no uniformity or industrial standard for testing and implementing dimensionality reduction across all applications; many researchers have developed area-specific dimensionality reduction algorithms and

techniques which has made comparison extremely difficult. Examples of such area or domain specific application are found in [3, 5, 6, 7, 16,17, 18, 19,20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, and 33] to mention but a few.

Most researchers make use of statistical illustrations and comparative graphs to compare dimensionality reduction and data mining techniques. Examples are found in [5,16]. Dimensionality reduction helps to make better statistical decisions that could lead to significant and concrete results in pipeline sensors data visualisations. This could be in the form of increased income or energising efficient processes. The future suggests that the choice of such an effective dimensionality reduction and data mining tool will depend on the expected return on the overall efforts put into it. It is therefore imperative to critically examine and assess the overall business situation in question and how the selected tool could effectively achieve the goals of dimensionality reduction and the data mining process. To help evaluation, some checklists have been compiled using the Cross Industry Standard Process for Data Mining (CRISP-DM).

The CRISP-DM is a six-phase process. The choice of tool however should be flexible thereby allowing selective changes to the entire data space as may be deemed necessary. The six stages involved are: Business understanding; Data understanding; Data preparation; Modelling; Evaluation and Deployment. The algorithms compared are the Principal Component Analysis (PCA), the Linear Discriminant Analysis (LDA), the Incremental Orthogonal Centroid (IOC) and the proposed High Performance Dimensionality Reduction algorithm (HPDR) on the datasets obtained from two source: The VVTA and the Turbulence Rheometer. The results obtained are presented in Table 3.

Table 3: Summary of the Result Obtained Comparing Four Dimensionality Reduction Algorithms

	(AOI) - SELECTED VARIABLES											
	<14			***15-30***			***31-45***			***46-60***		
	EPR %	CMCR	TT (s)	EPR %	CMCR	TT (s)	EPR %	CMCR	TT (s)	EPR %	CMCR	TT (s)
HPDR	10	0.20	2.0	15	0.15	2.2	15	0.22	3.3	25	0.14	3.6
PCA	5	0.452	2.26	8	0.389	3.11	10	0.315	3.15	15	0.217	3.25
LDA	7	0.429	3	7	0.43	3.01	5	0.602	3.01	5	0.602	3.01
IOC	4	0.50	1.99	4	0.50	1.99	8	0.25	2	10	0.20	2.0

Note: The CMCR is computed using the ratio TT/EPR

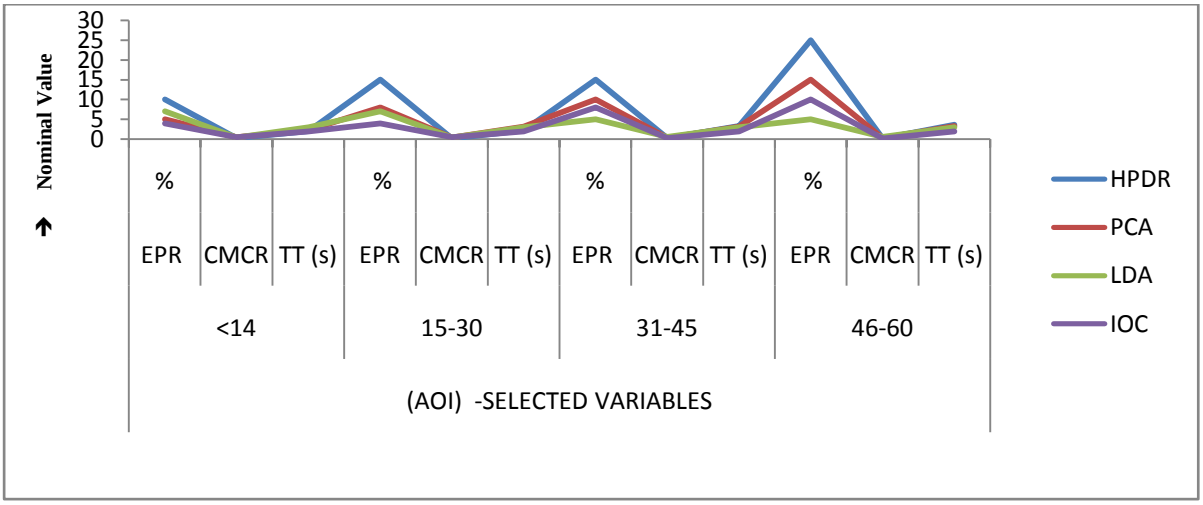


Fig. 1. Comparing Dimensionality Reduction Algorithms

The evaluation of the proposed method is designed as an assessment of the model proposed prior deployment when compared with existing and previously used techniques. The evaluation phase examines how the original data obtained from the sensors have been injected into the developed algorithm and how the results obtained is of any significance to the users of the system. However, this paper has been able to compare the dimensionality

reduction algorithms' efficiency when applied to reducing a five dimensional sensor data obtained from the velocity vane thermo-anemometer and the Turbulence Rheometer into one dimension. The parameters used for comparison are the Error in Prediction Ratio (EPR), the Covariance Matrix Convergence Ratio (CMCR) and the averaged Time Taken (TT) for the computation. Similar comparison methods are found in the works reported by [3, 5, 16, and 29].

From the graph in Figure 1, the HPDR shows a lot of promises for higher selection of AOI although this has not been tested beyond 15 rows of selected variables at any single time due to the limitations imposed by the renderer. As shown, the %EPR obviously promises to increase as the AOI selection increases. The HPDR algorithm also showed a better improvement when compared with the existing techniques that are currently being used. Figure 1 was generated automatically using the Microsoft Excel worksheet with the vertical axis representing the nominal value in terms of the algorithms' performances.

6. Conclusion

It was observed that as the number of variables begins to increase beyond the predefined set limit of 15 for each AOI, the IOC and the HPDR shows some similarities in terms of efficiency of time. In one of our recent publications, It was suggested that a synchronisation data steaming device could be used as a means of increasing the attributes and the variables without a compromise of data integrity but there are positions yet unclosed in this suggestion because it simply depended on heuristics. Here, the sensor datasets are non fuzzy, so heuristics has no part to play hence, it is advisable not to apply this streaming device for now until further researches proved otherwise.

However, looking at the example reported by [16], it could be stated that the modified algorithm may significantly be a good starting focus for predictions and fuzzy applications. In their example, they made use of the Penalised Independent Component Analysis on DNA microarray data whose results obtained justified this assertion.

When the attributes of the pipeline sensor datasets exceeds five with more excessively large amount of datasets beyond the Microsoft Excel native rows, there are no guidelines or rule to offer at the moment because of the limitations particular to the Microsoft Excel which is obviously outside the scope of this research. The future direction of this work is on the possibility applying the devices on data capture for the algorithm directly to further improve the depiction of certainty of the sensors' datasets visualisation as well as providing new algorithms for saving operational and hazards costs in pipelining.

Acknowledgements

This work is supported by the UTMViCubeLab, FSKSM, Universiti Teknologi Malaysia. We thank the Nigerian National Petroleum Corporation (NNPC) for the release of necessary data to test run the algorithms at various stages. Special thanks to (MoHE), Malaysia and the Research Management Centre (RMC), UTM, through Vot.No. Q.J130000.7128.00J57, for providing financial support and necessary atmosphere for this research.

References

[1]. C. Goodyer, J. Hodrien, W. Jason and K. Brodlie. (2009). "Using high resolution display for high resolution 3d cardiac data. The Powerwall"., University of Leeds – p. 5/16 . The Powerwall Built from standard PC components of 7computers.

[2]. D.S. Ebert, R.M. Rohrer, C.D. Shaw, P. Panda, J.M. Kukla, and D.A.Roberts (2000). "Procedural shape generation for multi-dimensional data visualisation". Computers and Graphics. Vol. 24, pp. 375–384.

[3]. S.Masashi (2007). "Dimensionality reduction of multimodal labeled data by local Fisher Discriminant analysis". Journal of Machine Learning Research, Volume 8, 2007, pp. 1027-1016.

[4]. D.L. Donoho (2000). "High-dimensional data analysis. The curses and blessings of dimensionality". Lecture delivered at the "Mathematical Challenges of the 21st Century" conference of The American Math. Society, Los Angeles, August 6-11.

[5]. J. Yan, Z. Benyu, L. Ning, Y. Shuicheng, C. Qiansheng, F. Weiguo, Y. Qiang, Xi Wensi, and C. Zheng (2006). "Effective and Efficient Dimensionality Reduction for Large-Scale and Streaming Data Preprocessing". IEEE Transactions on Knowledge And Data Engineering, Vol. 18, No. 3, March 2006. pp 320-333.

[6]. R. da Silva-Claudionor, A. Jorge, C. Silva and R.A. Selma (2008). "Reduction of the dimensionality of hyperspectral data for the classification of agricultural scenes". 13th Symposium on Deformation Measurements and Analysis, and 14th IAG symposium on geodesy for Geotechnical and structural Engineering, LNEC Libson May, 2008LBEC, LIBSON, May 12-15, pp. 1-10.

[7]. L. Giraldo, L.F Felipe,. and N. Quijano (2011). "Foraging theory for dimensionality reduction of clustered data". Machine learning, Vol 82, pp 71-90.

[8]. R.J. Vaccaro. (1991). "SVD and Signal Processing II: Algorithms, Analysis and Applications". Elsevier Science, 1991.

[9]. M. Artae, M. Jogan, and A. Leonardis (2002). "Incremental PCA for OnLine Visual Learning and Recognition". Proceedings of the 16th International Conference on Pattern Recognition. pp. 781-784.

[10]. J. Weng, Y. Zhang, and W.S. Hwang (2003). "Candid Covariance Free Incremental Principal Component

Analysis". IEEE Transaction on Pattern Analysis and Machine Intelligence. Vol. 25, pp. 1034-1040.
[11]. K. Hiraoka, K. Hidai, M. Hamahira, H. Mizoguchi, T. Mishima and S. Yoshizawa (2004). "Successive Learning of Linear Discriminant Analysis: Sanger-Type Algorithm". Proceedings of the 14th International Conference on Pattern Recognition. pp. 2664-2667.
[12]. M. Jeon, H. Park, and. J.B Rosen (2001). "Dimension Reduction Based on Centroids and Least Squares for Efficient Processing of Text Data". Technical Report MN TR 01-010, Univ. of Minnesota, Minneapolis, Feb. 2001
[13]. H. Park, M. Jeon, and J. Rosen, (2003). "Lower Dimensional Representationof Text Data Based on Centroids and Least Squares". BIT Numerical Math., vol. 43, pp. 427-448.
[14]. P. Howland and H. Park (2004). "Generalizing Discriminant Analysis Using the Generalized Singular Value Decomposition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26, pp. 995-1006.
[15]. J. Han and M. Kamber (2001). "Data Mining , Concepts and Techniques". Morgan Kaufmann Publishers.
[16]. K.V. Mardia, J.T. Kent, and J.M. Bibby (1995). "Multivariate Analysis. Probability and Mathematical Statistics". Academic Press.
[17]. J.H Friedrnan.,and Tibshiirani R. (2001). "Elements of Statistical Learning: Prediction". Inference and Data Mining. Springer.
[18]. A. Boulesteeix (2004). "PLS Dimension reduction for classification with microarray data". Statistical Applications in Genetics and Molecular Biology, Volume 3, issue 1, Article 33, 2004, pp. 1-30.
[19]. D.J. Hand (1981). "Discrimination and Classification". New York: .John Wiley.
[20]. J.R. Quinlan (1986). "Induction of decision trees". Machine Learning, Volume1, pp. 81-106,
[21]. J.R. Quinlan (1993)."Programs for Machine Learning". Morgan Kaufman.
[22]. T.F. Cox and M.A.A. Cox. (2001). "Multidimensional Scaling". Chapman and Hall, second edition.
[23]. H. Hoppe (1999). "New quadric metric for simplifiying meshes with appearance attributes". In: Proceedings IEEE Visualisation '99, IEEE Computer Society Press,
[24]. A. Hyvärinen (1999). "Survey on independent component analysis". Neural Computing Surveys, 2.94 128, 1999.
[25]. M., P., K. Levoy, B. Curless, S. Rusinkiewicz, D. Koller, L. Pereira, M. Ginzton, S. Anderson, J. Davis, J. Ginsberg, J. Shade, and D. Fulk (2000). "The Digital Michelangelo Project. 3D scanning of large statues". In: Proceedings of ACM SIGGRAPH 2000, Computer Graphics Proceedings, Annual Conference Series, ACM, pp. 131–144.
[26]. T.W. Lee. (2001). "Independent Component Analysis: Theory and Applications". Kluwer Academic Publishers.
[27]. M. Belkin and P. Niyogi (2002). "Using Manifold Structure for Partially Labelled Classification". Proceedings of the Advances in Neural Information Processing Conference. pp. 929-936.
[28]. A. Anthoniadis, S. Lambert-lacroix and F. Leblanc (2003). "Effective dimension reduction methods for tumor classification using gene expression data", Bioinformatics, Volume 19, no.5, 2003, pp. 563-570.
[29]. Li Lexin and Li Hongzhe (2004). "Dimension reduction method for microarrays with application to censored survival data". Bioinformatics, Volume 20, no.18, 2004, pp. 3406-3412.
[30]. P.H. Garthwaite (1994). "An interpretation of partial least squares". Journal of American Statistical Association, Volume 89, no. 425, pp. 122-127.
[31]. E. Kokiopoulou, J. Chen and Y Saad (2010)."Trace optimization and eigenproblems in dimension reduction methods". Numerical Linear Algebra with Application. John Wiley & Sons Ltd.
[32]. Liu Han and K. Rafal (2011). "Dimension Reduction of Microarray Data with Penalized Independent Component Analysis". White paper, from Computer Science Deapartment, University of Toronto, pp1-8.
[33]. T. Zhou , D. Tao and X. Wu (2011). "Manifold Elastic Net: A Unified framework for Sparse Dimension Reduction:. Data Mining and Knowledge Discovery Journal. Vol. 22. No. 3. Pp 340-371.

Olufemi A. Folorunso received the B.Sc. and M.Sc. degrees in Mathematics and C omputer Science from the Obafemi Awolowo University, Ile-Ife, and the University of Lagos, Nigeria in 1992 and 1997 respectively. He is a Senior Lecturer at the Yaba College of Technology, Lagos, Nigeria and just completed his Ph.D in Computer Science at the Universiti Teknologi Malaysia. His research interests include algorithms development, optimisations, signal processing, augmented reality and scientific visualization. He has published several articles in both local, international journals and leading conferences. He is a member of the Nigerian Computer society, the Computer Professional Registration Council of Nigeria and a member of vizNET, United Kingdom.

Mohd Shahrizal Sunar received the BSc degree in Computer Science majoring in Computer Graphics (1999) from Universiti Teknologi Malaysia and MSc in Computer Graphics and Virtual Environment (2001) from The University of Hull, UK. In 2008, he obtained his PhD from National University of Malaysia. His major field of study is real-time and interactive computer graphics and virtual reality. He is the head of computer graphics and multimedia department, Faculty of Computer Science and Information System, Universiti Teknologi Malaysia since 1999. He had published numerous articles in international as well as national journals, conference proceedings and technical papers including article in magazines. Dr. Shahrizal is an active professional member of ACM SIGGRAPH. He is also a member Malaysian Society of Mathematics and Science.

Filtration Of Artifacts In ECG Signal Using Rectangular Window-Based Digital Filters

Mbachu C.B [1], Idigo Victor [2], Ifeagwu Emmanuel [3],Nsionu I.I[4]

[1]Department of Electrical and Electronic Engineering, Anambra State University, Uli.
Anambra State(234),Nigeria

[2]Department of Electronic and Computer Engineering, Nnamdi Azikiwe University, Awka.
Anambra State(234),Nigeria

[3]Department of Electronic and Computer Engineering, Nnamdi Azikiwe University, Awka.
Anambra State(234),Nigeria

Abstract
Increased cases of cardiac arrests resulting from coronary heart disease (CHD) underscores the need for accurate equipment for monitoring and diagnosis of the health conditions of hearts of human beings so that proper medical treatment and advice can be given to people who suffer from heart or heart-related diseases. An electrocardiograph is an instrument or equipment for checking heath condition of hearts by measurement of the quality of electrocardiographic (ECG) signal. The ECG signal is usually corrupt by artifacts and these must be removed before the real shape of the signal can be determined. Filters can be used to realize this. This paper therefore presents designs of digital FIR low pass, high pass and notch filters necessary for the removal of the artifacts. Each filter is designed with a rectangular window. Noisy ECG signal is passed through each filter and results are obtained and presented.

Keywords: rectangular window, corrupting noises. Matlab and Simulink, elecroardiogram.

Introduction
The most vital informative signals used in the diagnosis of patients are the ECG signal, which is generated from the electrical activity of the heart; electromyographic (EMG) signal, which is generated from the electrical activity of the muscles; and the electroencephalographic (EEG) signal, which is generated from the electrical activity of the brain. ECG signals in practice and their natural forms present very small amplitudes of 1mV and frequency components below 100HZ. Due to these characteristics, recording ECG signal tends to be very sensitive to various interferences such as 50/60HZ power line, baseline wander, electromyogram, and electroencephagram. Baseline wander is signal generated due to respiration and the freqency is below 1Hz.

Different researchers have worked on the filtration of ECG signal. Abdel-Rahman-Qawasmi and Khaled Daqrouq suggested using discrete wavelet transform (DWT) [1] in filtering high and low frequencies in ECG signal. Mikhled Alfaouri and Khaled Daqrouq also suggested using wavelet transform thresholding (WTT) to process non-stationary signals such as ECG signals [2]. This is possible by applying the multi-resolution decomposing into subsignals. FC Chang et al presented the use of adaptive filters in removing power line interference in ECG [3]. In [4] Mahesh S. Chavan et al worked on FIR digital fillers for the removal

of baseline wander, powerline interference and encephalogram in ECG signal. In [5] Mateo et al made use of the madeline structure algorithm to remove baseline wander in ECG. This structure is based on a grown artificial neutral network (ANN) allowing for optimization of both the hidden layer number of nodes and the coefficient matrixes, and the matrixes are optimized following the Widrow-Hoff delta algorithm. Mahesh S. Chavan, RA Agarwala and M.D Uplane suggested that Kaiser window can be used to design and implement digital FIR filters for low frequency, high frequency and powerline interferences removal in ECG signal [6]. Mahrokh G. Shayesteh and Mahdi Mottagi-Kashtiban in [7] developed a new window, which is based on optimizing the coefficients of Hamming window using extended Kalman filter, and for designing FIR filters. This new window can be applied to the design of FIR filter for ECG signal denoising. In [8] Mahesh S. Chavan et al provided the use of rectangular window to design a 50- order FIR filters comprising low pass, high pass and
notch filters for reduction of high frequency, low frequency and power line interferences in ECG. In the work the authors applied the filters on the ECG signal in a real time manner using 711B add-on card. J.A Van Alste' and T.S Schilder worked on how the number of taps of FIR filter can be reduced without affecting its efficiency in the removal of baselne wander and powerline interference from ECG [9]. Mahesh et al provided the design and application of digital FIR equiripple filter in reduction of powerline interference in ECG. Fig. 1 is a normal standard ECG signal waveform.

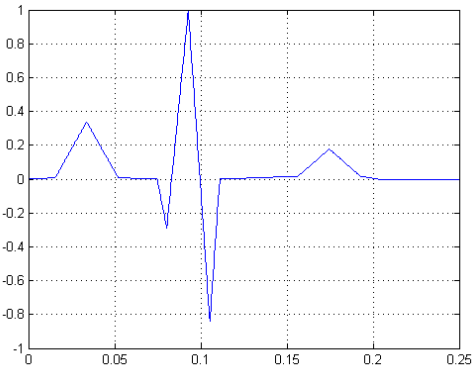


Fig. 1: Normal ECG waveform

2 Design of low pass filter

A rectangular window is used for the design. Fig. 2a shows the time domain amplitude response of a rectangular window function while Fig. 2b depicts the response in frequency domain [4, 11]. Low pass filter is used to remove the high frequency signals constituting noise in ECG. The cut- off frequency is 100Hz and sampling frequency is 1000Hz, while the order of the filter is 100. FDAtool of Matlab is used to carry out the design.

The impulse response, magnitude and phase responses of the filter are shown in fig. 3, fig. 4 and fig. 5 respectively

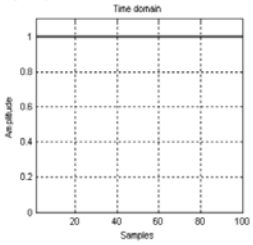


Fig. 2a: Time domain response of a rectangular window

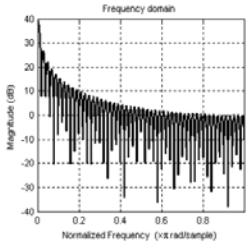


Fig. 2b: Frequency domain response of a rectangular window

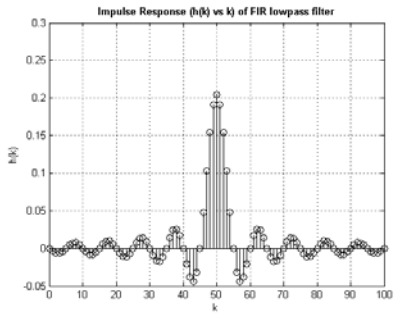


Fig. 3: Impulse response of the low pass filter.

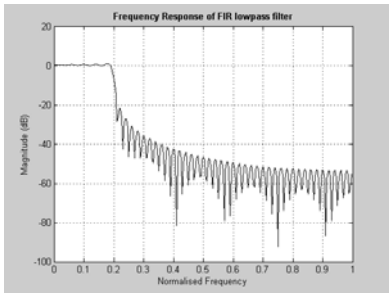


Fig. 4: Magnitude response of the low pass filter

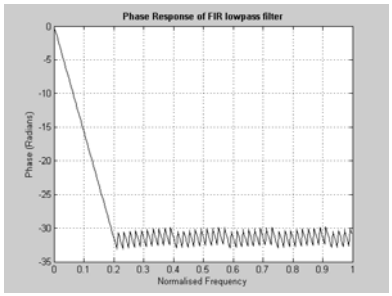


Fig. 5: Phase response of the low pass filter

3. Design of high pass filter.

The high pass filter is used for the removal of low frequency signals that constitute noise in ECG. The cut off frequency is 0.5Hz. The sampling frequency is 1000Hz and the order of the filter is 100. Rectangular window is applied as the weighting window. Fig. 6 depicts the impulse response of the filter while fig 7 and fig 8 provide the magnitude and phase response respectively.

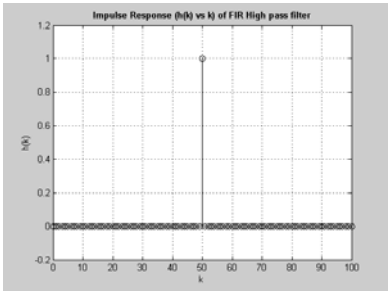


Fig. 6: Impulse response of the high pass filter.

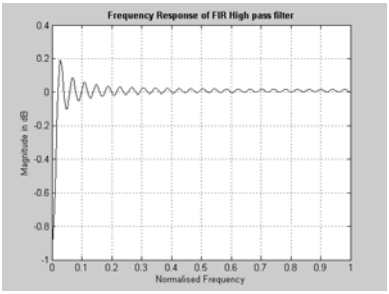


Fig. 7: Magnitude response of the high pass filter

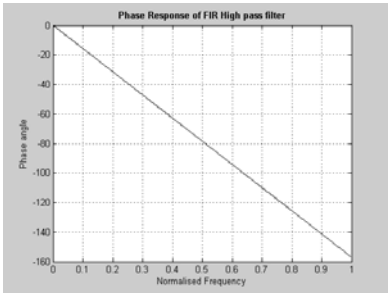


Fig. 8: Phase response of the high pass filter

4 Design of Notch Filter

The ECG signal is applied to the notch filter to remove the powerline interference in ECG. The powerline frequency here is 50Hz and the sampling frequency is 1000Hz. The order of the filter is 100 and rectangular window is the weighting window. The impulse response of the filter is shown in fig. 9 and the magnitude response, shown in Fig. 10. Fig. 11 shows the phase response of the filter

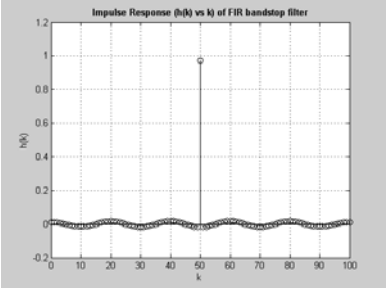


Fig. 9: Impulse response of the notch filter.

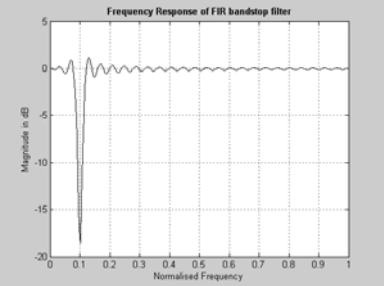


Fig. 10: Magnitude response of the notch filter

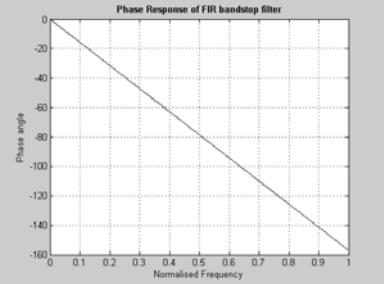


Fig. 11: Phase response of the notch filter

5 Results

The results of the implementation are divided into four groups: the result of the low pass filter, high pass filter, notch filter and the cascade of the three filters.

5.1 Results of the Implementation of the Low Pass Filter.

A raw noisy ECG signals contaminated with high frequency, low frequency and 50Hz powerline interference is shown in fig12. The frequency response of the raw ECG is shown in fig. 13. From fig. 13 the average power of ECG signal above 100Hz is (-52dB).

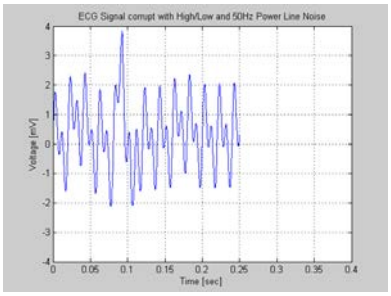


Fig. 12: ECG signal before application of low pass filter

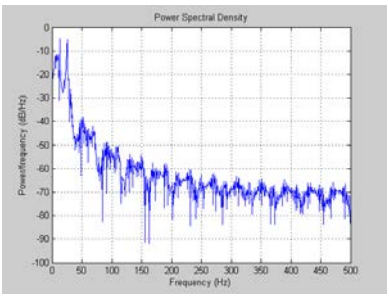


Fig. 13: Frequency response of ECG signal before application of low pass filter

Fig 14 shows the ECG signal after application of the low pass filter while fig 15 depicts the frequency response. From fig 15 it can be confirmed that the power of the signal above 100 Hz is reduced to (-60dB) which implies that the filter removes the high frequency noise from the raw ECG signal.

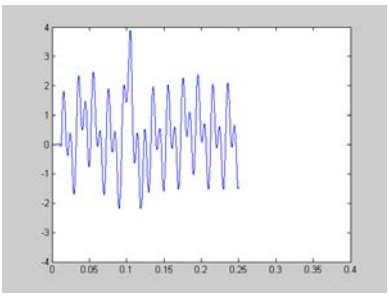


Fig. 14: ECG signal after application of low pass filter

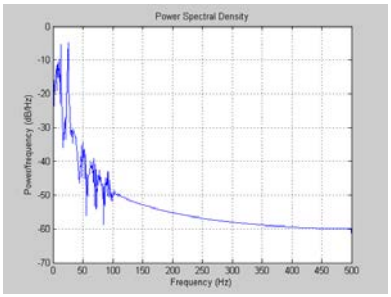


Fig. 15: Frequency response of ECG signal after application of low pass filter

5.2 Results of the implementation of the high pass filter.

From fig. 13, the average power of the raw ECG signal below 0.5Hz is approximately (-15.12dB). Fig 16 presents the ECG signal after application of the high pass filter, while fig 17 represents the frequency response. From fig 17 it can be seen that when the filter is used the power of the signals below 0.5Hz reduces to (-20.03dB). The power reduction implies that the high pass filter filters out the low frequency signal from the raw ECG signal.

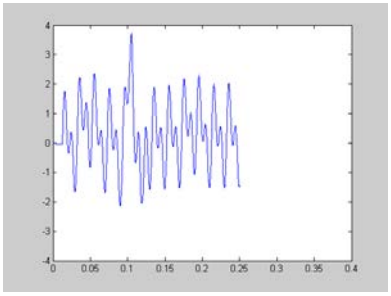


Fig. 16: ECG signal after application of high pass filter

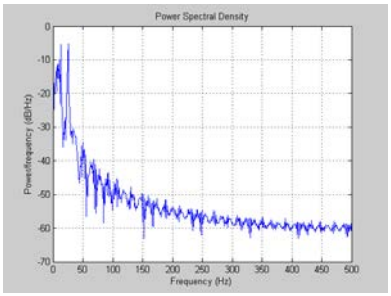


Fig. 17: Frequency response of ECG signal after application of high pass filter

5.3 Results of the Implementation of the Notch Filter.

From fig 13, the power of the raw ECG signal before filtration at 50Hz is (-38.25dB). Fig 18 shows the raw ECG signal after passing through the notch filter and fig 19 depicts the frequency response. From fig19 it can be seen that the power of the ECG signals after filtration with the notch filter drops to (-43dB), which is a confirmation that notch filter reduces power line interference.

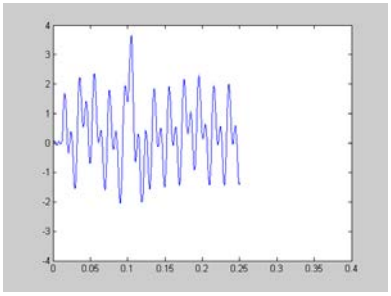


Fig. 18: ECG signal after application of notch filter

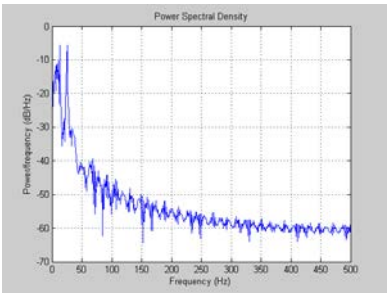


Fig. 19: Frequency response of ECG signal after application of notch filter

5.4 Result of Application of the Low pass, High Pass and Notch filters in Cascade

When the raw ECG signal of fig. 12 is passed through the three filters in cascade the signal appears as shown in fig.20. Comparing the result of fig.20, which is the result of the cascade filtering, with those of fig14, fig16 and fig18, which are outputs of the separate filtering, shows that fig. 20 represents an ECG signal filtered of its noises. The visible distortion appearing on it is because of Gibbs phenomenon associated with rectangular window. The summary of the results of the filtration is shown in table 1.

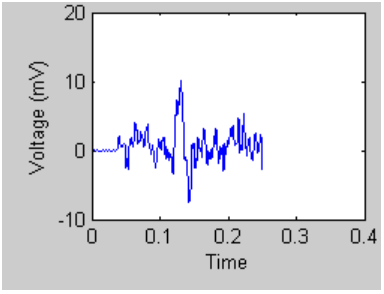


Fig. 20: ECG signal after filtration with the three filters in cascade.

Table 1: Filtration of ECG signal using Digital Filter Designed with Rectangular Window.

Type	Signal power before filtration in dB	Signal power after filtration in dB
Low pass filter, above 100 Hz	-52	-60
High pass filter, below 0.5 Hz	-15.12	-20
Notch filter, at 50 Hz	-3825	-43

Conclusion

The design of the filter shows that there are ripples in each of the filters but they have stable responses. The results show that each filter is able to remove the unwanted signals specifically designed for it to filter out. That is, the low pass filter is able to remove high frequency signals, the high pass filter is able to remove low frequency signals and the notch filter is also able to reduce power line interference. The distortions appearing in the cascade filtering output signal is because of the ripples associated with designs using rectangular windows. These distortions can be eliminated or reduced drastically by using other windows such as Kaiser Window, Hanning window or Hamming window instead.

References

1. Abdel-Rahaman AL- Qawasmi, and Khaled Daqrouq, “ECG signal enhancement using wavelet transform’’ WSEAS transactions on biology and biomedicine, issue 2, volume 7, pages 62- 72. April, 2010.

2. Mikhled Alfaouri and Khaled Daqrouq, “ECG signal denoising by wavelength transform thresholding”. American Journal of applied sciences, Vol. 5, issue 3 pages 276 – 281. 2008.

3. F.C Chang, CK Chang, K. Y. Chi and YD Lin, "Evaluation measures for adaptive PLI filters in ECG signal processing". Computers in cardiology, volume 34, pages 529-532. 2007.

4. Mahesh S. Chavan, R. A. Agarwala and M. D. Uplane, "Interference reduction in ECG using digital FIR filters based on rectangular window". WSEAS transactions on signal processing, issue 5 Vol. 4, pages 340 – 349. May, 2008.

5. J. Mateo, C. Sanchez, C. Vaya, R. cervigon and J. J. Rieta, "A new adaptive approach to remove baseline wander from ECG recordings using Madeline structure". Computers in Cardiology, Vol. 34, pages 533 – 536. 2007.

6. Mahesh S. Chavan, R. A. Agarwala and M. D. Uplane, "Use of Kaiser window for ECG processing". Proceeding of the 5th WSEAS international conference on signal processing, robotics and automation, Madrid, Spain, pages 285 - 289. February 15 – 17, 2006.

7. Mahrokh G. Shayesteh and Mahdi Mottaghi-Kashtiban; "FIR filter design using a new window function". 16th IEEE international conference on digital signal processing, pages 1- 6. 5 – 6 July, 2009.

8. Mahesh S. Chavan, R. A. Agarwala and M. D. Uplane, "Rectangular window for interference reduction in ECG". Proceeding of the 7th WSEAS transactions on signal processing (SIP08), Istanbul, Turkey, pages 110 – 114. May, 27–30. 2008.

9. J. A Van Alste, and T.S Schilder, "Removal of baseline wander and power-line interference from the ECG by an efficient FIR Filter with a reduced number of taps". IEEE transactions on biomedical engineering, Vol. BME-32, No 12, pages 1052- 1060. December, 1985.

10. Mahesh S. Chavan, R. A. Agarwala and M. D. Uplane, "FIR equiripple filter for reduction of powerline interference in ECG signal". Proceedings of the 7th international conference on signal processing, robotics and automation (ISPRA '08), University of Cambridge, UK, pages 147 – 150. February 20 – 22, 2008.

11. Sarkar N. (2003): Elements of digital signal processing. Khanna Publishers, Delhi, India

Study of Performance of the combined MIM0 MMSE VBLAST-OFDM for Wi-Fi (802.11n)

[1]S.FEROUANI, [2]G.ABDELLAOUI, [3]F.DEBBAT, [4]FT.BENDIMERAD
[1,2,3,4] Department of Electrical Engineering and Electronics, Tlemcen University,Algeria

Abstract

Wireless technologies such as WiFi and Bluetooth have transformed the world of networks, and this technical revolution is still in its infancy. Wi-Fi (802.11b and g) now offers a limited range. It is also very susceptible to interference originating DECT phones and other wireless units. Finally, the Wi-Fi in its current version is much slower in terms of flows, the good old Ethernet. All this should change over to succeed the 802.11g. The 802.11n standard expected to provide data rates higher than an Ethernet connection and double the range. The use of combination VBLAST-MMSE-OFDM with MIMO (multiple input-multiple outputs) can leverage the advantage of both methods: the robustness of the link on frequency selective channels for OFDM and robustness of uncorrelated channels for MIMO space-coded. This article discusses the impact of technology (MIMO) wireless transmission of type Wi-Fi (802.11n), in a context-VBLAST MMSE-OFDM.

We will carry out a study of performance in terms of BER for a Rayleigh channel that characterizes most communication systems without son.

Key words: *Multi antenna systems, spatial diversity, MIMO- OFDM channel capacity, WIFI.*

1. INTRODUCTION

In recent decades, the applications for wireless local area networks called WLAN(Wireless Local Area Networks) have become increasingly numerous, hence the need for greater transmission rates. The diversity techniques and especially the spatial diversity techniques [1], [2] are very effective in reducing the impact of these problems on system performance. The fact received signal is affected by multiple channels, we assume uncorrelated, induces a gain of diversity, properly operated can improve system performance in terms of quality of service (QoS: Quality of service) and transmission rate.

The implementation of multiple antennas for transmission and reception is made possible by recent advances in wireless technologies. To eliminate the selectivity of the channel and combat fading and interference, and advanced modulation interference, technique OFDM is applied.

The combination of MIMO MMSE-VBLAST -OFDM is an optimal solution to increase the transmission rate and improve the performance of signals received. The aim of this paper is to study the performance of MIMO systems by combining the OFDM with MMSE-VBLAST coding in a Rayleigh channel that characterizes most wireless transmissions.

In Section 2, we present the MIMO technology. Section 3 concerns the channel model considered. In Section 4, we describe the MIMO-OFDM model adopted for this work. Section 5concern the OFDM transmission technique. Section 6, we discuss the simulation results of our study performance. We end our paper with a conclusion.

2. MIMO technology

In multi antenna systems, the capacity increases linearly with the number of transmitting antenna, exceeding the theoretical limit of Shannon. These systems have an advantage because they more resistant to fading and interference. MIMO systems are considered as technology that can solve the problems of congestion and capacity limitations of wireless broadband. Because of these properties, MIMO is an important part of modern standards of wireless communication such as IEEE 802.11n (Wi-Fi), 4G, 3GPP Long Term Evolution, WiMAX and HSPA + [3] [4]. Figure 1 shows the overall architecture of a MIMO system.

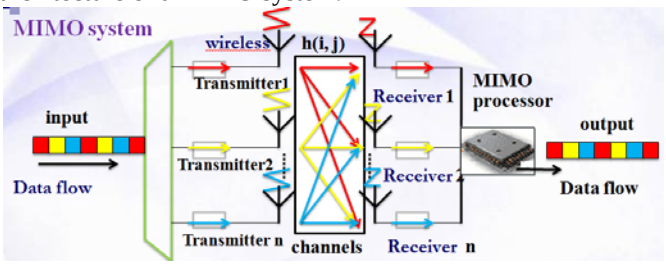


Fig.1: MIMO device

Channel capacity is given by:

$$C = \log_2 \det\left[I_N + \frac{\rho}{M} H.H^* \right] \quad (1)$$

H: the complex gain of the channel.
M: MIMO block size.
ρ: average SNR.
C: channel capacity in bps / Hz.

3. CHANNEL MODEL CONSIDERED

There are a multitude of models of the propagation channel. In this study we focus on the following types of channels [3]:

3.1 Canal with additive white Gaussian Noise:

The channel model with additive white Gaussian noise (AWGN) is the simplest of models. The received signal r (t) is the result of the signal s (t) with the addition of noise n (t) modeled by a function of Gaussian probability density defined as:

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-u)^2/2\sigma^2} \quad (2)$$

With x: random variable, μ: mean, σ: variance.

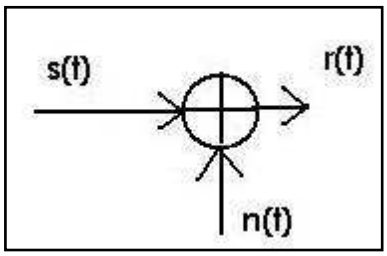


Fig.2: Model of a channel with white Gaussian noise

This channel is described by the equation:

$$r(t) = s(t) + n(t) \quad (3)$$

3.2 Channel with fading and additive white Gaussian noise (Rayleigh channel)

In this type of channel, which affect only the faint signals are taken into account. It is described by the equation:

$$r(t) = h(t, t_d) * s(t) + n(t) \quad (4)$$

The Rayleigh channel models both fading and AWGN is to say it includes the two channels described above. This channel can also theoretical model a frequency selective channel (and possibly time) for which we conducted a modulation / demodulation OFDM.

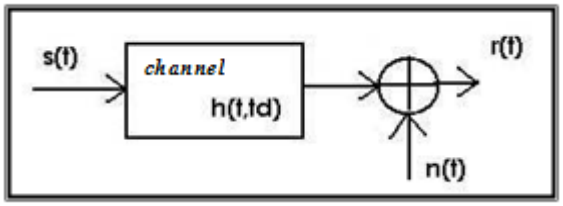


Fig.3: Model of a channel with fading and additive white Gaussian noise

Rayleigh fading:

If α i (t) have a Rayleigh probability density of type:

$$f_Z(Z) = (Z/\sigma^2)e^{-Z^2/2\sigma^2}U(Z) \quad (5)$$

With: α i (t): attenuation factor over time.
U (Z) is a random function.
σ^2 : Variance (α in this simulation).

4. MODELING OF A MIMO-OFDM:

We consider a MIMO system using OFDM modulation, system using OFDM modulation, where the transmitter and receiver are provided respectively with Nt antennas and Nr.
The antennas are disposed of the most commonly used, known in English Uniform Linear Array (ULA) [4], that is to say they are aligned and evenly spaced. The relative distance between two adjacent antennas is given by:

$$\Delta = 1 / 2\lambda \quad (6)$$

where λ is the wavelength.

Fig.4 describes the diagram of a MIMO / OFDM; in transmission we have the following stages:

- Serial parallel conversion of size P to obtain blocks of P symbols.
- Inverse Fourier transform of size P.
- Inserting a guard interval of size D at the beginning of the end block where the block is copied.

In reception, the dual operations are performed:

- Conversion parallel series.
- Conversion series parallel of size P to obtain blocks of $P + \Delta$ symbols.
- Remove the guard interval corresponding to the first samples of the block.
- Direct Fourier Transform of size P.
- Conversion parallel series.

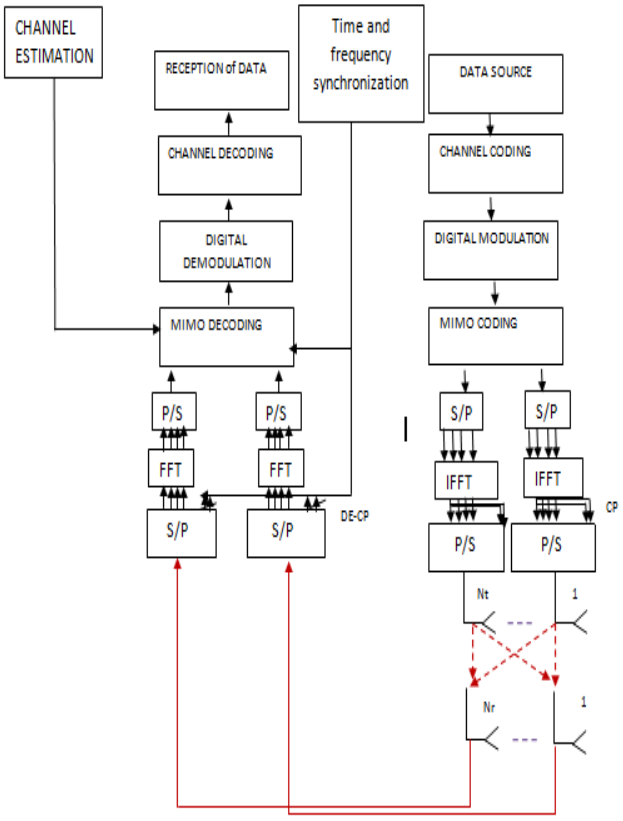


Fig.4: Model of MIMO-OFDM (Nt*Nr) system

5. OFDM modulation and demodulation

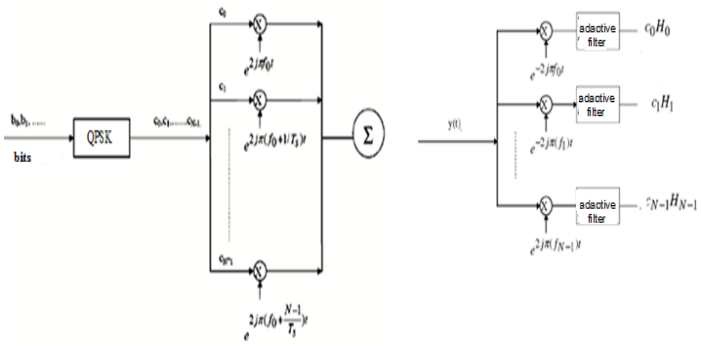


Fig.5: Diagram of OFDM modulation and demodulation

To distribute the data to be transmitted on the N carriers, symbols are grouped in bundles of N.

The complex numbers c_k are defined **[5]** from the bits by a constellation often QAM and PSK:

$$c_K e^{j2\pi f_K t} \quad (7)$$

The total signal s(t) is the set of N symbols reassembled in an OFDM symbol:

$$S(t) = \sum_{K=0}^{N-1} c_K e^{j2\pi f_K t} \quad (8)$$

The received signal is written over symbol duration T_s:

$$y(t) = \sum_{K=0}^{N-1} c_K H_K(t) e^{2j\pi(f_0 + K/Ts)} \quad (9)$$

$H_K(t)$ Is the channel transfer function around the frequency and time.

6. VBLAST-OFDM coding

The coding principle VBLAST is to transmit each Nt time [5].

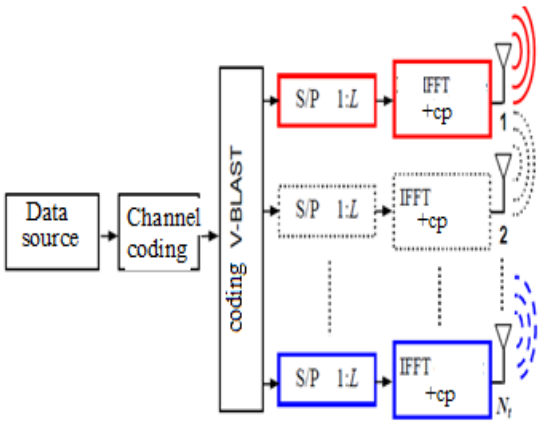


Fig.6: VBLAST-OFDM Transmitter

7. VBLAST-OFDM decoding

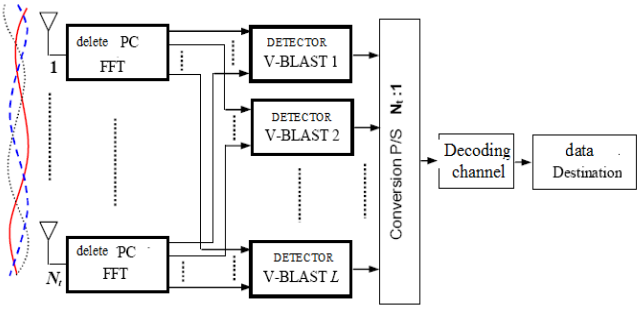


Fig.7: VBLAST-OFDM Receiver

Fig7 shows the block diagram of a receiver V-BLAST-OFDM. Each receiving antenna receives a signal for each of the L sub channels [6]. After the cyclic prefix is removed, each received signal is passed through a bloc FFT operation for demodulation.

The received signal after demodulation at the receiving antennas j for sub-channel l is given by:

$$y_{j,l} = \sum_{i=1}^{Nt} H_{j,i,l} x_{i,l} + n_{j,l}$$

(10)

where $h_{i,j,l}$ is the normal path complex gain of transmitting antenna i to receive antenna j at the frequency l, $x_{i,l}$ is the OFDM symbol transmitted from antenna i at frequency l, and the nj are independent Gaussian noise samples. The outputs of FFT blocks are passed through the L-VBLAST detectors, each with Nr inputs, and Nt outputs. The outputs of the VBLAST detectors are converted to sub-parallel streams into a single serial stream of data. Finally, the data is decoded by the channel decoder.

8. Study of Wi-Fi (802.11n):

Recent developments in telecommunications technology use multiple antennas spatial diversity. This principle already existed on the old generation wireless systems; in this case the addition of antennas to the receiver can, for example to make a selection on a criterion of antenna power to take advantage of transmission with the highest signal to noise ratio (SNR), and thus achieve the modulation with the best rate. Spatial diversity is now fully used in the modulator/ demodulator itself. Unlike the older generation WI-FI, MIMO uses multiple antennas technology to transmit different information on each antenna.

When the 802.11a and g wireless terminals used two antennas [7], the goal was to find and use the antenna that had the best power to communicate with other equipment. The access point then uses one of two antennas for communication. After various improvements had been made to the 802.11 a and g, a new generation of Wi-Fi has been designed taking into account two constraints: keep the same frequency spectrum and not to increase power. The MIMO then appeared. The latter, combined with WI-FI 802.11 a and g, gave birth to802.11n.

Unlike the previous generation terminals, the MIMO uses at least two antennas to send different information on each.

9. Performance of MIMO VBLAST-OFDM for the WI-FI:

9.1 Parameters simulation:

Our simulation is based on the following parameters:

- A band B of 2.4 GHz and 5 GHz;
- Number of subcarriers (N): 64 of which only 52 are used (the external carrier 12 are set to 0to reduce interference between adjacent channels) for the 2.4GHz band.
- Number of subcarriers (N): 128 of which only 108 are used (the external carrier 20 are set to 0 to reduce interference between adjacent channels) in the 5GHz band.
- OFDM symbol duration: 3.2 microseconds (312.5 kHz carrier).
- Modulation: QPSK.
- Cyclic prefix (averaging 800 ns max): 1/4 (Total symbol=4ms)
- Number of drivers: 4 for B=2.4GHz and 8 for B=5GHz.
- Nt and Nr are the Number of transmitting and receiving antennas.

1. Results of simulation:

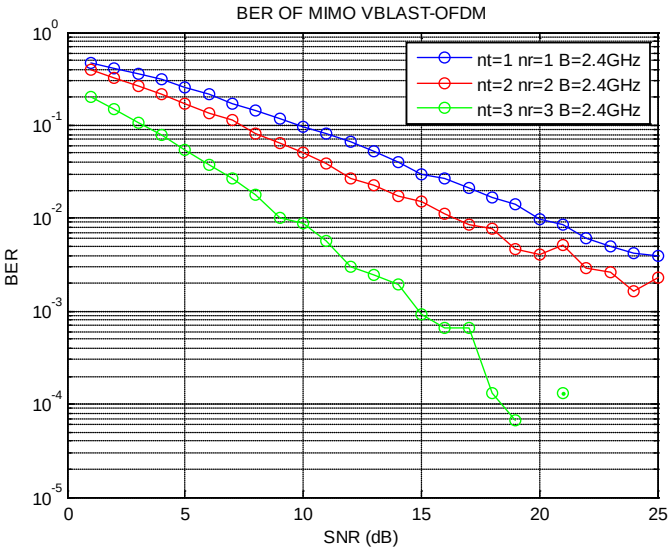


Fig.8: BER of MIMO VBLAST-OFDM for WI-FI (802.11n)
With B=2.4 GHz, N (number of subcarrier OFDM) =52

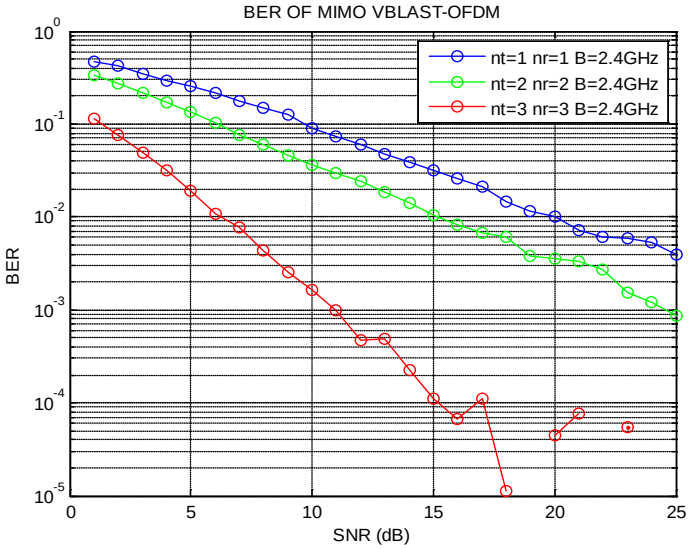


Fig.9: BER of MIMO VBLAST-OFDM for WI-FI (802.11n)
With B=5 GHz, N (number of subcarrier OFDM) =108

We note well the improvement of signal quality in terms of BER by increasing the number of transmitting and receiving antennas, which shows the usefulness of VBLAST MIMO-OFDM for wireless transmissions.

10. Conclusion

In this paper, we propose the MIMO-VBLAST-OFDM for Wi-Fi (802.11n). MIMO is a technique using multiple antennas to maximize throughput within the premises. Until recently, such environments pose serious problems for wireless networks. Waves transmitting the data tend to bounce off metal structures inside the furniture or wall and shot, to interfere with each other, resulting in performance degradation interfere with each other, resulting in performance degradation and a reduction in the scope. Other sources of interference such as cordless phones, microwave ovens, walkie-talkies and other nearby wireless networks also pose problems the standard Wi-Fi faces. It results in a reduction of the scope, or even intermittent loss of connections.

We used the encoding technique VBLAST OFDM to improve performance used the encoding technique VBLAST OFDM to improve performance and ensure the quality of the received signal.

The results of our simulations show the importance of the technology associated with the MIMO VBLAST-OFDM for wireless Wi-Fi (802.11n). The MIMO system show us a glimpse of what we can expect the Wi-Fi in the near future. With sustained higher flow, better coverage and greater resistance to interference, 802.11n will perhaps finally realize the dream of many users: namely, build a real home network entertainment, wireless.

References

[1]S.M. Alamouti, “A simple diversity technique for wireless communications ». IEEE Journal on Selected Areas in Communications, 16(8):1451–1458, October 1998.

[2] Olivier BERDER. “Optimization and allocation strategies of power transmissions multiple antennas” University of Western Brittany, 20 December 2002.

[3]K.MABROUK, “Design and implementation of a MIMO communications system with digital beam forming in reception calibration blind phase demodulator zero receiver and compared conventional two channels I end Q”, national school of Telecommunications 12 December 2008.

[4]B.Rihawi. « Analysis and Mitigation of Power Ratio radio systems communications multi- antennas », University of Rennes I, 20 March 2008.

[5] P GRUYER, S.PAILLARD, “Modeling OFDM modulator and demodulator”, University of Bretagne 12 December 2005.

[6] Gautier, G. Burel, J. Letessier, and O. Berder. “Blind estimation of scrambler offset using encoder redundancy”. In Proceedings of IEEE Asil omar Conference on Signals, Systems and Computers, volume 1, pages 626–630, Pacific Grove (CA), USA, 2002.

[7]Henri Happy, Hassan Benchikh, Gabriel Cognard, “Standard 802.11n”, University of Lille 1 , December 2006

Performance of MIMO VBLAST-OFDM in Ka-Band

[1]S.FEROUANI, [2]G.ABDELLAOUI, [3]F.DEBBAT, [4]FT.BENDIMERAD

[1,2,3,4]Departments of Electrical Engineering and Electronics, Tlemcen University, Algeria

Abstract

Technological advances have allowed the use of Ka-band for transmission in 30 GHz uplink and 20 GHz at downlink. The use of Ka-band has advantages over lower frequency bands; it allows a wider bandwidth and therefore a greater flow to pass. In addition, multibeams technology allows a wide reuse frequency, thereby reducing significantly the cost of the spectrum. In this Ka-band (20-30 GHz), quality of service can be degraded by the effects of effects of propagation through the atmosphere. We have proposed in this paper the MIMO OFDM-VBLAST technique for Ka-band application to improve performance. The simulation is based on the study of effect of the channel attenuation (rain and gas) in the Ka-band signal. The simulation results show the importance of using our proposal, where performance is determined in terms of bit error rate BER.

Keywords: *MIMO Systems, Ka band, OFDM Modulation, Satellite Transmission.*

1. INTRODUCTION

In the context of future satellite communications systems, deployment of Ka-band presents a necessity [1], particularly because of the saturation bands L, C and Ku. This operation offers the advantage of wider channels that support a greater number of users. It also reduces the size of the user terminal as well as those of the antenna. Nevertheless, the exploitation of the Ka-band is accompanied by certain disadvantages mainly associated with more severe propagation conditions. In comparison with the Ku band, the Ka-band signal received may be subject to strong attenuations result of weather disturbances. For example, Ka-band attenuation can exceed ten dB following heavy rainfall to overcome these problems; a MIMO diversity technique is used [2]. The transmission channel is the central problem that must be addressed in different transmission solutions proposed. When sending a symbol through the channel, it will be received in the form of delayed and attenuated versions super imposed, which can generate interference between symbols transmitted, for this we must use techniques such as transmission TDMA, FDMA, CDAMA, OFDM and MC-CDMA.

To maintain high flows and cancel the interference between symbols, we studied the association of the technology MIMO with OFDM Modulation, which consists of a parallel transmission of data with sufficiently long periods, and we also introduce the coding VBLAST to ensure the quality of transmission.

This technique is widely used in wireless networks such as WIFI (802.11n), WIMAX, and it can also be used for satellite transmission (DVB-S2, DVB-SH...).

Our study is based on first point on the architecture of MIMO-OFDM, in second point the encoding and decoding VBLAST applied to the signals, and in thirst point the characteristics of the Ka-band satellite transmission. Finally, simulation results of our study which clearly show the improvement provided by the MIMO-OFDM.

2. MIMO-OFDM SYSTEM

Increasing of the size modulations or the frequency band used are the only solutions to increase the flow of data in a single antennas system. In multi antenna MIMO, the capacity Increases linearly with the number of transmitting antenna, exceeding the theoretic limit of Shannon [3]. The systems have an additional advantage because they resist fading and interference.

MIMO-systems are considered as a technology capable of solving the problems of congestion and capacity limitations of wireless broadband networks.

2.1 MIMO device

Fig.1 shows the pattern of transmission MIMO (Multiple input Multiple output); it is to send packets to different antennas that emit at the same frequency. Upon receiving the signals are combined and processed by a processor MIMO containing very powerful calculate algorithms.

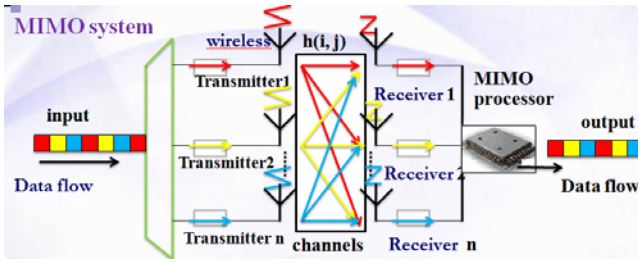


Fig.1: MIMO device

Channel capacity is given by:

$$C = \log_2 \det\left[I_N + \frac{\rho}{M} H.H^* \right] \quad (1)$$

H: the complex gain of the channel.
M: MIMO block size.
ρ: average SNR.
C: channel capacity in bps / Hz.

2.2 Modeling of a MIMO-OFDM:

We consider a MIMO system using OFDM modulation, where the transmitter and receiver are provided respectively with Nt and Nr antennas.
The antennas are disposed of the most commonly used, known in English Uniform Linear Array (ULA) [4], that is to say they are aligned and evenly spaced. The relative distance between two adjacent antennas is given by:

$$\Delta = 1 / 2\,\lambda \qquad (2)$$

Where λ is the wavelength.
Fig 2 describes the diagram of a MIMO / OFDM; in transmission we have the following stages:

- Serial parallel conversion of P size to obtain blocks of P symbols.
- Inverse Fourier transform of P size.
- Inserting a guard interval of D size at the beginning of the end block where the block is copied.

In reception, the dual operations are performed:

- Conversion parallel series.
- Conversion series parallel of P size to obtain blocks of $P + \Delta$ symbols.
- Remove the guard interval corresponding to the first samples of the block.
- Direct Fourier Transform of P size.
- Conversion parallel series.

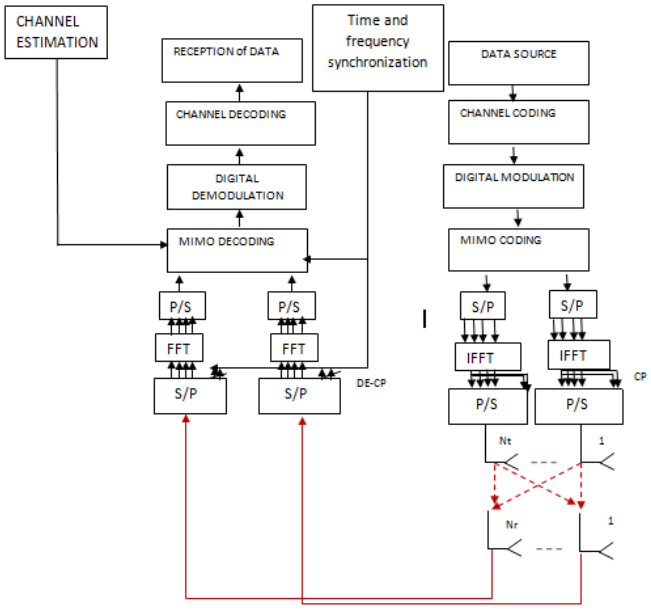


Fig.2: Model of MIMO-OFDM (Nt*Nr) system

3. VBLAST-OFDM Coding

The coding principle VBLAST is to transmit each Nt time [5].

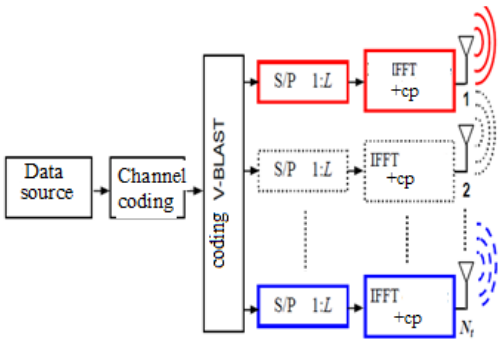


Fig.3: VBLAST-OFDM Transmitter

4. VBLAST-OFDM Decoding

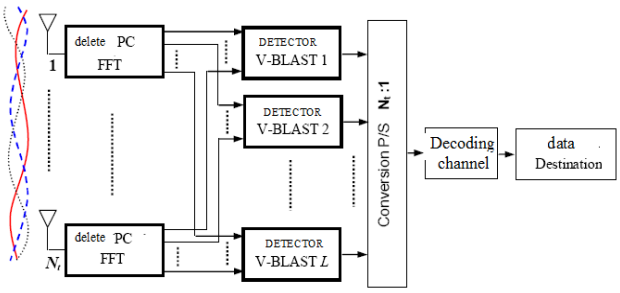


Fig.4: VBLAST-OFDM Receiver

Fig.4 shows the block diagram of a receiver V-BLAST-OFDM. Each receiving antenna receives a signal for each of the L sub channels [5]. After the cyclic prefix is removed, each received signal is passed through a bloc FFT operation for demodulation.
The received signal after demodulation at the receiving antenna j for sub-channel l is given by:

$$y_{j,l} = \sum_{i=1}^{Nt} H_{j,i,l} x_{i,l} + n_{j,l} \qquad (3)$$

where $h_{i,j,l}$ is the normal path complex gain of transmitting antenna i to receive antenna j at the frequency l, $x_{i,l}$ is the OFDM symbol transmitted from antenna i at frequency l, and the nj are independent Gaussian noise samples. The outputs of FFT blocks are passed through the L-VBLAST detectors, each with Nr inputs, and Nt outputs. The outputs of the VBLAST detectors are converted to sub-parallel streams into a single serial stream of data. Finally, the data is decoded by the channel decoder.

5. OFDM Modulation and demodulation

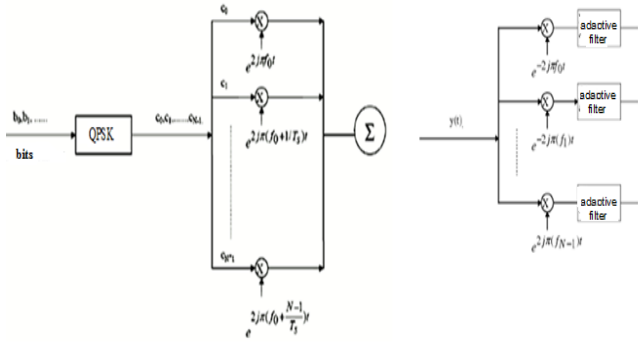


Fig.5: Diagram of OFDM modulation and demodulation

To distribute the data to be transmitted on the N carriers, symbols are grouped in bundles of N.
The complex numbers c_k are defined **[6]** from the bits by a constellation often QAM and PSK:

$$c_K e^{j2\pi f_K t} \quad (4)$$

The total signal s (t) is the set of N symbols reassembled in an OFDM symbol:

$$S(t) = \sum_{K=0}^{N-1} c_K e^{j2\pi f_K t} \quad (5)$$

The received signal is written over symbol duration T_s :

$$y(t) = \sum_{K=0}^{N-1} c_K H_K(t) e^{2j\pi(f_0 + K/Ts)} \quad (6)$$

$H_K(t)$ Is the channel transfer function around the frequency and time t.

6. KA-Band propagation

With the use of Ka band frequency, atmospheric phenomena frequency, related to wave propagation, become very important and severely limit system performance. These mitigations, contributing heavily to the degradation of satellite signals in the Ka band, are caused by atmospheric gases, clouds and precipitation [7]. The rain is the dominant phenomenon of attenuation in the frequency bands above 10GHz (ex Ku, Ka) causing significant decreases in signal quality.

6.1 KA-Band attenuation

The rain is the phenomena that most affects the quality of Ka-band signals which we studied its decay with the gas in this article:

6.1.1 Attenuation by rain

- Rain causes absorption and scattering that give rise to a weakening depends on the intensity of the precipitation and the frequency.
- The relationship between the linear attenuation γ_R (dB/Km) and intensity of rainfall R(mm/h) given by ITU-RP :

$$\gamma_R = a.R^b \quad (7)$$

Table I: a and b for different frequencies

F(GHz)	*a*	*b*
1	0.000038	0.912
10	0.0101	1.276
20	0.0751	1.099
30	0.187	1.021
40	0.350	0.939

Fig.6 shows an example of a time series of attenuation by rain in the Ka band (measured for a fixed point on the ground).

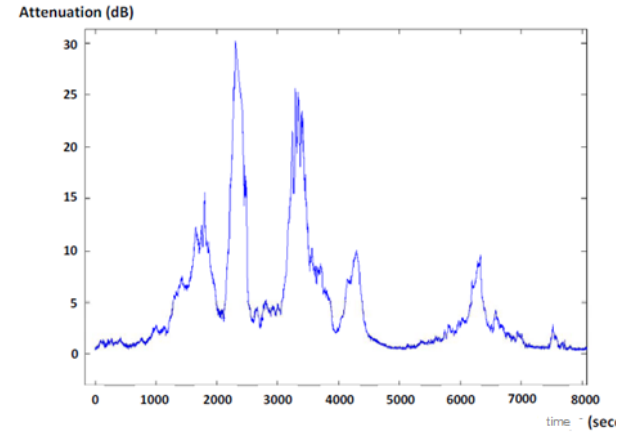


Fig.6: Example of time series of rain attenuation in Ka band.

6.1.2 Attenuation due to gas

Oxygen, in particular, is the gas component that most affects the quality of Ka band signal. Its effect is more substantial at low temperature.
For example, in a Europe a climate and at a frequency of 30 GHz the oxygen attenuation average is around 0.2 dB.

8. Simulations:

This section presents the results of simulations of the system proposed in this paper (MIMO-OFDM-BLAST); improvement brought by the use of this combination in the Ka band is shown in the figures in terms of BER.

8.1 Simulation Parameters

The main characteristics of a Ka band satellite channel are the atmospheric attenuations (gas end rain), that we introduce in our simulations.

We chose frequency of 29.7 GHz which is included in the interval [20, 30] GHz, an elevation of 10 °, latitude of 70 degrees, a concentration of water vapor from 20g/m3.

The evaluation of a communication system transmission quality is represented by two variables: the BER (Bit-Error-rate) and SNR (Signal to noise).

- The type of modulation used is QPSK,
- The number of OFDM subcarriers is 512
- The number of pilot OFDM is 32.
- Nt (number of transmit antennas) is equal to Nr (number of receive antennas) is 2 and 4 respectively.

8.2 Performance systems

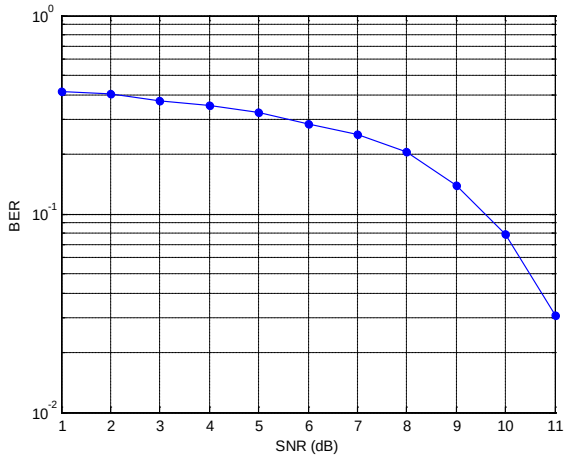


Fig.7: SISO-OFDM in Ka-band

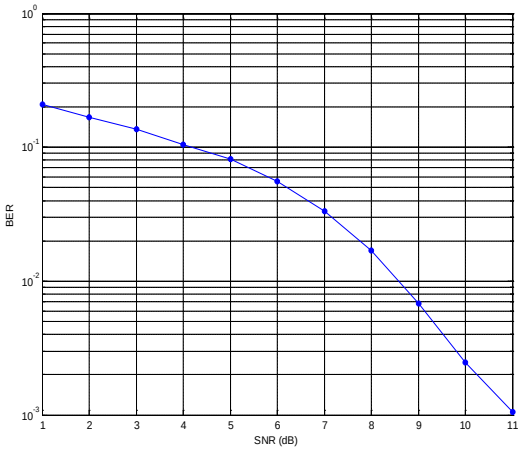


Fig.8: SIMO-OFDM in Ka band
nt=1, nr=2

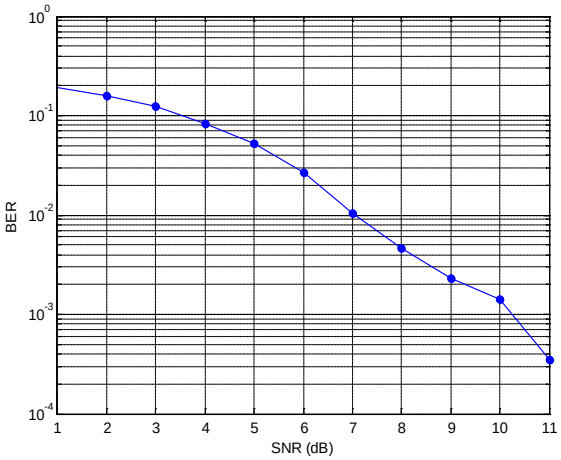


Fig.9: MIMO-OFDM in Ka band
nt=2, nr=2

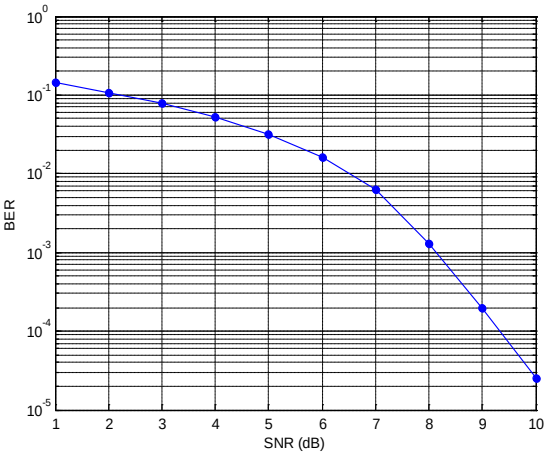


Fig.10: SIMO-OFDM in Ka band
nt=4, nr=4

The performance study of Ka band signal shows that the attenuation of the rain and gases affect the transmission quality, for example in the case of SISO-OFDM, BER reaches a value of approximately $10^{-1.5}$ for an SNR equal to 11dB.

- The transition to two reception antennas (SIMO-OFDM) allows for minimizing BER reaching the value 10^{-3} for an SNR equal to 11dB.
- The use of MIMO-OFDM technology allows a significant improvement in signal quality, Example: for two transmitting and receiving antennas respectively, BER is $10^{-3.5}$ for an SNR equal to 11dB and four transmitting and receiving antennas, it is $10^{-4.7}$.

9. Conclusion

The Ka-band signal undergoes attenuations degrading air quality. To remedy this problem, we have proposed in this paper, MIMO-OFDM-VBLAST systems.

This compensation technique allowed us to see the improvement of signal quality in terms of bit error rate, by increasing the number of transmitting and receiving antennas, and using a number of subcarriers equal to 512 and an encoding type to ensure transmission. The use of turbo codes is another issue for future investigation.

Reference

[1] H.O. Anh Tai, "Application techniques of multicarrier OFDM Systems for future satellite", University of Toulouse, 30 March 2009.

[2] B.Rihawi, "Analysis and mitigation of power ratio radio systems communications multi-antennas", University of Rennes I, 20 March 2008.

[3] S.M.Alamouti, "A simple diversity technique for wireless communications", IEEE Journal on Selected Areas in communications, 16(8):1451–1458, October 1998.

[4] I.Ouachani, "Performance analysis of communication systems wireless operator Micro and Macro Diversity", university of Paris XI Rosary, Discipline: Automatic and Signal Processing, June 28, 2005.

[5] P.W.Wolniansky, G.J. Foschini, G.D. Golden, and R.A. Valenzuela, "V-BLAST : An Architecture for Realizing Very High Data Rates Over the Rich-Scattering Wireless Channel", Bell Laboratories, Lucent Technologies, Crawford Hill Laboratory 791 Holmdel-Keyport RD., Holmdel, NJ07733.

[6] P. GRUYER and S. Paillard, "Modeling OFDM modulator and demodulator", University of Bretagne 12 December 2005.

[7] R. Chaggara, "The continuous phase modulation for the Design of a Waveform Adaptive Application to Future Multimedia Systems Ka-band satellite", National School of Telecommunication.

[8] N. Tao, "Study of Performance and Optimization of a Network Access Satellite Communications», 10 July 2009.

Energy Efficient Adaptive Protocol for Clustered Wireless Sensor Networks

K. Padmanabhan [1], Dr. P. Kamalakkannan [2]

[1]Department of Computer Applications
Muthayammal Engineering College
Tamilnadu, India.

[2]Department of Computer Science
Govt. Arts College (Autonomous), Salem
Tamilnadu, India.

Abstract

Wireless Sensor Networks (WSNs) is a network of an inexpensive low coverage, sensing, and computation nodes. The foremost difference between the WSN and the traditional wireless networks is that sensors are extremely sensitive to energy consumption. Energy saving is the crucial issue in designing the wireless sensor networks. Many researchers have focused only on developing energy efficient protocols for continuous-driven clustered sensor networks. In this paper, we propose a modified algorithm for Low Energy Adaptive Clustering Hierarchy (LEACH) protocol. Our modified protocol called "Energy-Efficient Adaptive Protocol for Clustered Wireless Sensor Networks (EEAP)" is aimed at prolonging the lifetime of the sensor networks by balancing the energy consumption of the nodes. EEAP makes the high residual energy node to become a cluster-head. The elector nodes are used to collect the energy information of the nearest sensor nodes and select the cluster-heads. We compare the performance of our EEAP algorithm with the LEACH protocol using simulations.

Keywords: ***E****nergy efficiency, LEACH, Wireless SensorNetworks.*

1. Introduction

Wireless Sensor Networks (WSNs) consists of spatially distributed autonomous devices using sensors to cooperatively monitor physical or environmental conditions such as temperature, sound, vibration, pressure, motion or pollutants, at different locations. The development of wireless sensor networks was originally motivated by military applications for battlefield surveillance. Therefore, wireless sensor networks are used in many civilian applications, including environmental and habitat monitoring, health-care applications, home automation and traffic control. This network contains a large number of nodes which sense data from an impossibly inaccessible area and send their reports toward a processing center which is called "sink". Since sensor nodes are power constrained devices, frequent and long-distance transmissions should be kept to minimum in order to prolong the network lifetime [1]. Thus direct communication between nodes and the base station are not encouraged. Several communications Protocols have been proposed to realize power-efficient communication in these networks. One efficient approach is to divide the network into several clusters, each electing one node as its cluster head. The cluster-head collects data from sensors in the cluster which will be fused and transmitted to the base station. Thus only some nodes are required to transmit data over a long distance and the rest of the nodes will need to do only short distance transmission. Then more energy is saved and overall network lifetime can be prolonged. Much energy efficient routing protocols are designed based on the clustering structure where cluster-heads are elected periodically [2]. To model energy consumption, three basic states of a node can be identified: sensing, data processing and data communication. Experimental measurements have shown that data transmission is very expensive in terms of energy consumption, while data processing consumes significantly less. Minimizing the number of communications by eliminating or aggregating redundant sensed data saves much amount of energy. In a homogeneous network, cluster head uses more energy than non cluster head nodes [3]. As a result, network performance decreases since the cluster head nodes go down before other nodes do. Thus dynamic, adaptive and energy efficient cluster head selection algorithm is very important issue in clustered WSNs.

Generally, there are three basic data delivery models, i.e., event-driven, query-driven, and continuous delivery models [4]. In continuous delivery model, the sink is interested in the conditions of the environment at all times and every node periodically sends data to the sink. In event-driven delivery model, the sink is only interested in hearing from the network when certain events occur. Query-driven data delivery model is similar to the event- driven model except that the data is pulled by the sink while the data is pushed to the sink in the event driven model. Configuring the network as event-driven is an attractive option for a large class of applications since it typically sends far fewer messages [5]. This is translated into significant energy saving, since message transmissions are much more energy intensive when compared to sensing and (CPU) processing. Also some existing energy-saving solutions take that into consideration

and switch some nodes off, leading the nodes to an inactive state, these are waken up only when interest matches the events "sensed" [6]. Therefore, event driven protocols are used to conserve the energy of the sensor nodes. Most research so far assumed that all nodes collect and send data at the same rate and network's energy consumption is uniform, so that they regulate the run-time of each round. However, in event-driven sensor network applications, events occur randomly and transiently, and accompanied by the burst of large numbers of data, therefore, network energy consumption is uneven. Energy-efficient Event Driven Clustering (EDC) [7] algorithm can decide which nodes will become cluster head nodes according to the maximum remainder energy of nodes.

Fig.1. Wireless Sensor Network Structure

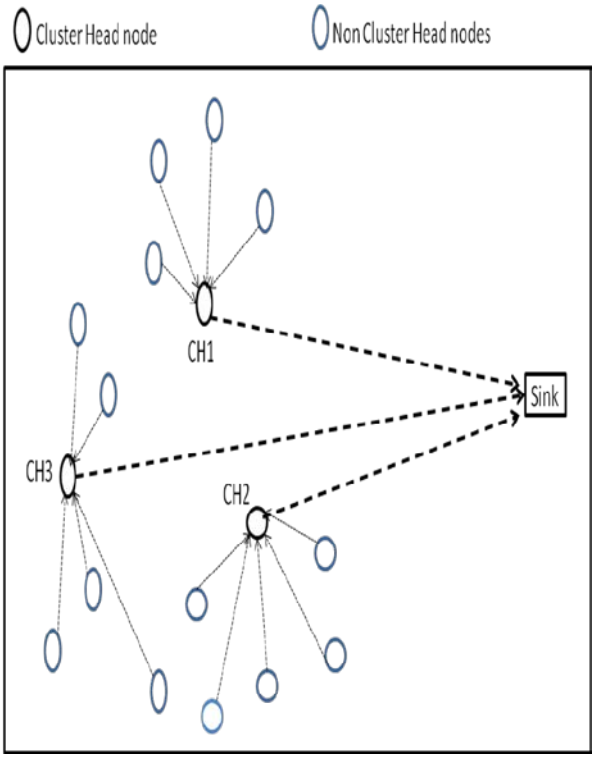


Fig.1. extracted from [8] describes the architecture of wireless sensor networks structure. In this paper, we focus on the energy efficient clustering algorithm for event-driven wireless sensor network. We propose a modified algorithm of LEACH called "Energy efficient Adaptive protocol for clustered Wireless sensor networks". Our proposed protocol facilitates the nodes with more residual energy have more chances to be selected as cluster head. In order to extend the lifetime of the whole sensor network, energy load must be evenly distributed among all sensor nodes so that the energy at a single sensor node or a small set of sensor nodes will not be drained out very soon.

In Section 2 related works for our protocol is described. Section 3 describes the proposed system model. Simulation results are shown in Section 4. Finally, we give concluding remarks in Section 5.

2. Related Works

Many clustering algorithms were developed and used in wireless sensor networks. A detailed survey of energy efficient clustering algorithms for wireless sensor networks is presented in [9]. Low Energy Adaptive Clustering Hierarchy (LEACH) is the first energy efficient routing protocol for hierarchical clustering. It reduces the energy significantly [10]. The LEACH protocol forms clusters in the sensor networks and randomly selects the Cluster-heads for each cluster. The non cluster-head nodes sense the data and transmit to the cluster-heads. The cluster-heads aggregate the received data and then forward the data to the sink. This aggregation process reduces the transmission of duplicate data.

There are two phases in LEACH protocol: i) Setup phase ii) steady-state phase. In the setup phase the clusters are formed and the cluster-heads are selected. In the steady-state phase, the data from non cluster heads are transmitted to the sink. The sensor nodes communicate to the cluster-heads using TDMA schedule. The nodes communicate to the cluster-head only in their allotted slots. It avoids collision. The cluster-heads are selected randomly for every round. The Power Efficient Gathering in Sensor information systems (PEGASIS) is a chain based power efficient protocol based on LEACH. The chain is formed on the basis greedy algorithm [11]. The chain starts from the farthest node to the nearest node to the sink. The node nearest to the sink is selected as a chain leader and aggregated and forwarded the received data to the base station. Each node in the chain selected as chain leader to balance the energy consumption.

Enhanced Low-energy Adaptive Clustering Hierarchy (E-LEACH) proposes a cluster head selection algorithm for sensor networks that have non-uniform starting energy level among the sensors. It also determines that the required number of cluster heads has to scale as the square root of the total number of sensor nodes to minimize the total energy consumption. LEACH-Centralized (LEACH-C) uses a centralized clustering algorithm and same steady-state protocol. During the set-up phase of LEACH-C, each node sends information about current location and energy level to base station (BS)[13]. The BS will determine clusters, CH and non-CHs of each cluster. The BS utilizes its global information of the network to produce better clusters that require less energy for data transmission. The number of CHs in each round of LEACH-C equals a predetermined optimal value.

Multi-hop LEACH (M-LEACH) modifies LEACH allowing sensor nodes to use multi-hop communication within the cluster in order to increase the energy efficiency of the protocol [13]. This work extends the existing solutions by allowing multi-hop inter-cluster communication in sparse

WSNs in which the direct communication between CHs or the sink is not possible due to the distance between them. Thus, the main innovation of the solution proposed here is that the multi-hop approach is followed inside the cluster and outside the cluster. CHs can also perform data fusion to the data receive, allowing a reduction in the total transmitted and forwarded data in the network. Among the hierarchical routing protocols, LEACH is the most popular cluster-based routing protocol. A node becomes a CH for the current rotation round if the number is less than the following threshold:

$$T(n) = \frac{p}{1-p}[r \bmod(1/p)], \quad n \in G = 0, \text{otherwise}$$

Where p is the percentage of nodes that Can become CHs, r is the current round and G is the set of nodes that have not served as cluster head in the past 1/p rounds [8].

As long as optimal energy consumption is concerned, it is not desirable to select a cluster head node randomly and construct clusters. However, repeating round can improve the total energy dissipation and performance in the sensor network. There are some problems with the LEACH protocol. The main problem is the residual energy of a node is not considered for cluster formation. The nodes are started with the same initial energy. The cluster-heads are randomly selected rotationally. The proposed protocol selects the cluster-heads based on their residual energy.

3. EEAP System Model

The proposed system assumed with the following properties:

- The sink located very far from the cluster-heads.
- All the sensor nodes are stationary with limited energy.
- All the sensor nodes are equipped with power control capabilities to vary their transmitting power.
- The network is assumed to be continuous data delivery model.

Assume that there are N sensor nodes randomly deployed into M x M region. It is assumed that M=100 and the base station locates very far from the sensing area. Then the distance from the cluster nodes to the base station is very long. Also, we assumed that clusters are equally sized. Thus there are average N/k nodes per clusters and (N/k)-1 non cluster head nodes. The EEAP involves three main phases: the Initial phase, the clustering phase, and the data transmission phase. The initial phase is performed only once in the beginning of network operation.

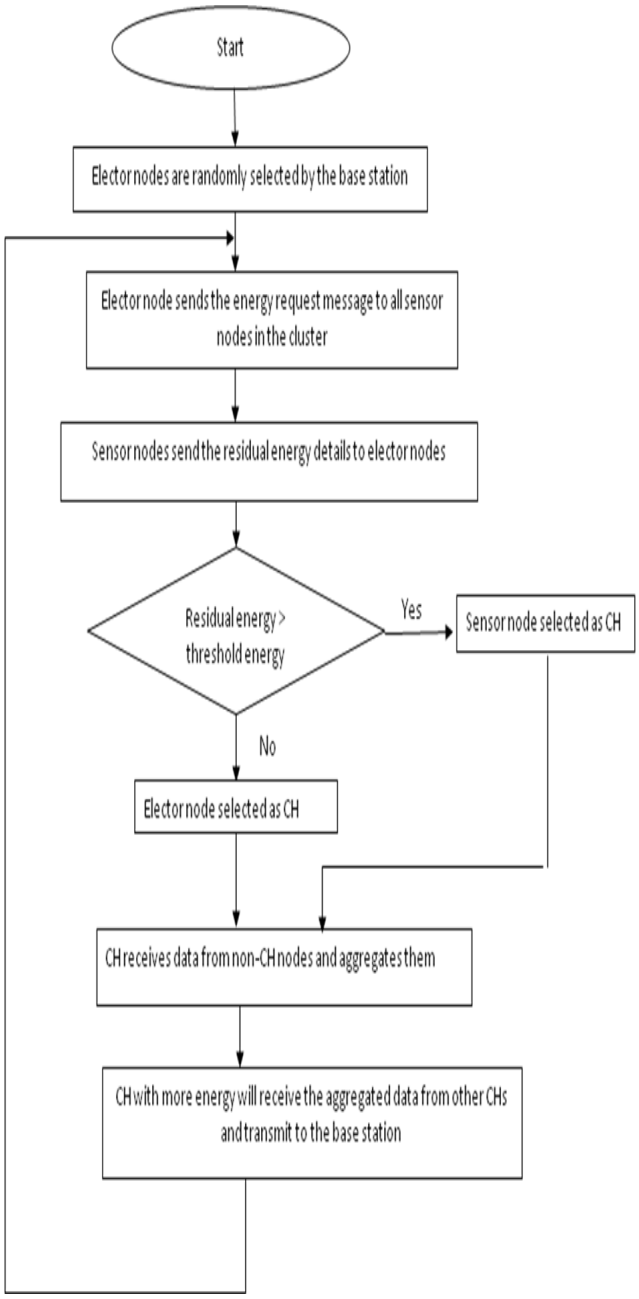


Fig.2. Flow chart for proposed protocol

As in the LEACH protocol, the EEAP is divided into rounds, where each round consists of the clustering phase and the data transmission phase. Each round begins with the clustering phase when the clusters are organized, followed by a data transmission phase when data are transferred from the nodes to cluster head and on to the base station.

3.1 Initial Phase

The sink selects most separated k optimal number of elector nodes, then the sink broadcasts and elector advertisement message in initial phase.

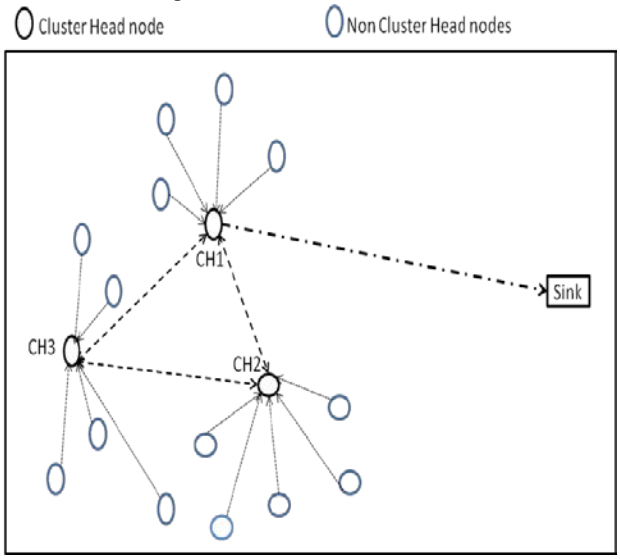
3.2 Clustering Phase

The clustering phase involves the cluster formation and cluster head selection. In the LEACH protocol, the cluster head is selected randomly. But in the proposed algorithm, the elector node selects the cluster head based on the residual energy of each sensor node within the cluster. Elector nodes take responsibility for collecting nearest sensor's information and energy is greater than others. After the cluster head node is selected by elector node, the cluster head node broadcasts a cluster head advertisement message containing cluster head ID. Non – cluster head nodes then select the most relevant cluster head node according to the signal strength of the advertisement message from the cluster head nodes. Each member node transmits a join request message.

3.3 Data transmission phase

The cluster head node collects and aggregates the data from the non-cluster head nodes. Once the cluster head node receives the join request message from member nodes, the cluster head setup a TDMA schedule according to their active member nodes. In LEACH approach, when cluster heads have aggregated data from the nodes, they send it to the Base Station. In our proposed protocol, the cluster heads of all clusters exchanging their aggregated data between them. On rotational basis, one cluster head collects all the data from other cluster heads and transmit to the base station. Each cluster head takes the responsibility of transmitting data to the base station as shown in Fig.3. Once the data transmission phase ends, network reforms the cluster head selection procedure in a new round.

Fig.3. Data Transmission from CH1



In Fig.3, the Cluster-head CH1 received the data from the cluster-heads CH2 and CH3. CH1 will aggregate the data received from other cluster-heads and the sensor nodes in its cluster, and transmit the aggregated data to the sink which far from the cluster-heads. Instead of all the cluster-heads sending the aggregated to the sink, one cluster-head collects, aggregates, and transmit the data the sink. Many research works proved that the transmission will consume more power than the sensing and reception. This approach will reduce the battery usage and saves extend the life time of network.

Fig.4. Data Transmission from CH2

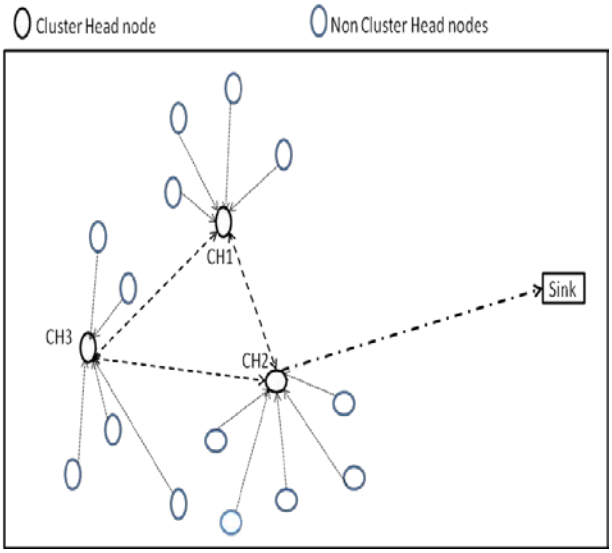


Fig.4. shows that the cluster-head CH2 collects the data from the cluster-heads CH1 and CH3 and from its own sensor nodes. It will aggregate the collected data and transmit to the sink.

Fig.5. Data Transmission from CH3

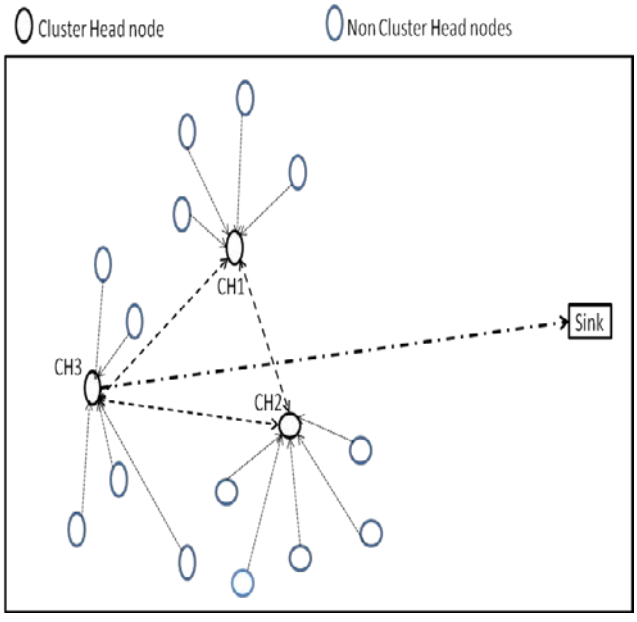


Fig.5. shows that the cluster-head CH3 collects the data from the nodes in its cluster and also from the cluster-heads CH1 and CH2. The collected data is aggregated and transmitted to the sink.

4. Performance Analysis

4.1 Energy Consumption and Simulation parameters

In this simulation, energy is decreased whenever a node transmits or receives data and whenever it performs data aggregation. The carrier sense operation consumes less energy. Table.1. shows simulation parameters.

Table.1. Simulation parameters

Description	***Value***
Simulation Area	100x100
Number of sensor nodes	150 and 300
Sink node location	Far from the network
Initial energy	0.2 J
Energy of data aggregation	5nJ/bit
Data packet size	500bytes
Optimal cluster number	10

4.2. Simulation Result and Analysis

In the simulation, we compared the performance of our proposed EEAP algorithm and with LEACH protocol in under the continuous delivery model. We simulated the model for the equal initial energy (0.2J) in each node. The size area was considered with small (100x100) situation. Our performance criteria are total residual energy per round in the network and total network lifetime. Network lifetime is the number of round from the start of operation until the death of the last alive node. The network connectivity which depends on the time of the first node failure is a meaningful measurement in the sense that a single node failure can make the network partitioned and further services be interrupted. When a sensor node is depleted of energy, it will die and be disconnected from the network which may impact the performance of the application significantly.

When the nodes start with the same initial energy and the total number of nodes in a network is 150 and 300, the number of living nodes per round is shown in Fig.6 and Fig.7 respectively. Fig.6 shows that the total network lifetime of our algorithm is longer than that of LEACH. During most of the network lifetime, our algorithm EEAP runs with much more living nodes than LEACH. Fig.7 is the case for the number of nodes in a network of 300. The result is similar with Fig.6. In case of LEACH protocol, the round first dead node appears is very fast and linearly decreases until last round. EEAP shows the round of first dead node and network lifetime is almost same regardless with the increase of network size. Simulation shows our algorithm can balance the energy consumption of the entire network compared to LEACH protocol.

Fig.6. Number of Living nodes in each round with same initial energy

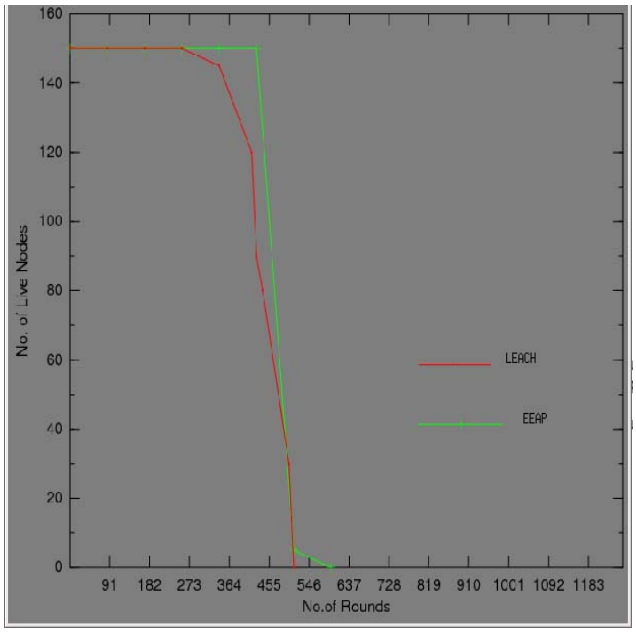
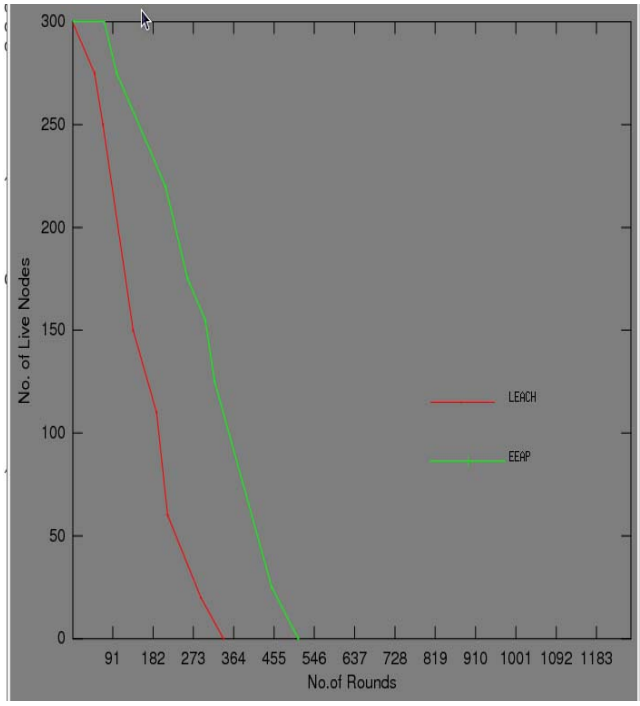


Fig.7. Number of Living nodes in each round with same initial energy with increased senor nodes



5. Conclusion

Wireless sensor networks are increasingly being used for health care, transportation, manufacturing, and much more. In this paper, we proposed an "Energy Efficient Adaptive Protocol for Clustered Wireless Sensor Networks" (EEAP), to extend the lifetime of a sensor network by balancing energy usage of the nodes. Our algorithm ensures that the nodes with more energy should be cluster heads more often than the nodes with less energy. We showed that in many cases our algorithm is more energy efficient than LEACH. The results show that the proposed algorithms can maintain a balanced energy consumption distribution among nodes in a sensor network and thus prolong the network lifetime.

References

[1] J.N Al-Karaki and A.E. Kamal, "Routing techniques in wireless sensor networks: a survey", IEEE wireless communications, vol.11, no.6, 6-28, 2004.

[2] M.J.Handy, M.Haase, and D.Timmermann, "Low energy adaptive clustering hierarchy with deterministic cluster-head selection", In proceeding of the IEEE international conference on mobile and wireless communications networks, Stockholm, Sweden, 368-372, 2002.

[3]Otgonchimeg Buyanjargal, Youngmi Kwon, "Adaptive and Energy Efficient Clustering Algorithm for Event-Driven Application in Wireless Sensor Networks", Journal of networks, vol. 5, no. 8, august 2010

[4] S.Tilak, N.Abu-Ghazaleh and W.Heinzelman, "A taxonomy of wireless micro-sensor network communication models, "ACM Mobile Computing and Communication Review(MC2R), June 2002.

[5] L.B. Ruiz, I.G. Siqueira, L.B. Oliveira, H.C. Wong, J.M.S. Nogueira, A.A.F. Liureiro, "Fault management in event-driven wireless sensor networks," in Proceedings of MSWIM' 04

[6] C.Intanagonwiwat, R.Govindan, and D.Estrin, "Directed diffusion: A scalable and robust communication paradigm for sensor networks," in Proceedings of 6th ACM/IEEE MOBICOM, 2000.

[7] Zheng Zeng-wei, Wu Zhao-hui and Lin Huai-zhong, "An event driven clustering routing algorithm for wireless sensor networks", in the Proceedings of IEEE/RSJ, Oct'2004

[8] Vinh Tran Quang and Takumi Miyoshi, "Adaptive routing protocol with energy efficient and event clustering for wireless sensor network," IEICE Trans.Commun., vol E91- B, No.9, Sep 2008

[9]Dali Wei, Shaun Kaplan and Anthony Chan, "Energy efficient clustering algorihtms for wireless sensor networks", in Proceedingd of the ICC 2008 Workshop.

[10] W. Heinzelman, A. Chandrakasan and H. Balakrishnan, "Energy efficient communication protocol for wireless microsensor networks," in the Proceedings of HICSS '00, Jan 2000.

[11] S. Lindsey and C.S. Raghavendra, "PEGASIS: Power efficient gathering in sensor information systems," in Proceedings of IEEE Aerospace Conference, vol 3, pp. 3-1125–3-1130, Mar 2002.

[12] Christopher Ho, Katia Obraczka, Gene Tsudik, and Kumar Viswanath, "Flooding for reliable multicast in multi-hop ad hoc networks", In Proceedings of the 3rd International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications (DIAL-M'99), 1999, pp. 64–71

[13] S.K. Singh, M.P. Singh, and D.K. Singh, "A survey of Energy-Efficient Hierarchical Cluster-based Routing in Wireless Sensor Networks", International Journal of Advanced Networking and Application (IJANA), Sept.–Oct. 2010, vol. 02, issue 02, pp. 570–580.

[14] Tung-Jung Chan, Ching-Mi Chen, Yung-Fa Huang, Jen- Yung Lin and Tair-Rong Chen, "Optimal cluster number selection in ad-hoc wireless sensor networks," WSEAS Transaction on Communications, Issue 8, vol 7, Aug 2008

Dr. P. Kamalakkannan received his B.Sc and MCA degrees in 1988 and 1991 from University of Madras, India. He has obtained his Ph.D degree in Computer Science in the year 2008. His research interest includes Distributed Systems, Pervasive Computing, and Wireless Adhoc networks, Wireless sensor networks.

K. Padmanabhan is working toward his Ph.D degree at the Computer Applications department of Anna University of Technology, Coimbatore. His research interest includes Wireless sensor networks and Wireless Adhoc Networks.

Encrypted IT Auditing and Log Management on Cloud Computing

[1] Rajiv R.Bhandari, [2] Nitin Mishra

[1] M-TECH *,Department of Information Technology,
NRI Institutions, University of RGPV,Bhopal(MP)
India

[2] Prof, Department of Information Technology,
NRI Institutions, University of RGPV,Bhopal(MP)
India

Abstract

In this paper we are conducting the investigation studies over the IT auditing for assuring the security for cloud computing. During this investigation, we are implementing working of IT auditing mechanism over the cloud computing framework in order to assure the desired level of security. In the IT auditing mechanism, the concept of checklists are prepared for the cloud computing application and their lifecycle. Those checklists are prepared on the basis of models of cloud computing such as deployment models and services models. With this paper our main concern is to present the cloud computing implications for large enterprise applications like CRM/ERP and achieving the desired level of security with design and implementation of IT auditing technique. As results from practical investigation of IT auditing over the cloud computing framework, we claim that IT auditing assuring the desired level of security, regulations, compliance for the enterprise applications like CRM and ERP. Another problem in cloud computing is that huge amount of logs make the system administarator hard to analyse them.In this paper we proposed the method that enables cloud computing system to achive both effectiveness of system resource and strength of security service without trade-off between them.

Keywords*:* Customer Relationship Management,Enterprises Resource Planning

1.Research Background

Recently, all over the world mechanism of cloud computing is widely acceptable and used by most of the enterprise businesses in order increase their productivity. However there are still some concerns about the security provided by the cloud environment are raises.

The top concern of Cloud adption- Security[2]

Cloud computing is most probability of collection such as service oriented topic, as well as on-centric concept and good practices techniques.

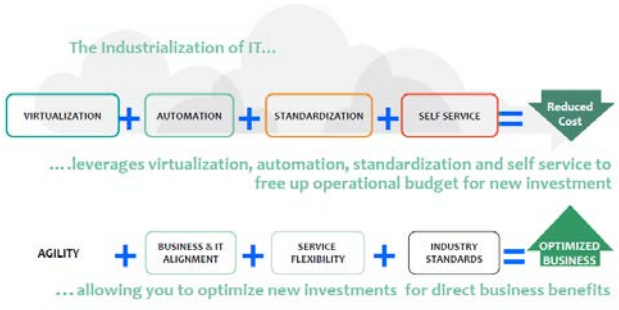


Fig1.Cloud Computing

Cloud computing gives benefit of provisioning resources and application of services to customer. Customer is needs to subscribe its related services. This service is to depend upon its development, infrastructure, and storage capacity. These services also provide of two types of computing services software and desktop services. In the cloud computing is the thin client interaction with remote cloud using operating system. It give the virtual desktop in virtual local operating .this operating system is access the virtual data storage. This o.s executes application at anytime & anywhere. When IBM Watson claimed the world is needed only five machine. It is back all the all things.

Now a day's why IT is reaching a critical point of view. In storage total growth is 54% of Explosion of information. Large scientific calculation such as medicine, forecast, and healthcare is most energetic and faster processing capacity. In reality near about 85% computing capacity is idle. Average of IT budget as 70%.It is specifically managed by IT infrastructure added by new things. Many technologies are different than cloud computing such as parallel computing, virtual computing, architecture of services oriented, and autonomic computing. All computing are advancing computing in unusual pace. Connectivity is additional part of the keeps falling. Cloud can be depicted based on web application through internet, this application are standard application.

People can understand without most of period of knowledge's, training section, and they understand to operating system as well as basic thing such as hardware maintenance it can be accomplished of their work done easily and properly. Consumers are purchase on demand for cloud computing capacity but they are not concerned used in underlying technologies. Typically computing data & resources can be accesses by own. They are access by third party provider. It is not copulation to locate nearly. They are potentially beyond state in physical boundary country. Those applications can be moving there its own infrastructure to cloud. It has shifted in house control to a third party.[2][11]

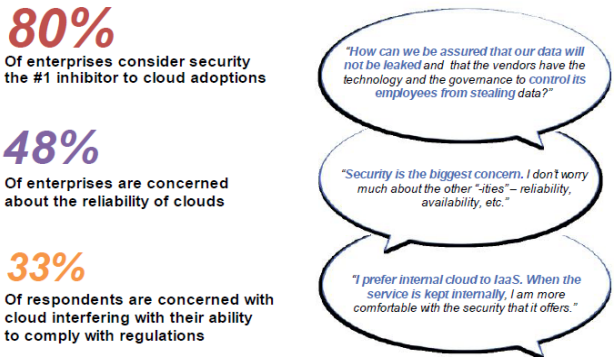


Fig2. Checklist for Cloud

1.1 Checklist for public cloud:

Government is kept to use public cloud to take the advantages of the cost effective by the providing public useful information in cloud. It can mention the cloud concepts to the integrated computing resources.there are different departments in the manageable pool. It auditing in public cloud can have different type which is based on different type of services model. There are address two popular service models in this topic, Infrastructure as a Service(IaaS) and Software as a Service (SaaS).[1]

1.2 Checklist for Private Cloud

Checklist for a private cloud is a very practical approach and attractive option to many security sensitive enterprises. The private cloud gives not only the self control but also the benefits of cloud computing, it is mainly sharing computing resources including processing power and storage capacity among different departments within an enterprise. Traditionally, department computing resources are not shared due to data sensitivity, self control and different business nature of departments. Private cloud could remove or blur these boundaries

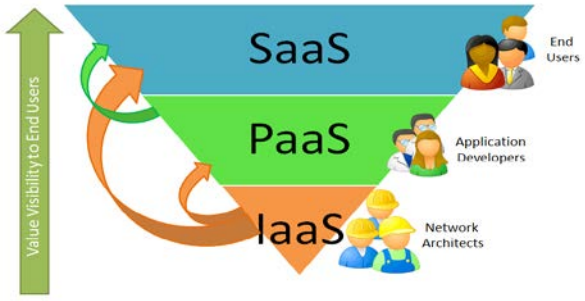


Fig3. Cloud Computing Layers

Each department is allocated by computing resources from the pool by provisioning need on demand. From the department point of view, the computing resource is unlimited. Therefore achieving a task faster or making a task not achievable before due to computing power constrain.

In most cases, a private cloud could cut IT cost down; it is increase the flexibility and scalability also, make available 24x7 and even do applications that are impossible before the cloud. Private cloud certainly poses a great management challenging as well as auditing challenging. . It virtualizes of all computing resources from the different departments into the computing resource pool.[1]

2. Research Methodologies

Two research methodologies such as qualitative research methods and quantitative research methods:

Qualitative Research Methodologies:

For the collection of data, qualitative research methods used the observations, interviews as well as may include the surveys, case studies, document and historical analysis. Survey researches as well as case study are commonly used methods for the data collection in various researches. Case study and survey research are also often considered methods on their own. [8]

In order to use this research method for determining the research problem, researcher must need to raise some questions.

2.1 Methods of Data Collection

Interviews:.Interview is nothing but the conversation form in which the main purpose is for the researcher to collect the data which address overall research study questions and goals.This method is directly intercative and frequently used.[8] [9]

Document and Artifact Analysis: In this method things are following roughly into document categories as well as artifact analyses, however, overlapping with other methods. In our case, we identified various artifacts over the SPR, its history, with some kind of analysis with

simple experiments of SPR constructions in order to answer the proposed research questions. [8] [9]

2.2 Methods of Data Analysis

The data which is collected using above qualitative data collection methods is nothing but just the rough materials that researchers gathering from different aspects of world related to their research problems and questions. Qualitative data is collected in different forms like objects, photos, video recordings of behaviors, choices patterns in computer materials. But words are frequently are raw materials which are further analyzed by qualitative researchers using the different techniques of data analysis. There many methods are available for the researchers to analyze the qualitative data depending on qualitative researcher basic philosophical approach. According to Huber man and Mile, the process of qualitative data analysis is made up of three parallel flows of activities such as data display, data reduction, as well as conclusion verification or drawing. Hence most of the qualitative analysis researchers use the technique data reduction method for the analysis of collected data in order to seek the correct meaning of it for particular research. [9] [10]

2.3 Quantitative Research Methodologies

Using the quantitative research, we can obtain three various classifications of numbers like customer profiles, attitudinal data and market measures. For the data collection in case of quantitative research, there are different ways:
By asking questions related to the research problem either using telephone, face to face, postal or computer medium.
By observing the things like person, diaries or instruments.
WHAT the people think and WHY must be determined by asking different questions.
Thus, in this research we frequently used the market analysis tool and sampling mechanism for some measurements related to Peer to Peer Networks and Security study. [2]

3. Implementation Environment

3.1 Introduction to the application:

Checklist generation for incident management system is an application which is used for online problem solving which can be encountered in the computer and mobile devices.
The basic idea of the application is that the problem solving can be done online and the admin can keep the record of the process that is done as a checklist or in a log format.

Incident management system is a part of corrective software, handles an any event which may cause an stoppage to a service or decrease in the quality of service. In incident user and customer May reporting by email by telephone, by chat services, voice mail, by letter, and visits.
Incidents are define by IT infrastructure library, incident are classified into

- Software failure
- Hardware failure.
- Service request.

In addition to incidents the system should assets service desk or help desk to handle problems, change request product orders, and development ideas effectively following the time scale defines in service level agreement (SLA).
Incident management is very tool oriented process. Service desk workers use various application while searching solution to the incident such as incident management tools, email application remote desk of application, office application and communication application, remote connection to server ,and internet search tools.

3.2 Checklist generation:

Here the admin can view the detailed report of the daily events created by the users of this application. These details are stored in the form of the checklist that can be viewed on admin page.

3.3 Encryption and Decryption:

Here the customer registration information is in the encrypted form and it can be decrypted by the user.

3.3.1 Encryption

In the Encryption technique, if new user is register his information in this application, this in information is saved in encrypted form in table. So this private information is not known to any other user .so it is better security to our application.

3.3.2 Decryption

Also we have provided the decryption technique for the user to read the information is correct or not by showing it on below the encryption table. This data is known to only that user which is login.

4.Log Management

Log data can provide valuable security and operations insight into enterprises applications like CRM/ERP.Many companies with limited IT staffing will find that outsourcing log management can bring them more value from their log data than they could attain on

their own—without all the expenditures in hardware,staffing and product management. If in-cloud providers can deliver prompt, secure, reliable service, cloud-based log management could be a growth sector over the next few years, particularly for the CRM/ERP Applications.

One of the big questions in making a decision between internal and in-cloud log management is how much time can be allocated to monitoring and upgrading the system internally to meet the business needs of the organization. If an organization's primary goal is regulatory compliance or to minimize IT staff requirements, then outsourcing log management to cloud application providers will probably be suitable.[X][XI]

Organizations that decide to outsource their log management should be careful to select flexible services that allow for expanded correlation and use of the log data for organizational benefit. This is also true of internally-developed log management systems, which today are experiencing interoperability issues that make data normalization and correlation difficult for organizations of all sizes.

Cloud Computng system checks user behaviour everyday and decreases risk point if user uses cloud computing service more than one hour.so many people use Cloud Computing service so the huge logs arises from transaction between systems,user information update,mass data processing and so on therefore it is very difficult to analyse using log in emergency.to make analysing log better i proposed the method that devides log priority according to security level.

The auditing priority of logs is also decided by anomaly level of users.it means log generated by who hav most high anomaly level are audited with top priority and log of low level users are audited at last

5. Conclusion

Cloud computing technology provides human to advantages such that enables cloud computing system to achive both effectiveness of system resource and strength of security service without trade-off between them and manages users logs.

References

[1] IT Auditing to Assure a Secure Cloud Computing 2010 IEEE 6th World Congress on Services

[2] Enterprises new Dimension in computing[WAVV 2011] Kemp Little – 'Hot Topic' Article for PLC on Cloud Computing 19 February 2009

[3] NIST Definition of Cloud Computing v15, accessed on 4/15/2010,

[4] http:/csrc.nist.gov/groups/SNS/cloudcomputing/cloud-def-v15.docWill Forrest, Clearing the Air on Cloud Computing, Discussion Document from McKinsey and Company, March 2009

[6] Security Guidance for Critical Areas of Focus in Cloud Computing V2.1, by Cloud Security Alliance, December 2009.

[7] Gerard Briscoe, Alexandros Marinos: Digital Ecosystems in the Clouds: Towards Community Cloud Computing, IEEE Digital Ecosystems and Technologies DEST (2009), online access a http://arxiv.org/PS_cache/arxiv/pdf/0903/0903.0694v3.pdf

[8] Rajkumar Buyyaa, Chee Shin Yeoa, Srikumar Venugopala, James Broberga, and Ivona Brandicc,Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility,Future Generation Computer Systems, Volume 25, Issue 6, June 2009, Pages 599-616.

[9] Michael Armbrust, et al, Above the Clouds: A Berkeley View of Cloud Computing, UC Berkeley Reliable Adaptive Distributed Systems Laboratory, Feb, 2009.

[10] Cloud Computing Architecture and Strategy Gerd Breiter IBM Distinguished Engineer gbreiter@de.ibm.com

[11 Multi-level Intrusion Detection System and log management in Cloud Computing Jun-Ho Lee ; Min-Woo Park ; Jung-Ho Eom Tai-Myoung Chung ; Sch. of Inf. Commun. Eng., Sungkyunkwan Univ., Suwon, South Korea Advanced Communication Technology (ICACT), 2011 13th International Conference.

Emotion Recognition using Dynamic Time Warping Technique for Isolated Words

N. Murali Krishna[1] **P.V. Lakshmi[2]** **Y. Srinivas[3]** **J. Sirisha Devi[4]**

1. Dept of CSE, GITAM University
Vishakhapatnam, 530045, Andhra Pradesh, India

2. Dept of IT, GITAM University
Vishakhapatnam, 530045, Andhra Pradesh, India

3. Dept of IT, GITAM University
Vishakhapatnam, 530045, Andhra Pradesh, India

4. Dept of CSE, CMRCET, JNTU (H)
Hyderabad, 501401, Andhra Pradesh, India

Abstract

Emotion recognition helps to recognize the internal expressions of the individuals from the speech database. In this paper, Dynamic time warping (DTW) technique is utilized to recognize speaker independent Emotion recognition based on 39 MFCC features. A large audio of around 960 samples of isolated words of five different emotions are collected and recorded at 20 to 300 KHz sampling frequency. Training and test templates are generated using 39 MFCC features. In the proposed work, we have extracted the MFCC coefficients from the speech database and DTW is used to store a prototypical version of each word in the vocabulary and compute incoming emotion with each word. For the classification of emotions SVM is used. The experimental results are provided using MFCC, Delta Coefficients (ΔMFCC) and Delta Delta Coefficients(ΔΔMFCC) . It is proposed that higher recognition rates can be achieved using MFCC features with DTW which is useful for different time varying speech utterances.

Keywords: Dynamic Time warping (DTW), MFCC (Mel frequency cepstral coefficient), Feature extraction, SVM

I. Introduction

Emotion recognition(ER) has made great strides with the development of digital signal processing hardware and software. But despite of all these advances, machines cannot match the performance of their human counterparts in terms of accuracy and speed, especially in case of speaker independent emotion recognition. Emotion identification [1] provides useful information in other sound source identification applications, such as speaker recognition and speech recognition. Here our approach is to classify emotions using MFCC features [6] and DTW. Recognition accuracy for MFCC feature is considered as it mimics the human ear perception. So ER recognition using MFCC features is illustrated in this paper.

A central topic in our paper is the emotion recognition using isolated words. In section-2 we describe the basics of the proposed system, in section-3 we discuss regarding feature extraction, in section-4 we present the classification procedure with DTW and section -5, concludes the paper.

2. Proposed Work:

2.1 Database collection:

In this paper we ,have considered five Emotions, namely Sad, Happy, Angry, Surprise and Neutral . We have collected 150 isolated words of each person with different time varying constraint at 20 to 300 KHz sampling frequency. Total 960samples are collected for our experimentation. Out of 960samples 480 samples are considered for training and 480 samples are considered as testing samples. Training templates are generated using 39 MFCC features.

2.2 Principal of Emotion recognition:

The two basic tasks in emotion recognition [3], [5] are pre-processing of speech signals and then classification part. In pre-processing, we analyse the speech signal before extracting the required features from it. Different operations are performed on the input speech signal such as removal of silence part, reemphasis, segmentation and framing, windowing, Mel Cestrum analysis and recognition (Matching) of the isolated words. The two phases of emotion recognition algorithms are testing and training phase. The block schematic is given in the figure 1.

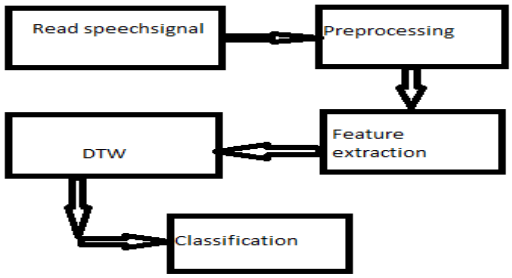


Fig 1: Block Diagram- schematic of emotion recognition

2.3 Removal of silence part:

By reducing the dimensions of feature vector we can improve the time complexity, which can be done by the removal of the silence part of speech. One of the best techniques to remove silence part is considering the Energy feature. Energy of each frame is calculated. Based on threshold value of energy the silence part is removed. The energy of each frame is given by

$$E(i) = \frac{1}{N}\sum_{n=1}^{N}|xi(n)|2 \quad (1)$$

We used Silence part removal algorithm which says divide the signal into number of frames, calculate energy for each frame, calculate threshold using median and compare energy of each frame with threshold. If energy of frame is greater than threshold then consider it otherwise it is silence part of signal and eliminates it. Some of the results of silence part removal for some speech signals are shown in figures- 2a, 2b, 2c, 2d.

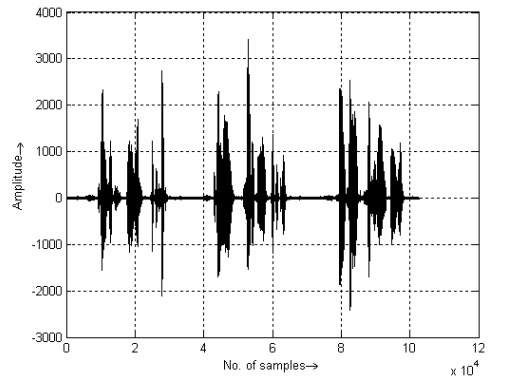


Figure-2a Speech signal in sad emotion before silence removal

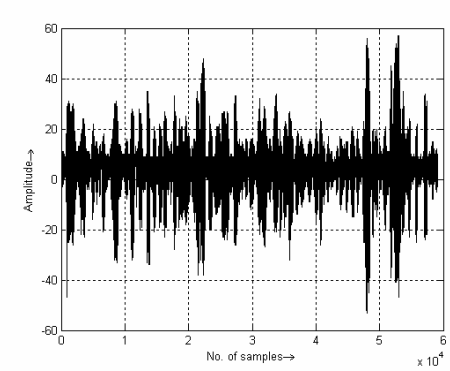


Figure-2b Speech signal in sad emotion after silence removal

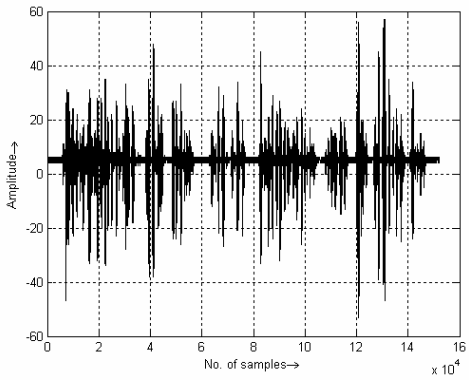


Figure-2c Speech signal in happy emotion after silence removal

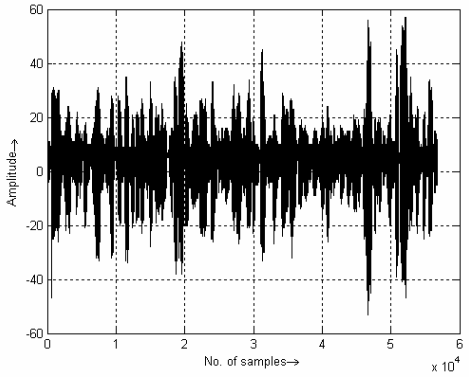


Figure-2d Speech signal in happy emotion after silence removal

3. Feature extraction:

This is the vital part of emotion recognition based on which the remaining part of classification and recognition depends on of acoustic model of speech signals. MFCC feature works better for Emotion recognition than LPC, LPCC with different experimentation [11], [12], [13]. There is no standard number of MFCC coefficients for recognizing the sound in any literature. MFCC is the way of representing the spectral information of a sound in compact form. In our paper we carried out the experimentation on MFCC to finalize the number of coefficients. It is proposed that 8-14 number of MFCC coefficients is sufficient to recognize the emotions [4], [10]. The algorithm for getting MFCC feature is as follows:

1) pre-emphasis,
2) hamming windowing,
3) FFT to obtain power spectrum,
4) log of FFT,
5) Mel filter bank,
6) DCT for decorrelation,
7) Δ MFCC (optional),
8) and ΔΔMFCC coefficients.

The mathematical details of each step are briefly described below.

Step 1: Pre–emphasis

This process will increase the energy of signal at higher frequency. It enables the passing of each speech signal through a first order FIR filter which emphasizes higher frequencies. The first order FIR filter equation is used is

$$Y[n] = X[n] - 0.95\,X[n] \quad (2)$$

Step 2: Framing

Each speech signal is divided into frames of 36 ms(milliseconds) and most of spectral characteristics remain the same in this duration, with 50 % of overlapping.

Step 3: Windowing

To remove edge effects, each frame is shaped with hamming window. Hamming window works better than other windows. The hamming window is represented by

$$w(n) = 0.54 - 0.46\cos\frac{2\pi n}{n-1} \quad (3)$$

where $0 \leq N \leq n\text{-}1$

Step 4: Fast Fourier Transformation (FFT)

FFT is used to get log magnitude spectrum to determine MFCC. We have used 1024 point to get better frequency resolution.

Step 5: Mel Filter Bank Processing

The 20 Mel triangular filters are designed with 50% overlapping .From each filter the spectrum are added to get one coefficient each, in this way we have considered the first 13 coefficients as our features. These frequencies are converted to Mel scale using following conversion formula.

$$f(mel) = 2595 * \log 10\,(1 + \frac{f}{700}) \quad (4)$$

we have considered 13 MFCC coefficients because, of the fact it gives better recognition accuracy than other coefficients.

Step 6: Discrete Cosine Transformation(DCT)

DCT of Each Mel frequency Ceptral are taken for de-correlation and energy compaction is called as MFCC. The set of coefficient are called MFCC Acoustic Vectors. Therefore, each input speech signal is transformed into a sequence of MFCC Acoustic Vector from which reference templates are generated.

Step 7: Delta Energy and Delta Spectrum

Features related to the change in cepstral features over time are represented by 13 delta features (12 cepstral features plus one energy feature), and 13 double delta or acceleration features. Each of the 13 delta features represents the change between frames, while each of the 13 double delta features represents the change between frames in the corresponding delta features. In similar fashion all the total 39 MFCC feature are calculated for every frame which constitute feature vector. Mel filter bank generated is shown in figure 3.

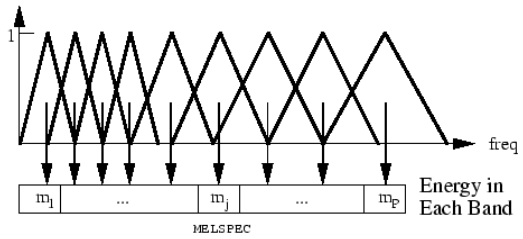


Figure 3: Mel filter bank.

4. Classification

After generating reference template i.e., generating the Training Set, the classification is done using SVM classifier [2]. For best matching, each input signal feature vector is calculated using MFCC as shown in fig 1. For every word DTW score is determined with reference template (Test sample). The best matching sample is classified using SVM classifier

4.1 Dynamic time Warping (DTW):

DTW [7], [8] is one of the approaches to emotion recognition to store a prototypical version of each word in the vocabulary and compute incoming emotion with each word. First a template of sequence of feature vectors is taken. The template is a single utterance of the word selected to be typical by some process. Then the comparison can be achieved by pair-wise comparison of the feature vectors. DTW achieves its goal by finding an optimal match between two sequences of features vectors which allows for stretched and compressed sections of the sequence. The DTW scores are calculated using above algorithm. Based on score SVM[4] is applied to find minimum DTW distortion. Based on minimum DTW distance the emotions are recognized. The DTW works better for different time varying /duration signal of same speech signal. This is an advantage over usual Euclidian distortion measure.

5. Conclusions:

We have conducted experiments on five human emotions which are happy, sad, angry, surprise and neutral using MFCC and DTW algorithms. Algorithms are implemented in MATLAB [9] environment, to identify the emotions. The results obtained after classification are tabulated and shown in Table 1 ,against each emotion.

	Recognition accuracy in % for individual emotion				
Features	Happy	Sad	Angry	Surprise	Neutral
MFCC(13)	83	82	85	81	78
MFCC + ΔMFCC (26)	84	86	91	87	82
MFCC+ ΔMFCC + ΔΔMFCC(39)	92	94	91	89	90

Table 1: Recognition accuracy rates

The Recognition accuracy for MFCC,ΔMFCC,ΔΔMFCC are obtained and presented in the Table- 1.The results are considered for 39 features (MFCC+ΔMFCC+ΔΔMFCC).

From the above Table-1, it can be seen thqat the emotions are clearly recognized and the recognition rates are above 78%, this is recognition rates are due to the consideration of using 39 MFCC features . Dynamic time warping is simply time alignment method which works better for time varying signals. Every distortion measure should be based on DTW for better recognition accuracy. Here recognition system considers only one specific measure of a sound i.e. the MFCCs, and yet still achieves quite accurate results. Also the results can be further improved using Hidden Markov Model (HMM), Gaussian Mixture Model (GMM).

References

[1] E. Mower, M. Matari´c, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," IEEE Trans. on Audio, Speech, and Language Processing, Accepted for Publication.

[2] B. Schuller, G. Rigoll and M. Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing, vol.1, pp. I-577-580, 17-21 May, 2004.

[3] V.A. Petrushin. Emotion recognition in speech signal: Experimentalstudy, development, and application. ICSLP, 2000.Conf. on Machine Learning and Cybernetics, vol. 8, pp. 4898–4901, Aug. 2005.

[4] E. Osuna, R. Freund, and E Girosi, "Support vector machines: Trainingand applications," A.I. Memo 1602, MIT A. I. Lab., 1997.

[5] O. Kwon, K. Chan, J. Hao, and T. Lee, "Emotion recognition by speechsignals," in Proc. 8th Eur. Conf. Speech Commun. Technol., 2003, pp.32–35.

[6] Ezzat T., Tomaso Poggio T. Discriminative Word-Spotting Using Ordered Spectro-Temporal Patch Features, Interspeech 2008

[7] E. Keogh and M. Pazzani, Scaling up Dynamic Time Warping for Datamining Applications, In Proc. of the Sixth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, pp.285-289, Boston, Massachuseetts, 2000.

[8] H. Sakoe, and S. Chiba, Dynamic Programming Algorithm Optimization for Spoken Word Recognition, In IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-26, 1978.

[9] Introduction to Computer Programming With MATLAB Department of Phonetics and Linguistics, Univ. College Landon [Online]. Available:www.phon.ucl.ac.uk/courses/spsci/matlab/, 2004.

[10] YI-LIN LI, Gang Wei, "Speech emotion recognition based on HMM and SVM", Proceedings of the FourthInternational Conference on Machine Learning andCybernetics, Vol.8, 18-21 Aug. 2005, pp.4898 – 4901

[11] Zhongzhe, X,Dellandrea, E.: Weibei D eal.Features Extraction andSelection for Emotional Speech Classification IEEE.(2005)411-416

[12] Ververidis, D., Kotrropoulos, K.: Emotional speech recognition:Resource, features, and methods. Speech Communication 48(2006) 1162-1181

[13] Prof .Sujata Pathak, Prof .Arun Kulkarni (2011) ,"Recognising emotions from speech".

Decentralized Lifetime Maximizing Tree with Clustering for Data Delivery in Wireless Sensor Networks

Deepali Virmani[1] and Satbir Jain[2]

[1]Department of CSE, BPIT, GGSIPU, Delhi, India, 110085.

[2]Department of CSE, NSIT, DU, Delhi, India, 110003.

Abstract

A wireless sensor network has a wide application domain which is expanding everyday and they have been deployed pertaining to their application area. An application independent approach is yet to come to terms with the ongoing exploitation of the WSNs. In this paper we propose a decentralized lifetime maximizing tree for application independent data aggregation scheme using the clustering for data delivery in WSNs. The proposed tree will minimize the energy consumption which has been a resisting factor in the smooth working of WSNs as well as minimize the distance between the communicating nodes under the control of a sub-sink which further communicate and transfer data to the sink node.

Keywords: *Lifetime, Tree, Clustering, Wireless Sensor Networks, HyMac.*

1. Introduction

The current technological advancement has already come to terms with immense potential of Wireless Sensor Network, Which consists of tiny sensor nodes scattered in a region communicating with each other over well defined protocols and transferring information of temperature, humidity etc between each other. To exploit the potential of WSNs various studies have focused on Data aggregation approach which requires data to be collected and processed at a single node prior to its transfer to the parent node. Realization of this data aggregation approach has been a major concern owing to the limited battery life of the sensors which limits the lifetime for which a sensor node remains active. The limited battery leads to disruption in the connections from the network. These disruptions suggest that the design must incorporate topological changes. Since the dense deployment of sensors nodes leads to the detection and transmission of data from the nearby nodes upon receiving a single stimulus. Thus, idea is not to allow the direct transmission of to interested users upon event detection instead aggregating them to remove redundancy. The application domain of WSN is still expanding therefore it is important to support the data aggregation scheme from multiple nodes for simultaneous and fast processing of the data.

In this paper, we focus on the construction of decentralized life time maximizing tree based on clustering. Our scheme consists of three parts, namely, the clustering of nodes by using the Expectation-maximization (EM) [9] algorithm, construction of decentralized life maximizing tree within the cluster [8] , and aggregating the data collected from the WSN nodes by applying a cluster scheduling approach to transfer it , which uses HyMac[1] mechanism.

The rest of the paper is organized as follows. Section 2 describes some related works. Section 3 elucidates our approach. Section 4 has the proposed algorithm. Section 5 we explain our approach with example, section 6 shows the simulation results. Finally, concluding remarks are provided in Section 7.

2. Related Works

In recent literature many studies have achieved data aggregation using several approaches namely mobile sink , LEACH(Low-Energy Adaptive Clustering Hierarchy), Directed Diffusion [9].All these schemes have tried to prolong the network lifetime and reduce the energy consumption The mobile sink scheme increases the network lifetime to four times as compared to the network in which sink is static but it suffers from serious shortcomings , it leads to an increased physical delay owing to the slow physical mobility of the sink then the wireless communication. It exhausts the battery life unnecessarily. LEACH is a self-organizing, adaptive clustering protocol. To have minimum energy consumption, nodes in LEACH are grouped into a number of clusters based on their battery usage. Each cluster has a cluster head, which communicates with every node of that cluster. The sink aggregates data, transmitted by cluster heads, from other nodes. Since a cluster head loses energy due to repeated transmissions, the cluster head is re-selected based on the residual energy, as a consequence it prolongs the network

lifetime. Directed Diffusion involves two types of messages, namely, the "Interest" message and the actual data messages. To aggregate data by using Directed Diffusion, the sink node broadcasts an "Interest" message that consists of a time-to-live value, and also the addresses of the source and destination nodes. The destination node on receiving the request transmits appropriate data message to the source having the sensed data. If the downstream nodes cannot be reached by the "interest" message from the current source then the current destination becomes the source node by changes its address, reduces the time–to-live value and rebroadcasts the "Interest "message.

3. Proposed Approach

We have discussed before that our main focus has been on distance minimization between the nodes, minimization of energy utilization and efficient utilization of bandwidth. Considering the problem sensor network is divided into clusters using an EM [9] algorithm based on the close proximity of the nodes and a decentralized life maximizing [8] tree is constructed within the cluster choosing a parent closest to the sink node to serve as sub-sink. Once the sub sink is chosen, scheduling of cluster is done using FDMA approach. That is assigning a range of frequencies from the available one to the sub-sink. The frequency ranges can be re-allotted to new clusters from the free pool or assigning a half frequency to a sub-sink from a low data rate transferring cluster. The sink will broadcast a topology packet containing information of the network as which source nodes are attached to which sub-sink node as per their location [4]. By making use of the hybrid TDMA/FDMA channel access technique [1], the sink node broadcasts a schedule packet informing others about their time slots as well as their channel frequencies for exchanging messages. But, our new idea lies in single sleep awake concept in which the source nodes wake up only once to listen and to transmit and rest of the time, they will remain in sleep state. We incorporate a concept of LPL (low power listening),the nodes are in LPL [2] state all the time to gauge topology changes and if there is a topology packet coming their way they wake up and make necessary changes. The tree construction follows a decentralized method [8]. To overcome the problems in [1]and to decrease the interference respectively we propose this new method. We will assign specific frequency slots based on attributes of the sensor nodes [6], with fixed interference ranges so that they can send their data in scheduled time in slotted frequency. Once the sender finishes sending, same frequency can be assigned to some other source accounting for the increase slots needed and also minimizing interferences.

To preserve the functional lifetime of all sources and efficient utilization of the energy of the source nodes, a Decentralized life maximizing tree construction algorithm [8] was studied, the DLMT [8] constructs a tree by selecting highest residual energy parent node to act as a center of data aggregation. The DLMT [8] construction algorithm arranges all nodes in a way that each parent will have the maximal-available energy resources to receive data from all of its children. Such arrangement extends the time to refresh the tree and lowers the amount of data lost due to a broken tree link before the tree reconstructions. The DLMT [8] algorithm can be further improved by considering distance also as a factor. In the proposed method we also include distance between the sensor nodes. Transmission distance has a major impact on the working of sensor network because the required power of wireless transmission is proportional to the square of the transmission distance. We follow the approach of clustering of nodes based on EM [9] algorithm. The EM algorithm includes minimizing the sum of the squares of the distances between nodes and cluster centroids. Therefore, we use the EM [9] algorithm to group the WSN nodes into K clusters on the basis of distance. We apply the concept of EM [9] algorithm initially and then use a new form of decentralized life maximizing tree, DLMT [8] algorithm accordingly. The cluster formed using the EMD algorithm goes through our proposed algorithm called Decentralized Lifetime Maximizing Tree using Clustering based energy and distance (DMLTC), which creates trees within the clusters already created .The choice of the tree is based on the minimum distance of the sub-sink from the sink.

The tree that we get after application of both the algorithms is efficient in terms of distance as well as energy, now to improve the bandwidth utilization we apply a method of FDMA_SINK ( ) that allots range of frequencies to the sub-sink for data transfer and also checks for the efficient utilization of bandwidth by assigning half frequency range from a low data transfer cluster to a new one.

Finally, after frequency allotment we implement HyMac [10] algorithm which provides fixed time slots to nodes to transmit sensed data, the sub-sink remains in a LPL state and listens for receiving data from the children, the child nodes awake once and start the synchronous data transfer to the sub-sink which further sends sensed data to the sink in the same assigned time slot.

4. PROPOSED ALGORITHM

4.1 EM Algorithm

The following algorithm that we use divides the network into K clusters. It has been renamed from EM [9] to EMD (Expectation Maximization on Distance) Algorithm.

K : The number of clusters

πk: The mixing coefficients of the kth cluster

μk: The 2-dimensional vector indicates the mean of the kth cluster

Σk: The 2 × 2 covariance matrix of the kth cluster

Algorithm 1: EMD Algorithm

The mobile sink node groups all nodes into K clusters by using the EMD algorithm in the following manner.

1: Initialize μ, Σ ,π and the convergence criterion θEM , and evaluate the initial value of P:

$$P = \sum_{n=1}^{N} ln\left\{\sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)\right\}$$

Where N is the number of nodes.

The above presented algorithm groups the sensor nodes into clusters based on their distance and hence ensure that the nodes in a cluster have a close proximity which will lead to minimization of data transfer delays. The algorithm can be iteratively used to account for any new nodes coming up in the network.

4.2 Decentralized Life Maximizing Tree with Clusters: Algorithm

After the nodes have been grouped into clusters using EMD algorithm, our next step is to apply DLMT [8] in a new form, DLMTC which further takes into consideration both the energy and the distance factors. It constructs trees within the clusters considering energy factor but the final choice of the tree is done by taking into account the distance of the sub-sink or parent from the sink node. The algorithm can be applied iteratively to compensate for the broken links or when a new node is added to the existing network. The presented tree construction leads to minimum delays in data transfer as well as judicious use of the available energy of the nodes.

4.3 Proposed DLMTC Algorithm:

This algorithm compares the previous tree constructed based on their distance from the sink and chooses the suitable one with minimum distance from the sink.

BestDLMTC (DLMTCi, DLMTCj, di, dj)

1. if rows in (DLMTCj) > rows in (DLMTCi)
2. return true
3. if [rows in (DLMTCj) = rows in (DLMTCi)] and [DMLTEj > DLMTC_Ei]
4. return true
5. if [rows in (DLMTCj) = rows in (DLMTCi)]and [DMTLEj = DLMTC_Ei] and [tree depth of DLMTCj < tree depth of DLMTCi }
6. return true
7. if [rows in (DLMTCj) = rows in (DLMTCi)] and [DMLTEj = DLMTC_Ei] and [tree depth of DLMTCj = tree depth of [DLMTCi] and [ej > ei] and [dj < di]

8. return true

9. if [rows in(DLMTCj) = rows in(DLMTCi)] and [DMLTEj = DLMTC_Ei] and [tree depth of DLMTCj = tree depth of DLMTCi] and [ej = ei] and j < i

10. return true Else if return false.

4.4 Applying FDMA_SINK on the DLMTC Tree

Scheduling using FDMA System Model:

All the frequencies available in the band are divided in frequency ranges(R) based on the number of clusters (K), initially, all sub-sinks send a synchronous message to sink requesting allotment of the frequencies, sink then checks for the availability of the frequency ranges. If it is available it is assigned. Otherwise sink may withdraw a frequency from some other cluster if its data transmission rate is low and assign that to some cluster. We propose an algorithm FDMA_SINK () with parameters (K, number of clusters, Fi, frequency band, i ,Cluster),SSi is subsink of ith cluster.

Algorithm 3: FDMA_SINK (DLMTC Treei,Ki,Fi,Ci)

1. Calculate the range Ri = Fi/Ki

2. For each S Si $\in$ Ci where i $\leftarrow$ 1 to k

3. Assign Ri

4. If SSi receives the last packet then

Withdraw the frequency and add it to the free pool.

5. End for

6. Request for new Fi allotment

7. If available from free pool, Assign from free pool

Else withdraws half the frequency range from a low data rate transfer cluster and assign to the subsink.

4.5 Cluster Scheduling Using HyMac

Scheduling algorithm is applied on the DLMTC tree having the base node as its root. As each node Ni is traversed by DLMTC, it is assigned a default time slot and a frequency using FDMA_SINK () function discussed before. Then the possibility of having an interference with any of its same-height previously-visited one-hop AND two-hop neighbours is checked. If a conflicting neighbour Nj is found for Ni, the algorithm checks whether Ni and Nj are siblings. If so, Ni will be assigned a different time slot than that of Nj. If they are not siblings then Ni will be assigned a different frequency than that of Nj , allowing both Ni and Nj to send messages to their parents at the same time slot but in different channels. When DLMTC is about to start a new level (height) of nodes the default time slot number will be increased by one. Once all nodes are processed according to the above heuristic, the entire time slot assignments will be inverted such that the slot number assigned to every node is smaller than that of its parent. This inversion is done as following:

$$t_{new} = t_{max} - t_{current} + 1 \quad (1)$$

Algorithm 4: HyMAC with clusters Scheduling Algorithm

```
Require: A Graph of Sensor Network Topology
Ensure: An scheduled Tree of the Given Network
1: ENQUEUE (Q, S)
2: while Q is not empty do
3:      v ← DEQUEUE (Q)
4:      timeSlot[v] ← currentT imeSlot
5:      FDMA_SINK () /* assigning channels */
6:      for all Visited same-height 1-2-hop nbr n of v do
7:          if parent[n] == parent[v] or #Channel >= available chnls then
8:          if timeSlot[v] = timeSlot[n] then
9:          timeSlot[v] ← timeSlot[n] + 1
10:         end if
11:         else
12:         if timeSlot[v] =timeslot[n] and channel[v]=channel[n] then
13:         channel[v] ← channel[n] + 1
14:          end if
15:         end if
16:     end for
17:     for all unexplored edge e of v do
18:         let w be the other unvisited endpoint of edge e
19:         parent[w] ← v
20:         height[v] ← height[w] + 1
21:     end for
22: end while
```

Where t new is the new inverted assigned slot, tcurrent is the current slot number assigned to the node and tmax is the total number of assigned slots. Note that such an assignment allows the data packets to be aggregated and propagated in a cascading manner to the base station in

a single TDMA cycle. The complete steps of the overall process are presented in algorithm 4.

FINAL ALGORITHM:

Require: Set of sensor nodes

Step 1: Apply EMD () /* Creates clusters */

Step 2: Creating decentralised Maxlife tree using DLMTC ()

Step 3: Choosing the life maximising tree using Best DMLTC ()

Step 4: Scheduling Using HYMAC () which calls FDMA_SINK for scheduling based on time slots and frequency range.

5. An Illustration of our approach:

After the cluster formation using EMD algorithm, we apply DLMTC [9] on the given cluster to create decentralized MaxLife tree (DMLTree) based on their energy. After applying this algorithm we get several trees with different parents, choice of which depends on the energy of the node and distance of the node from the sink. When a node with maximum energy and minimum distance is found the node is taken as SubSink node where data aggregation is then performed. The concept is shown by assuming the trees formation in figure 1-4.

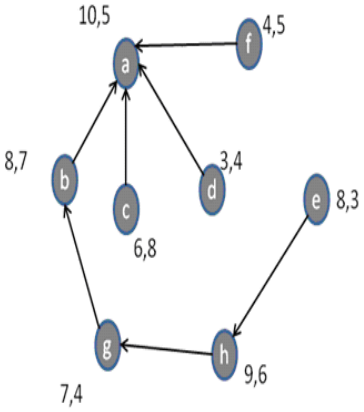


Fig.1 Tree construction within cluster C1 with a as parent

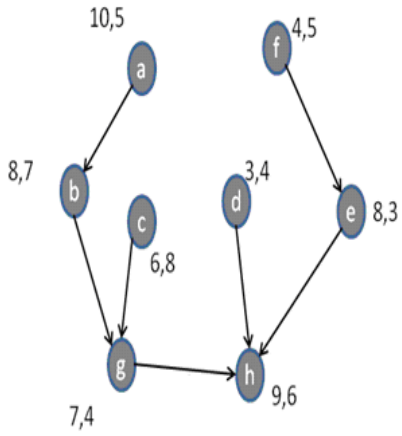


Fig.2 Tree construction within cluster C1 with h as parent

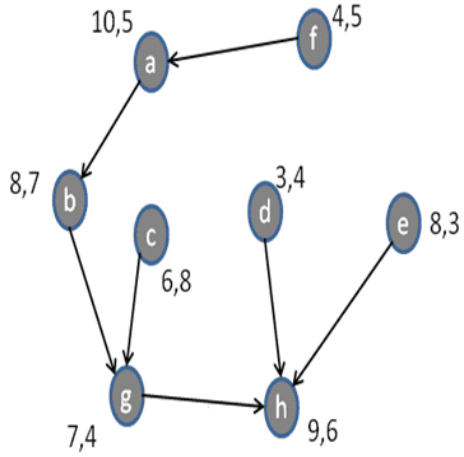


Fig. 3 Tree construction within cluster C1 with g as parent

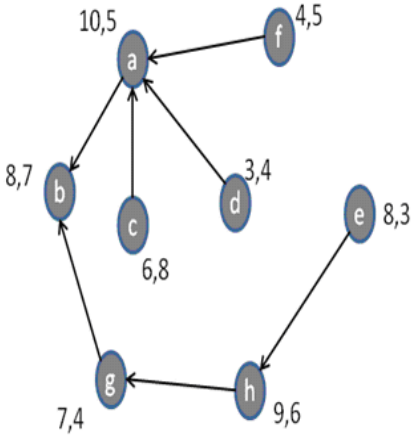


Fig. 4 Tree construction within cluster C1 with b as parent

The best tree found here is the one with node 'a' as the parent (fig.1); hence the node is taken as a SubSink for the cluster C1.

6. Simulations Parameters

We implemented our tree construction modules on top of Forwarded Diffusion in the J-Sim network simulator (the J-Sim comes with diffusion support). In all of our experiments, a square sensor field with each side measuring X meters is being considered. A number of N identical nodes, ranging from 50 to 300 in the increment of 50, are randomly deployed in this sensor field such that the average node density is kept at $\lambda = 55/1652$ nodes per meter square, a parameter which we borrowed from Forwarded Diffusion [10, 11]. Furthermore, there are five sinks randomly deployed in the field and sources are randomly chosen among the nodes, subject to the conditions that SR=10% of N and the sources have to be interconnected to each other (to model a single stimulus). Each node is assumed to have a radio range of 45 meters. We considered an event-driven data sensor network throughout all our experiments. To model the periodic transmissions, each source generates random data reports of size fixed at 138 bytes in constant intervals of DR = 1 packet/second. To introduce some randomness, data start to be generated only after a time randomly chosen between t = 0 to 5 seconds. The data are collected at the root, if they exist, and are sent to the sinks. We assign each source with an initial energy that is randomly chosen between 10 to 18 Joules in order to keep the total simulation time at a reasonable limit. In all of our experiments, all other nodes are given an initial energy that is greater than that of any event source such that their absence in the network, due to energy depletion, does not affect the functionalities of any participating sources during data collection. Lastly, the idle time power, receive time power and transmit power dissipation are set at 40, 400 and 680 mW respectively. We assume a negligible energy cost to process and aggregate incoming data reports. To trace the energy, an application that logs the residual energy of each node in constant intervals of 550 ms is employed. The J-Sim simulator implements a 1.6 Mbps 802.11 MAC layer. Since Forwarded Diffusion is chosen as our routing platform,

6.1 Average Dissipated Energy (ADE): ADE measures the average amount of energy consumed throughout the entire simulation. This metric computes the average work done in delivering periodic data to the sink/ Root over a simulation run. As shown in fig. 5 energy conservation is more with DLMTC as compared to DLMT that is with clusters we are able to preserve more energy.

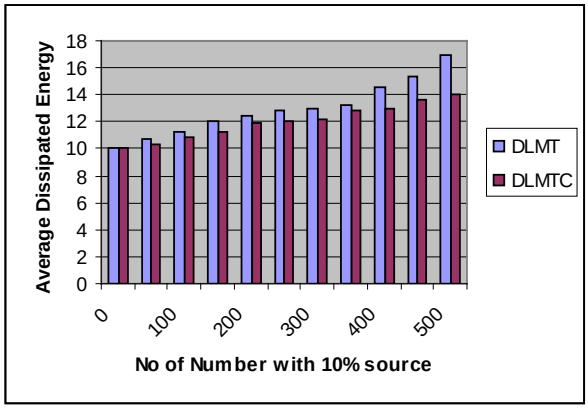


Fig. 5 Average dissipated energy

6.2 Average Network Lifetime (ANLT):

In order to study the impact of DLMT, CLMT and E-Span on the lifetime-savings, we measure the node lifetime of each source as a function of network size for DLMT, CLMT and E-Span respectively. Each node is assigned number of with an initial energy that is randomly chosen between 12 to 18 Joules so as to limit the total simulation time at a controllable range. Reference [13] shows that DLMT enhances the network lifetime in comparisons with other trees constructed for the same purpose. But now these simulation results shown in fig. 6 we are able to prove that including clusters in the DLMT makes the node alive for enhanced time.

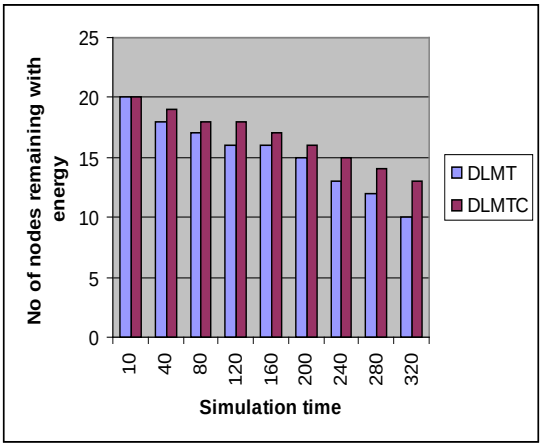


Fig.6 Average network lifetime

6.3 Average Delay

Average delay measures the delay between transmitting data from each source to each of the sinks. This is the

basic limitation of DLMTC average delay is maximized as extra time is required in setting cluster head and other parameters. Once the clusters are formed the process speeds up and energy is preserved. Fig. 7 shows the results that at start of simulation delay is maximum as the process of formation of clusters is going on , once the cluster formation is done then the process speeds up minimizing the delay.

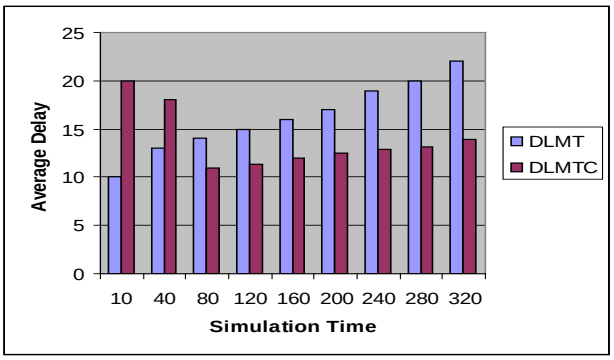


Fig. 7 Average delay

6.4 Bandwidth Utilization

As bandwidth allocation with DLMTC is based on allotment of frequency from the free pool of frequencies or by withdrawing half frequency from cluster with low data transfer rate. Results as shown in fig. 8 prove that bandwidth utilization is almost double for DLMTC as compared to DLMT.

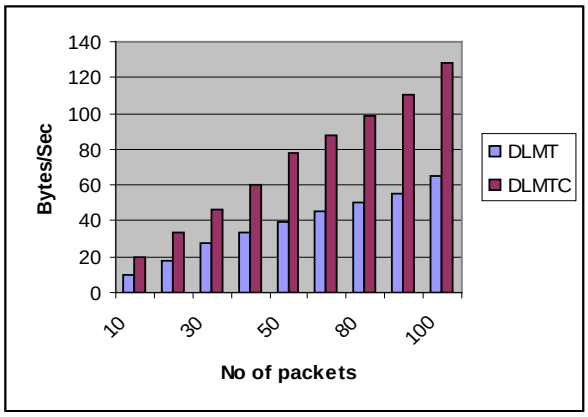


Fig. 8 Bandwidth Utilization

7. Conclusions

In this paper we proposed the Decentralized Lifetime Maximizing Tree with clustering construction algorithm. Clustering on the basis of distance ensures close proximity of the nodes and thus leads to reduction in data transfer delays. Energy conservation by waking nodes once instead of twice leads [5] to further reduction in data transfer delays, thus utilizing the energy available effectively. Efficient utilization of bandwidth achieved through allotment of frequencies from the free pool or withdrawing half frequency from cluster with low data transfer and assigning it some other. Efficient utilization of energy is achieved by using HyMAC technique. Simulation results prove enhancement if network lifetime, reduction in energy consumption and minimization of average delay.

References

[1] HYMAC: Hybrid TDMA/FDMA medium access control protocol for WSN By Mastooreh Salajegheh, , Hamed Soroush, IEEE, 2007.

[2] An Energy efficient pre-schedule scheme for Hybrid CSMA/TDMA MAC in WSN By Wei Wang, Honggang Wang, IEEE, 2006.

[3] Energy aware routing in Cluster based networks By Mohamed Younis, IEEE, 2002.

[4] PEDAMACS: Power Efficient and Delay Aware Medium Access Protocol for Sensor Network By Sinem Coleri Ergen, IEEE, 2006.

[5] Energy efficient wake up scheduling for data collection and Aggregation By Yanwei Wu, Member, IEEE, Xiang-Yang Li, Senior Member, IEEE, YunHao Liu, Senior Member, IEEE, and Wei Lou .

[6] Attribute aware data aggregation using dynamic routing in WSN By Jiao Zhang , Fengyuan Ren, Tao He , Chuang Lin ,2010 IEEE.

[7] A Scheduling algorithm for TDMA-based MAC protocol in WSN By Yan Zhang, Shijue Zheng, IEEE 2009.

[8] Construction Of Decentralized Lifetime Maximizing Tree for Data Aggregation in Wireless Sensor Networks By Deepali Virmani , Satbir Jain ,PWASET VOLUME 40 APRIL 2009 ISSN 2070-3740 .

[9] An Efficient Data Aggregation Scheme Using Degree of Dependence on Clusters in WSNs By Tetsushi Fukabori, Hidehisa Nakayama, Hiroki Nishiyama, Nirwan Ansari, and Nei Kato Graduate School of Information Sciences, Tohoku University, Sendai, Japn Tohoku Institute of Technology, Sendai, Japan Advanced Networking Lab., ECE Department, New Jersey Institute of Technology, Newark, NJ, USA.

[10] Wireless Sensor Networks– Technology, Protocols and Application ebook By Kazem Sohraby, Daniel Minoli, Taieb Znati.

[11] Data Aggregation Techinques in Sensor Networks By Karthikeyan Vaidyanathan, Sayantan Sur, Sundeep Narravula, Prasun Sinha.

[12] AIDA: Application independent data aggregation in wireless sensor networks. In ACM Transactions on Embedded Computing System Special issue on Dynamically Adaptable Embedded Systems, 2003 By T. He, B. M. Blum, J. A. Stankovic, and T. F. Abdelzaher.

[13] Deepali Virmani and Satbir Jain "Performance analysis of lifetime maximizing trees for data aggregation in wireless sensor networks" International Journal on Computer Science and Engineering, Vol .3 No.1 Jan 2011, ISSN: 0975-3397, pp 276-285.

Proposing Cluster_Similarity Method in Order to Find as Much Better Similarities in Databases

Mohammad-Reza Feizi-Derakhshi[1] and Azade Roohany[2]

[1]Department of Computer, University of Tabriz, Tabriz, Iran

[2] Department of Computer, Shabestar Branch, Islamic Azad University, Shabestar, Iran

Abstract

Different ways of entering data into databases result in duplicate records that cause increasing of databases' size. This is a fact that we cannot ignore it easily. There are several methods that are used for this purpose. In this paper, we have tried to increase the accuracy of operations by using cluster similarity instead of direct similarity of fields. So that clustering is done on fields of database and according to accomplished clustering on fields, similarity degree of records is obtained. In this method by using present information in database, more logical similarity is obtained for deficient information that in general, the method of cluster similarity could improve operations 24% compared with previous methods.

Keywords: *Clustering, cluster similarity, Record similarity, Field similarity*

1. Introduction

Nowadays the size of databases is increasing by developing of information and advancing of technology and requirement for proper and accurate restoration of necessary information has become an important issue in this field. One of the matters that are introduced in most sources is the lack of compatibility of identical data in databases that despite of same meaning, they stored in different shapes that is resulted from improper entering of data such as type errors, the way of speaking, abstraction and etc.

Duplicate records are unfavorable [1]. The main fact is that how we can eliminate similar records. Such fact is called record linkage or record matching. This is the task of accurate labeling of pair records that are related to same entity from different sources [2]. In other words the aim of one record linkage algorithm is to detect records which do not have complete matching, but they have some similarities. By finding similar records, we can combine them that help to decrease the size of databases. In the next sections of present paper, first we will discuss about finding of similar records and previous methods and in the next section we will describe the steps of proposed algorithm and finally we will present conclusion.

2. Finding similar records

We can identify and combine similar records by using some methods in order to minimize the size of databases. So first, field matching algorithms take two fields as an input and then they return their similarity in the numerical format between one and zero. After that the detection of similar records among records of database is done according to obtained numbers and finally clustering is done based on obtained similarities of records.

One of the previous algorithms that we need is *Jaro* algorithm that is used for field similarity. One clustering algorithm is also needed in order to use single linkage method that will be described.

2.1 Jaro Distance Metric

Jaro introduced a string comparison algorithm that was mainly used for comparison of last and first names. The basic algorithm for computing the Jaro metric for two strings S1 and S2 includes the following steps:

1. Compute the string lengths |S1|and |S2|.
2. Find the "common characters" c in the two strings; common are all the characters S1[j] and S2[j] in (1)

$$|i-j| <= 1/2 \min \{ |S1| , |S2| \} \quad (1)$$

3. Find the number of transpositions t; the number of transpositions is computed as follows: We compare the ith common character in S1 with the ith common character in S2. Each non matching character is a transposition. The Jaro comparison value is calculated by equation (2).

$$Jaro(|S1|,|S2|) = \frac{1}{3}\left(\frac{c}{|S1|} + \frac{c}{|S2|} + \frac{c - \frac{t}{2}}{c}\right) \quad (2)$$

From the description of the Jaro algorithm, we can see that the Jaro algorithm requires O(|S1| *|S2|) time for two strings of length |S1|and |S2|, mainly due to Step 2, which computes the "common characters" in the two strings. Winkler and Thibaudeau modified the Jaro metric to give higher weight to prefix matches since prefix matches are generally more important for surname matching [1,3].

2.2 Single linkage method

Single linkage method is one of the oldest and simplest clustering methods and also is one of the hierarchical and individual clustering methods. We can also call it nearest neighbor and either connectedness method or minimum method. That by assuming that B and A are two clusters, according to figure 1, distance d(A, B) equals to at least the distance between correspondent patterns of B and A that is calculated by equation (3).

$$d(A,B) = \min_{i \in A,\, j \in B} d(i.j) \quad (3)$$

In this method the nearest distance between two clusters is considered. Clustering based on this distance is one of the most common methods in clustering. Since this algorithm is hierarchical, when clusters are combined in order to form new clusters, it erases correspondent rows and columns in the adjacency matrix [4].

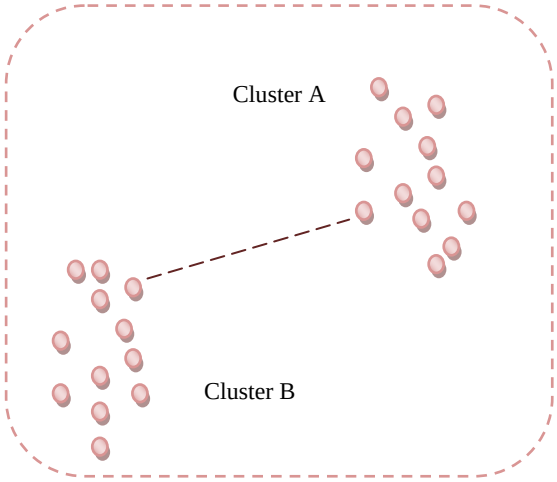


Fig. 1 Single linkage method.

3. Proposed method

In previous methods, field similarities were concerned directly and other operations were done on this obtained method, but in this method (Cluster Similarity Method) we will use cluster similarity of field in order to find the duplicate records. Initially the similarity of field is evaluated, and then clustering is done on them. In other words, each field is converted to some clusters, then similarity of records is calculated according to clusters of each group of fields and finally the last clustering is done in order to determine which records are placed in a cluster and which of them are similar.

3.1 Cluster similarity

Type errors and other cases result in duplicate records. Some of these defects are related to deficient entering of data, so that for example if one two-part word is in the form of AB and it is entered either A or B that their field similarity becomes like figure 2, in fact when we consider field similarity, AB has 0.5 similarity with A, but A does not have any similarity with B that is AB, when we consider cluster similarity, A and B are in one database, the fields of A and B are located in one cluster and in the next step, the more meaningful record similarity is gained and in fact this option decreases the rejection of error. In fact we fine some similarities by using present data in database that it seems that they do not have any similarity but they are the same.

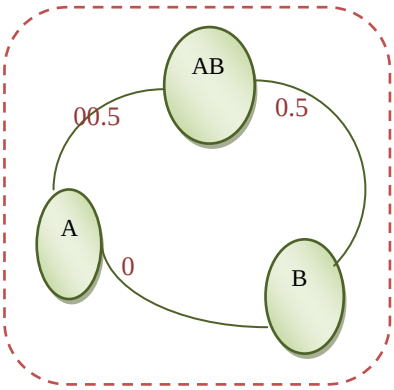


Fig. 2 Schematic presentation of field similarity.

In "Cluster Similarity Method", the cluster similarity is used in order to fine proper and meaningful similarity, so that it maximizes the accuracy and could find records which do not have more similarity and put them in one cluster. Clustering is done on each field, that each field is divided into a set of clusters. It may be the number of created clusters differs from one field into another field, for example that may become the name of 10 clusters and family name of 18 clusters. After finishing the clustering of fields, it may have the form of figure 3.

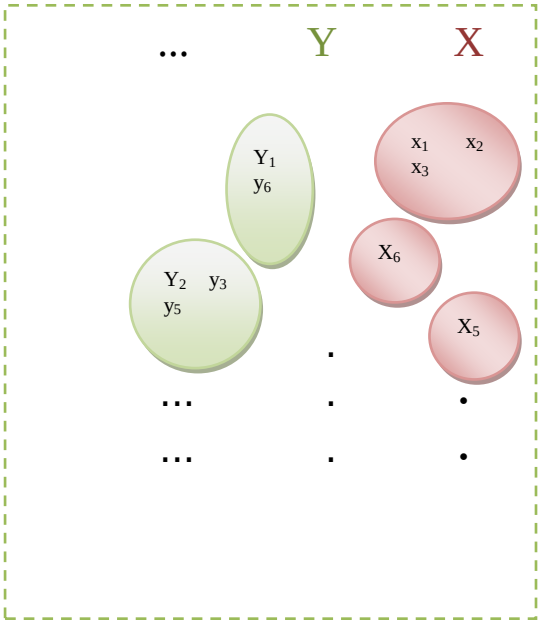


Fig.3 The way of fields' clustering.

In figure 3 the parameters of X and Y is the representative of the present fields, example: X= name field and Y= last name field. Also x1, x2, x3 … related to X show correspondent records of that field, by assuming that x1 means the value that is in the field x and in record 1 and so on.

In "Cluster Similarity Method", we have three types of clustering: invalid cluster, empty cluster and valid clusters. If the present value in one field is invalid, it locates in invalid cluster. Invalid value means for example there is value of 125 in the name field that is invalid and or if there is the value of "ssss" in the field of national code that is invalid. If the field is empty and no value was entered, it locates in empty cluster. Other values that have proper format will locate in valid clusters that in this group the values are clustered based on their similarities.

3.2 Record similarity

After accomplishing the clustering on each field, in this step decision making is done according to cluster similarity about record similarity and the extent of similarity among records is specified.

The extent of record similarity is given by equation (4) that is a number between one and zero that one means completely same and zero means there is not any similarity.

$$Sr = \sum_{i=1}^{f} \frac{1}{K} Sc_i Df_i \qquad (4)$$

where Df shows the importance of fields that is calculated according to equation (7) and Sc show the degree of cluster similarity that is a number between one and zero that zero means these two clusters do not have any similarity and one means these two valued are in one cluster and also f means the number of present fields in a record. The value of K is determined according to the number of fields that are in invalid, empty and valid clusters and in other words, it is determined based on the validation of fields and it is calculated according to equation (5).

$$K = \sum_{i=1}^{f} k_i v_i \qquad (5)$$

f shows the number of present fields in one record and vi equals to the weight that is given to each field and the range of ki in this equation is between zero and one that is initialized according to the type of cluster by using equation (6).

$$k_i = \begin{cases} 0 & \text{Invalid} \\ 0 & \text{Empty} \\ 1 & \text{Valid} \end{cases} \quad (6)$$

The second component of equation (4) is Df that shows the importance extent of fields that is given by equation (7)..

$$Df_i = k_i v_i \quad (7)$$

That in this equation ki is obtained by equation (6) and vi equals to the weight that is given to each field. Since different fields have different values, we retain this balance by putting weight. For example in a database the value of family name may be more than address, so we give more weight to family name.

3.3 Record similarity

After finding the degree of similarity among records, clustering on records is done, so that similar records are located in one cluster. Selective clustering is a single linkage clustering that has better accuracy compared with other methods.

4. Conclusions

We introduce 3 criteria for evaluation that include R (Recall), P (Precision) and F1 (F-measure). Since final evaluation has been done on clustering, criteria also discuss both on the number of proper clusters and improper clusters, so that P and R are given by equations (9) and (10), respectively. And finally the value of F1 is calculated based on P and R like equation (10) [5].

$$R = \frac{\text{Number Of Common Clusters}}{\text{Number Of Manual Clusters}} \quad (8)$$

The number of common clusters means how many clusters are there to make them same either manually or programmatic. The number manual clusters shows that these records were clustered in how many clusters manually.

$$P = \frac{\text{Number Of Common Clusters}}{\text{Number Of Program Clusters}} \quad (9)$$

The number which is considered to the number of program clusters equals to the number of clusters that program form after final clustering.

$$F1 = \frac{2 * P * R}{P + R} \quad (10)$$

In hierarchical clustering, initially each data is put into one cluster and in each step the near clusters are combined in order to reach one unit cluster. Here we want to combine clusters to some extent that this amount shows that to somewhat similar clusters will be combined. In other words, it shows the maximum extent of similarity that clusters should have in order to combine. The proper selection of extent value has great influence on results, so one other parameter that is introduced here is Te and it shows the stop condition for single linkage clustering algorithm. The range of this value is between zero and one. We can obtain the proper value of Te for clustering by giving different values to Te and final testing of R, P and F1 that the results of these tests were shown in figures 4 and 5.

According to figure 4, as the value of Te increases up to 0.8, the value of F1 increases, so that in Te=0.8, the value of F1 reaches its highest value and when the value of Te becomes more that 0.8, the value of F1 decreases gradually. Therefore for “Cluster Similarity Method” Te=0.8 was set. Also according to figure 5, the value of Te in previous method in 0.8, obtained the highest value of F1, so for this method Te=0.8 was selected too.

We clustered database records in two ways, both with “Cluster Similarity Method” and with one of the previous methods which was described in section 2 and we presented the results in table 1. As table shows, the “Cluster Similarity Method” has better accuracy and F1 compared with previous method that approximately 24 percent improvement was reached.

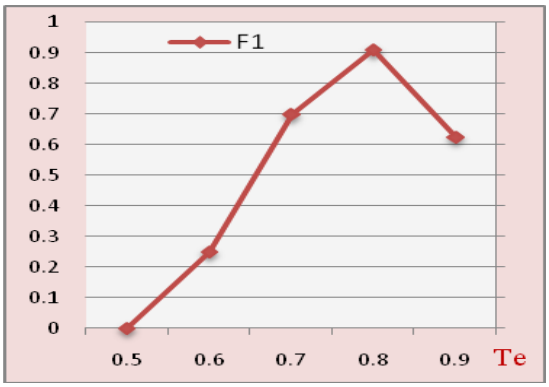


Fig. 4 Determination of stop condition's value (Te) for "Cluster Similarity Method".

Fig. 5 Determination of stop condition's value (Te) for previous method.

Table 1: The results of methods' comparison

	F1	*P*	*R*
Cluster Similarity Method	0.91	0.89	0.94
Previous method	0.65	0.62	0.67

5. Conclusions

In this method, cluster similarity was used; cluster similarity use present data in database in order to find similar fields and it is not based on direct similarity of fields. So it resulted in more logical clustering of similar records. Also the accuracy of calculation was increased, so that the accuracy of 80 to 95 percent was gained.

References

[1] K. Elmagarmid, P. G. Ipeirotis and V. S. Verykios, "Duplicate record detection: A survey" IEEE Trans. on Knowledge and Data Engg., vol. 19, no. 1, pp. 1–16, 2007
[2] F. Maggi, "A Survey of Probabilistic Record atchingModels, Techniques and Tools", Scienti_c Report TR-2008.
[3] A. Furer, "Combining Runtime and Static Universe Type Inference" Master Project Report, Software Component Technology Group Department of Computer Science ETH Zurich, 2007.
[4] F. Mali, S. Mitra, "Clustering of symbolic data and its validation", Advances in Soft Computing, 2002.
[5] J. B. Santos, C. A. Heuser, V. P. Moreira and L. K. Wives, " Automatic threshold estimation for data matching applications", Elsevier, information sciences, 2010.

Mohammad-Reza Feizi-Derakhshi: was born in 1975. He received the B.Sc. degree in Computer Engineering from Isfahan University, Isfahan, Iran, in 1997, the M.Sc. degree in Computer Engineering from Iran University of Science and Technology, Tehran, Iran, in 2000, the Ph.D. degree in Computer Engineering from Iran University of Science and Technology, Tehran, Iran, in 2007. Since 2007, he is an assistant professor in the University of Tabriz, Tabriz, Iran.

Aazde Rohany: was born in 1980. She received the B.Sc. degree in Computer Engineering from Islamic Azad University, shabestar branch, Shabestar, Iran, in 2008, the M.Sc. degree in Computer Engineering from Islamic Azad University, shabestar branch, Shabestar, Iran, in 2011.

Survey on Power Optimization for Disk Based Systems

G. Ravikumar[1] and Dr.N. Nagarajan[2]

[1]Coimbatore Institute of Engineering and Technology, Coimbatore.

[2]Coimbatore Institute of Engineering and Technology, Coimbatore.

Abstract

Energy optimization has become a growing concern in the present world. Energy optimization can influence the overall system design and reliability. Power can greatly influence the performance of the disk, as power dissipation generates heat that affects stability and reliability of the component, particularly for large server systems. Hence, developers concentrate on the configuration of disk arrays which can deliver extremely high performance. Though, there are several significant techniques for tackling disk power for laptops and workstations, using them in a server environment are a considerable challenge, especially under stringent performance needs. Excessive power consumption is a major barrier to the market acceptance of hard disks in mobile electronic devices. Studying and reducing power consumption, however, often comprises running time intensive disk traces on real hardware with specialized power-monitoring equipment. Most of the conventional energy optimization techniques are based on architectural level techniques and is found to be effective only in certain scenarios. This paper proposes a survey on the disk energy optimization techniques. This paper analyses the functionalities, advantages and the disadvantages of the various techniques for the disk power consumption.

Keywords: *Power Dissipation, Traditional Power Management (TPM), Finite-Element (FE) Model, Flash Memory, Data Prefetching*

1. Introduction

The performance has been the main aim when designing hardware and software. Disk drive [23] [25] designers try to construct faster hard disks with denser storage capacity, while software designers adjust their functions for peak performance on a given hardware platform.

Recently, power dissipation has become a rising anxiety. Though, conventionally regarded as a problem for mobile, battery-operated systems, power also creates disputes in terms of electricity charges and overall system design and reliability.

Studies and investigations on disk power management can be an annoying procedure, as important disk traces take long time to run, and researchers need access to expensive power-monitoring equipment.

In majority of the single-disk applications, Traditional Power Management (TPM) consists of two steps. TPM initially identifies appropriate idle periods, then spins down [27] the disk to a low-power standby mode when the power management approach forecasts that doing so will save energy. But, spinning the disk back up from this standby mode when the system receives an I/O request incurs additional latency and power costs.

Organizing power in single-disk systems such as laptops and desktops has been an extensive area of research [19, 20]. There are various conventional techniques proposed in the literature for disk power consumption. Most of the traditional techniques are based on architectural mechanism like spinning down idle disks [21] [22] or rotating disk with reduced speed. Even though, these implementing them still presents challenges that need additional research. Excessive power consumption is a main blockade to the market acceptance of hard disks in mobile electronic devices.

Thus, due to the importance of the power consumption and its influence in the current market, there are several approaches developed by various researchers for the energy optimization [26] of the disks. This paper deals with the analysis of the various existing approaches for disk energy optimization.

2. Literature Survey

Power utilization of disk systems is a significant issue in systematic computing where data-intensive applications exercise disk storage comprehensively. At the same time one can turn inoperative disks [24] when idleness is identified, turning up them takes many cycles and utilizes additional power. Consequently, it can be extremely helpful when put into practice to enhance the re-usage of disk, specifically using the same set of disks as much as feasible. When it is possible to put into practice, the unexploited disks can be detained in the so called spin-down mode for maximum length of time, and this assists in enhance power savings.

Kandemir et al., [1] suggested a method for minimizing the disk power utilization by enhancing re-usage of disk. The proposed technique reorganized a particular application code by taking only the disk boundaries of the datasets it manipulate into consideration. The disk layout-conscious technique executed inside an openly-accessible compilation structure and evaluated it against a traditional data reuse optimization technique that is also executed by utilizing the same compiler using six scientific applications that carry out disk I/O. The consequences gathered until now specifies that this layout-conscious technique and the traditional data reuse optimization technique decreases the disk energy utilization by 25.3% and 10.3%, correspondingly, typically, over the case where no disk power optimization is functional. The equivalent savings in overall energy consumption together with CPU, memory and network energies are approximately 6.5% for the traditional technique and 16.5% for this disk layout-conscious technique. The experimental estimation also proves that the savings achieved are reliable with varying number of disks and any other disk layouts.

Disk subsystem is recognized to be a most important contributor to the entire power consumption of high-end parallel systems. Earlier researches recommended numerous architectural-level approaches to diminish disk power by captivating benefit of idle time period experienced by disks. Even though such approaches have been identified to be very successful in some cases, they all have a general disadvantage: they function in a reactive way, i.e., they manage disk power by examining previous disk movement (for instance, idle and active periods) and approximating future ones. As a result, they can fail to notice the chances for consuming power and gain noteworthy performance consequences because of inaccuracies in forecasting idle and active time periods. Inspired by this examination, Seung Woo Son et al., [2] proposes and estimates a compiler-driven technique to reducing disk power utilization of array-based scientific applications accomplishing on parallel systems. The proposed technique exposes disk layout data to the compiler, permitting it to obtain the disk access pattern, specifically the arrangement in which parallel disks are accessed. This technique also reveals two uses of this data. The first use is that is easy to put proactive disk power management into practice, to be precise, choose the most suitable power-saving approach and disk-preactivation approach according to the compiler-predicted future idle and active periods of parallel disks. The second use is that it is simple to reorganize the application code to enhance the time-span of idle disk periods, which shows the way to improved utilization of available power-saving capabilities. Both these approaches are employed in this technique inside an optimizing compiler and tested their efficiency with the help of a set of benchmark codes from the Spec 2000 suite and a disk power simulator. Experimental consequences show that the compiler-driven disk power management is a very potential method. The experimental outcome also exposes that, even though proactive disk power management is very efficient, code restructuring for disk power accomplishes further energy savings across all the benchmarks examined, and these savings are extremely near to best possible savings that can be achieved through an integer linear programming (ILP)-based method.

Enhancing security and reducing power utilization are essential for large-scale data storage organizations. Even though numerous studies have been concentrated on data protection and energy efficiency, the majority of the available techniques have concentrated on just one of these two metrics. Shu Yin et al., [3] presented a novel technique to incorporate power optimization with security services to boost the security of energy-efficient large-scale storage organizations. In this approach, the dynamic speed control for power management procedure is used, or DRPM, to preserve energy in protected storage systems. The author had given two manners of incorporating privacy services with the dynamic disk speed control method. The first method - security aggressive in nature - is mostly concentrated on the enhancement of storage system security with fewer importances on energy preservation. The second method provides advanced precedence to energy preservation as different to the security optimization. The experimental outcome shows that the energy-aggressive method offers better energy savings than the security-aggressive method. On the other hand, the superiority of security realized by the security-aggressive method is advanced than that of the energy-aggressive method. Furthermore, the observed result shows that energy savings yielded by the two methods turn out to be more distinct when the data size is larger. The result demonstrates that the response time of the security-aggressive method is more responsive to data size than that of the energy-aggressive method.

Storage finds an essential role in the performance of a lot of applications. Numerous applications, particularly those that run on servers, are I/O concentrated and so need better performance storage systems. These high-end storage systems use a large quantity of power, the most quantity of which is because of the disk drives. Optimizing disk structural designs is a design time with run time concern and necessitates balancing between performance and power. There are dissimilar figures of advantage, for instance performance and energy, and a huge space of design and runtime "knobs" that can be utilized to optimize disk drive performance. Specified such a huge space, it is desirable to have an organized method to optimally set these knobs to assure the figures of advantage as resourcefully as achievable. Sankar et al., [4] present the sensitivity-based optimization technique for disk architectures (SODA), which influences results before obtained in digital circuit design optimization circumstances. By means of comprehensive models of the electro-mechanical manners of disk drives and a suite of practical workloads, SODA can assist in design and runtime optimization of disk drive structural designs.

The dynamic voltage and frequency scaling method in CPUs is an illustration of regulating a device's control variable to exchange power consumption and performance. This inspiration of energy optimization by means of speed control has been then applied to additional components of electronic systems such as disk drives and wireless transceivers. The energy-optimal speed profile (a function of time) of a common device that has to carry out a specified task in a certain time is obtained systematically. Ravishankar Rao et

al., [5] proposed technique is technique to devices with either distinct or continuous-speed sets. The most significant improvement of the technique is that for discrete-speed sets, the environment of the fundamental continuous power-speed association does not require to be known. The discrete power-speed data points just require convincing a W-convex relation: a discrete analog of a convex function. According to the observation that the majority of devices have W-convex power-speed associations, it is exposed that the optimal speed profile utilizes at most one speed for permanent speeds or two speeds. Additionally, each device has a built-in speed (self-sufficient of the task) uc at which it uses the smallest amount of energy per unit work completed. It is revealed that this speed can be computed directly from measured values of power-speed data points (for distinct-speed sets) or by an investigational line search process where each step engrosses determining a power-speed data point for continuous-speed sets. In whichever case, no curve fit or information of analytical power models is essential. The most favorable speed profile was revealed to be either uc or the smallest feasible speed(s) for the particular task, with the option depending on the energy overheads and task constraints.

Significant performance, high reliability and energy-proficient storage systems are very vital for mobile data-intensive applications such as remote surgery and mobile data center. Mobile disk-array-based storage systems are more liable to disk malfunctions than with traditional stationary storage systems. This is mainly because of their complicated application environments. Moreover, Mobile disk-array-based storage has very inadequate power supply. Hence, data reconstruction techniques, which are carried out in the existence of disk malfunctions, for mobile storage systems must be performance-driven, reliability-aware, and energy-efficient. Existing reconstruction approaches cannot accomplish the three objectives concurrently as they mostly overlooked the information that mobile disks have much superior failure rates than stationary disks. In addition, they generally disregard energy-saving. In this paper, Tao Xie et al., [6] proposed a novel reconstruction approach, called Multi-level Caching-based Reconstruction Optimization (MICRO), which can be used to RAID-structured mobile storage systems to obviously cut down reconstruction times and user response times while saving energy. MICRO collaboratively uses storage cache and disk array controller cache to lessen the number of physical disk accesses produced by reconstruction. The simulation results reveal that MICRO technique lessens reconstruction times on average 20.22% and 9.34% when compared with the approaches like DOR and PRO. Moreover, it saves energy no less than 30.4% and 13%, respectively.

Kyungtae Kim et al., [7] investigate the vibrant features of slim optical disk drives and the modification of their structural dynamics to decrease vibration using a simplified Finite-Element (FE) model. The FE model was generated by means of a basic geometry and valid element types that efficiently reflect the dynamic characteristic features. Experimental Modal Analysis (EMA) is used to verify FE system. Design parameters were taken out and chosen to adapt the structural dynamics using design of experiments, topology optimization, and modal strain energy distribution. A prototype of the modified model was formed and its anti-vibration significance was evaluated using EMA.

Energy consumption has turned out to be a vital concern in the design of battery-operated mobile devices and complicated data centers. The storage hierarchy, which comprises memory and disks, is a key energy consumer in such systems; particularly for high-end servers at data centers. Majority of the research has focused on energy control techniques for storage systems that transition a device into a low power mode when a certain usage function goes beyond a particular threshold. These techniques are complicated to apply in real systems, since designers must carefully and manually tune threshold values; its performance is still very low. In order to tackle these drawbacks, Xiaodong Li et al., [8] developed three algorithms: 1) a performance guarantee approach that designers can use with any fundamental energy-control algorithm 2) a performance-directed control technique that occasionally allocates a static configuration to various devices by solving an optimization problem 3) Additional performance-directed control technique that dynamically self-tunes based on an optimal set of thresholds.

A video player may prefetch video frames into buffer to allow disk to go into standby mode, which comprises entire spindown of the spindle motor. Frequent spindowns, but, influence disk long life, so it is very vital to reduce the number of times that disk enters standby mode. Minseok Song et al., [9] proposed the design and implementation of a data prefetching approach that lessens disk power consumption for a limited number of disk spindowns. A data prefetching system is presented that entirely exploits the available buffer space and examine how power consumption is influenced by the bit-rates of the frames in the buffer. Then the issue is devised that decides when the disk should enter standby mode and offer an optimal solution using dynamic programming. The proposed approach is implemented in MPlayer running on the Linux 2.6. The simulation results reveal that it minimizes disk energy consumption by up to 59%.

Minimizing energy consumption is a vital issue for data centers. Storage is one of the main consumers of energy among various components of a data center. Earlier researches have revealed that the average idle period for a server disk in a data center is very little compared to the time taken to spin down and spin up. This greatly limits the effectiveness of disk power management schemes. Qingbo Zhu et al., [10] in this paper proposes several power-aware storage cache management techniques that offer more chances for the fundamental disk power management approaches to save energy. More particularly, an off-line power-aware greedy algorithm is proposed that is better energy-efficient than Belady's off-line algorithm (which minimizes cache misses only). An online power-aware cache replacement algorithm is also proposed in this paper. The trace-driven simulations reveal that, the proposed algorithm saves 16% more disk energy when compared with LRU and offers 50% better average response time for OLTP I/O workloads. The effects of four storage cache write policies on disk energy consumption are also examined.

Portable media players are vastly using Hard Disk Drives (HDD) to meet their storage needs, but HDDs consume a

considerable amount of energy. Hence video frames are prefetched into Dynamic Random Access Memory (DRAM) to facilitate the disk to go into low-power mode; but majority of the mobile systems have limited DRAM, so only very few energy is actually saved in this way. Jaewoo Kim et al., [11] propose two new energy saving approaches: one enhances the use of DRAM in storing prefetched frames, and the other widens this technique by using auxiliary flash memory. The experimental results reveal that deploying a realistic amount of auxiliary flash minimizes disk energy consumption by up to 86% when compared with traditional prefetching techniques.

In recent times, the requirement for micro hard disk drive that offers high-capacity detachable storage for handheld electronic devices is mounting quickly. The most important issue in the design of seek servo controller in micro disk drives is to diminish power utilization. The input power sent to the seek servo system is used by the transistors of power amplifier and motor coil resistance. Chang-Ik Kang et al., [12] proposed a novel seek servo controller for diminishing the power utilization. In this technique, Fourier decomposition and constrained nonlinear programming are used to find out the optimum seek profile that diminishes the power utilization.

To maintain the huge storage necessities, consumer electronics for video playback are progressively more being outfitted with hard disk drives (HDD) that use a considerable amount of energy. A video player possibly will prefetch several frames to provide a chance to disk to go to standby mode, however this might cause playback to be unclear or blocked if appropriate power mode transitions are not built-in. Jaedoo Go et al., [13] provided the design, implementation and estimation of a data prefetching method for energy-aware video data retrieval for portable media players (PMP). A difficulty is formulated when the prefetching is used for variable-bit-rate (VBR) streams to diminish disk energy utilization and then developed a novel energy-aware data retrieval scheme that prefetches video data in a quick way in order to raise the period in which disk reside in standby mode while promising the real-time service. This method is implemented in the legacy video player known as Mplayer that is characteristically used for Linux-based consumer machines. Experimental observation shows that it saves energy to the extent that 51% compared with traditional methods.

A hybrid hard disk drive that makes use of a non-volatile memory as a cache is increasing attractiveness because of its enhanced consistency and performance. Wanhyung Ryu et al., [14] proposed the design and implementation of a data prefetching method that makes use of flash memory to lessen disk energy utilization in media players. According to the estimated time used to prefetch data into flash memory, selectively decide when to spin up or down the disk in an appropriate way with the intention of reducing disk energy utilization while offering real-time video playback. This method is implemented in MPlayer operating under the platform of the Linux 2.6. Experimental observation shows that the disk energy utilization can be reduced between 35% and 63%, when a sensible amount of flash memory to a small DRAM can be added.

More power utilization of high-performance systems show the way to consistency, survivability, and cooling related difficulties. Inspired by this examination, numerous modern efforts concentrated on minimizing disk power utilization with the help of hardware, OS and compiler based approaches. Seung Woo Son et al., [15] developed a new technique to minimize disk power utilization of large-scale, array-concentrated scientific applications. It recommends and estimates a compiler-based technique that utilizes two complementary approaches: data reorganization and disk mapping. The data reorganization approach finds out an appropriate layout for data in the array space, while the second approach disk mapping, chooses the related layout in the disk space. The objective of data reorganization and disk mapping is to guarantee that data (from the various disk-resident arrays) that are accessed within the equivalent loop iteration are co-located in the similar set of disks. In this approach, the disk inter-access times (idle periods of disks) can be increased and this consecutively permits enhanced utilization of the basic hardware mechanisms used for minimizing power. The experimental results also confirms that both the components of this method are extremely significant since applying any of these components alone does not produce large savings for the majority of the applications.

Disk subsystem is recognized to be a most important contributor to the entire power budget of large-scale parallel systems. The majority of scientific applications at the moment rely greatly on disk I/O for out-of-core computations, check-pointing, and revelation of data. To diminish surplus energy utilization on disk system, previous studies proposed numerous hardware or OS-based disk power management methods. At the same time as such methods have been identified to be efficient in some cases, they possibly will miss opportunities for enhanced energy savings because of their reactive nature. Whereas compiler based methods can create more precise decisions on a specified application by extracting disk access patterns statically, the shortage of runtime details on the condition of shared disks possibly will lead to incorrect decisions when multiple applications makes use of the similar set of disks concurrently. Seung Woo Son et al., [16] recommended a runtime system based approach that offers more efficient disk power management. In this method, the compiler presents vital information on the future disk access patterns and favorite disk speeds from the perception of individual applications, and a runtime system uses this information together with current state of the shared disks to formulate decisions that satisfy to all applications. To test the performance of proposed technique runtime system support within PVFS2, a parallel file system is examined. The experimental outcome with four I/O-intensive scientific applications specify large energy savings: 19.4% and 39.9% over the formerly-proposed pure software and pure hardware based methods, correspondingly.

Bircher et al., [17] proposes the use of microprocessor performance contradictions for online calculation of entire system power utilization. This technique takes the benefit of the “trickle-down” result of performance measures in microprocessors. Though it has been well-known that CPU power utilization is correlated to processor

performance, the utilization of recognized performance-related events within a microprocessor for instance cache misses and DMA transactions to approximate power utilization in memory and disk and other subsystems external of the microprocessor is new. By means of amount of definite systems running scientific, business and productivity workloads, power models for six subsystems (CPU, memory, chipset, I/O, disk and GPU) on two platforms (server and desktop) are implemented and validated. These models are revealed to possess an normal error of under 9% per subsystem across the considered workloads. With the utilization of these models and existing on-chip performance event counters, it is feasible to approximate system power utilization without the requirement for power sensing hardware.

Flash memory has several smart characteristics like little size, low-power utilization, shock resistance, and elevated performance. Because of these high-quality characteristics, flash memory has been extensively used in the mobile consumer devices for example portable media players (PMPs) and smart phones. On the other hand, the expenditure of flash memory is more to accommodate constantly-growing mobile applications and multimedia contents. With the help of flash memory and mobile disk collectively as secondary storage is a different solution to offer large storage capacity with reasonable cost. Since heterogeneous storage devices are collectively used, the system requires a buffer management strategy that is responsive to different I/O characteristics of buffered blocks based on which devices they belong to. Particularly, power consumption price of all storage devices must be taken into consideration in the design of an effective buffer management strategy because battery restriction of mobile systems is essential. Hyojung Kang et al., [18] developed an innovative buffer management strategy for mobile systems that has heterogeneous storage devices like flash memory and mobile disk. With the consideration of different power-consumption rates all storage media in addition to I/O operation type and reference potential of buffered blocks, this strategy minimizes storage power utilization considerably and also enhances I/O performances.

APPROACHES	FUNCTIONALITIES
[1]	Disk layout-conscious approach restructures a given application code considering the disk layouts of the datasets it manipulates
[2]	Exposes disk layout information to the compiler, allowing it to derive the disk access pattern. Disk power management Application code to increase the length of idle disk periods
[3]	Dynamic speed control for power management tech or DRPM.
[4]	Sensitivity-based optimization methodology for disk architectures (SODA.
[5]	Energy-optimal speed profile (a function of time) of a generic device that has to execute a given task in a given time is obtained analytically
[6]	A novel reconstruction strategy, called multi-level caching-based reconstruction optimization (MICRO). MICRO collaboratively utilizes storage cache and disk array controller cache to diminish the number of physical disk accesses caused by reconstruction.
[7]	A simplified finite-element (FE) model. Constructed using simplified geometry and valid element types
[9]	Data prefetching scheme that minimizes disk power consumption for a limited number of disk spindowns
[10]	An off-line power-aware greedy algorithm is proposed that is better energy-efficient
[11]	Two new energy saving schemes: one improves the utilization of DRAM in storing prefetched frames, and the other extends this approach by making use of auxiliary flash memory.
[12]	Fourier decomposition and constrained nonlinear programming are used to find out the optimum seek profile that diminishes the power utilization.
[13]	Design, implementation and estimation of a data prefetching method for energy-aware video data retrieval.
[14]	Use of flash memory to lessen disk energy utilization in media players.
[15]	A compiler-based technique that utilizes two complementary approaches: data reorganization and disk mapping.
[16]	Compiler provides vital information on the future disk access patterns and favorite disk speeds from the perception of individual applications.
[17]	Cache misses and DMA transactions to approximate power utilization in memory and disk and other subsystems external of the microprocessor.
[18]	Buffer management strategy for mobile systems that contains flash memory which has several smart characteristics like little size, low-power utilization, shock resistance, and elevated performance.

3. Problems and Directions

Various disk power consumption techniques have been available in the literature. Previous researches on power

consumption mostly focused on the architectural level techniques. These approaches takes advantages of idle periods experimented by disks. Existing disk power consumption techniques suffer from various drawbacks. The most common limitation of most of the existing energy optimization techniques is that, the techniques operate in a reactive manner. Moreover, the system can miss the power saving opportunities. Moreover, there are various performance degradations because of the inaccuracies in predicting the idle and active time. With subject to different kinds of disturbances that yield unresolved issues and uncertain consequences in different disk energy optimization problems. With such limitations, it is difficult to save power in the disk systems effectively by these kinds existing techniques. Therefore, other types of modern techniques are necessary for effective energy optimization.

- Better disk scheduling techniques are needed for better performance.
- Need to identify better and significant disk idle and active time.
- Moreover, better optimized compilers are necessary for better power consumption of the disks.

4. Conclusion

This review is undertaken to explore and analyze the existing disk power consumption techniques present in the literature which are very much required to maintain significant power saving capability. This paper attempts to present major studies of disk energy optimization techniques, such as adopting architectural mechanisms such as spinning down idle disks, Traditional Power Management (TPM) etc which is available in the literature. These conventional energy optimization techniques form the basis for the new innovation of the effective power consumption approaches. Most of the stabilization techniques available in the literature operate in reactive manner, and moreover there is significant performance penalities. By contrast, the advantages of PSO techniques convinced and encouraged many researchers to apply these techniques to solve the problems of power system control.

5. References

[1] Kandemir, M.; Seung Woo Son; Karakoy, M.; “Improving disk reuse for reducing power consumption”, 2007 ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), Page(s): 129 – 134, 2007.

[2] Seung Woo Son; Guangyu Chen; Kandemir, M.; “A compiler-guided approach for reducing disk power consumption by exploiting disk access locality’, 2006

[3] Shu Yin; Alghamdi, M.I.; Xiaojun Ruan; Nijim, M.; Tamilarasan, A.; Ziliang Zong; Xiao Qin; Yiming Yang; “Improving Energy Efficiency and Security for Disk Systems”, 2010 12th IEEE International Conference on High Performance Computing and Communications (HPCC), Page(s): 442 – 449, 2010

[4] Sankar, S.; Yan Zhang; Gurumurthi, S.; Stan, M.R. “Sensitivity-Based Optimization of Disk Architecture”, IEEE Transactions on Computers, Page(s): 69 – 81, 2009.

[5] Ravishankar Rao; Sarma Vrudhula, “Energy-Optimal Speed Control of a Generic Device”, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Volume: 25 , Issue: 12, Page(s): 2737 – 2746, 2006.

[6] Tao Xie and Hui Wang; “MICRO: A Multilevel Caching-Based Reconstruction Optimization for Mobile Storage Systems”, IEEE Transactions on Computers, Volume: 57, Issue: 10, Page(s): 1386 – 1398, 2008.

[7] Kyungtae Kim; Seungho Lim; No-Cheol Park; Young-Pil Park; Kyoung-Su Park; Ik-Joo Cha; “Structural Dynamics Modification of Slim Optical Disk Drive”, IEEE Transactions on Magnetics, Volume: 45 , Issue: 5 , Part: 2, Page(s): 2209 – 2212, 2009.

[8] Xiaodong Li; Zhenmin Li; Pin Zhou; Yuanyuan Zhou; Adve, S.V.; Kumar, S.; “Performance-directed energy management for storage systems”, IEEE Micro, Volume: 24 , Issue: 6, Page(s): 38 – 49, 2004.

[9] Minseok Song; Wanhyung Ryu; Jeong Seop Sim; “Reducing disk power consumption in portable media players”,2010 8th IEEE Workshop on Embedded Systems for Real-Time Multimedia (ESTIMedia), Page(s): 81 – 89, 2010.

[10] Qingbo Zhu; David, F.M.; Devaraj, C.F.; Zhenmin Li; Yuanyuan Zhou; Pei Cao; “Reducing Energy Consumption of Disk Storage Using Power-Aware Cache Management”, Page(s): 118, 2004.

[11] Jaewoo Kim; Ahron Yang; Minseok Song; “Exploiting flash memory for reducing disk power consumption in portable media players”, IEEE Transactions on Consumer Electronics, Volume: 55 , Issue: 4, Page(s): 1997 – 2004, 2009.

[12] Chang-Ik Kang; Sang-Eun Baek; Jun-Seok Shim; “A new seek servo controller for minimizing power consumption in micro hard disk drives”, IEEE Transactions on Magnetics, Volume: 40 , Issue: 4 , Part: 2, Page(s): 3127 – 3129, 2004.

[13] Jaedoo Go; Minseok Song; “Adaptive disk power management for portable media players”, Consumer Electronics, IEEE Transactions on, Volume: 54, Issue: 4, Page(s): 1755 – 1760, 2008.

[14] Wanhyung Ryu; Minseok Song; “Design and implementation of a disk energy saving scheme for media players which use hybrid disks”, IEEE Transactions on Consumer Electronics, Volume: 56 , Issue: 4, Page(s): 2382 – 2386, 2010.

[15] Seung Woo Son; Kandemir, M.; “Integrated Data Reorganization and Disk Mapping for Reducing Disk Energy Consumption”, Seventh IEEE International Symposium on Cluster Computing and the Grid, 2007. Page(s): 557 – 564, CCGRID 2007.

[16] Seung Woo Son; Kandemir, M.; "Runtime system support for software-guided disk power management", 2007 IEEE International Conference on Cluster Computing, page(s): 139 – 148, 2007.
[17] Bircher, W.; John, L.; "Complete System Power Estimation using Processor Performance Events", IEEE Transactions on Computers, 2011.
[18] Hyojung Kang; Junseok Park; Hyokyung Bahn; "LBM: a low-power buffer management policy for heterogeneous storage in mobile consumer devices", IEEE Transactions on Consumer Electronics, Volume: 56 , Issue: 4, Page(s): 2387 – 2392, 2010.
[19] T. Heath et al., "Application Transformations for Energy and Performance-Aware Device Management," Proc. Int'l Conf. Parallel Architectures and Compilation Techniques (PACT 2002), IEEE CS Press, 2002, pp. 121-130.
[20] Y-H. Lu et al., "Quantitative Comparison of Power Management Algorithms," Proc. Design Automation and Test in Europe, ACM Press, 2000, pp. 20-26.
[21] Helmbold, D. P., Long, D. D. E., Sconyers, T. L., and Sherrod, B. Adaptive disk spin-down for mobile computers. Mobile Networks and Applications 5, 4 (2000), 285-297.
[22] Krishnan, P., Long, P. M., and Vitter, J. S. Adaptive disk spindown via optimal rent-to-buy in probabilistic environments. In Proceedings of the Twelfth International Conference on Machine Learning (1995), pp. 322-330.
[23] Aboutabl, M., Agrawala, A. K., and Decotignie, J.-D. Temporally determinate disk access: An experimental approach. In Measurement and Modeling of Computer Systems (1998), pp. 280{281.
[24] Douglis, F., Krishnan, P., and Bershad, B. Adaptive disk spin-down policies for mobile computers. In Proceedings of the Second USENIX Symposium on Mobile and Location Independent Computing (April 1995), pp. 121{137.
[25] Douglis, F., Krishnan, P., and Marsh, B. "Thwarting the power-hungry disk", In Proceedings of the Winter USENIX Conference, pp. 292-306, 1994.
[26] Farkas, K. I., Flinn, J., Back, G., Grunwald, D., and Anderson, J.-A. M. "Quantifying the energy consumption of a pocket computer and a java virtual machine", In Proc. International Conference on Measurement and Modeling of Computer Systems (2000), pp. 252{263.
[27] Helmbold, D. P., Long, D. D. E., Sconyers, T. L., and Sherrod, B. Adaptive disk spin-down for mobile computers. Mobile Networks and Applications Vol. 5, No. 4 285-297, 2000.

G.Ravikumar received his M.Tech., degree and B.E., degree in Computer Science and Engineering from Bharathidasan and Sastra University respectively. He is currently working as assistant professor in department of Computer Science and Engineering at Coimbatore Institute of Engineering and Technology, Coimbatore. His research interests accumulate in the area of Optimization in parallel disk based system towards energy and power.

Dr.N.Nagarajan received his B.Tech and M.E. degrees in Electronics Engineering at M.I.T Chennai. He received his PhD in faculty of information and communication engineering from Anna University, Chennai. He is currently working as Principal in Coimbatore Institute of Engineering and Technology at Coimbatore. He is member of board of study of faculty of information Technology at Anna University, Coimbatore. His specialization includes optical, wireless Adhoc and sensor networks. He is guiding assorted research scholars in power and energy optimization.

Bandwidth Estimation in Mobile Ad-hoc Network (MANET)

Rabia Ali[1] and Dr. Fareeha Zafar[2]

[1]Department of Computer Science
Kinnaird College for Women
Lahore, Pakistan

[2] Department of Computer Science
GC University Lahore
Lahore, Pakistan

Abstract

In this paper we presents bandwidth estimation scheme for MANET, which uses some components of the two methods for the bandwidth estimation: 'Hello Bandwidth Estimation' & 'Listen Bandwidth Estimation'. This paper also gives the advantages of the proposed method. The proposed method is based on the comparison of these two methods.

Bandwidth estimation is an important issue in the Mobile Ad-hoc Network (MANET) because bandwidth estimation in MANET is difficult, because each host has imprecise knowledge of the network status and links change dynamically. Therefore, an effective bandwidth estimation scheme for MANET is highly desirable.

Ad hoc networks present unique advanced challenges, including the design of protocols for mobility management, effective routing, data transport, security, power management, and quality-of-service (QoS) provisioning. Once these problems are solved, the practical use of MANETs will be realizable.

Keywords: *Bandwidth Estimation, Mobile Ad Hoc NETwork (MANET), "Hello" Bandwidth Estimation Method, "Listen" Bandwidth Estimation Method, QoS.*

1. Introduction

Bandwidth estimation is a basic function that is required to provide QoS in MANETs [1]. It is a way to determine the data rate available on a network route. It is of interest to users wishing to optimize end-to-end transport performance, overlay network routing, and peer-to-peer file distribution [1].

Techniques for accurate bandwidth estimation are also necessary for traffic engineering and capacity planning support [1]. Having information existing can help to develop better methods for e.g. gateway selection, channel selection, routing, etc. [2].

Literally, ad-hoc means in Latin, ad-hoc means is "for this," meaning "for this special purpose". An ad-hoc network is a local area network (LAN) that is built spontaneously as devices connect [3] and autonomous self-organized wireless and mobile networks [4]. They do not require any fixed infrastructure for instance a base station to work. The nodes themselves address topology changes due to the mobility, the entrance or the exits of nodes. These networks use a radio medium [4].

MANET is a group of two or more devices or nodes or terminals with wireless communications and networking competence that communicate with each other without the help of any centralized administrator also the wireless nodes that can form a network to exchange information according to their need at that time [5], [6] and [7]. It is an independent system in which mobile hosts connected without wire and are free to move dynamically and sometimes they act as routers at the same time [5], [6] and [7].

There are two types of mobile network namely Mobile IP and MANET [8]. MANET consists of nodes that are cable to communicate wirelessly among themselves [7] and [8]. MANETs consist of a group of wireless mobile nodes which dynamically exchange data among themselves [7] without the reliance on a fixed base station or a wired backbone network [6].

MANET nodes are typically differentiated by their limited power, processing, and memory resources as well as high degree of mobility [6]. In MANETs, the wireless mobile nodes may dynamically enter in the network as well as leave the network. Because of the limited transmission range of wireless network nodes, multiple hops are generally required for a node to exchange information with any other node in the network [6].

Multipath routing permits the formation of multiple paths between one source node and one destination node. It is basically proposed in order to enhance

the reliability of data transmission (i.e., fault tolerance) or to provide load balancing [6].

Available bandwidth estimation techniques can be divided in two major approaches [2]:

1. Intrusive Bandwidth Estimation Techniques:

   The intrusive approaches techniques are based on end-to-end probe packets to estimate the available bandwidth along a path.

2. Passive Bandwidth Estimation Techniques:

   The passive approaches techniques uses local information on the used bandwidth and that may exchange this information via local broadcasts**.**

Till date much of the research work is targeted at finding a possible path from a source to a destination without considering current network traffic or usage requirements. Such QoS support can be accomplished by either finding a path to fulfill the application requirements or offering network response to the application, when the requirements cannot be met. This paper is also about a QoS-aware routing protocol that incorporates a feedback scheme and an admission control scheme to meet the QoS requirements (provides better than best-effort service) of real-time applications using IEEE 802.11. The novel work of this QoS-aware routing protocol is the use of the approximate bandwidth estimation to response to the network traffic.

The rest of the paper is organized as: the Section II contains the Literature Review Discussion is in Section III and the Conclusion is in Section IV.

2. Literature review

In an ad hoc network, a host's available bandwidth refers to amount of bandwidth available to the node to send packets to the network [5]. Whole channel will not be used for packet transmission. Bandwidth estimation can be done using various methods; for example, bandwidth estimation is a cross-layer design of the routing and MAC layers and the available bandwidth is estimated in the MAC layer and is sent to the routing layer for admission control. Therefore, bandwidth estimation can be carried out in various network layers [1].

Present bandwidth estimation tools measure one or more of three related metrics: capacity, available bandwidth, and bulk transfer capacity [9]. Currently available bandwidth estimation tools utilize a various strategies to measure these metrics [9].

The issues of multipath routing in MANETs were specifically examined [6]. They also discuss the application of multipath routing to support application constraints such as reliability, load-balancing, energy-conservation, and QoS [6].

An improved mechanism was proposed to estimate the available bandwidth in IEEE 802.11-based ad hoc networks [10]. In 802.11-based ad hoc networks, few works deal with solutions for bandwidth estimation [4].

In a distributed ad hoc network, a host's available bandwidth cannot decided only by the raw channel bandwidth, but also by its neighbour's bandwidth usage and interference caused by other sources, each of which reduces a host's available bandwidth for transmitting data. Therefore, applications cannot properly optimize their coding rate without knowledge of the status of the entire network [1].

An incorporating QoS into routing, and introduce bandwidth estimation by propagating bandwidth information through "Hello" messages [11] and [12]. A cross-layer approach, including an adaptive feedback scheme and an admission scheme to give information to the application about the network position, are implemented [11] and [12].

According to the simulations show that their QoS-aware routing protocol can improve packet delivery ratio greatly without impacting the overall end-to-end throughput, while also decreasing the packet delay and the energy consumption significantly [11].

The problem in available bandwidth estimation was rethink in IEEE 802.11 based ad hoc networks [12]. According to them estimation accuracy is increased by improving the calculation accuracy of the probability for two adjacent nodes idle period to overlap.

All the information of MANET which include the History of ad hoc, wireless ad hoc, wireless mobile approaches and types of MANETs, and then they present more than 13 types of the routing Ad Hoc Networks protocols were proposed [7]. They give description of routing protocols, analysis of individual characteristics and advantage and disadvantages to collect and compare, and present all the applications or the Possible Service of Ad Hoc Networks [7].

2.1 Characteristics of MANET

The intention of the MANET is to standardize IP routing protocol functionality is appropriate for the wireless routing application within both dynamic and static topologies with raised dynamics because of node motion and other factors:

- **Dynamicity**: Every host can randomly change position. The topology is generally unpredictable, and the network status is imprecise.
- **Non-centralization:** There is no centralized control in the network and, thus, network resources cannot be assigned in a predetermined manner.
- **Radio properties:** The wireless channel can suffer fading, multipath effects, time variation, etc.

With these constraints, Hard QoS (e.g., guaranteed constant bit rate and delay) is difficult to achieve. The reasons are as follows [11] and [12]:

- To support QoS the end host should have knowledge of the worldwide position of the network. The dynamic nature of MANETs makes it difficult for hosts to determine information about their local neighborhood, much less the global status of the network.
- It is hard to establish cooperation between neighboring hosts to determine a transmit schedule for guaranteed packet delivery without centralized control. In MANETs, each host's transmissions will interfere with neighboring hosts' transmissions.
- The wireless channel's main deficiency is its unreliability caused by various reasons such as fading and interference.

Thus if the topology changes too frequently, the source host cannot detect the network status changes and cannot make the corresponding adjustment to meet the specific QoS requirements. Therefore, combinatorial stability must first be met before we can consider providing QoS to real-time applications. Solution is a QoS-aware routing protocol that either **provides feedback about the available bandwidth** to the application (feedback scheme), or admits a flow with the requested bandwidth **(admission scheme)** [11]**.** Both the feedback scheme and the admission scheme require knowledge of the *end-to-end bandwidth* available along with the *route from the source to the destination*. Thus, bandwidth estimation is important to support QoS.

2.2 Bandwidth Estimation Methods

Estimating accurate available bandwidth allows a node to make optimal decision before transmitting a packet in networks. It is therefore clear that the available bandwidth estimation enhances the QoS in wired and wireless Networks. Measuring available bandwidth in ad hoc networks is challenging issue in MANET and calculating the residual bandwidth using the IEEE 802.11 MAC is still a challenging problem, because the bandwidth is shared among neighboring hosts, and an individual host has no knowledge about other neighboring hosts' traffic status. Two methods for estimating bandwidth are used below [11] and [12]:

1. **"Listen" bandwidth estimation:** For hosts to listen to the channel and estimate the available bandwidth every second based on the ratio of free and busy times. The IEEE 802.11 MAC utilizes both a physical carrier sense and a virtual carrier sense [via the network allocation vector (NAV)], which can be used to find out the free and busy times. The MAC detects that the channel is free when the following three requirements are met [11] and [12]:
   - NAV's value is less than the current time;
   - Receive state is idle;
   - Send state is idle.

   The MAC declares that the channel is busy when one of following occurs:
   - NAV sets a new value;
   - Receive state changes from idle to any other state;
   - Send state changes from idle to any other state.

   $\Rightarrow$ Channel BW*free time/over all time

   Weight factor

2. ***"Hello" bandwidth estimation:*** The sender's current bandwidth consumption as well as the sender's one-hop neighbours' (from its two-hop neighbours) current bandwidth consumption is piggybacked onto the standard "Hello" message. Each host estimates its available bandwidth based on the information provided in the "Hello" messages and knowledge of the frequency reuse design [11] and [12].

   The second neighboring host's information was proposed by using hop relay to propagate [11]. **AODV** uses the

"Hello" messages to update the neighbor caches. The "Hello" message used in AODV only keeps the address of the host who initiates this message. Modify the "Hello" message, including two fields. The first field includes host address, consumed bandwidth, timestamp, and the second field includes neighbor's addresses, consumed bandwidth, timestamp, as shown in Figure 1. Each host finds out its used bandwidth by monitoring the packets it supplies into the network. This value is recorded in a bandwidth-consumption register at the host and is updated periodically.

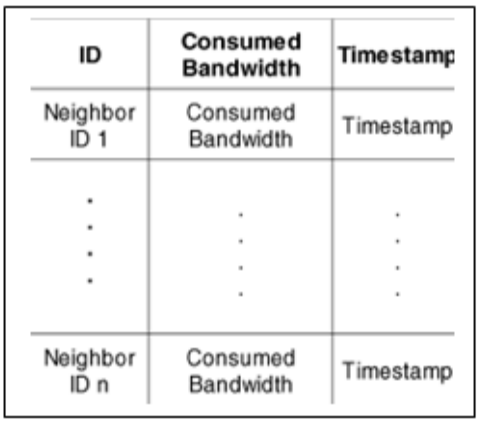


Figure 1. Hello structure [11]

3. Discussion

By using the "Listen" method the host cannot release the bandwidth immediately when a path breaks, because it does not know how much bandwidth each node in the broken path consumes. The time interval between claiming a path break and setting up the path is only several milliseconds. In such a small time interval, it is almost impossible for the hosts to automatically and correctly update their bandwidth registers in the "Listen" bandwidth estimation method because the consumed bandwidth estimation is based on averaging bandwidth consumption every 1s interval and the hosts in the broken path were transmitting data in the previous second [11].

If the topology is static Hello or Listen be used e.g. listen or Hello. But the problem is when the topology is not static that is mobile topology. But bandwidth estimation is difficult, because each host has imprecise knowledge of the network status and links change dynamically. Therefore, an effective bandwidth estimation scheme is highly desirable [1].

Therefore, the "Listen" bandwidth estimation approach has difficulty correctly estimating the residual bandwidth. Even if some forced update schemes can be adopted, the hosts still cannot release the bandwidth correctly; since the hosts do not know how much bandwidth each node in the broken path consumes [11].

The "Hello" bandwidth estimation method and the "Listen" bandwidth estimation method in [11] and [12] we compare these two methods which are summarized as follows in the form of Table 1.

Table 1. Comparison of the "Hello" bandwidth estimation & the "Listen" bandwidth estimation methods

	Listen Bandwidth Estimation	**Hello Bandwidth Estimation**
Counts	It counts the used bandwidth	It counts the transmitted packets only
Performance	The host cannot release the bandwidth immediately when a path breaks, because it does not know how much bandwidth each node in the broken route consumes.	It is better when releasing the bandwidth immediately is important.
Performance in mobile topology	It performs better in term of packet delivery ratio.	It performs better in term of end-to-end throughput
Performance in Static topology	The "Hello" and "Listen" schemes work equally well, using large weight factors to reduce the congestion and minimize the chance of lost "Hello" messages incorrectly signalling a broken route.	
Overhead	It does not add extra overhead	It adds overhead by attaching neighbor's bandwidth consumption information.

3.1.1 ADVANTAGE OF LISTEN METHOD:

The Listen Method does not add an extra overhead by attaching neighbor's bandwidth consumption information.

3.1.2 DISADVANTAGES OF LISTEN METHOD:

In this method the host cannot release the bandwidth immediately when a path breaks, because it does not know how much bandwidth each node in the broken path consumes.

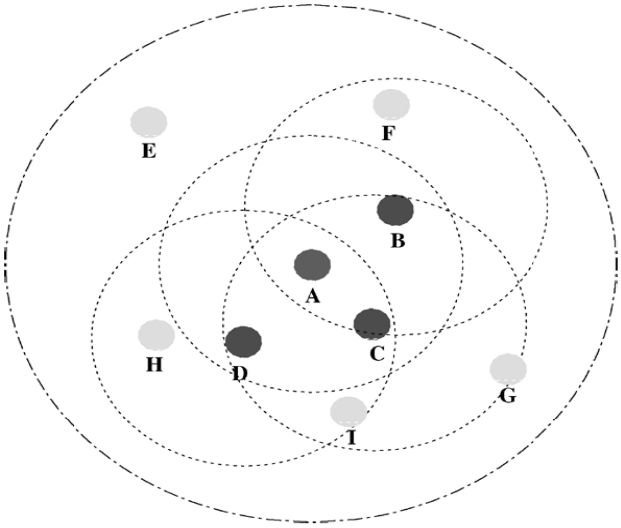


Fig. 2 Hidden node scenario. The big circle indicates host A's interference range. The small circles show host A and its first neighboring hosts' transmission range. Hosts B, C, and D are A's first neighbours of host's A and hosts F, G, H, and I are host A's second neighbors. Host E is in host A's interference range [11] and [12].

3.2.1 ADVANTAGE OF HELLO METHOD:

The "Hello" bandwidth estimation method can solve this problem easily by using the forced update scheme [11].

This approach avoids creating extra control messages by using the "Hello" messages to propagate the bandwidth information.

The first neighbouring hosts' information can be obtained directly, but there is no way to get the second neighbouring hosts' bandwidth information directly Figure 2. There are several ways to get the second neighbouring hosts' information, such as propagating the host bandwidth information using higher transmission power to reach the two-hop neighbourhood, setting up a separate signalling channel to broadcast the bandwidth information.

3.2.2 DRAWBACKS OF HELLO METHOD:

Drawbacks of Getting Second Neighbouring Hosts' Information are [11] and [12]:

1. Imprecise Information about the Hidden Hosts [11] and [12] as shown in Figure 2.
2. Overhead caused by attaching neighbour's bandwidth consumption information.
3. using higher power to propagate information consumes much more power.
4. It destroys the frequency reuse pattern and causes much more interference.
5. Using a separate channel to propagate the bandwidth information needs an additional control that is an intense burden for an ad hoc network in terms of bandwidth consumption and hardware support.

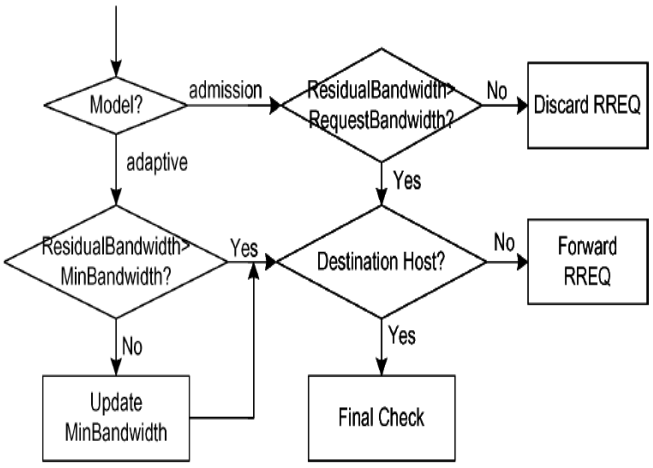


Figure 3. Host's working procedure after receiving RREQ in AODV [11]

Host's working procedure after receiving reply request RREQ in AODV is explained in Figure 3.

If SOURCE ADAPTIVE

If B.W > min B.W

ALLOW Destination host ,

Else

Reject.

If SOURCE ADMISSION

If B.W > Requested B.W

ALLOW Destination host ,

Else

Reject.

After completing this checking procedure, it is not sufficient to say that the current network can offer the min-bandwidth indicated in the RREQ packet. The reason is that if the route is chosen, the chosen hosts will bring mutual interference into the network during transmission. Therefore, one final check procedure is required before transmiting the RREP packet back to the source host. We directly use the relation of the end-to-end throughput with the number of hops and the bottleneck bandwidth in the route as follows (the details can be found in [11] and [13].

If (hop num =1)

Min B.W =Min B.W

Else if (hop num =2)

Min B.W =Min B.W / 2

Else if (hop num =3)

Min B.W =Min B.W /3

Else if (hop num =4)

Min B.W =Min B.W/4

.

.

.

Else if (hop num =n)

Min B.W =Min B.W / n

This equation offers the upper limit of the available bandwidth. A more accurate estimation is studied in [14] and [15], where the interflow contention is accounted for by using the contention counter. Finally, the destination host sends the RREP with a changed header ***(min-bandwidth, AODV RREP header)*** to the source host. Once intermediate hosts receive the RREP, they enable the path and also record the min-bandwidth in their routing table, which is useful for path maintenance of QoS-aware routing with "Hello" bandwidth estimation.

3.3 PROPOSED METHOD:

The proposed approach uses some components of both: the Listen bandwidth estimation and the Hello bandwidth estimation method. The proposed method is used for the routing in MANET using ADOV protocol. The proposed is described below:

1) It estimates the bandwidth by counting the used bandwidth, as in the Listen bandwidth estimation method.
2) If there is a route break then it uses the update scheme used in Hello bandwidth estimation method to immediately release the bandwidth when the route is broken.
3) Then it reply request back to whom who send request then it sends according to the Listen bandwidth estimation method it does not add an extra overhead as it does not need to attach neighbour's bandwidth consumption information.

3.3.1 ADVANTAGES OF PROPOSED METHOD:

1) In the proposed method the host can release the bandwidth immediately when a route breaks, because it uses the Update Scheme used in the "Hello" Bandwidth estimation method.
2) No Overhead which is caused by attaching neighbour's bandwidth consumption information as in the Hello method.

3.3.2 EXAMPLE**:**

If there is no route break then it can estimate the bandwidth normally. The problem is when the route is broken. So, the following example has shown the case of the broken route.

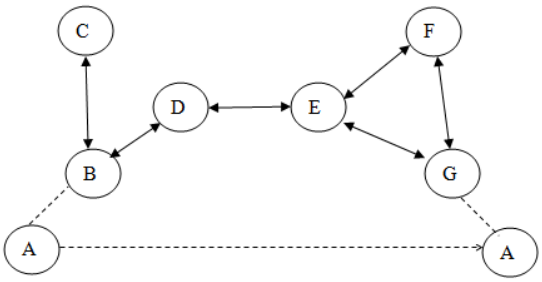


Figure 4.

Suppose that in Figure 4. 'A' moves away from 'B' towards 'G', and has active sessions with 'C' and 'D'. The following actions occur:

- 'B' notices that its link to 'A' is broken.
- 'B' checks its routing table, and finds that its link to 'A' was actively in use by 'C' and 'D'.
- 'B' unicasts ∞metric route update, with an incremented destination sequence number, to 'C' and 'D'. 'C' may subsequently issue a new RREQ (Route Request) for 'A'.
- 'D' also notes that its route to 'A' was actively in use, and forwards the ∞-metric route update to 'E'.

- The ∞-metric route update for 'A' may also be included in the next hello message issued by 'B'
- 'E' may subsequently issue a new route request for 'A'.
- Any subsequent route request for 'A' which is satisfied by a RREP (route reply) through 'B' may cause 'B' to update its route table.

The symbol of infinity '∞' means that route does not exist or the route is broken.

IV. Conclusion

In this paper we proposed a new method after comparing the "Hello" Bandwidth estimation method and the "Listen" bandwidth estimation method. The proposed method removes the problems caused by these two methods. The proposed method, immediately releasing bandwidth when the route breaks as in the "Listen" method and replace this with an "update technique" used in the "Hello" method. The proposed method also does not cause an overhead which was in the "Hello" bandwidth Estimation method due to attaching neighbours bandwidth usage information.

References

[1] Lei Chen and Wendi Heinzelman, "Network Architecture to Support QoS in Mobile Ad Hoc Networks".

[2] Cheikh Sarr, Claude Chaudet, Guillaume Chelius, "Improving Accuracy in Available Bandwidth Estimation for 802.11-based Ad Hoc Networks", 2006.

[3] Downloaded in June, 2011: http://searchmobilecomputing.techtarget.com/definition/ad-hoc-network

[4] Cheikh Sarr, Claude Chaudet, Guillaume Chelius amd Isabelle Gu´erin Lassous, "A node-based available bandwidth evaluation in IEEE 802.11 ad hoc networks", International Journal of Parallel, Emergent and Distributed Systems Vol. 00, No. 00, July 2005, 1–21.

[5] K. Mohideen Vahitha Banu, "Improving Ad Hoc Network Performances by Estimating Available Bandwidth", (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 08, 2010, 2589-2592

[6] Stephen Mueller, Rose P. Tsang, and Dipak Ghosal, "Multipath Routing in Mobile Ad Hoc Networks:Issues and Challenges" .

[7] Saleh Ali K.Al-Omari, Putra Sumari, "An Overview of Mobile Ad hoc Networks for the Existing Protocols And Applications", International Journal on Applications of Graph Theory in Wireless Ad hoc Networks and Sensor networks (Graph-Hoc), Vol.2, No.1, March 2010.

[8] Ali, Ahmed and Abdul Latiff, Liza and Fisal, Norsheila (2004) *Indoor Location Tracking in Mobile Ad Hoc Network (MANET) using RSSI.* In: RFM 2004, Oct 5-6, 2004, Subang, Selangor, Malaysia., "Indoor Location Tracking in Mobile Ad Hoc Network (MANET) using RSSI".

[9] R. S. Prasad, M. Murray, C. Dovrolis, K. Claffy, "Bamdwidth Estimation: Metrics, Measurement Techniques and Tools"

[10] Cheikh Sarr, Claude Chaudet, Guillaume Chelius, and Isabelle Gue´ rin Lassous, "Bandwidth Estimation for IEEE 802.11-Based ad Hoc Networks", IEEE Transactions on Mobile Computing, Vol. 7, No. 10, October 2008.

[11] Lei Chen, *Student Member, IEEE,* and Wendi B. Heinzelman, *Member, IEEE,* "QoS-Aware Routing Based on Bandwidth Estimation for Mobile Ad Hoc Networks", IEEE Journal on Selected Areas in Communications, Vol. 23, No. 3, March 2005.

[12] Deepak Vidhate, Anita Patil and Supriya Sarkar, "Bandwidth Estimation Scheme for Mobile Adhoc Network", Communications in Computer and Information Science, Volume 70, 2010, DOI: 10.1007/978-3-642-12214-9_23, 130-135.

[13] J. Li, C. Blake, D. D. Couto, H. Lee, and R. Morris, "Capacity of ad hoc wireless networks," in "*Proc. 7th ACM Int. Conf. Mobile Comput. Netw. (MobiCom'01)*", 2001, pp. 61–69.

[14] Y. Yang and R. Kravets, "Contention-aware admission control for ad hoc networks," Univ. Illinois at Urbana–Champaign, Urbana–Champaign, IL, Tech. Rep. 2003-2337, 2003.

[15] K. Sanzgiri, I. Chakeres, and E. Belding-Royer, "Determining intra-flow contention along multihop paths in wireless networks," in "*Proc. Broadnets 2004 Wireless Netw. Symp.*", Oct. 2004, pp. 611–620.

Autonomic Management for Multi-agent Systems

Nadir K.Salih[1#] Tianyi Zang[1@] G.K.Viju[2*] Abdelmotalib A.Mohamed[1&]
[1]School of Computer Science and Engineering, Harbin Institute of Technology, China
[2]Department of Computer Science, Karary University, Khartoum, Sudan

Abstract-

Autonomic computing is a computing system that can manage itself by self-configuration, self-healing, self-optimizing and self-protection. Researchers have been emphasizing the strong role that multi agent systems can play progressively towards the design and implementation of complex autonomic systems. The important of autonomic computing is to create computing systems capable of managing themselves to a far greater extent than they do today. With the nature of autonomy, reactivity, sociality and pro-activity, software agents are promising to make autonomic computing system a reality. This paper mixed multi-agent system with autonomic feature that completely hides its complexity from users/services. Mentioned Java Application Development Framework as platform example of this environment, could applied to web services as front end to users. With multi agent support it also provides adaptability, intelligence, collaboration, goal oriented interactions, flexibility, mobility and persistence in software systems.

Keywords: Autonomic, Multi-agent System, Web Services

I. Introduction

A new computational framework called Agent Oriented Programming (AOP), which can be viewed as a specialization of object oriented programming. It is relatively a new software paradigm that brings concepts from the theories of artificial intelligence into the mainstream realm of distributed systems. AOP essentially models an application as a collection of components called agents that are characterized by, among other things, autonomy, proactivity and an ability to communicate. Being autonomous they can independently carry out complex, and often long-term, tasks [1]. Intelligent autonomic agent must build and maintain a model of the external environment and of its own components. Atop-level executive component makes decisions based on the models and its current emotional state. A planner component is used to create multiple step scripts or sequences of actions necessary to achieve the high-level goals being pursued by the executive [2]. Agents have the capability to move from one environment to another see fig.1.In the agent design, using FraMaS “advanced behavior" wrappers (like autonomously search according to the agent knowledge of the user or planning strategy to arrive to the target point) [3]. The autonomy of each agent and the messaging interface are useful in most of flexible and extensible systems. Because agents are not directly linked to others, then it is easy to take one out of operation or add a new one while the others are running [4]. Multi-agent development has emerged as a viable approach to meet the autonomic system requirements-autonomy, adaptability, intelligence, goal-oriented interaction, collaboration, and flexibility. Using multiagent approach, real-world problems can be modeled in the form of autonomous, interacting agent components [8].

An autonomic system is an autonomous computing environment that completely hides its complexity. Complexity hiding from users/services means that autonomic computing will provide users with a computing environment that allows them to concentrate on what they want to do without worrying about how it has to be done [12]. The characteristics of Multiagent Systems (MASs) are that (1) each agent has incomplete information or capabilities for solving the problem and, thus, has a limited viewpoint; (2) there is no system global control; (3) data are decentralized; and (4) computation is asynchronous [13].

The paper is organized as follows: Section II reviews Related Work Section III focuses on the Programming Language and Tools. Section IV describes JADE and the Agents Paradigm. Finally, Section V takes The Utility of Agent and WEB Service Integration before concluding the paper.

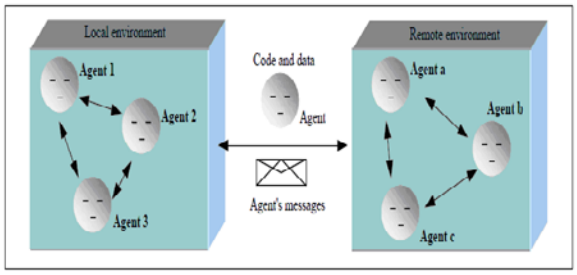


Fig.1 Multi Agent System Model

II. Related Work

The Unity system components are implemented as autonomic elements— individual agents that control resources and deliver services to humans and to other

autonomic elements. Every Unity component is an autonomic element. This includes: computing resources (e.g. databases, storage systems, servers), higher-level elements with some management authority (e.g. workload managers or provisioners), and elements that assist other elements in doing their tasks (e.g. policy repositories, sentinels, brokers, or registries) [9]. Each autonomic element is responsible for its own internal autonomic behavior, namely, managing the resources that it controls, and managing its own internal operations, including self-configuration, self-optimization, self-protection and self-healing. MAACE emphasis on self-organization and self-healing of application services and it is an open and extensible computing environment to allow heterogeneous agent to join it. By the cooperation of agent federation system, agent mediate system and agent monitor system, MAACE lead to automated control and management of a wide range of network centric applications and services [10].The Bean Generator is a tool that supports agent engineers in creating message content ontologies compliant with the JADE support. The tool is a plug-in for Protege, which is a commonly used ontology editor that enables engineers to graphically model ontologies. Furthermore, additional functionality and storage formats can be 'plugged in' to the system. Another advantage of the Protege tool is that other ontologies can be imported. Repositories of existing ontologies ranging from biological domains to market place product and service descriptions can be found at the Protege community page and at the DAML site1. The languages used to represent these ontologies can be XML, RDF, DAML-OIL, XMI, SQL or UML [5]. The jademx JADE add-on, which provides two major capabilities: the ability to interface JADE agents with Java JMX (Java Management Extensions) and the ability to unit test JADE agents using JUnit. Jademx is available for download from the third-party software area of the JADE website. Everyday, useful software systems rarely exist in isolation. Indeed, one of the strengths of JADE is that the full capabilities of the Java environment are available when creating a software agent application. JMX is the Java technology for management and monitoring of software systems; it was originally part of the Java EE enterprise platform (formerly known as J2EE), but as of Java 5 it is available as part of the standard J2SE environment. Furthermore, unit testing is an important technique for the development of robust software and Junit is a standard methodology for the unit testing of applications written in Java. Jademx was developed for an industrial software agent effort requiring management using Java EE and to be unit-testable. A jademx agent can be configured either programmatically or by using XML [1]. The Java Sniffer is a stand-alone Java application, developed by Rockwell Automation, Inc., that can remotely connect to running JADE systems and is intended as an alternative to JADE's built-in sniffer. The tool receives messages from all agents in the system, reasons about the information, and presents it from different points of view (see Fig.2) [6]. We observed Jadex Belief Desire Intention BDI reasoning engine that allows development of rational agents using mentalistic notions in the implementation layer. In other words, it enables the construction of rational agents following the BDI model. In contrast to all other available BDI engines, Jadex fully supports the two-step practical reasoning process (goal deliberation and means–end reasoning) instead of operationalizing only the means–end process. This means that Jadex allows the construction of agents with explicit representation of mental attitudes (beliefs, goals and plans) and that automatically deliberate about their goals and subsequently pursue them by applying appropriate plans. The reasoning engine is clearly separated from its underlying infrastructure, which provides basic platform services such as life-cycle management and communication. Hence, running Jadex over JADE combines the strength of a well-tested agent middleware with the abstract BDI execution model. For the programming of agents, the engine relies on established techniques, such as Java and XML and, to further simplify the development task, Jadex includes a rich suite of run-time tools that are based upon the JADE administration and debugging tools. It also includes a library of ready-to-use generic functionalities provided by several agent modules (capabilities) [7]. OMACS is a metamodel for agent organizations. It defines the required organizational structure that allows multiagent teams to autonomously reconfigure at runtime, thus enabling them to cope with unpredictable situations in a dynamic environment [11]. MAGE, an agent-oriented programming environment, with complete tools to support agent-based requirement analysis, design, development and deployment, is a powerful development environment for autonomous computing [14].

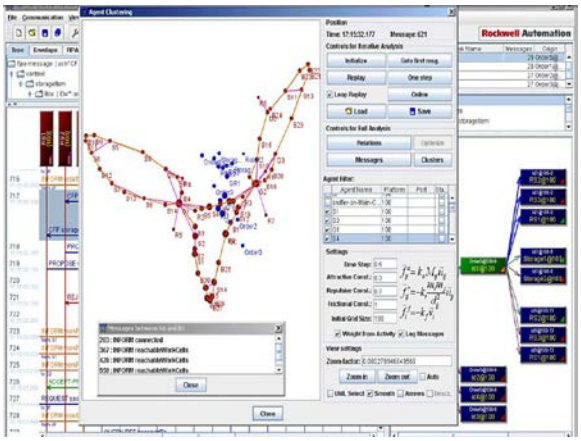


Fig 2 JavaSniffer user interface

III. Programming Language and Tools

Multi-agent systems programming languages, platforms and development tools are important components that can affect the diffusion and use of agent technologies across different application domains. In fact, the success of multi-agent systems is largely dependent on the availability of appropriate technology (i.e. programming languages, software libraries and development tools) that allows relatively straightforward implementation of the concepts and techniques that form the basis of multi-agent systems. Multi-agent systems can be realized by using any kind of programming language. In particular, object-oriented languages are considered a suitable means because the concept of agent is not too distant from the concept of object. In fact, agents share many properties with objects such as encapsulation, and frequently, inheritance and message passing. However, agents also differ from objects in several key ways; they are autonomous (i.e. they decide for themselves whether or not to perform an action on request from another agent); they are capable of a flexible behavior; and each agent of a system has its own thread of control. An important characteristic that multi-agent systems should provide is the capability to support interoperability among legacy software systems. Therefore, the availability of software tools for their integration with other common technologies can be a key to their success. The Internet is one the most important application domains and the most important communication means that multi-agent systems can use to provide interoperability among legacy software systems; therefore, a lot of current research and development work is oriented towards providing suitable techniques and software tools for the integration of multi-agent systems with Web technologies such as, for example, Web services and Semantic Web technologies.

IV. JADE and the Agents Paradigm

JADE is a software platform that provides basic middleware-layer functionalities which are independent of the specific application and which simplify the realization of distributed applications that exploit the software agent abstraction. A significant merit of JADE is that it implements this abstraction over a well-known object-oriented language, Java, providing a simple and friendly API. The following simple design choices were influenced by the agent abstraction**.** An Agent is Autonomous and Proactive: An agent cannot provide call-backs or its own object reference to other agents in order to mitigate any chance of other entities coopting control of its services. An agent must have its own thread of execution, using it to control its life cycle and decide autonomously when to perform which actions. The System is Peer-to-Peer each agent is identified by a globally unique name (the Agent Identifier, or AID, as defined by FIPA). It can join and leave a host platform at any time and can discover other agents through both white-page and yellow-page services (provided in JADE by AMS and the DF agents as defined also by the FIPA specifications). An agent can initiate communication with any other agent at any time it wishes and can equally be the object of an incoming communication at any time. On the basis of these design choices, JADE was implemented to provide programmers with the following ready-to-use and easy-to-customize core functionalities:-

• A fully distributed system inhabited by agents, each running as a separate thread, potentially on different remote machines, and capable of transparently communicating with one another, i.e. the platform provides a unique location-independent API that abstracts the underlying communication infrastructure.

• Efficient transport of asynchronous messages via a location-transparent API. The platform selects the best available means of communication and, when possible, avoids marshalling/unmarshalling java objects. When crossing platform boundaries, messages are automatically transformed from JADE's own internal Java representation into proper FIPA-compliant syntaxes, encodings and transport protocols.

• Support for agent mobility. Both agent code and, under certain restrictions, agent state can migrate between processes and machines. Agent migration is made transparent to communicating agents that can continue to interact even during the migration process.

• A set of graphical tools to support programmers when debugging and monitoring. These are particularly important and complex in multi-threaded, multi-process, multi-machine systems such as a typical JADE application.

• Integration with various Web-based technologies including JSP, servlets, applets and Web service

technology. The platform can also be easily configured to cross firewalls and use NAT systems.

• An in-process interface for launching/controlling a platform and its distributed components from an external application.

V. The Utility of Agent and WEB Service Integration

Integrating Web services and software agents brings about an obvious benefit: connecting application domains by enabling a Web service to invoke an agent service and vice versa. However, this interconnection is more than simply cross-domain discovery and invocation; it will also allow complex compositions of agent services and Web services to be created, managed and administered by controller agents. To the users of Web services, whether human or computational, agents can be a powerful means of indirection by masking the Web service for purposes of, for example, redirection, aggregation, integration and administration. Redirection describes the case where a Web service may no longer be available for some reason, or the owner of the Web service wishes to temporarily redirect invocations to another Web service without removing the original implementation. Aggregation allows several Web services to be composed into logically interconnected clusters, providing patterned abstractions of behavior that can be invoked through a single service interface. Integration describes the means of simply making Web services available to consumers already using, or planning to use, agent platforms for their business applications, and administration covers aspects of automated Web service management where the agent autonomously administers one or more Web services without necessary intervention from a human user.

VI. Conclusion

Many researchers in the MAS community have recognized the advantages of an agent based approach to building deployable solutions in the number of application domains comprising complex, distributed systems. Autonomic Computing is providing new vistas in reducing the complexity incurred in today's distributed systems. It minimizes human intervention and reduces the administration cost of enterprise IT systems. With multi agent support it also provides adaptability, intelligence, collaboration, goal oriented interactions, flexibility, mobility and persistence in software systems.

In this paper, we mentioned JADE was implemented to provide programmers with the ready-to-use and easy-to-customize core functionalities .An Agent is Autonomous and Proactive. In addition we have recommended an agent-Web service that has the features of both the agent technology as well as the Web services technology and is managed by an autonomic system based on multi-agent support. This can help to develop enterprise IT systems that are optimal, highly available.

References

[1] John Wiley, Sons Ltd, 2007, Developing Multi-Agent Systems with JADE

[2] J. P. Bigus D. A. Schlosnagle, A toolkit for building multiagent autonomic systems, IBM Systems Journal, Vole 41, NO 3, 2002

[3] Henri Avancini, Analía Amandi. A Java Framework for Multi-agent Systems, SADIO Electronic Journal of Informatics and Operations Research, vol. 3, no. 1, pp. 1-12 (2000).

[4] Fatemeh Daneshfar, Hassan Bevrani. Multi-Agent Systems in Control Engineering: A Survey, Hindawi Publishing Corporation Journal of Control Science and Engineering Volume 2009, Article ID 531080, 12 pages.

[5] Alessio Bosca, Dario Bonino. Ontology Exploration through Logical Views in Protégé, 18th International Workshop on Database and Expert Systems Applications, IEEE, 2007.

[6] Pavel Vrba, Pavel Tich. Rockwell Automation's Holonic and Multiagent Control Systems Compendium, IEEE Transactions On Systems, MAN, And Cybernetics—PART C: Applications And Reviews, Vole. 41, NO. 1, January 2011

[7] Frank Chiang, Robin Braun, Autonomic Service Configuration for Telec-munication MASS with Extended Role-Based GAIA and JADEx. 2005 IEEE

[8] Gilda Pour, Multi-Agent Autonomic Architectures for Quality Control Systems, San Jose State University San Jose, CA, U.S.A.

[9] Gerald Tesauro, David M. Chess. A Multi-Agent Systems Approach to Autonomic Computing, AAMAS'04, July 19-23, 2004, New York, New York, USA

[10] Jun W, JI Gao, Bei-Shui Liao, Jiu-Jun Chen. Multi - Agent System Based Autonomic Computing Environment, Proceedings of the Third International Conference on Machine Laming and Cybemetics, Shanghai, 26-29 August 2004.

[11] Walamitien H. Oyenan and Scott A. DeLoach. Design and Evaluation of a Multiagent Autonomic Information System. International Conference on Intelligent Agent Technology 2007 IEEE/WIC/ACM.

[12]Huaglory Tianfield. Multi-Agent Autonomic Architecture and Its Application in E- Medicine, Proceedings of the IEEE/WIC International Conference on Intelligent Agent Technology, 2003.

[13] Katia P. Sycara. Multiagent Systems, This Is a Publication of the American Association for Artificial Intelligence, 1998

[14] Zhongzhi Shi, Haijun Zhang, Yong Cheng. MAGE: An Agent-Oriented Software Engineering Environment, Proceedings of the Third IEEE International Conference on Cognitive Informatics, 2004.

Application of Cluster Analysis In Expert System – A Brief Survey

Mamta Tiwari[1], Dr. Bharat Mishra[2]

[1] Dept. of Computer Application, U.I.E.T., C.S.J.M. University Kanpur, U. P., India

[2] Dept. of Physical Sciences, M.G.C.G. Vishwavidyalaya, Chitrakoot. (M.P.) India.

Abstract

This is era of knowledge and information. One very major task that has been evolved now a day is to mine a knowledge base. On the other hand expert systems are used extensively in many domains. There are many applications of expert systems for predicting and finding a feasible solution for any particular problem. Various tools also have been evolves for upgrading and modifying the existing expert systems and making them more useful in their intended purposes. The current paper explains the expert systems that use cluster analysis as a tool and briefly discusses few such expert systems.

Keywords: *Clustering methods, Expert systems, MovieGEN, Illiad, Tourist Expert systems.*

1. Introduction

There is virtually an explosion of information these days but unfortunately this tremendous progress of mankind in every walk of life is largely concentrated in the urban limits. In a country like India and the countries of the so-called third world, the importance of good expert system is beyond description.

Expert systems are intelligent computer programs that are designed to simulate the problem-solving behavior of a human being, who is an expert in a narrow domain or discipline. Since the human race is still facing plethora of problems directly related to their life and livelihood involving poor living conditions, poor medical facilities, lack of educational facilities and recreational facilities etc. Expert systems can be proved a boon in disguise for such sufferers. There are many applications of expert systems ranging from agriculture, finance, education, medicine to military science, process control, space technology and engineering.

Extracting knowledge from existing sources of information is a key development area to unlock yet unknown relationships between specific data point and data mining can be proved as great help in this regard.

Before discussing applications of various tools of data mining especially cluster analysis in expert system, let us have a look on clustering and various methods and techniques used for clustering.

The process of grouping or making sets of nearly similar type of physical or abstract objects is known as clustering. The groups thus formed are known as clusters. Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters [1]. We can compare the clusters with classes as in object-oriented programming paradigm. The slight difference between clusters and class is that, in class every object of it is exactly identical in properties whereas in cluster, every object is similar to other objects of its cluster and dissimilar to the objects of other cluster based on some particular properties.

When data mining is concerned, clustering is having an edge over classification. In data mining, we have to mine a large set of data, clustering saves us from a costly overhead of collection and labelling of a large set of training tuples or patterns, which the classifier uses to model each group. Clustering is sometimes also called as data segmentation in some applications because clustering partitions large data sets into groups according to their similarity.

There are several clustering techniques available and those are organized into the following categories as partitioning methods, hierarchical methods, density-based methods,

grid-based methods, model-based methods, methods for high-dimensional data and constraint-based clustering.

2. Clustering Methodology

We here briefly present various methods of clustering techniques [1].

2.1 Partitioning Method

Given a database of *n* objects or data tuples, a partitioning method constructs *k* partitions of the data, where each partition represents a cluster and $k \leq n$. That is, it classifies the data into *k* groups, which together satisfy the following constraints:

(1) Each group must contain at least one object.
(2) Each object must belong to exactly one group.

Some popular techniques as k-mean and k-medoids are placed in this category.

2.2 Hierarchical Method

A hierarchical method creates a hierarchical decomposition of the given set of data objects. It can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed.

The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one or until a termination condition holds.

The divisive approach which is also known as the top-down approach starts with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds. Chameleon and BIRCH are some techniques of this kind.

2.3 Density-based Methods

Most partitioning methods cluster objects, based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty at discovering clusters of arbitrary shapes. That is why some other clustering methods have been developed those are based on the notion of density. The general idea is to continue growing the given cluster as long as the density (number of objects or data points) in the neighbourhood exceeds some threshold; that is, for each data point within a given cluster, the neighbourhood of a given radius has to contain at least a minimum number of points.

DBSCAN and its extension, OPTICS, are typical density-based methods.

2.4 Grid-based Methods

Grid-based methods quantize the object space into a finite number of cells that form a grid structure. All of the clustering operations are performed on the grid structure (i.e., on the quantized space). The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space. STING is a typical example of a grid-based method.

2.5 Model-based Methods

Model-based methods hypothesize a model for each of the clusters and find the best fit of the data to the given model. A model-based algorithm may locate clusters by constructing a density function that reflects the spatial distribution of the data points.

EM is an algorithm that performs expectation-maximization analysis based on statistical modelling. COBWEB is a conceptual learning algorithm that performs probability analysis and takes concepts as a model for clusters. SOM (or self-organizing feature map) is a neural network-based algorithm that clusters by mapping high dimensional data into a 2-D or 3-D feature map,

2.6 High - Dimensional Data-based Methods

Many applications require the analysis of objects containing a large number of features or dimensions. As the number of dimensions increases, the data become increasingly sparse so that the distance measurement between pairs of points become meaningless and the average density of points anywhere in the data is likely to be low. CLIQUE, pCluster and PROCLUS are some such techniques.

2.7 Constraint-based Method

This is a clustering approach that performs clustering by incorporation of user-specified or application-oriented constraints. A constraint expresses a user's expectation or describes "properties" of the desired clustering results, and provides an effective means for communicating with the clustering process. Various kinds of constraints can be

specified, either by a user or as per application requirements.

3. Various Application of Clustering in Expert Systems

We have chosen for review, three expert system from three different areas of utilization and interest. The one we first pick for the review is Iliad [2]. Iliad was designed to solve a broad variety of internal medical problems. The others are MovieGEN [3] and an expert system for tourism. MovieGEN, an expert system for movie recommendation. The system uses machine learning and cluster analysis based on a hybrid recommendation approach. The system takes in the users' personal information and predicts their movie preferences using well-trained support vector machine (SVM) models [3].

A brief review of the above written three types of expert systems is given next, starting with the Iliad.

3.1 Iliad

Iliad is a medical expert system designed to run on Macintosh computers. Iliad is a expert system whose medical knowledge is organized by disease into "frames" that each contains multiple findings that may be expected in that disease [2].

Most medical expert systems contain two essential components: an "inference engine" and a "knowledge base". Iliad's knowledge base was frame-oriented and was upgraded by including clustered knowledge frames. Clusters are Boolean decision frames that contain finding those are conditionally dependent and describe pathophysiologic concepts.

Frames in Iliad, contain multiple findings that may be likely to present in that disease. These findings are processed sequentially, using Bayes' Theorem, when knowledge about the patient becomes available. Iliad incorporates newly designed knowledge frames called "clusters".

The clustered model of knowledge representation was developed for Iliad because the previous, non-clustered model caused Iliad to produce inaccurate, overconfident diagnoses. Since Iliad's inference engine uses Bayes' theorem, all patient findings used in a case must be, conditionally independent but the assumption of conditional independence required by Bayes' theorem is often violated in medicine. For instance, the patient findings "fever" and "chills" commonly co-occur. Such conditionally dependent findings, if used together in a Bayesian frame, bias the inference engine towards overconfident diagnoses.

A new, "clustered" knowledge model was devised to solve the problem of conditional dependence. Clusters are knowledge frames containing highly conditionally dependent findings. These groups of findings often have pathophysiological meaning. Most clusters are decided according to Boolean logic. Boolean frames may be decided with various levels of certainty (e.g., "definite" "probable" "unlikely" or "absent"). Decisions made about clusters are passed to Bayesian frames, much as procedures in Pascal pass variable results to the main program.

The cluster analysis was performed using the BMDP8M (BioMeDical Proprietary package version 8M) cluster analysis programs on the University of Utah's Sperry Univac 1100 mainframe computer. Although the goal of cluster analysis is to generate groups of highly similar findings, this goal can actually be reached most easily by starting with a matrix of dissimilarities between the findings. Groups of highly similar findings can be defined as having low dissimilarities. The elements of this dissimilarity matrix can be based on Pearson correlations $(1 - r^2)$, conditional probabilities [1 - p(fi/fj)], or Euclidean distances $(fi - fj)^2$.

The final step in the cluster analysis program is assembly of a "results" matrix. This is a two-dimensional matrix of findings by dissimilarity scores. The findings dimension must be rank ordered by average dissimilarity. The BMDP8M program has several strategies to accomplish this goal. One easily understandable way is to locate the most dissimilar finding and set it aside. This process can be performed recursively until all the findings are sorted by dissimilarity. This process produces a results matrix that is rank-ordered along the findings dimension.

There is evidence from psychological research that humans naturally employ hierarchically structured, clustered knowledge models. Cluster analysis is a technique that can be used to discover and validate clustered knowledge concepts. Since clustered knowledge models are natural for humans, a clustered expert system may be able to provide better explanations for its diagnoses.

3.2 MovieGEN

MovieGEN, an expert system for movie recommendation. The system was implemented using machine learning and cluster analysis based on a hybrid recommendation approach. The system takes in the users' personal

information and predicts their movie preferences using well-trained support vector machine (SVM) models [3].

Based on the SVM prediction it selects movies from the dataset, clusters the movies and generates questions to the users. Based on the users' answers, it refines its movie set and it finally recommends movies for the users.

Recommendation systems are special types of expert systems in the sense that they combine the knowledge of the expert in a given domain (for the product type being recommended) with the user's preferences to filter the available information and provide the user with the most relevant information. Personalization of the recommendations works by filtering a candidate set of items (such as products or web pages) through some representation of a personal profile. Two main paradigms for the filtering are content-based approach and collaborative approach. Most recommendation systems use a hybrid approach, which is a combination of these two approaches. A content based recommendation system uses the user's past history to recommend new items where as a collaborative approach uses the preferences of other people with similar tastes for recommending items to the user.

The MovieGEN, a movie recommendation system was developed that recommends movies to users based on their personal information and their answers to questions based on movies. A user model was created using SVM based learning techniques. Using this model it can be predicted the genres and the period of the movies that the user prefers based on the user's personal information.
This incorporates the collaborative approach i.e. the user's choices are predicted based on the choices of other similar users. A variation of the content based approach was implemented by taking into consideration the user choices not based on the user's past history but based on the answers he gives to the questions asked by the system. The system has been developed in Java and currently uses a simple console based interface. Machine learning constitutes an essential step in this approach. For any machine learning model, the data sets are composed of two parts, the input and the output. The output is usually the subjects of interest, in other words, the part that we want to predict or classify, while the input is the set of elements that might have impacts upon the output. Machine learning attempts to correlate the output and the input, by approximating functions in between whose formulations are unknown.

K-means is used in this approach as the cluster analysis tool. K-means is one of the most widely-used partitioning methods in the data mining community, and has been studied and applied in a wide range of domains, including bioinformatics (Guralnik and Karypis, 2001; Zhong et al., 2005), pattern recognition (Estlick et al, 2001; Saegusa and Maruyama, 2007; Filho et. Al, 2003), text classification (Steinbach, 2000), etc.

Support Vector Machine (SVM), an effective and efficient machine learning tool that has been extensively studied within the machine learning community, is utilized in this expert system as the machine learning algorithm. SVM is incorporated in this system to establish a
correlation analysis between personal particulars of a user and his or her personal preference for movies. For each set of user input, a SVM is trained based on a predefined set of training samples, which increases in size after each time the system is used.

In this system, there are two sets of data, the training data set and the testing data set. The training data set comprises of different training samples, each of which is a combination of an input vector and an output vector. The testing data set comprises of different testing samples, each of which contains only an input vector, while the output vector is to be predicted by the machine learning.

Once the genres and period the user prefers based on his/her personal information using SVM, has been predicted, this information can be used to select movies from the dataset, generate questions about these movies and finally return a refined movie recommendation to the user.

3.3 An Expert System for Summer Tourism

This study had an aim to support tourism sector in Turkey by using an expert system [4]. Thus, tourists will be able to select the most suitable holiday places for themselves. Before the tourists go to a holiday place which they have not visited before, they make a research about this place. Also, some surprises in this place are learnt before the tourists go and many tourists do not like this situation.

Therefore, an operation of text mining was preferred in this study. Thus, tourists do not need a research about the holiday places. All that the expert system will return will be a decision according to users' preferences. The expert system had an aim to return more decisions than one. When a tourist uses the system; only one place is not returned, sorted places from the most suitable place to the least suitable one are given. Therefore, a clustering structure was needed. After the system decides the most suitable place for the tourist; the cluster where this suitable place locates finds and the all holiday places in this cluster

are recommended in order from the most suitable to the least suitable.

There are lots of features as attributes from text collection although there will be a low number of holiday places; thus, a large dataset was obtained. Therefore, K-Means clustering algorithm as both simple and fast clustering algorithm was preferred. However, K-Means has problem about deciding the space of clusters, because K-Means can give a different space of clusters with same dataset at each working. The cause of this situation is that K-Means starts clustering with random initial centre points. Therefore, K-Means++ clustering was used as a new approach to K-Means without random initial centre points and with consistent result spaces.

This study had four steps briefly. Firstly, the most preferable places for summer holiday in Turkey were decided. According to a research on web pages of Cultural and Tourism Ministry of Turkey about tourism, the most important places are Alanya, Ayvalık, Bodrum, Çeşme, Datça, Didim, Dikili, Fethiye, Kaş, Kuşdası, Marmaris, Side and Yalova. These places are preferred by both foreign and regional tourists a lot because of both common and unique features of these places. Therefore, secondly, the features must be determined. For this step, a research with rich documents about these places was done on web and these documents were collected in a text file for each place. These text files would be used for text mining operations in the next steps.

Thirdly, a dictionary was created for each place from the collection of text files. These dictionaries are too large to process, because these dictionaries content stop-words and unnecessary words for tourism. Therefore, some words were determined to be deleted from the dictionaries and they were deleted; thus, the satisfactory dictionaries were obtained for each holiday places. A data warehouse must be needed for mining operations on these dictionaries.

Therefore, pre-processes with vector space model were needed; thus, a dataset was obtained with tuples and their attributes. In last step, this dataset was used by K-Means++. It gave a space of clusters where there were the places. Finally, an expert system was ready to use and holiday places were recommended according to these clusters and the expectations of tourists.

4. Summary and Future Scope

We strongly believe that clusters and cluster analysis should be a part of expert systems because they can improve the accuracy of Bayesian decision systems [1]. But clusters are also important because they model innate human knowledge structures. Strong evidence supports our assertion that cluster-like knowledge structures are natural human mechanisms for organizing information. While mechanisms of human knowledge organization are incompletely understood, it is clear that humans must employ simplifying heuristics in complex situations. The cluster heuristic allows data to be combined into useful patterns that may lead to better decisions.

In present scenario the application of expert system has already gained momentum, still there are lot of areas where a great deal of efforts is still required. The knowledge engineers and information scientists have done tremendous work if form of expert systems now there is an immense need to upgrade them and make them more useful. We believe that various data mining approaches and techniques such as k-mean, pCluster and STING etc. are going to play a vital role in this mega job. The validation and improvement of expert system, which in turn is already a very complex phenomenon, is the demand of time and we hope that we can see equal efforts in this direction also.

5. References

[1] Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques 2nd Ed. - Morgan Kaufmann Publishers

[2] Michael J Lincoln, Charles Turer, Brad Hesse MS, Randolph Miller - A Comparison of Clustered Knowledge Structures in Iliad and in Quick Medical Reference.

[3] Tilak Gaurangi, Eyrun A. Eyjolfsdottir, Nan Li MovieGEN: A Movie Recommendation System.

[4] Yunus Doğan and Alp Kut - An Expert System for Summer Tourism in Turkey by Using Text Mining and K-Means++ Clustering ,ICT Innovations 2010 Web Proceedings ISSN 1857-7288

Mamta Tiwari completed her Master of Technology in Computer Science from U.P. Technical University, Lucknow (U.P.) in 2006. Earlier she had completed Master of Computer Applications from Rani Durgavati University, Jabalpur (M.P.) in 2001 and Master of Science from Kanpur University, Kanpur (U.P.) in 1994. She has put up over 10 years of teaching experience to engineering students. Presently she is working as a lecturer in the Dept. of Computer Application in University Institute of Engineering and Technology, C.S.J.M. University, Kanpur. Presently she is perusing her Doctoral Program from Mahatma Gandhi Gramodaya Viswavidyalaya, Satna (M.P.) India. Her research interest includes Data Mining, Artificial Intelligence and Software Engineering.

A Framework for Picture Extraction on Search Engine Improved and Meaningful Result

Anamika Sharma[1], Sarita Sharma[2]

[1] Research Scholar, Singhania University, Rajasthan, India

[2] Computer Science, MDU Rohtak, DAVIM Faridabad, Haryana, India

Abstract

Searching is an important tool of information gathering, if information is in the form of picture than it play a major role to take quick action and easy to memorize. This is a human tendency to retain more picture than text. The complexity and the occurrence of variety of query can give variation in result and provide the humans to learn something new or get confused. This paper presents a development of a framework that will focus on recourse identification for the user so that they can get faster access with accurate & concise results on time and analysis of the change that is evident as the scenario changes from text to picture retrieval. This paper also provides a glimpse how to get accurate picture information in advance and extended technologies searching framework. The new challenges and design techniques of picture retrieval systems are also suggested in this paper.

Keywords: *Picture Retrieval, CBIR, Standard Query, Image Searching, Online study.*

1. Introduction

The world has been dependent on the searching and is going to depend on it in the future. Searching is base of Learning and Learning is a never-ending process. Searching normally does on search engine in the form of text, images, news, maps, web sites etc. Learning will be more effective if it will be in the form of picture. Picture can be of text, graph, and images. Why only image searching? Simple answer is this because learning is more interactive and interesting in comparison with text. Human tendency to retention text is 20% as well as for images retention is 80%. On-line collections of images are growing larger and more common, and tools are needed to efficiently manage, organize, and navigate through them. Image searching is very helpful in every field and of any age group, leading fields are researching, learning and education and its users have been able to search the required image content with the help of the search engines like Goggle. But as the human race is moving towards the future, changes are taking place, even in the nature of the searching and the formats used. Content Base Image searching has shown ways to cope up with this change. It has helped to find the required information in easy way of learning, like images, graphs, pictures etc.

2. Image Searching

Today a number of search engines are available that give search facility for online database. Categories of these search engine includes web search engine, selection based search engine, Meta search engine, desktop search engine, web portal and vertical market web sites. Search Engines are information retrieval system designed to help to find information stored in online database. There are different ways of image searching. Some are based on simple searching of embedded annotations, metadata, textual context etc, while other complicated methods may include image classifications for searching that are based on color, texture, false color concepts etc. Sometimes for better precision of the resultant images, this type of search requires associating meaningful storage methods or semantic techniques that can be used further for all images of the database.

There are many types of search engines but this study is focused on query based image search engines. Before going to image search, first to find the need of image search. As per [1] there are different varieties of images available.

- Explosive growth of online image/video
- 5 billion images on web /31 million hours of TV program each year.
- Successful service like you Tube and Flickr
- Image/Video search exciting opportunity

2.1 Image Segmentation

Each image most likely contains multiple objects, or an object and a background. Therefore, extracting features globally is not appropriate. For this reason, as per [2] start by splitting each image into regions of similarity, using an image segmentation algorithm, with the intuition that each of these regions is a separate object in the image. Image segmentation is a well-studied problem in computer vision.

This segmentation algorithm partitions an image into similar regions using a graph-based approach. Each pixel is a node in the graph with undirected edges connecting its adjacent pixels in the image. Each edge has a weight encoding the similarity of the two connected pixels. The partitioning is done such that two segments are merged only if the dissimilarity between the segments is not as great as the dissimilarity inside either of the segments.

2.2 Image Annotations

Manual image labeling, known as manual image annotation, is practically difficult for exponentially increasing image database. As per [3], most of those images are not annotated with semantic descriptors, it might be a challenge for general users to find specific images from the Internet. Image search engines are such systems that are specially designed to help users find their intended images.

3. Crucial Concept In Image Searching

The biggest issue for image searching system is to incorporate versatile techniques so as to process images of diversified characteristics and categories. Many techniques for processing of low-level cues are distinguished by the characteristics of domain-images. As per [4], The performance of these techniques is challenged by various factors like image resolution, intra-image illumination variations, non homogeneity of intra-region and inter-region textures, multiple and occluded objects etc.The major difficulty is a gap between mapping of extracted features and human perceived semantics. The dimensionality of the difficulty becomes adverse because of subjectivity in the visually perceived semantics, making image content description a subjective phenomenon of human perception, characterized by human psychology, emotions, and imaginations. The image retrieval system comprises of multiple inter-dependent tasks performed by various phases. Inter-tuning of all these phases of the retrieval system is inevitable for over all good results.

4. Query Base Image Searching

QBIS is the primary mechanism for retrieving information from a database and consists of questions presented to the database. In query base image searching gives number of images matches with words present in query. To search for images, a user may provide query terms such as keyword, image file/link, or click on some image, and the system will return images "similar" to the query. The similarity used for search criteria could be meta tags, color distribution in images, region/shape attributes, etc. Most commercial image search engines fall into this category. On the contrary, collection-based search engines index image collections using the keywords annotated by human indexers. As per [5],Different approaches of QBIS includes text base,content base,context base,keybase and semantic base image searching.

5. Different Approaches of QBIS

- Text base query image searching
- Content base query image searching
- Context base query image searching
- Key based query image searching
- Semantic base query image searching

5.1. Text-Based Query Image Search Engines

Index images using the words associated with the images. Depending on whether the indexing is done automatically or manually, image search engines adopting this approach may be further classified into two categories: Web image search engine or collection-based search engine. Web image search engines collect images embedded in Web pages from other sites on the Internet, and index them using the text automatically derived from containing Web pages .

5.2 Content -Based Query Image Search

Content-based query image searching was initially proposed to overcome the difficulties encountered in keyword-based image search in 1990s. As per [6] Image meta search - search of images based on associated metadata such as keywords, text, etc. Content-based image Retrieval (CBIR) – the application of computer vision to the ***image search***.

5.3 Context Base Query Image Searching

In context base query , where searching query processe on the thesaurus meaning of words present in query .For example if the query is to find a “CAR” then ihe images can be toy car for children,brand product of car

company,image of any electonic circuit etc.It depend in what context user wants the result.

5.4 Key Based Query Image Serching

A Key base image searching is based on key words include in query required to transform knowledge of a passage into effective query strings in order to retrieve images from keyword based .As per example [7] consider the following passage in a child story book:

"I see three lemurs jumping around and screaming. The snake scares them. However, a sloth is still soundly sleeping. Around the corner, many children are watching a shark swimming swiftly."

The subjects and objects in this passage include "snake", "lemur" and "sloth". During the first execution of the image retrieval process, the query string is formulated as "snake & lemur & sloth".

In the case the response from the image archives indicates that there is no image annotated with all these terms, there is a need for a second query. Presumably, one reasonable strategy is to find images of a place where the "snake", "lemur", and "sloth" could all possibly appear

5.5 Semantic Base Query Image Search

The ideal CBIR system from a user perspective would involve what is referred to as semantic serching, where the user makes a request like "find pictures of dogs" or even "find pictures of Abraham Lincoln". This type of open-ended task is very difficult for computers to perform - pictures of chihuahuas and Great Danes look very different, and Lincoln may not always be facing the camera or in the same pose. Current CBIR systems therefore generally make use of lower-level features like texture, color, and shape, although some systems take advantage of very common higher-level features like faces. Although search engine is really a general class of programs, the term is often used to specifically describe systems like Google, Alta Vista and Excite that enable users to search for documents on the World Wide Web and USENET newsgroup.

According to [8] in systems extract visual features from images and use them to index images, such as color, texture or shape. Color histogram is one of the most widely used features. It is essentially the statistics of the color of pixels in an image.

Web images have rich metadata such as filename, URL and surrounding text for indexing and searching, different from traditional text-based approach, no manually labeling work is needed in current Web image search engines. The success of text-based image search engines has shown the power of textual information associated with Web images.

6. Shifting From Text To Image Searching

Searching is based today either on video image or text as per the requirement of the user. The Searching scenario has been totally changed as compared to the previous or only text based searching. In Image or video searching planning and thinking is important, especially for constantly changing and dynamic modules that are industry-linked and practice-oriented. Image searching is the next step in the evolution of video searching. However, Image searching offers a high degree of learning than the text searching. Now video searching, image searching and text searching comes under the umbrella of information searching.

7. Image Searching Issues

Searching through still images presents an interesting challenge to search engine developers. The way most search engines operate today is by appending text descriptions to the video clips and/or images, so that the searches are based on the text. As per [9] Video and photo searching is something that is still being developed and explored. For example, being able to search for an image of the Taj Mahal would be very challenging for developers, as this would require a query by image content. Basically, this means that the search engine or tool would have to be sophisticated enough to recognize an image of the Taj Mahal and differentiate it from all other possible images. Instead, one of the approaches for searching images is to search for distinctive features of an image. For example, the search tool could look for images with Taj Mahal's features such as a architecture, doors or different types of work done on the walls of Taj Mahal. Another approach would be to search for distinctive colors of the known image. In this case, the search engine could look for the distinctive fading white color of the buildings.

8. Challenges In Image searching

This would present an interesting challenge for users and developers to establish a good interface for a searching tool of this type.As per [10], Efficiently searching video is even more complex than still images because now that search engines or tools have to be sophisticated enough to handle movement, lighting, and different camera angles. The searching of a video or film would have to be more sophisticated than to simply search a video frame by frame for the desired result. Users may also want to search for specific scenes in video or for zooming in and out. As we all know, user can use any kind of query (text, content, context, keywords, semantic) for image

searching, so the need arises for the structured query content, so that it can be give the resultant images according to the type of query. Following problems can arises when the user give the query in text box of search engine:

1. No rules for writing of query for images.
2. Difficulties to identifying important words
3. Content base query systems have not any standard format of query image
4. Cannot mention in what context user want what type of image.
5. No Typing limitation but it takes only few important words.
6. Cost of spending lots of time scrolling through image searching result.
7. Speed of accessing the web.
8. Unwanted links / irrelevant data on searching.
9. Non-compatible sites of web (product searching sites) only.
10. Difficulties to narrowing down the semantic gap.

8.Implementation Technique Of Image Searching

Because of number of difficulties, we need to develop a framework that will focus on recourse identification for the user so that they can get faster access with accurate & concise results on time.

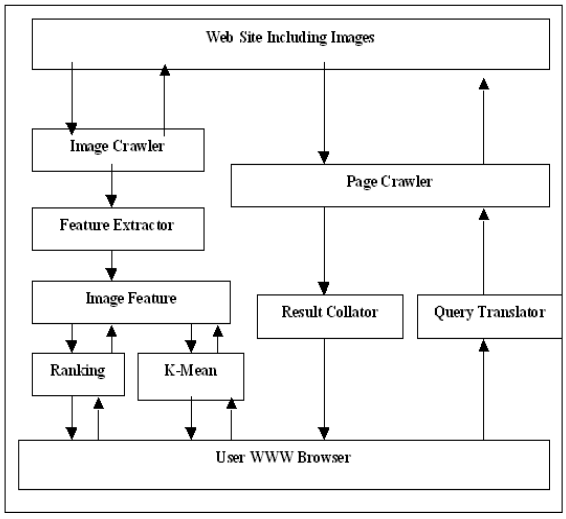


Fig.1. Framework of Picture Retrieval

Development of Framework for user to find the recourses on the bases of type of query given. Now the Proposed framework of implementation technique is to put the query on web browser. This query will translate through as

1.User input a text Query to the browser.

2.The Query Translator extracts the query from HTTP request, then translate the query into the input format for each text based image search engine.

3. The page crawler sends the query to the each search engine and collects the HTML file containing the URL of image retrieved by the search engines, then parses the HTML file to obtain the individual URL.

4. The result collator merges the result and shows the first page of retrieval image.

5. Using the URLs the image crawler retrieves the image from the Internet to construct the initial image set.

6. Feature extractor computes the image content feature vector for all images in the initial image.

7.Based on user's new request, cluster the image set using the feature vector and K-mean algorithms.

8.Based on the feedback images selected by the user, compute the distance of feature vector between the feedback image and the image in the initial image set. Re-rank images according to the distance and display the re-ranked image to the user.

9.Conclusion

This paper has presented an up coming wave of image searching and force for shifting from Text searching to Image searching with acceptance of new technology and directions. This includes the embedded software tools that help in online searching of content include text, image and video based on query. The related researches are also presented in this paper. With the advent of the Internet, information from all over the world is available to the people. Since there is so much information out there, people require an automated method to search through all of it. The search engines available today provide users with this ability, but primarily for text based searching. As the Internet moves further and further into being able to support multimedia, users and corporations will need to take advantage of new searching techniques. Some companies have even begun to hire "Web Specialists" to assist them in becoming aware of what searching facilities.

The idea of being able to search images, video, or audio based on the content is possible since the Internet is an electronic medium, but something that is still in development. As more and more users begin to understand the concepts behind searching these media, the need to do so will rise. However, today, most of the work being done to allow users to search this media is still in the developmental phases. The approaches described above are simply ideas and theories for ways in which this type of searching could be possible. As per [11] The idea of generating a storyboard from a video, or searching a video based on a moving sketch, or searching audio based on

content and colored wave files are simply ways in which searching multimedia may become a reality.

10.Future Scope

Image searching provides some unique and interesting challenges for developers to come up with some sort of automated way of searching through video and/or still images. One method that is currently in development is to generate a storyboard out of a video. Storyboards typically consist "of a series of sketches showing each shot in each scene as it will be filmed, and possibly some indication of the action-taking place e.g., an arrow showing the direction of movement. A 'shot' is defined as a section of action during which the camera films continuously without interruption." As per [12] A storyboard is typically used by writers and directors while making a movie to plan the action of the shot, to review camera angles, and provide a summary of the film. Essentially, the proposed approach would be to take a finished video product and generate the storyboard based on the finish video. "In order to reverse-engineer a storyboard from the finished video sequence, it is necessary to identify three properties of each shot in the sequence. These are: (1) the start point of the shot, (2) the end point of the shot and (3) the picture that best represents the shot as a whole." As per [12] Once the storyboard has been generated, it will be easier to search for video sequences, especially in large video libraries.

Another approach to video searching is the search actual video using video cues. At Columbia University, a system called VideoQ is being developed that does this. The theory behind this system is to have the user actually draw out an animated scene as the query. "In an animated sketch, motion and temporal duration is the key attributes assigned to each object in the sketch in addition to the usual attributes such as shape, color, and texture. Using the visual palette, we sketch out a scene by drawing a collection of video objects." As per [13] The VideoQ system will then search its video library for videos that match the animated sketch. The VideoQ system is intended to be on the Internet and use various Java applets to allow the user to create these animated sketches.

References

[1] Chang, Shih-Fu, June –2007," Recent Advances And Open Issues Of Digital Image/Viedo Search" ,Digital Vedio And Multimedia Lab.
[2]Pedro F Felzenswalb,Daniel P . Huttenlocher,"Efficient Graph-Based Image Segmentation "Artificial Intelligence Lab & Computer Science.
[3]Wang, X. J., Zhang, L., Jing, F., And Ma, W. Y. . In CVPR, 2006." Annotation Search : Image Auto-Annotation By Search"
[4] Nida Aslam , Infanullah " Limitation And Challenges: Image/Video Search & Retrieval "Doi:10.4156/Jdcta.Vol3.Issuel.Aslam
[5]Christopher C .Yang " Content Based Image Retrival : A Comparision Between Query By Example And Image Browsing Map Approaches " The Chinese Univerity Of Hong Kong,7Jan 2004.
[6] Abby A. Goodrum (2000) , Image Information Retrieval : An Overview of Current Research, SHFLDO ,VVXH RQ
[7]Sheng-Hao Hung, Pai-Hsun Chen , Context-Based Image Retrieval:A Case Study In Background Image Access For Multimedia Presentations "Iadis International Conference WWW/Internet 2007, Vila Real : Portugal (2007)"
[8] Dr. Fuhui Long, Dr. Hongjiang Zhang and Prof. David Dagan Feng, fundamentals of Content-based image retrieval.
[9]Oscar Palma . "Text And Multimedia Searching: Current Issues And Possibilities For The Future"CIS 447: Human Computer Interface Final Term Paper
[10]Shih-Fu Chang, John R. Smith," Finding Images/Video In Large Archives"Columbia's Content-Based Visual Query Project, D-Lib Magazine, February 1997 ,ISSN 1082-9873.
[11]Prakash Panday,Uday Pratap ,Sanjeev Jain ." Categorization And Searchng Of Color Image Using Mean Algo",L N College Of Technology ,Bhopal.
[12]Macer, Peter, Peter J. Thomas, Nouhman Chalabi, John F. Meech "Finding The Cut Of The Wrong Trousers: Fast Video Search Using Automatic Storyboard Generation" Communications Of The ACM, 1996.
[13]Christopher C Yang."Content –Based Image Retrival:A Comparision Between Query By Example And Image Browsing Map Approaches",The Chinese University Of Hong Kong,7th Jan 2004.

Ms. Anamika Sharma did her Master in computer Application from Gurukul University Haridwar in 1998 ,M.Tech From Allahabad Agriculture University,M.Phil from Vinayka Mission University Tamil Nadu and pursuing Ph.D from Singhania University Rajasthan. She is having about 13 years of teaching experience of postgraduate courses. She has guided more than hundred students in their project and published number of papers in national/International journals. She is a member of Computer Society of India .Her main Area of research include Computer Graphics, Data mining, Software testing and Quality assurance and object oriented Analysis and Design. At Present she is working in DAV Institute of Management, Faridabad as Associate Professor in Computer Science Deptt.

Ms. Sarita Sharma did her Master in Computer Application from IGNOU, New Delhi, India, She did her M.phil (Computer Science) from Ch. Devi Lal University,Sirsa , India and is pusuing Ph.D from Singhania University,Rajasthan, India. She is presently working as Associate Professor in Deptt. Of Computer Science, DAV Institute of Management , India. She has guided more than 90 students in their Projects and has published a number of papers in National/International journals. She is a member of Computer Society of India. Her areas of interest include Software Engineering, Data Mining, Relational Databases, Computer Languages etc. She has about 15 years of teaching experience. with current employment; association with any official journals or conferences

Analysis of Stemming Algorithm for Text Clustering

N. Sandhya[1], Y. Sri Lalitha[2] , V.Sowmya[3] , Dr. K. Anuradha[4] and Dr. A. Govardhan[5]

[1] Associate Professor, CSE Department, Gokaraju Rangaraju Institute of Engineering and technology, Hyderabad, Andhra Pradesh, 500 072, India

[2] Associate Professor, CSE Department, Gokaraju Rangaraju Institute of Engineering and technology, Hyderabad, Andhra Pradesh, 500 072, India

[3] Associate Professor, CSE Department, Gokaraju Rangaraju Institute of Engineering and technology, Hyderabad, Andhra Pradesh, 500 072, India

[4]Professor and Head, CSE Department, Gokaraju Rangaraju Institute of Engineering and technology, Hyderbad, Andhra Pradesh, 500 072, India

[5] Principal, JNTUH College of Engineering Jagityal, Andhra Pradesh, 500 501, India

Abstract

Text document clustering plays an important role in providing intuitive navigation and browsing mechanisms by organizing large amounts of information into a small number of meaningful clusters. In Bag of words representation of documents the words that appear in documents often have many morphological variants and in most cases, morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of clustering applications. For this reason, a number of *stemming Algorithms*, or *stemmers*, have been developed, which attempt to reduce a word to its *stem* or root form. Thus, the key terms of a document are represented by stems rather than by the original words. In this work we have studied the impact of stemming algorithm along with four popular similarity measures (Euclidean, cosine, Pearson correlation and extended Jaccard) in conjunction with different types of vector representation (boolean, term frequency and term frequency and inverse document frequency) on cluster quality. For Clustering documents we have used partitional based clustering technique K Means.

Performance is measured against a human-imposed classification of Classic data set. We conducted a number of experiments and used entropy measure to assure statistical significance of results. Cosine, Pearson correlation and extended Jaccard similarities emerge as the best measures to capture human categorization behavior, while Euclidean measures perform poor. After applying the Stemming algorithm Euclidean measure shows little improvement.

Keywords: *Text clustering, Stemming Algorithm, Similarity Measures, Cluster Accuracy.*

1. Introduction

With ever increasing volume of text documents, the abundant texts flowing over the Internet, huge collections of documents in digital libraries and repositories, and digitized personal information such as blog articles and emails are piling up quickly every day. For text documents, clustering has proven to be an effective approach and an interesting research problem. Clustering of text documents plays a vital role in efficient Document Organization, Summarization, Topic Extraction and Information Retrieval. Initially used for improving the precision or recall in an Information Retrieval System [1,2], more recently, clustering has been proposed for use in browsing a collection of documents [3] or in organizing the results returned by a search engine in response to user's query [4] or help users quickly identify and focus on the relevant set of results. Customer comments are clustered in many online stores, such as Amazon.com to provide collaborative recommendations. In collaborative bookmarking or tagging, clusters of users that share certain

traits are identified by their annotations. Document clustering has also been used to automatically generate Hierarchical clusters of documents [5]. The automatic generation of taxonomy of Web documents as the one provided by Yahoo! (www.yahoo.com) is often cited as a goal.

This paper is organized as follows. Section 2 describes the document representation used in the experiments, section 3 deals with the related work in finding stem of a word and an insight into clustering algorithms, Section 4 discusses the similarity measures and their semantics. Section 5 presents the K-means clustering algorithm and Section 6 explains experiment settings, evaluation approaches, results and analysis and Section 7 concludes and discusses future work.

2. Document Representation

The representation of a set of documents as vectors in a common vector space is known as the vector space model. Despite of its simple data structure without using any explicit semantic information, the vector space model enables very efficient analysis of huge document collections. The vector space model represents documents as vectors in *m*-dimensional space, i.e. each document *d* is described by a numerical vector of terms. Thus, documents can be compared by use of simple vector operations.

There are three document encoding methods namely, *Boolean, Term Frequency and Term Frequency with Inverse Document Frequency.*

The simplest document encoding is to use binary term vectors, i.e. a vector element is set to one if the corresponding word is used in the document and to zero if the word is not. Using Boolean encoding the importance of all terms is considered as similar. To improve the performance, term weighting schemes are used, where the weights reflect the importance of a word in a specific document of the considered collection. Large weights are assigned to a term that are used frequently in relevant documents but rarely in the whole document collection [11] and is represented by the term frequency (TF) vector:

$$d_{tf} = [tf_1, tf_2, \ldots, tf_D] \quad (1)$$

Where, tf_i is the frequency of term i in the document, and D is the total number of unique terms in the text database.

Terms that occur in few documents are helpful to discriminate the documents from the rest of the collection. The inverse document frequency term weighting is used to assign higher weights to the more discriminative words. IDF is defined via the fraction N/n_i, where, N is the total number of documents in the collection and n_i is the number of documents in which term i occurs.

Due to the large number of documents in many collections, this measure is usually squashed with a log function. The resulting definition IDF is thus:

$$\mathrm{idf}_i = \log\left(\frac{N}{n_i}\right) \quad (2)$$

Combining term frequency with IDF results in a scheme known as tf-idf weighting.

$$w_{i,j} = tf_{i,j} \times \mathrm{idf}_i \quad (3)$$

Thus, the tf–idf representation of the document d is:

$$d_{tf-idf} = [tf_1 \log(n/df_1), tf_2 \log(n/df_2), \ldots, tf_D \log(n/df_D)] \quad (4)$$

To account for the documents of different lengths, each document vector is normalized to a unit vector (i.e., $\|d_{\text{df-idf}}\|=1$). In the rest of this paper, we assume that this vector space model is used to represent documents during the clustering. Given a set C_j of documents and their corresponding vector representations, the centroid vector c_j is defined as:

$$c_j = \frac{1}{|C_j|} \sum_{d_i \in C_j} d_i \quad (5)$$

where each d_i is the document vector in the set C_j, and j is the number of documents in Cluster C_j. It should be noted that even though each document vector d_i is of unit length, the centroid vector c_j is not necessarily of unit length. In this paper we experimented with all the three representations of Vector Space Model (VSM).

3. Related Work

In Bag of words representation of documents the words that appear in documents often have many morphological variants and in most cases, morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of clustering applications. For this reason, a number of *stemming Algorithms*, or *stemmers*, have been developed, which attempt to reduce a word to its *stem* or root form. Thus, the key terms of a document are represented by stems rather than by the original words.

Stemming refers to the process of removing affixes (prefixes and suffixes) from words. In the information retrieval context, stemming is used to conflate word forms to avoid mismatches that may undermine recall. As a simple example, consider searching for a document entitled “How to write”. If the user issues the query “writing” there will be no match with the title. However, if

the query is stemmed, so that "writing" becomes "write", then retrieval will be successful. In many languages stemming increases the number of documents retrieved by between 10 and 50 times. Nonetheless, stemming has shown to produce reliable retrieval improvement [15]. Furthermore, affixes often carry information such as part of speech, plurality, and/or tense that is crucial for the development of more sophisticated information systems. For efficient clustering of related documents we require a high precision stemmer as a preprocessing step [12].

The most widely cited stemming algorithm was introduced by Porter (1980). The Porter stemmer applies a set of rules to iteratively remove suffixes from a word until none of the rules apply. The Porter stemmer has a number of well-documented limitations. The words like "fisher", "fishing", "fished", etc. gets reduced to it's stem word "fish". The Porter stemmer follows a strategy of suffix stripping. Like many existing stemmers it ignores prefixes completely, so "reliability" and "unreliability" remain as unrelated tokens. The Lovins stemmer [16] is similar in mechanism but has a larger set of suffixes (each of which may include multiple morphemes) and does not apply its rules iteratively. While it tends to be more conservative than the Porter stemmer still suffers from over conflation and non-word stems.

Many clustering techniques have been proposed in the literature. Clustering algorithms are mainly categorized into Hierarchical and Partitioning methods [2, 3, 4, 5]. Hierarchical clustering method works by grouping data objects into a tree of clusters [6]. These methods can further be classified into agglomerative and divisive Hierarchical clustering depending on whether the Hierarchical decomposition is formed in a bottom-up or top-down fashion. K-means and its variants [7, 8, 9] are the most well-known partitioning methods [10].

Hierarchical clustering is often portrayed as the better quality clustering approach, but is limited because of its quadratic time complexity. In contrast, K-means and its variants have a time complexity which is linear in the number of documents, but are thought to produce inferior clusters.

There are a number of Partitional techniques, but we shall only describe the K-means algorithm which is widely used in document clustering. K-means is based on the idea that a center point can represent a cluster. In particular, for K-means we use the notion of a centroid, which is the mean or median point of a group of points. Note that a centroid almost never corresponds to an actual data point. The algorithm is discussed in detail in section 5.

4. Similarity Measures

Document clustering groups similar documents to form a coherent cluster. However, the definition of a pair of documents being similar or different is not always clear and normally varies with the actual problem setting. For example, when clustering research papers, two documents are regarded as similar if they share similar thematic topics. When clustering is employed on web sites, we are usually more interested in clustering the component pages according to the type of information that is presented in the page. For instance, when dealing with universities web sites, we may want to separate professor's home pages from student's home pages, and pages for courses from pages for research projects. This kind of clustering benefits further analysis and utilize the dataset such as information retrieval and information extraction, by grouping similar types of information sources together.

Accurate clustering requires a precise definition of the closeness between a pair of objects, in terms of either the pair wise similarity or distance. A variety of similarity or distance measures have been proposed and widely applied, such as cosine similarity, Jaccard coefficient, Euclidean distance and Pearson Correlation Coefficient.

4.1 Cosine Similarity Measure

For document clustering, there are different similarity measures available. The most commonly used is the cosine function. For two documents d_i and d_j, the similarity between them can be calculated

$$cos(d_i , d_j) = \frac{d_i . d_j}{|| d_i || || d_j ||} \quad (6)$$

Since the document vectors are of unit length, the above equation is simplified to:

$$cos (d_i, d_j) = d_i . d_j \quad (7)$$

When the cosine value is 1 the two documents are identical, and 0 if there is nothing in common between them (i.e., their document vectors are orthogonal to each other).

4.2 Jaccard Coefficient

The Jaccard coefficient, which is sometimes referred to as the Tanimoto coefficient, measures similarity as the intersection divided by the union of the objects. For text document, the Jaccard coefficient compares the sum weight of shared terms to the sum weight of terms that are

present in either of the two documents but are not the shared terms.

The Cosine Similarity may be extended to yield Jaccard Coeff. in case of Binary attributes

$$\text{Jaccard Coff (A,B)} = \frac{\sum_i A_i \,.\, B_i}{\sum_i \|A_i\|^2 + \sum_i \|B_i\|^2 - \sum_i A_i * B_i} \quad (8)$$

$$\text{Jaccard Index (A, B)} = \frac{A \cap B}{AU\, B} \quad (9)$$

4.3 Euclidean Similarity

This is the most usual, "natural" and intuitive way of computing a distance between two samples. It takes into account the difference between two samples directly, based on the magnitude of changes in the sample levels. This distance type is usually used for data sets that are suitably normalized or without any special distribution problem.

$$\text{Euclidean Distance (A, B)} = \sqrt{\sum_i (A_i - B_i)^2} \quad (10)$$

$$\text{Euclidean Similarity (A, B)} = 1 - \sqrt{\sum_i (A_i - B_i)^2} \quad (11)$$

4.4 Pearson Correlation Coefficient

This distance is based on the Pearson correlation coefficient that is calculated from the sample values and their standard deviations. The correlation coefficient 'r' takes values from –1 (large, negative correlation) to +1 (large, positive correlation). Effectively, the Pearson distance dp is computed as dp = 1 - r and lies between 0 (when correlation coefficient is +1, i.e., the two samples are most similar) and 2 (when correlation coefficient is -1).

$$SIM_P(\vec{t_a}, \vec{t_b}) = \frac{m\sum_{t=1}^{m} w_{t,a} \times w_{t,b} - TF_a \times TF_b}{\sqrt{[m\sum_{t=1}^{m} w_{t,a}^2 - TF_a^2][m\sum_{t=1}^{m} w_{t,b}^2 - TF_b^2]}}$$

$$\text{where } TF_a = \sum_{t=1}^{m} w_{t,a} \text{ and } TF_b = \sum_{t=1}^{m} w_{t,b}. \quad (12)$$

Where t_a and t_b are m-dimensional vectors over the term set $T = \{t_1, \ldots, t_m\}$.

The Euclidean distance is a distance measure, while the cosine similarity, Jaccard coefficient and Pearson coefficient are similarity measures. We apply a simple transformation to convert the similarity measure to distance values. Because both cosine similarity and Jaccard coefficient are bounded in [0, 1] and monotonic, we take D = 1 − SIM as the corresponding distance value. For Pearson coefficient, which ranges from −1 to +1, we take D = 1 − SIM when SIM ≥ 0 and D = |SIM| when SIM < 0.

5. Clustering Algorithm

For our analysis, we have chosen K-means algorithm to cluster documents. This is an iterative Partitional clustering process that aims to minimize the least squares error criterion [6]. As mentioned previously, Partitional clustering algorithms have been recognized to be better suited for handling large document datasets than Hierarchical ones, due to their relatively low computational requirements [7, 8, 9]. The standard K-means algorithm works as follows. Given a set of data objects D and a pre-specified number of clusters k, k data objects are randomly selected to initialize k clusters, each one being the centroid of a cluster. The remaining objects are then assigned to the cluster represented by the nearest or most similar centroid. Next, new centroids are recomputed for each cluster and in turn all documents are re-assigned based on the new centroids. This step iterates until a converged and fixed solution is reached, where all data objects remain in the same cluster after an update of centroids. The generated clustering solutions are locally optimal for the given data set and the initial seeds. Different choices of initial seed sets can result in very different final partitions. Methods for finding good starting points have been proposed [10]. However, we will use the basic K-means algorithm because optimizing the clustering is not the main focus of this paper.

The K-means algorithm works with distance measures which basically aims to minimize the within-cluster distances. Therefore, similarity measures do not directly fit into the algorithm, because smaller values indicate dissimilarity.

1. Select K points as the initial centroids.
2. Assign all points to the closest centroid.
3. Recompute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroids don't change.

5.1 Porter Stemming Algorithm

The Porter Stemmer is a conflation Stemmer developed by Martin Porter at the University of Cambridge in 1980. The stemmer is a context sensitive suffix removal algorithm. It is the most widely used stemmer and implementations are available in many languages. This stemmer is a linear step stemmer divided into a five linear steps that are used to produce the final stem. Within each step, if a suffix rule matched to a word, then the conditions attached to that rule are tested on what would be the resulting stem, if that suffix was removed, in the way defined by the rule. For example such a condition may be, the number of vowel

characters, which are followed be a consonant character in the stem (Measure), must be greater than one for the rule to be applied. The resultant stem being returned by the Stemmer after control has been passed from step five. See Porter Stemmer figure 1. However a number of definitions regarding the stemmer need to be made before the steps can be explained. The following definitions are presented in [17].

A *consonant* is a letter other than A, E, I, O or U and other than Y preceded by a consonant. For example in the word *boy* the consonants are B and Y, but in *try* they are T and R. A *vowel* is any letter that is not a consonant. A list of consonants greater than or equal to length one will be denoted by a *C* and a similar list of vowels by a *V* [17].

Any word can therefore be represented by the single form;

$$[C]\ (VC)^m\ [V]$$

Where the *m* denotes m repetitions of VC and the square brackets *[]* denote the optional presence of their contents [17]. The value *m* is called the *measure* of a word and can take any value greater than or equal to zero, and is used to decide whether a given suffix should be removed. All such rules are of the form; (*condition*) S1 → S2 which means that the suffix S1 is replaced by S2 if the remaining letters of S1 satisfy the *condition* [17].

The first step of the algorithm is designed to deal with past participles and plurals. This step is the most complex and is separated into three parts in the original definition, 1a, 1b and 1c. The first part deals with plurals, for example *sses* → *ss* and removal of *s*. The second part removes *ed* and *ing*, or performs *eed* → *ee* where appropriate. The second part continues only if *ed* or *ing* is removed and transforms the remaining stem to ensure that certain suffices are recognized later. The third part transforms a terminal *y* to an *i*, this part is inserted as step 2.
The remaining steps are relatively straightforward and contain rules to deal with different order classes of suffices, initially transforming double suffices to a single suffix and then removing suffices providing the relevant conditions are met [17].

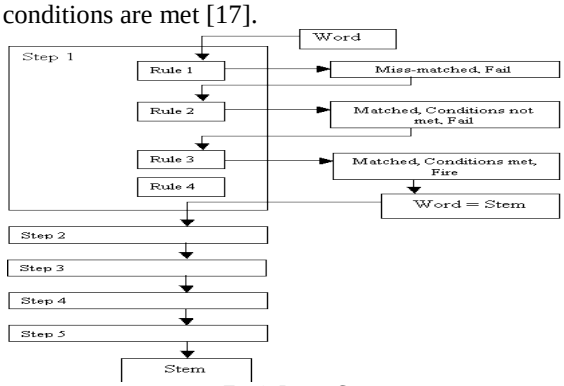


Fig 1: Porter Stemmer

6. EXPERIMENT

It is very difficult to conduct a systematic study comparing the impact of similarity measures on cluster quality with and without preprocessing the documents, because objectively evaluating cluster quality is difficult in itself. In practice, manually assigned category labels are usually used as baseline criteria for evaluating clusters. As a result, the clusters, which are generated in an unsupervised way, are compared to the pre-defined category structure, which is normally created by human experts. This kind of evaluation assumes that the objective of clustering is to replicate human thinking, so a clustering solution is good if the clusters are consistent with the manually created categories. However, in practice datasets often come without any manually created categories and this is the exact point where clustering can help. The rest of this section first describes the characteristics of the datasets, then explains the evaluation measures, and finally presents and analyzes the experiment results.

6.1 Dataset

This work experiments with one bench mark dataset Classic dataset collected from uci.kdd repositories. Classic dataset consists of four different collections CACM, CISI, CRAN and MED. We have considered 800 documents of the total 7095 documents.

In this datasets, some of the documents consists single word only, so it is meaningless to take such documents for document dataset. For eliminating these invalid documents we apply file reduction on each category, which returns the documents that supports mean length of each category. For file reduction we construct the Boolean matrices of all documents by category wise and calculate mean length of each category and removed the documents from the dataset which doesn't support mean length. By this we got valid documents. From these valid documents we have collected 800 documents of four categories each. From classic dataset 200 documents of each category again totaling to 800 documents.

6.2 Pre-Processing

Preprocessing consists of steps that take as input a plain text document and output a set of tokens (which can be single terms or n-grams) that are to be included in the vector model. In this work we performed removal of stop words and after taking users choice to perform stemming and built vector space model. We have pruned words that appear with very low frequency throughout the corpus with the assumption that these words, even if they had any discriminating power, would form too small clusters to be useful. Words which occur frequently are also removed. In

this work we have compared the performance of kmeans algorithm on documents without stemming with the documents with stemming.

6.3 Evaluation

For clustering quality evaluation are using entropy as a measure of quality of the clusters (with the caveat that the best entropy is obtained when each cluster contains exactly one data point). Let CS be a clustering solution. For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the "probability" that a member of cluster j belongs to class i. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula

$$E_j = -\sum_i p_{ij} \log(p_{ij}) \quad (13)$$

where the sum is taken over all classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster:

$$E_{CS} = \sum_{j=1}^{m} \frac{n_j * E_j}{n} \quad (14)$$

where n_j is the size of cluster j, m is the number of clusters, and n is the total number of data points.

6.4 Results Analysis

The seed points are statically chosen, but efficiency can be improved if seeds selected are random or run the code more than once to check the efficiency. As shown in tables 3, 4 Euclidean performs worst without applying stemming algorithm. As shown in Tables 1, 2 and Tables 3, 4 Euclidean distance performs worst with and without preprocessing the data. We also observe from tables 3, 4 that Jaccard Measure performs well after applying the stemming algorithm. We observe that Pearson performs the best with and without preprocessing of the data. From our results it is observed that Boolean representation with Pearson measure, Frequency count with Cosine and Euclidean also has non-zero clusters when we do not apply the stemming algorithm. Hence the overall entropy representation table for Boolean, Frequency Count and Term frequency and Inverse Document Frequency shows NaN values for other measures as some of the clusters are empty. On an average, the Jaccard and Pearson measures are slightly better in generating more coherent clusters, which means the clusters have lower entropy scores. Tables 5,6 shows one partition as generated by the Boolean Pearson measure using Reuter's dataset, and Tables 7,8 shows one partition as generated by the TF-IDF Jaccard Coefficient measure using Classic dataset which has the lowest entropy value.

Table 1: Entropy Results of Different Vector Space Representations Using Classic dataset without Porter stemming algorithm

Entropy	Cosine	Jaccard	Euclidean	Pearson
Boolean	NaN	NaN	NaN	0.08
Frequency Count	NaN	0.20	NaN	0.08
TF-IDF	0.16	0.13	NaN	0.08

Table 2: Entropy Results of Different Vector Space Representations Using Classic dataset with Porter stemming algorithm

Entropy	Cosine	Jaccard	Euclidean	Pearson
Boolean	NaN	NaN	NaN	0.08
Frequency Count	0.25	0.17	0.44	0.07
TF-IDF	0.08	0.11	0.44	0.07

We see from tables 1 and 2 that the Euclidean distance is again proved to be an ineffective metric for modeling the similarity between documents. But after applying Porter there is little improvement in the Euclidean measure. But Cosine tends to perform well in TF-IDF representation after applying porter algorithm. The Pearson's coefficient tends to outperform all the measures before and after stemming of the documents.

Table 3: TF-IDF Entropy Results using Classic dataset without Porter stemming

	Cosine	Jaccard	Euclidean	Pearson
Clusters[0]	0.31	0.0	0.41	0.03
Clusters[1]	0.01	0.23	0.28	0.04
Clusters[2]	0.06	0.10	0.10	0.07
Clusters[3]	0.13	0.15	NaN	0.17

Table 4: TF-IDF Entropy Results using Classic dataset with Porter stemming

	Cosine	Jaccard	Euclidean	Pearson
Clusters[0]	0.05	0.01	0.30	0.01
Clusters[1]	0.01	0.08	0.30	0.04
Clusters[2]	0.06	0.07	0.30	0.07
Clusters[3]	0.13	0.11	0.00	0.10

Here we see in tables 3 and 4 Jaccard measure performs well after applying porter algorithm.

Table 5: Clustering Results from Boolean Pearson Correlation Measure using Classic dataset without porter

	CACM	CISI	CRAN	MED
Cluster[0]	1	1	2	198
Cluster[1]	2	2	195	2
Cluster[2]	12	188	2	5
Cluster[3]	185	1	9	3

Table 6: Clustering Results from Boolean Pearson Correlation Measure using Classic dataset with porter

	CACM	CISI	CRAN	MED
Cluster[0]	0	0	3	189
Cluster[1]	4	1	193	4
Cluster[2]	7	186	1	5
Cluster[3]	189	13	3	2

Table 7: Clustering Results from TFIDF Jaccard Measure using Classic dataset without porter

	CACM	CISI	CRAN	MED	Label
Cluster[0]	0	0	0	164	**MED**
Cluster[1]	18	6	198	31	**CRAN**
Cluster[2]	10	166	1	2	**CISI**
Cluster[3]	172	28	1	3	**CACM**

Table 8: Clustering Results from TFIDF Jaccard Measure using Classic dataset with porter

	CACM	CISI	CRAN	MED	Label
Cluster[0]	0	1	0	166	**MED**
Cluster[1]	8	5	199	30	**CRAN**
Cluster[2]	3	166	0	4	**CISI**
Cluster[3]	189	28	1	0	**CACM**

We can see from the above tables 7 and 8 that the cluster accuracy with porter is 90% and of without porter is 87.5%. Hence applying stemming will improve cluster quality.

The Clustering accuracy r is defined as

$$r = \frac{\sum_{i=1}^{4} a_i}{n} \qquad (15)$$

where a_i is the number of instances occurring in both cluster i and its corresponding class and n is the number of instances in the dataset.

7. Conclusions and Future Work

In this study we found that all the measures have significant effect on Partitional clustering of text documents before and after applying the stemming algorithms. Of course the Euclidean distance measure performs worst. Pearson correlation coefficient is slightly better as the resulting clustering solutions are more balanced and is nearer to the manually created categories. The Jaccard and Pearson coefficient measures find more coherent clusters. The Jaccard Measure works better after applying stemming algorithm. Considering the type of cluster analysis involved in this study, we can see that there are four components that affect the final results—representation of the documents, applying the stemming algorithms, distance or similarity measures considered, and the clustering algorithm itself. In our future work our intension is to apply semantics knowledge to the document representations to represent relationships between terms and study the effect of these stemming algorithms exhaustively.

REFERENCES

[1] C. J. Van Rijsbergen, (1989), Information Retrieval, Buttersworth, London, Second Edition.
[2] G. Kowalski, Information Retrieval Systems – Theory and Implementation, Kluwer Academic Publishers, 1997.
[3] D.R. Cutting, D.R. Karger, J.O. Pedersen, and J.W. Tukey, Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, SIGIR '92, Pages 318 – 329, 1992.
[4] O. Zamir, O. Etzioni, O. Madani, R.M. Karp, Fast and Intuitive Clustering of Web Documents, KDD '97, Pages 287-290, 1997.
[5] D. Koller and M. Sahami, Hierarchically classifying documents using very few words, Proceedings of the 14th International Conference on Machine Learning (ML), pp. 170-178, 1997.
[6] G. Salton. Automatic Text Processing. Addison-Wesley, New York, 1989.
[7] M. Steinbach, G. Karypis, and V. Kumar. A Comparison of Document Clustering Techniques. In KDD Workshop on Text Mining, 2000.
[8] D. R. Cutting, J. O. Pedersen, D. R. Karger, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In Proceedings of the ACM SIGIR, 1992.
[9] B. Larsen and C. Aone. Fast and Effective Text Mining using Linear-time Document Clustering. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999.

[10] D. Arthur and S. Vassilvitskii. k-means++ the advantages of careful seeding. In Symposium on Discrete Algorithms, 2007.
[11] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.
[12] Krovetz, R. (1993). Viewing morphology as an inference process. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 191-202.
[13] Xu, J. and Croft, B. (1998). Corpus-Based Stemming using Co-occurrence of Word Variants. ACM Transactions on Information Systems, 16 (1).
[14] Arampatzis, A, van der Weide, Th.P., Koster, C.H.A., and van Bommel, P. (2000). Linguistically-motivated Information Retrieval. Encyclopedia of Library and Information Science, published by Marcel Dekker, Inc. - New York – Basel. To appear.
[15] Hull, D. (1996). Stemming algorithms: A case study for detailed evaluation. Journal of the American Society for Information Science, 47(1), 70-84.
[16] Lovins, J. B. (1968). "Development of a Stemming Algorithm", Mechanical Translation and Computational Linguistics, 11.
[17] Porter, M.F. (1980) *An Algorithm for Suffix Stripping*, Program, 14(3): 130-137

N.Sandhya B.Tech, M.Tech (Ph.D). I passed B.Tech in 2000 and M.Tech in 2007. Registered Ph.D in 2008. Has 11 years of experience in teaching. Working in GRIET. My areas of interest are Databases, Data Mining , Information Retrieval and Text Mining.

Y.Srilalitha M.Tech (Ph.D). I completed M.Tech in 2001. Registered Ph.D in 2008. Has 16 years of experience in teaching. Working in GRIET. My areas of interest are Information Retrieval, Data Mining and Text Mining.

V.Sowmya M.Tech (Ph.D). I completed M.Tech in 2009. Registered Ph.D in 2011. Has 6 years of experience in teaching. Working in GRIET. My areas of interest are Information Retrieval, Data Mining and Text Mining.

Dr.K.Anuradha M.Tech, Ph.D. I completed Ph.D in 2011. Working as professor and Head of the CSE Dept in GRIET. My areas of interest are Information Retrieval, Data Mining and Text Mining.

Dr.A.Govardhan M.Tech, Ph.D.
Working as professor and Principal of JNTU, Jagityal. Has an experience of 20 years in teaching. My areas of interest are Information Retrieval, Databases, Data Mining and Text Mining.

An Adaptive Notch Filter For Noise Reduction and Signal Decomposition

Mr. Vikas Mane, Mrs. Amrita Agashe.

[1] Dept. of Electronics, Walchand College of Engineering, Sangli, Maharashtra, India.

[2] Dept. of Electronics, Walchand College of Engineering, Sangli, Maharashtra, India.

Abstract

Detection, estimation and filtering of desired signal in the presence of noise are some of the most common and practical problems in the analysis of time series. The analysis in the time domain often involves the comparison of two different signals and stochastic signals are usually more profitably analyzed in the time domain. Extraction of signal are considerably improves quality of signal. In many signal processing applications is to decompose an original signal into its primitive or fundamental constituents and to perform simple operations separately on each component, thereby accomplishing extremely sophisticated operations by a combination of individually simple operations. An adaptive notch filter is able extract the desired signal from noisy signal. In this paper proposed filter is parallely connected adaptive notch filter which is able to extract the fundamental frequency from the noise corrupted signal and each filter able decompose 'n' sinusoid which is harmonically related to its constituent component.

Keywords: *Adaptive notch filter, signal decomposition, estimation of fundamental frequency, steady state performance.*

1. Introduction

The term *estimator* or *filter* is commonly used to refer to a system that is designed to extract information about a prescribed quantity of interest from noisy signal. Fundamental frequency estimation has many practical applications in various branches of engineering. Among these applications are active noise and vibration control in helicopters, cancellation of periodic disturbances in control system, elimination of power line noise in ECG signals.

Notch filter is able to extract the desired sinusoid component of a given signal provided that the signal frequency remains constant. The method of suppressing a sinusoid interference which corrupting an information bearing signal is to use a fixed *notch filter* tuned to the frequency of interference. A fixed notch filter may eliminate the noise when it is centered exactly at the frequency for which the filter was designed. But what if the notch is required to be very sharp and the interfering sinusoid is known to drift slowly? Clearly, then, we have a problem that calls for an adaptive solution. When the signal is non stationary that is fundamental frequency of the signal varies with time to time then the notch filter fails to remove noise from the signal. If any change in the input frequency then an adaptive notch filter which is capable of changing the notch frequency accordingly by tracking the frequency variations of the input signal. By such a system we mean one that is self designing in that the adaptive filter relies for its operation on a recursive algorithm, which makes it possible for the filter to perform satisfactorily in an environment where complete knowledge of the relevant signal characteristics is not available.

2. Motivation

The proposed adaptive notch filter in this paper is originated from Regali's algorithm. He proposed lattice-based discrete time adaptive IIR notch filter [1]. He proposed novel and efficient second order lattice structure. The filtered error and regressor signal necessary for an adaptive implementation which is available from single, numerically robust. The algorithm features favorable convergence properties, such as unbiased frequency estimation, improved tracking behavior compared to competitive designs, and improved reliability for extrimal input frequencies.

Regalia's algorithm was later adopted for continuous-time by Bodson [2]. He proposed two algorithms for rejection of sinusoidal disturbances with unknown frequency. The first is an indirect algorithm where the frequency of the disturbance is estimated, and the estimate is used in another adaptive algorithm that adjusts the magnitude and phase of the input needed to cancel the effect of disturbance. The second is direct algorithm that uses the concept of a phase locked loop is also presented in which

frequency estimation and disturbance cancellations are performed simultaneously.

The modified version of Bodson's algorithm [3] was proposed by Hsu *et al.* He solved the problem of global frequency estimation. The proposed algorithm is a scaled and normalized ANF. The new ANF was analyzed in terms of its stability, convergence and tuning. Despite a quite formidable complexity of its dynamical behavior, some conclusive results were established. Some preliminary results about the integration of the new ANF in noise cancellation systems are reported.

Hsu's stability analysis is guaranteed only when the forcing signal is pure sinusoidal. In some practical applications, such as periodic disturbances in active noise control, and voltage and current harmonics in the presence of nonlinear loads in power systems. The strength of this analysis is that it permits the presence of harmonic components in the input signal and is not limited to pure sinusoidal signals.

3. Frequency estimation and harmonic extraction

The dynamic behavior of the ANF of is characterized by the following set of differential equations:

$$\ddot{x} + \theta^2 x = 2\zeta(\theta^2 y(t) - \theta\dot{x})$$
$$\dot{\theta} = -\gamma x(\theta^2 y(t) - \theta\dot{x}) \qquad (1)$$

In equation (1), $y(t)$ is the input signal, θ represents the estimated frequency, ζ real positive number which determines "depth of notch", γ real positive number which determines "adaptation speed". In the modification in the given equations are the input signal $y(t)$ is scaled by θ instead of θ^2. Thus the ANF equation (1) changes to,

$$\ddot{x} + \theta^2 x = 2\zeta\theta(y(t) - \dot{x})$$
$$\dot{\theta} = -\gamma x\theta(y(t) - \dot{x}) \qquad (2)$$

Very often, when it exhibits some periodicity, a signal is modeled by a single or a combination of multiple sinusoids given by,

$$y(t) = \sum_{i=1}^{n} A_i \sin(i\omega_0 t + \varphi_i) = \sum_{i=1}^{n} y_i(t) \qquad (3)$$

where A_i, φ_i and ω_0 are real unknown parameters. In such a modelling, signal characteristics may also vary with time. Estimation of frequency and extraction of the individual sinusoidal components of such a signal provide useful information about the signal and therefore, provide means for signal analysis. Introducing an algorithm which capable of decomposing such a signal into its individual frequency components. A close observation of the filter dynamics in equation (2) is a resonator $\ddot{x} + \theta^2 x = 0$ that is forced with the error signal $e(t) = y(t) - \dot{x}$. The regressor signal $x(t)$ and the error signal $e(t)$ incorporate in the θ update law in equation (3). The term θ in both equations is for scaling. Also, the steady state error signal tends to zero and $\dot{x}$ is an estimate for the input signal $y(t)$. These key ideas are used for a new structure to estimate the fundamental frequency of periodic signal and extract its individual constituting harmonics as follows.

The ith componentof the signal in equation (3) satisfies,

$$\ddot{x}_i + i^2\omega_0^2 x_i = 0$$

Thus, a filter dynamic to extract ith component may be proposed as,

$$\ddot{x}_i + i^2\theta^2 x_i = 2\zeta_i\theta[y(t) - \textstyle\sum_{l=1}^{n} \dot{x}_l], \quad i = 1,2,\dots n \qquad (4)$$

where, θ is an estimate of ω_0. The error signal which forms the force function is redefined as,

$$e(t) = y(t) - \sum_{l=1}^{n} \dot{x}_l$$

The ω_0 is the frequency of the first component, i.e. x_1. Therefore, x_1 is used as the regressor signal and the update law for frequency estimation is proposed as,

$$\dot{\theta} = -\gamma x_1\theta[y(t) - \textstyle\sum_{l=1}^{n} \dot{x}_l] \qquad (5)$$

Rewriting the equation set (4) and (5) in terms of the redefined error signal $e(t)$ yields the resultant equations for the proposed signal analysis method as follows:

$$\ddot{x}_i + i^2\theta^2 x_i = 2\zeta_i\theta e(t), \qquad i = 1,2,\dots n$$
$$\dot{\theta} = -\gamma x_1\theta e(t) \qquad (6)$$

In equation (6), ζ_i and γ are real positive numbers and they determine the behavior of the ith filter and the θ update law in terms of (steady-state) accuracy and (transient) convergence speed. Figure. 1. and Figure. 2. shows that a general configuration of the proposed dynamics of equation (6). The detailed implementation block diagram of the ith filter path is shown in Figure. 2. where the error signal $e(t)$ is applied to each filter and the update law of equation (6) is employed to force the error signal to zero. Corresponding output of the ith filter is $\dot{x}_i$. Both hardware and software implementation of the proposed parallel structure are feasible and can take advantages of both pipelining and parallel computing. For the ith filter dynamic and in the steady-state, the output is that is the ith component of the input signal.

$$\dot{x}_i = A_i \sin(i\omega_0 t + \varphi_i)$$

This means that the Figure. 1. separates the input signal components in a manner that the ith component is made available by the ith filter. This feature desirable for a variety of real-time applications, for example, extraction and subsequent elimination of one or multiple harmonic components of the input signal in active noise cancellation schemes or active power filtering applications.

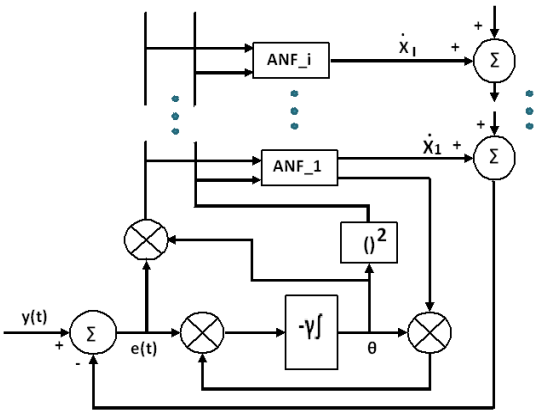


Figure.1. Block diagram of proposed algorithm.

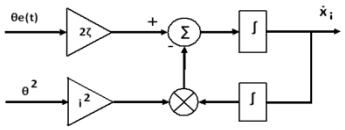


Figure. **2**. Details of the ith parallel filter.

4. Performance Evaluation

The performance of ANF based analysis simulated on computer using Matlab/Simulink software. The initiatory performance and tracking features are examined in the simulation. A structure shown in Figure 1 with $n = 8$ is simulated and constructed in Matlab/Simulink that is, to extract the first eight constituting components of the signal as well as the fundamental frequency ω_0. A set of values are $\zeta_i = 0.5, i = 1,2, \ldots 8, \gamma = 250$ are choosen for simulatons. The input signal consists of fundamental frequency with third, fifth and seventh harmonics as,

$$y(t) = \sin(\omega_0 t + \phi_1) + 0.5\sin(3\omega_0 t + \phi_3) + 0.2\sin(5\omega_0 t + \phi_5) + 0.4\sin(7\omega_0 t + \phi_7) \quad (1)$$

in which $\omega_0 = 30\, rad/sec$ and the initial phase angle ϕ_i's are selected randomly between 0 and $2\pi\, rad$. The initial conditions of all the integrators are set to zero except the one associated with the fundamental frequency which is set to its nominal value, i.e., $\omega_0 = 30\, rad/sec$. The responses of system to the input signal of (7) are shown in Figure . 3 . The eight extracted constituting components are shown in Figure . 4. a. to Figure 4. h respectively. The error signal $e(t) = y(t) - \sum_{n=1}^{6} y_i(t)$ is shown in Figure . 5. a. and the estimated fundamental frequency is shown in Figure. 5. b.

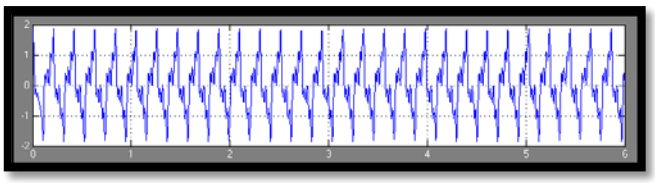


Figure. **3**. Sketch of input signal.

a) Extracted signals:

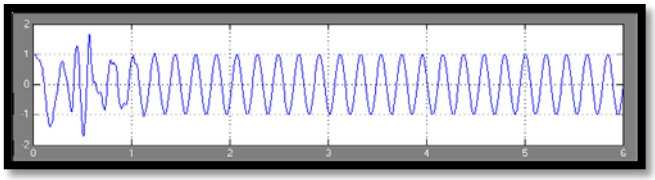


Figure. **4**. $\boldsymbol{a}$. Fundamental frequency.

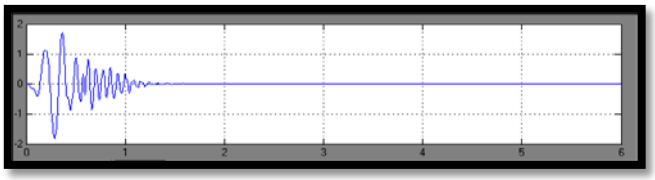


Figure. **4**. $\boldsymbol{b}$. Second component.

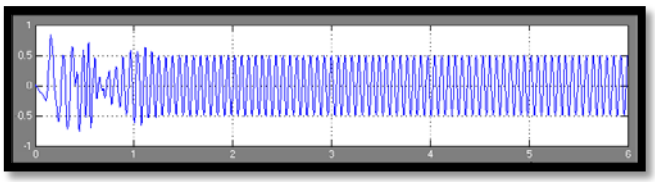


Figure. **4**. $\boldsymbol{c}$. Third component.

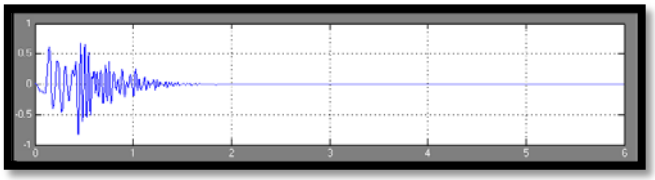


Figure. **4**. $\boldsymbol{d}$. Fourth component.

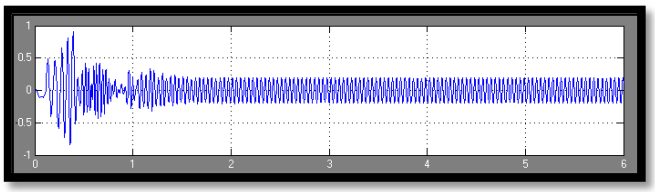


Figure. **4**. $\boldsymbol{e}$. Fifth component.

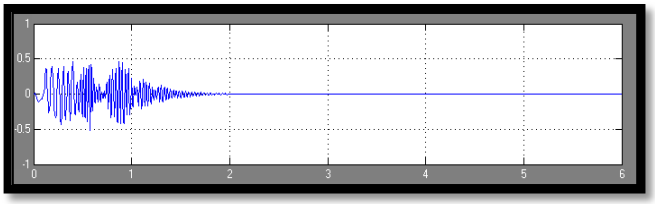


Figure. 4. *f*. Sixth component.

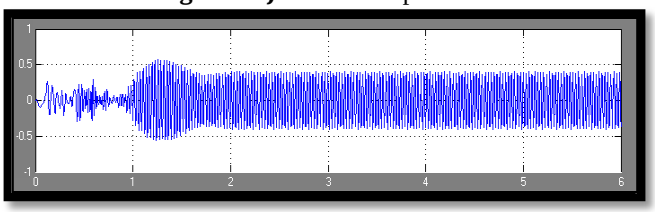


Figure 4. *g*. Seventh Component.

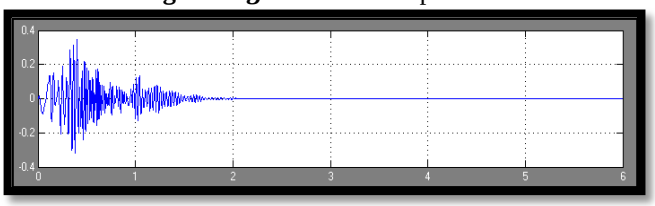


Figure 4. *h*. Eighth Component.

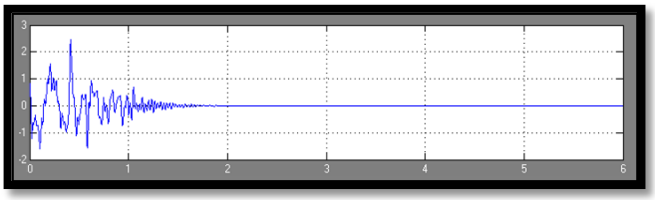


Figure. 5. *a*. Error signal.

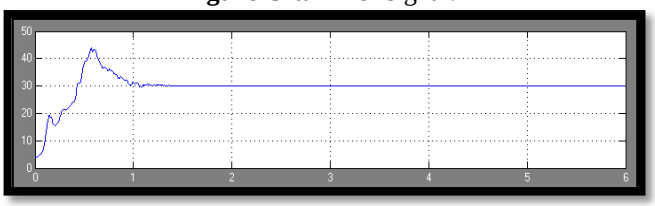


Figure. 5. *b*. Estimated frequency.

b) Tracking Performance:

Tracking performance of proposed algorithm is examined. If a step change from 30 *to* 40 *rad/sec* in the fundamental frequency of the input signal, then the step change is faithfully detected. The error signal, the actual and the estimated signals are shown in Figure. 6.

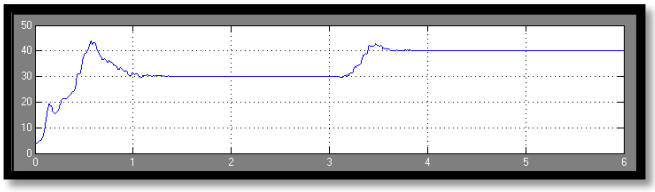


Figure. 6. *a*. Tracking: estimated signal.

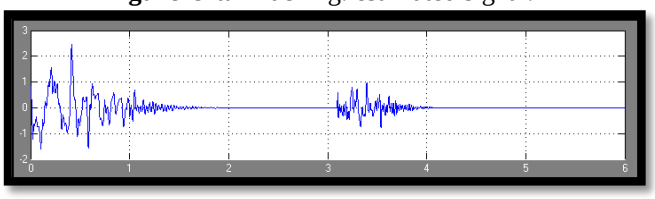


Figure. 6. *b*. Error signal.

c) Tracking multiple changes in parameters of input signal:

The third case study of this section verifies the capability of the presented method in tracking multiple simultaneous changes in the parameters of the input signal. The changes are following,

1) The amplitude of the fundamental component changes from 1 to 0.8
2) The amplitude of the third component changes from 0.5 to 0.3
3) A fourth component with amplitude 0.2 (and random phase) is introduced.
4) The amplitude of the fifth component changes from 0.2 to 0.4
5) The amplitude of the seventh component changes from 0.4 to 0.6
6) The fundamental frequency changes from 30 to 28 rad/s.

The extracted constituting components are shown in Figure 7 and they confirm desired tracking of the variables. The estimated signal and the error signal are also shown in Figure 8.

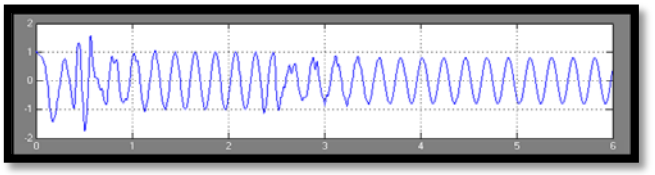


Figure 7.a. Fundamental frequency.

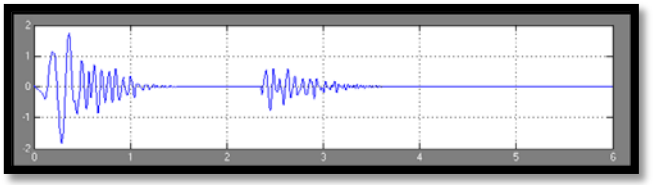


Figure 7.b Second Component.

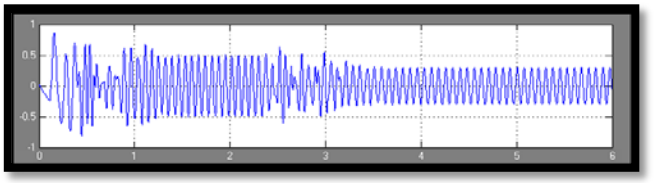


Figure 7.c Third Component.

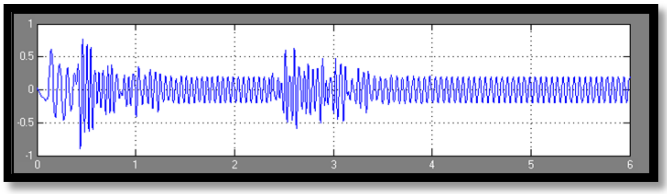


Figure 7.d Fourth Component.

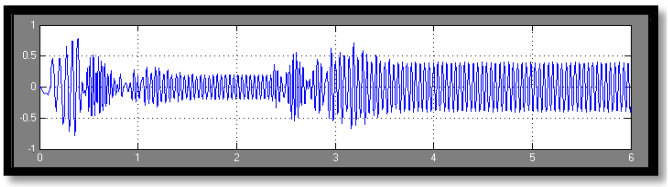


Figure 7.e Fifth Component.

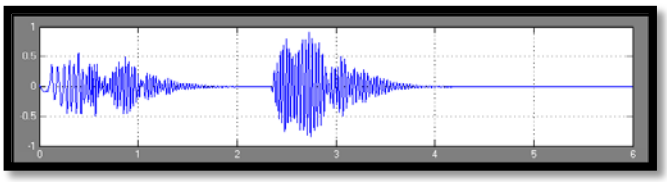


Figure 7.f Sixth Component.

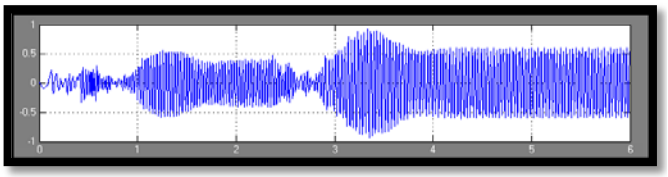


Figure 7.g Seventh Component.

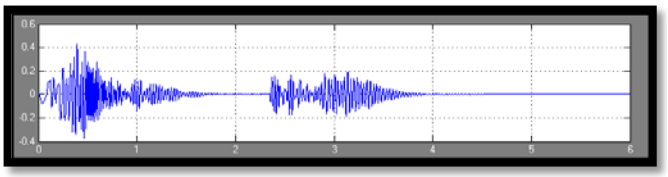


Figure 7.h Eighth Component.

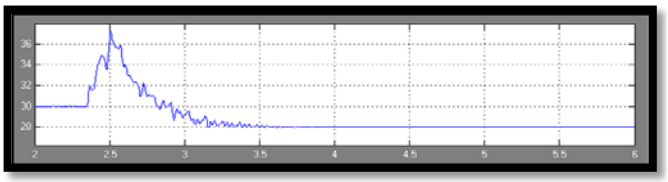


Figure 8.a Estimated Signal.

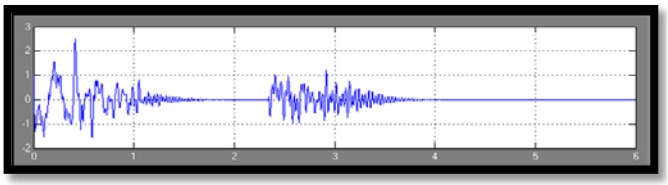


Figure 8.b Error Signal.

d) Noise characteristics:

To conduct a thorough simulation study for noise in the Matlab/Simulink environment, dynamics of the antialiasing filter must also be included. We consider a digital antialiasing filter which is implemented at a high sampling rate (such as 10 kHz). The input signal is sampled at 10 kHz and is passed through this antialiasing filter before being down-sampled to 1 kHz. Then, the low-frequency sampled signal (at 1 kHz) is forwarded to the proposed system. Figure 9 shows this mechanism.

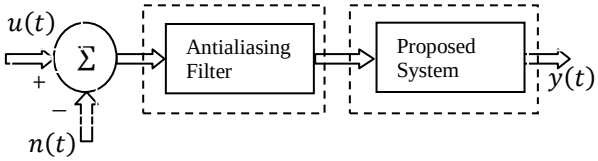


Figure 9 Block Diagram for Noise Study.

The input signal comprises a fundamental component at 30 rad/s with unity amplitude which is corrupted with a white Gaussian noise with a variance of $\sigma^2 = 0.5$ and zero mean. This signal, shown in Figure 5.12.a, is applied to the digital antialiasing filter, as shown in Figure 9. The output of the antialiasing filter, shown in Figure 5.12.b. This signal is down-sampled and forwarded to the proposed system. The output of the proposed system is shown in Figure5.12.c. This is confirms the desirable noise rejection of the presented system.

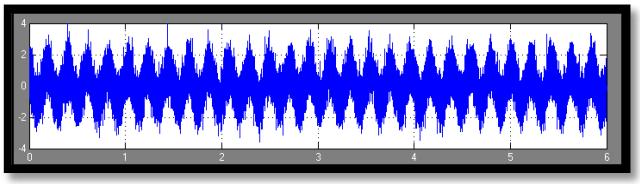


Figure 5.12.a Input Signal Corrupted by Gaussian Noise.

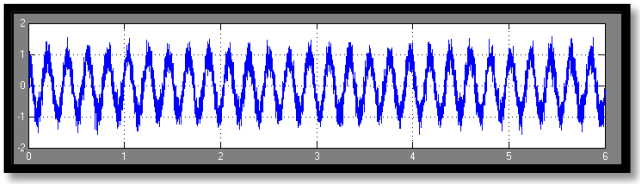


Figure 5.12.b Output of Anti-aliasing Filter.

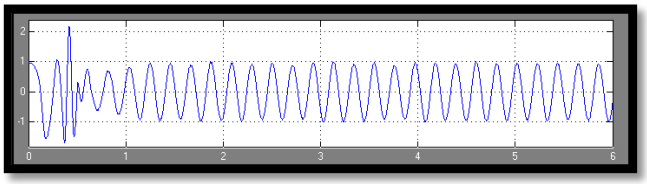


Figure 5.12.c Output of Presented System.

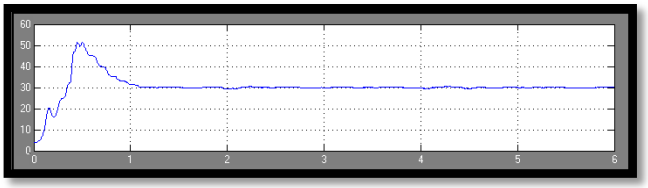


Figure 5.12.d Estimated Frequency.

5. Conclusions

An algorithm presented in this paper is capable for estimating the fundamental frequency of input signal which is composed of n harmonically related sinusoids and for separating or extracting input signal into its constituent components. The structure of the presented algorithm is composed of n second-order Notch Filters and each of which extracts one constituting component. An update law estimates the fundamental frequency of input signal and forwards it to the notch filters. All simulations are performed on computer in Matlab/Simulink software. The desirable initiatory performance, tracking features, and noise characteristics of the presented algorithm are examined. In the initiatory performance we set the filter parameters for optimizing damping and adaptation speed. Then we observe the filter responses for the input signal along with noise. The result of initiatory performance shows that the filter can extracts the fundamental frequency and harmonically related components from the input signal. Tracking performance results are indicates that, if fundamental frequency changes then presented Adaptive Notch Filter tracks faithfully the change in the fundamental frequency. In the noise characteristic case study, fundamental frequency corrupted by white Gaussian noise, the presented filters is capable of extracting fundamental frequency from the noise.

References

[1] P.A.Regalia,"An improved lattice-based IIR notch filter", *IEEE Trans. Signal process.* vol.39, no.9, pp. 2124-2128, Sept.1991.

[2] M. Bodson and S. C. Douglas, "Adaptive algorithms for the rejection of sinusoidal disturbances with unknown frequency," *Automatica*, vol.33, no.12, pp. 2213–2221, 1997.

[3] L. Hsu, R. Ortega, and G. Damm, "A globally convergent frequency estimator," *IEEE Trans. Autom. Control*, vol. 44, no. 4, pp. 698–713, Apr. 1999.

[4] Mohsen Mojiri, Masoud Karimi-Ghartemani, and Alireza Bakhshai, "Time-Domain Signal Analysis Using Adaptive Notch Filter" *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, VOL. 55, NO. 1, JANUARY 2007.

[5] John G. Proakis, Dimitris G. Manolakis, Digital Signal Processingm, 3rd Edition *Pearson Education.*

[6] Simon Hykin, "Adaptive Filter Theory", 3rd Edition *Prentice Hall.*

[7] Bernard Widrow, Samuel D. Sterns, "Adaptive Signal Processing" *Prentice Hall.*

Mr. Vikas S. Mane was born in Mumbai, India in 1986. He received the B.E. (Electronics) degree in Dr. D.Y. Patil College of Engineering and Technology, Kolhapur in 2009. He received M.Tech in Electronics engineering in Walchand College of Engineering, Sangli in 2011. He joined Sanjeevan Engineering and Technology Institute, Panhala, Kolhapur.

Dr. Amrita A. Agashe has received B.E., M.E., Ph.D. (Electronics Engineering) from Shivaji University, Kolhapur, Maharashtra, India. Currently she is working in Walchand College of Engineering, Sangli, Maharashtra, India. Her research interests are Signal Processing and Communication Engineering.

An Efficient I-MINE Algorithm for Materialized Views in a Data Warehouse Environment

[1]T.Nalini, , [2]Dr. A.Kumaravel, [3]Dr.K.Rangarajan
Dept of CSE, Bharath University,
173, Agaram Road, Selaiyur, Chennai – 600 073, India.

***Abstract*—**The ability to afford decision makers with both accurate and timely consolidated information as well as rapid query response times is the fundamental requirement for the success of a Data Warehouse. Selecting views to materialize for the purpose of supporting the decision making efficiently is one of the most significant decisions in designing Data Warehouse. Selecting a set of derived views to materialize which minimizes the sum of total query response time & maintenance of the selected views is defined as view selection problem. Therefore, to select an appropriate set of a view is the major target that diminishes the entire query response time and also maintains the selected views. Selecting a suitable set of views that minimizes the total cost associated with the materialized views is the key objective of data warehousing. However, these views have maintenance cost, so materialization of all views is not possible. In this paper we are taking into consideration of query frequency, query processing cost and space requirement. In order to find the frequent queries, we make use of I-mine mining techniques from which the frequently user accessible queries will be generated. Then, an appropriate set of views can be selected to materialize by minimizing the total query response time and/or the storage space along with maximizing the query frequency. These can be utilized by the users to obtain the quicker results once a set of views is materialized for the data warehouse.

Keywords : materialization view, data warehousing, selection cost, I-mine item set index, FP growth

I.INTRODUCTION

Data warehouse (DW) can be defined as subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decision [4]. It can bring together selected data from multiple database or other information sources into a single repository [6]. To avoid accessing from base table and increase the speed of queries posed to a DW, we can use some intermediate results from the query processing stored in the DW called materialized views. Therefore, materialized view selection involved query processing cost and materialized view maintenance cost. Selecting views to materialize for the purpose of supporting the decision making efficiently is one of the most significant decisions in designing Data Warehouse [5]. Selecting a set of derived views to materialize which minimizes the sum of total query response time & maintenance of the selected views is defined as view selection problem. Therefore, to select an appropriate set of a view is the major target that diminishes the entire query response time and also maintains the selected views. So, many literatures try to make the sum of that cost minimal.[3-15]

In order to find the frequent queries, we make use of I-Mine techniques from which the frequently user accessible queries will be generated. Then, an appropriate set of views can be selected to materialize by minimizing the total query response time

and/or the storage space along with maximizing the query frequency. These can be utilized by the users to obtain the quicker results once a set of views is materialized for the data warehouse. Given a set of queries *Q* and a quantity *S* (available storage space), the view selection problem is to select a set of views *M* to materialize, that under the multiple objectives constraint that the total space occupied by *M* is less than *S*. [1-2]
In this paper we are consideration three things to improve the query response time, space constraints, query frequency. First we describe I-mine index for materialized view to finding query frequency. Second we are finding query response time and space constraint.[1] The threshold is frequency of query is high and query cost and space constraint is low which query is meet the threshold that particular query is create the materialize view. [2]

The paper is organized as follows. In Section 2, we describe a related work of materialized view and propose work of selection view. In section 3, we describe Terminology and methods using in selection of materialized view query. In section 4, w explain experimental setup and results. In section 5, we describe concluded the paper and section 6 will provide the references.

2. RELATED WORKS

The problem of finding views to materialize to answer queries has traditionally been studied under the name of view selection. Its original motivation comes up in the context of data warehousing.

Harinarayan et al. [21] presented a greedy algorithm for the selection of materialized views so that query evaluation costs can be optimized in the special case of "data cubes". However, the costs for view maintenance and storage were not addressed in this piece of work. Yang et al. [8] proposed a heuristic algorithm which utilizes a Multiple View Processing Plan (MVPP) to obtain an optimal materialized view selection, such that the best combination of good performance and low maintenance cost can be achieved. However, this algorithm did not consider the system storage constraints.
Himanshu Gupta and Inderpal Singh Mumick [9] developed a greedy algorithm to incorporate the maintenance cost and storage constraint in the selection of data warehouse materialized views.
Amit Shukla et al. [12] proposed a simple and fast heuristic algorithm, PBS, to select aggregates for precomputation. PBS runs several orders of magnitude faster than BPUS, and is fast enough to make the exploration of the time-space tradeoff feasible during system configuration.
Himanshu Gupta and Inderpal Singh Mumick [6] developed algorithms to select a set of views to materialize in a data warehouse in order to minimize the total query response time under the constraint of a given total view maintenance time. They have designed approximation algorithms for the special case of OR view graphs.
Chuan Zhang and Jian Yang [4] proposed a completely different approach, Genetic Algorithm, to choose materialized views and demonstrate that it is practical and effective compared with heuristic approaches.
Sanjay Agrawal et al. [11] proposed an end-to-end solution to the problem of selecting materialized views and indexes. Their solution was implemented as part of a tuning wizard that ships with Microsoft SQL Server 2000.
Chuan Zhang et al. [4] explored the use of an evolutionary algorithm for materialized view selection based on multiple global processing plans for queries. They have applied a hybrid evolutionary algorithm to solve problems.
Elena Baralis, Tania Cerquitelli, and Silvia Chiusano,developed a the I-Mine index, a general and compact structure which

provides tight integration of item set extraction in a relational DBMS.[1]

The primary intent of this research is to develop a framework for selecting views to materialize so as to achieve finer query response in low time by reducing the total cost associated with the materialized views. The proposed framework exploits materialize the candidate views by taking into consideration of query frequency, query processing cost and space requirement. In order to find the frequent queries, we make use of I-Mine techniques from which the frequently user accessible queries will be generated. [11] Then, an appropriate set of views can be selected to materialize by minimizing the total query response time and/or the storage space along with maximizing the query frequency. The outcome can be directly utilized by the users to obtain the quicker results once a set of views is materialized for the data warehouse.[11-14]

3. TERMINOLOGY AND METHODS

This section explains the proposed cost effective framework for materialized view selection. We materialize the candidate views by taking into consideration of query frequency, query processing cost and space requirement. In order to find the frequent queries, we make use of I-mine techniques which generates the frequently user accessible queries.[11,12]

3.1 I-Mine(Item set-Mine index) indexes in materialized view

The I-Mine index [1] is a general and compact structure which provides tight integration of itemset extraction in a relational DBMS. Since no constraint is enforced during the index creation phase, I-Mine provides a complete representation of the original database. Data access as well as itemset extraction go in parallel in to reduce the I/O cost. The I-Mine index structure can be efficiently exploited by different itemset extraction algorithms. In particular, I-Mine methods currently support the (FP-growth and LCM v.2 algorithms), but they can straightforwardly support the enforcement of various constraint categories. [16]

Experiments, run for both sparse and dense data distributions, show the efficiency of the proposed index and its linear scalability also for large datasets. Itemset mining supported by the I-Mine index shows performance always comparable with, and often (especially for low supports) better than, state of the art algorithms accessing data on flat file.

The I-Mine index (is a novel data structure that provides a compact and complete representation of transactional data supporting efficient item set extraction from a relational DBMS. It is characterized by the following properties:

1. It is a covering index. No constraint (e.g., support constraint) is enforced during the index creation phase. Hence, the extraction can be performed by means of the index alone, without accessing the original database.
2. The I-Mine index is a general structure which can be efficiently exploited by various item set extraction algorithms.
3. The I-Mine physical organization supports efficient data access during item set extraction.
4. I-Mine supports item set extraction in large data Sets

The index performance has been evaluated by means of a wide range of experiments with data sets characterized by different size and data distribution. The execution time of frequent item set extraction based on I-Mine is always comparable with, and often (especially for low supports).

Mapping relational queries into an item set

A materialized view is a table on disk that contains the result set of a query.

Materialized views are most often used in data warehousing / business intelligence applications where querying large fact tables with thousands of millions of rows would result in query response times that resulted in an unusable application.

Keeping the materialized views under control we need to create materialized views as forms of aggregate tables, or as copies of frequently executed queries, this can greatly speed up the response time of any end user application.
Let Q = { q_1 , q_2 ,.. q_n } be a set of finite number of queries accessing T = { T_1 , T_2 ,.. T_m }, set of finite number of tables having attributes a_i belongs to any table in T. An P-Tree associated to relation R is actually a forest of prefix-trees, where each tree represents a group of transactions all sharing one or more items. In order to make 'item sets' based on queries, we consider the user log for query usage and construct the relation R with the transaction of occurrences of subset of Q as a row in R. Hence any algorithm meant for index selection can be dealt with Q for better performance.

An effective way to compactly store transactional records is to use a prefix-tree. Trees and prefix-trees have been frequently used in data mining and data warehousing indices, including cube forest, FP-tree, H-tree, Inverted Matrix , and Patricia-Tries . Our current implementation of the I-Tree is based on the FP-tree data structure, which is very effective in providing a compact and lossless representation of relation R.

3.2 Fp-Growth

The initial phase of FP-growth is the construction of a memory structure called FP-tree. FP-tree is a highly compact representation of the original database, which is assumed to fit into the main memory (a scalable, disk-based version of FP-tree has also been proposed). FP-tree contains only frequent items, each transaction has a corresponding path in the tree, and transactions having a common prefix share the common starting fragment of their paths. The procedure of creating an FP-tree requires two database scans: one to discover frequent items and their counts, and second to build the tree by adding transactions to it one by one.

After an FP-tree is built, the actual FP-growth procedure is recursively applied to it, which discovers all frequent itemsets in a depth-first manner by exploring projections (conditional FP-trees) of the tree with respect to frequent prefixes found so far. It should be noted that after the FP-tree is created, the original database is not scanned anymore, and therefore the whole mining process requires exactly two database scans.[22]

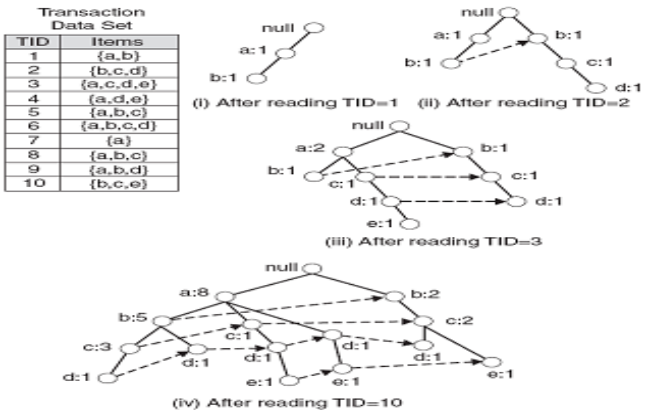


Figure : 1

- Nodes correspond to items and have a counter
- FP-Growth reads 1 transaction at a time and maps it to a path
- Fixed order is used, so paths can overlap when transactions share items (when they have the same prefix).
- In this case, counters are incremented
- Pointers are maintained between nodes containing the same item, creating singly linked lists (dotted lines)

- The more paths that overlap, the higher the compression. FP-tree may fit in memory.
- Frequent itemsets extracted from the FP-Tree.

Algorithm

Input: database *D*, minimum support threshold *minsup*

Output: the complete set of frequent patterns

Method:

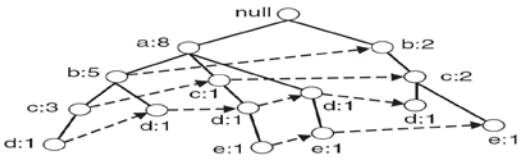
1. scan *D* to discover frequent items and their counts
2. create the root of *FP-tree* labeled as *null*
3. scan *D* and add each transaction to *FP-tree* (omitting non-frequent items)
4. call *FP-growth*(*FP-tree*, *null*)

procedure *FP-growth*(*FP-tree*, α) {
if *FP-tree* contains a single path *P*
then for each combination β of nodes in *P* **do**
generate frequent itemset βUα
with *support*(βUα,*D*)= min support of nodes in β;
else for each a_i in header table of *FP-tree* **do** {
generate frequent itemset β = a_iUα
with *support*(β,*D*) = *support*(a_i,*D*);
construct β's conditional pattern base and
β's conditional *FP-tree*β;
if *FP-tree*≠Ø**then** *FP-growth*(*FP-tree*β);
}

3.3 Frequent Itemset Generation

FP-Growth extracts frequent itemsets from the FP-tree.

- Bottom-up algorithm _ from the leaves towards the root
- Divide and conquer: first look for frequent itemsets ending in e, then de, etc. . . then d, then cd, etc. . .
- First, extract prefix path sub-trees ending in an item(set). using the linked lists.



↑ Complete FP-tree

Example: prefix path sub-trees

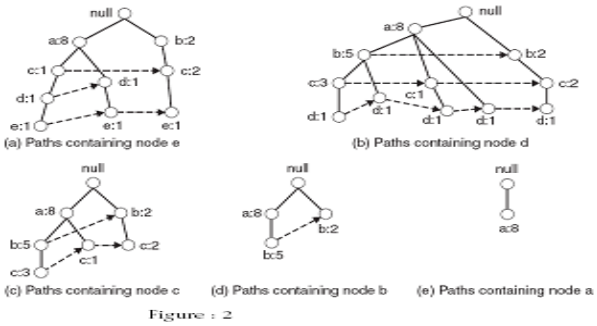


Figure : 2

Each prefix path sub-tree is processed recursively to extract the frequent itemsets. Solutions are then merged.

E.g. the prefix path sub-tree for e will be used to extract frequent itemsets ending in e, then in de, ce, be and ae, then in cde, bde, cde, etc.

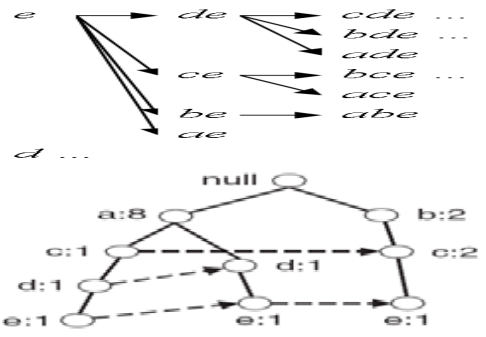


Figure : 3

Prefix path sub-tree ending in e.

3.4 Illustration of frequent itemset Generation

Let minSup = 2 and extract all frequent itemsets containing e.

1. Obtain the prefix path sub-tree for e:
2. Check if e is a frequent item by adding the counts along the linked list (dotted line). If so, extract it.

Yes, count =3 so {e} is extracted as a frequent itemset.

3. As e is frequent, find frequent itemsets ending in e. i.e. de,ce, be and ae. i.e. decompose the problem recursively.

To do this, we must first to obtain the conditional FP-tree for e.

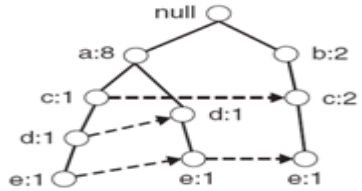


Figure : 4

3.4 Mining rules from the generated frequent itemsets

While association rule mining over FP-Growth without constraints is trivial, when constraints are in play, it is *not* trivial.

To calculate the confidence of a rule {A, B} ⇒ {C} (where {A, B} is called the rule *antecedent* and {C}is called the rule *consequent*), one must use the following formula:

a = rule.antecedent
c = rule.consequent
f = a UNION c = frequent itemset
confidence(a ⇒ c) = support(a UNION c) / support(a) = support(f) / support(a)

So to calculate the confidence of a rule, one needs two values: the support of the entire (frequent) itemset of which the rule consists (support(f)) and the support of the antecedent (support(a)). When the resulting confidence is smaller than minConf, the candidate association rule is dropped, otherwise it is added to the result. [22]. This procedure is implemented based on the following steps.

1) Step through header table from end to start (least common single attribute to most common single attribute). For each item.
a) Count support by following node links and add to linked list of supported sets.
b) Determine the "ancestor trails" connected to the nodes linked to the current item in the header table.
c) Treat the list of ancestor itemSets as a new set of input data and create a new header table based on the accumulated supported counts of the single items in the ancestor itemSets
d) Prune the ancestor itemSets so as to remove unsupported items.
e) Repeat (1) with local header table and list of pruned ancestor itemSets as input.

3.5 Computation of cost selecting queries constructing materialized views

Given space restrictions and, if available, a set of frequently users' queries, these algorithms select an appropriated set of views to materialize in order to achieve a good performance in the query processing of data warehousing environments.

For finding the selection cost S_Q of the every query, the query frequency cost Q_f, query storage cost Q_s and Query processing cost Q_p are computed using the following formulae,

$$Q_f = \frac{f_Q}{\underset{i}{\text{Max}}\ f_Q^{(i)}}$$

$$Q_p = \frac{P_Q}{\underset{i}{\text{Max}}\ P_Q^{(i)}}$$

$$Q_s = \frac{S_Q}{\underset{i}{\text{Max}}\ S_Q^{(i)}}$$

Where the parameters are defined as,
$f_Q \rightarrow$ frequency of query Q
$P_Q \rightarrow$ Processing cost of query Q
$S_Q \rightarrow$ Storage of cost S_Q

Using these parameters such as, Q_f, Q_s and Q_p, the selection cost S_Q is computed using the designed formulae that maximize the query frequency and minimize the spatial cost and query processing cost.

$$S_Q = \alpha * Q_f + \beta * (1 - Q_p) + \delta * (1 - Q_s)$$

Where, α, β *and* δ are Weights such that sum of α, β *and* δ equals 1. Moreover,

Q_f represent Query frequency cost,

Q_s represent query storage cost, and

Q_p represent Query processing cost respectively.

Then, the set of queries whose cost is implemented in less than the minimum threshold (T_M) is selected to build the materialized views.

$$T_M = \sum_{i=1}^{M} \frac{S_Q}{M}$$

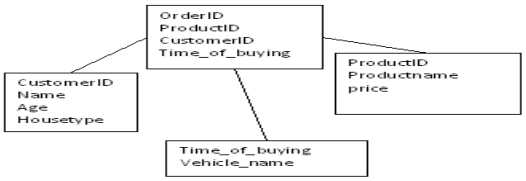
Thus, the selected views to materialize can be achieved the best combination of good query response, low query processing cost and low storage space.

4. EXPERIMENTIAL SETUP AND RESULT

We set up the environment with a purchase-order processing context having four physical table and applying fifty sample queries. The program has been written in Java and the backend is used SQL server8.Then we consolidate the user profile for using those queries by one thousand users as follows:

The input to the proposed approach is data warehouse model, D_W and a user's table (U_T) that contains the list of queries used by the number of users. For the constructing materialized view, the queries that are mostly used by the users should be selected but, at the same time, the query processing cost should be less as w discussed in previous section.

The schema of the data warehouse used in the proposed approach is represented with four various tables such as *customer* (T_1), *order* (T_2), *product* (T_3) and *vehicle* (T_4). Here, '*order*' (T_2) is a target table, which consists of four field records such as *OrderID, ProductID, CustomerID* and *Time of buying* where, *ProductID* and *CustomerID* are two foreign key relations. The *order* table contains one tuple for each new order, and its key is *OrderID*. The *customer* table contains details about the customer and its field records are *customerID, Name, Age, Housetype* and *City*.



Schema design

U_T is used to find the frequency of every queries for computing the query frequency cost. U_T- consisting of 'm' columns signifies the set of queries used by the corresponding users and 'n' rows signifies the number of users who are used the data warehouse to find the important information by posing the queries.

we apply I-Mine algorithm to user's query table U_T for finding the frequent queries and their corresponding support value. Details are given in section 3. After mining the frequent queries we find selection cost for materialized views. Details of Methodology are shown in section 4.

By considering these multi-objective, at first we sort the queries in a descending order based on frequency and at the same time, for other objectives, the queries are sorted in a ascending order according to the storage cost and query processing cost. Then, we select the top 'k' queries from the every sorted list so that the queries that are satisfying multiple objectives can be possibly selected. [13,14]

Sample queries

1. select customer.cid,customer.name from dbo.Customer where cid in (select vehicle.cid from dbo.Vehicle where Vehicle.model='model-19' and Vehicle.color='color-1')
2. SELECT Customer.name, Orders.oid from dbo.Customer INNER JOIN dbo.Orders ON Customer.cid = Orders.cid
3. SELECT Customer.name, Customer.age from dbo.Customer WHERE (((Customer.age)>30))
4.SELECT Customer.age, Vehicle.model from dbo.Customer INNER JOIN dbo.Vehicle ON customer.cid = Vehicle.cid WHERE (((Customer.age)='DISTINCT') AND ((Vehicle.model)='DISTINCT'))
5.SELECT Product.price, Product.quantity from dbo.Product WHERE (((Product.price)>7000) AND ((Product.quantity)>20))

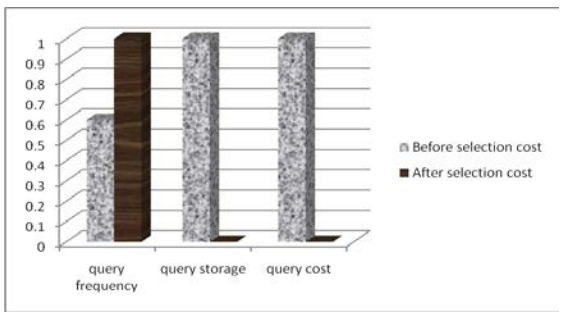
Sample U_T

Query	Qs	Qf	Qc	sq
1	0	0.7	0.2	8.2
2	0	0.8	0.2	8.4
3	0	0.7	0.2	8.2
4	0	0.7	0.2	8.2
5	0	0.8	0.2	8.4
6	0	0.6	0.2	8
7	0	0.6	0.2	8
8	0	0.6	0.2	8
...n (168)	0	0.6	0	8.7

1000 users are access 46 queries.
FP tree generate 2355 nodes.
FP mining generate from 2355 nodes to 168 frequent sets.

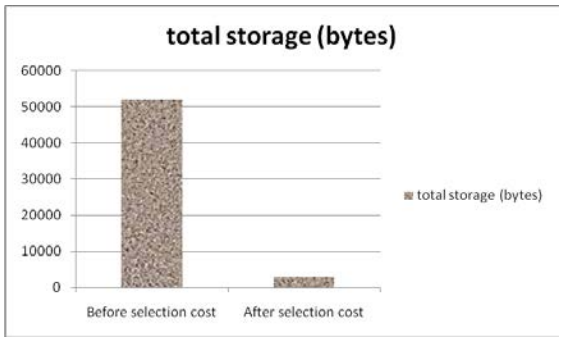
Query	Qs	Qf	Qc	sq
85	0	1	0	9.9
57	0	0.9	0	9.7
49	0	0.9	0	9.7
44	0	1	0.1	9.7
159	0	0.9	0.1	9.5
105	0	0.9	0	9.5
131	0	0.9	0	9.5
86	0	0.9	0	9.4
63	0	0.8	0	9.3
101	0	0.8	0	9.3
167	0	0.8	0	9.3
28	0	0.8	0	9.3
140	0	0.8	0	9.3
125	0	0.8	0	9.3
75	0	0.8	0	9.3
50	0	0.8	0	9.3
150	0	0.8	0	9.2
55	0	0.8	0	9.1
78	0	0.8	0	9.1
130	0	0.8	0	9.1
142	0	0.8	0	9.1
141	0	0.8	0	9.1
88	0	0.8	0	9.1
89	0	0.8	0	9.1
116	0	0.8	0	9.1
41	0	0.8	0	9.1
27	0	0.8	0	9.1
31	0	0.8	0	9.1
29	0	0.8	0	9.1
132	0	0.8	0	9.1
90	0	0.8	0	9.1
25	0	0.8	0	9.1
26	0	0.8	0	9.1
21	0	0.8	0	9.1
16	0	0.9	0.2	9.1
24	0	0.8	0	9.1
96	0	0.8	0	9.1

From the frequent set we apply the multiple objective to find the selection cost of queries. From that we select top 'k' queries.



Before selected the query to materialize storage cost and query cost are high

After satisfy multiple objectives selected queries are materialized.



Total storage is taken less space after select cost of query**.**

7. CONCLUSIONS

The selection of views to materialize is one of the most important issues in designing a data warehouse. The view selection problem has been addressed in this paper by means of taking into account the essential constraints for selecting views to materialize so as to achieve the best combination of low storage cost, low query processing cost and high frequency of query. It can be utilized by the users to obtain the quicker results once a set of views is materialized for the data warehouse. For experimentation, the proposed approach is executed on the simulated data warehouse model and the query list to find the efficiency of the proposed approach in maintaining of materialized view.

In addition to, the choice of algorithm is a major concern in finding the frequent queries for further reducing the time complexity. By considering these, we make use of the I-Mine algorithm, Index Support for Item Set Mining to mine the frequent queries. The advantage of the I-Mine algorithm is that it can mine the frequent queries with less computation time due to its I-Mine index structure compared with the traditional algorithms like, Apriori and FP-Growth.

As further extensions of this work, Incremental update of the index. Currently, when the transactional database is updated, the I-Mine index needs to be rematerialized. A different approach would be to incrementally update the index when new data become available. Since no support threshold is enforced during the index creation phase, the incremental update is feasible without accessing the original transactional database. Also work can be extended with various granular sizes of query i.e. a query can be charactered with the tables, records, attributes levels to meet accurate requirements instead considering only the queries as items in our itemset.

8. REFERENCES

[1] Elena Baralis, Tania Cerquitelli, and Silvia Chiusano," I-Mine: Index Support for Item Set Mining" IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 4, april 2009

[2] B.Ashadevi, R.Balasubramanian," Cost Effective Approach for Materialized Views Selection in Data Warehousing Environment", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.10, October 2008

[3] Y. Zhuge, H. Garcia-Molina, J. Hammer, and J. Widom, "View Maintenance in a Warehousing Environment." In Proceedings of the ACM SIGMOD Conference, San Jose, California, May 1995.

[4] C. Zhang, X. Yao, and J. Yang. An evolutionary Approach to Materialized View Selection in a Data Warehouse Environment. IEEE Transactions on Systems, Man and Cybernetics, vol. 31, no.3, pp. 282–293, 2001.

[5] H. Gupta, I.S. Mumick," Selection of views to materialize under a maintenance cost constraint", In Proc. 7th International Conference on Database Theory (ICDT'99), Jerusalem, Israel, pp. 453–470, 1999.

[6] J.Yang, K. Karlapalem, and Q. Li. "A framework for designing materialized views in data warehousing environment". Proceedings of 17th IEEE International conference on Distributed Computing Systems, Maryland, U.S.A., May 1997.

[7] S. Agrawal, S. Chaudhuri, and V. Narasayya, "Automated Selection of Materialized Views and Indexes in SQL Databases," Proceedings of International Conference on Very Large Database Systems, 2000.

[8] P. Kalnis, N. Mamoulis, and D. Papadias, "View Selection Using Randomized Search," Data and Knowledge Eng., vol. 42, no. 1, 2002.
[9] Gupta, H. & Mumick, I., Selection of Views to Materialize in a Data Warehouse. IEEE Transactions on Knowledge and Data Engineering, 17(1), 24-43, 2005.
[10] M. Lee and J. Hammer, Speeding up materialized view selection in data warehouses using a randomized algorithm,
International Journal of Cooperative Information Systems, 10(3): 327–353, 2001.
[11] Gang Gou; Yu, J.X.; Hongjun Lu., "A* search: an efficient and flexible approach to materialized view selection Systems," IEEE Transactions on Man, and Cybernetics, Part C: Applications and Reviews, Vol. 36, no. 3, May 2006 pp: 411 - 425.
[12] A. Shukla, P. Deshpande, and J. F. Naughton, "Materialized view selection for multidimensional datasets," in Proc. 24th Int. Conf. Very Large Data Bases, 1998, pp. 488–499.
[13] T.Nalini,S.K.Srivatsa,K.Rangarajan," International Journal of Advanced Research in Computer Engineering(IJARCE),"Method of ranking in indexes on materialized view for database workload" Vol.4,No.1,pp 157-162
[14] T.Nalini,S.K.Srivatsa,K.Rangarajan," International journal of computer science, systems engineering and information technology(IJCSSEIT)," Efficient methods for selecting materialized views in a data warehouse"Vol.3,No.2, pp 305-310
[15] R. Agrawal and R. Srikant, "Fast Algorithm for Mining Association Rules," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), Sept. 1994.
[16] R. Agrawal, T. Imilienski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. ACM SIGMOD '93, May 1993.
[17] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," Proc. ACM SIGMOD, 2000.
[18]A. Savasere, E. Omiecinski, and S.B. Navathe, "An Efficient Algorithm for Mining Association Rules in Large Databases," Proc. 21st Int'l Conf. Very Large Data Bases (VLDB '95), pp. 432-444, 1995.
[19] H. Toivonen, "Sampling Large Databases for Association Rules," Proc. 22nd Int'l Conf. Very Large Data Bases (VLDB '96), pp. 134-145, 1996.
[20] M. El-Hajj and O.R. Zaiane, "Inverted Matrix: Efficient Discovery of Frequent Items in Large Datasets in the Context of Interactive Mining," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD), 2003.
[21] V. Harinarayan, A. Rajaraman, and J. Ullman. "Implementing data cubes efficiently". Proceedings of ACM SIGMOD 1996 International Conference on Management of Data, Montreal, Canada, pages 205--216, 1996.
[22] Frequent Pattern Growth (FP-Growth) Algorithm An Introduction, Florian Verhein , January 2008.

Mrs. T.Nalini received M.Sc from the Karanataka university, M.Tech from Bharath University in 2000, 2007 respectively. Now, she is pursuing Ph.D. in Bharath University. She was a Lecturer between 2000 and 2006. Currently she is an Assistant Professor in the Department of CSE. She has published more than 4 research papers in international journals. She also presented the paper in 15 national conferences and 2 international conferences. She is a life member of many professional bodies like ISTE, CSI, IEEE.

Dr. A.Kumaravel received PG in Computer Science in Applied Sciences from the MIT Chennai , Ph.D in theoretical computer science from Anna university in 1988,1992 respectively. He was worked as a professor in CS for more than 20 years in Singapore. He has published more than 10 papers, and he also presented 20 papers in national and international conferences. He has organized workshops, national and international conferences, seminars in various organizations. Now, he is working as Dean of Computing Studies. He had received senior fellowship in UGC. He is a life member of many professional bodies like ISTE, CSI, IEEE.

Dr.K.Rangarajan is a Professor and Dean of Science & Humanities, Bharath University, TN, India. He has about 60 research publications and guiding many research scholars. His research areas include Automata Theory, Formal Languages, Petri Nets and Graph theory.

A new method for classification of Brachiopods based on the radon transformation

Youssef Ait khouya[1] and Noureddine Alaa[2]

[1] UFR Metrology Automatic and systems Analysis , Laboratory of Applied Mathematics and Computer Science (LAMAI), Faculty of Science and Technology, BP 549, Marrakech, Morocco

[2] Department of Mathematics, Laboratory of Applied Mathematics and Computer Science (LAMAI), Faculty of Science and Technology, BP 549, Marrakech, Morocco

Abstract

Brachiopods have a lateral outline which is quite important in systematic studies. It is often assessed by a qualitative evaluation and linear measurements, which are not clear enough and precise for describing the shape of the shell and its changes

In this paper we propose a new method for classification of fossils based on the radon transform from their greyscale image. We take the case of brachiopods which has Complex shapes. We use an adaptation of Radon transform called R-transform which is invariant to common geometrical transformations. Each shape is described by R_{3D} transform. We consider the grayscale image as a set of cuts obtained from successive binarization for each gray level in image, and for each segmentation we compute the R-transform then we obtained the R_{3D} transform. The advantages of the proposed method are robustness to noise, and invariant to common geometrical transformations scale, translation and rotation.

Keywords: Brachiopod, *Fossil, Classification, Radon transform, R-transform, Grayscale image, R_{3D} transform, Shape recognition.*

1. Introduction

Technological advances and development gigantic of storage capacity push geologists thinking of automating the task of classification of fossil species. That an important interest in palaeontological studies. On the one hand, they make it possible to understand the biodiversity in its morphological dimension. On the other hand, they highlight the morphological transformations undergone during the biological evolution. Historically, the form of was encircled by a purely descriptive approach based on the qualitative evaluations of the morphological change starting from simple images. This approach was replaced gradually by the biometric methods having leads to the populated design of the fossil species. The variables used in such methods are linear dimensions, angles, surfaces and ratio or combination of dimensions. But, these biometric descriptors are insufficiently informative since they give only one approximate quantitative representation of the form and its changes [1], [2], [3], [4].Then we used the Fourier analysis which consists in approximating the shape by a goniometrical function defined by a sum of terms of sine and cosine [5], [6], [7], [8], [9], [10]. This function is broken up into a series of amplitude of harmonics and phases or into a series of coefficient of Fourier being useful like variables for the quantitative analyses. But this method is valid right for the forms regular, indeed when morphologies become complex, it is not more possible to use such descriptors. In this paper, we propose a new method for classification of Brachiopods based on the Radon transform [11], [12], [13], [14] from their grayscale image. We use an adaptation of Radon transform called R-transform [15], [16] which is invariant to common geometrical transformations. Each shape is described by R_{3D} transform. We considered the grayscale image as a set of cuts obtained from successive binarization for each gray level in image, and for each segmentation we compute the R-transform then we obtained the R_{3D} transform. The advantages of the proposed method are robustness to noise, and invariant to common geometrical transformations scale, translation and rotation. This article is organized in the following way: In the first section we defined the concept of the Radon transform. The R-transform and 2D R-transform are described in section 3. The R_{3D} transform is defined in section 4 and the measure of similarity between two shapes (Brachiopod) is defined in section 5.

Cyclothyris vespertilio

Anathyris ezquerai

Cheirothyris fleuriausa

Fig.1 Images of some Brachiopod families

2. Radon transform

The Radon transform of an image $f(x,y)$ is determined by a set of projections of the image along lines taken at different angles θ. The resulting projection is the sum of the intensities of the pixels in each direction. The result is a new image $R_f(\rho,\theta)$. Therefore, the Radon transform can be written as [17]:

$$R_f(\rho,\theta)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\delta(\rho-x\cos(\theta)-y\sin(\theta))dxdy \qquad (1)$$

Where $\delta(.)$ is the Dirac delta-function ($\delta(t)$ = 1 if t = 0 and 0 elsewhere), $\theta\in[0,\pi[$ and $\rho\in]-\infty,+\infty[$. In other words, R_f is the integral of f over the line $L_{(\rho,\theta)}$ defined by $\rho=x\cos(\theta)+y\sin(\theta)$. As show in the figure (2).

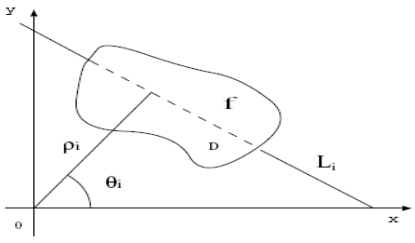


Fig. 2 Radon transform

There are two distinct Radon transforms. The source can either be a single point or it can be an array of sources as shown in (Figure 3). The method discussed in this paper uses an array of sources.

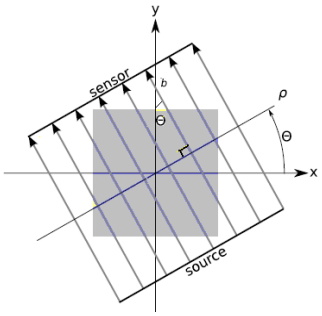


Fig. 3 Radon transforms uses an array of sources

The source and sensor are rotated about the center of the object. For each angle θ the density of the pixel the ray from the source passes through is accumulated at the sensor. This is repeated for a given set of angels, usually from $\theta\in[0,180[$. The angel 180 is not included since the result would be identical to the angel 0. The equation of the summation line is given as $y=ax+b$. As can be seen by using trigonometry, the inclination is $a=-\cos(\theta)/\sin(\theta)$ and the intersection with the y axis is $b=\rho/\sin(\theta)$. These parameters are determined for each combination of θ and ρ. In order to reduce the number of calculations necessary the maximum and minimum of either x or y are determined. ρ_{max} is the size of the diagonal of the image. The value of x can be real, to resolve the problem we used a linear interpolation. So the algorithm is:

For k from 1 to 180
  $\theta=k\pi/180$, Compute $a=-\cos(\theta)/\sin(\theta)$
  For ρ from 1 to ρ max
  Compute $b=\rho/\sin(\theta)$,Determine y_{min} and y_{max}
    For y from y_{min} to y_{max}
      Compute $x=(y-b)/a$, $R_f(\rho,\theta)+=f(x,y)$
    End
  End
End

We tested our algorithm on a binary image shown in Figure (4). The figure (5) present the sinogram of Radon transform used an array of sources. We compared our implementation with the result obtained using the Radon

transform found in MATLAB as show in figure (6). This is nearly identical the difference is presumably due to using a different interpolation.

Fig. 4 Binary image

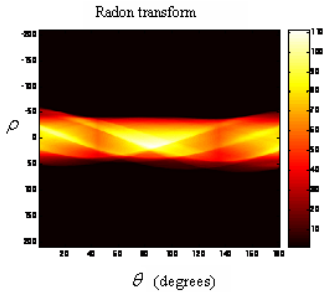


Fig. 5 result obtained with our implementation

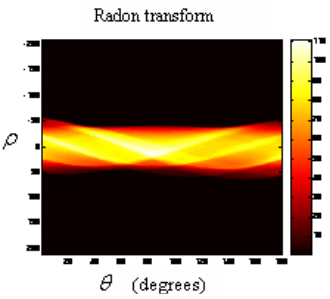


Fig. 6 result obtained with Radon transform found in MATLAB

Radon transform has some interesting properties relating to the application of affine transformations. We can compute the Radon transform of any translated, rotated or scaled image, knowing the Radon transform of the original image and the parameters of the affine transformation applied to it. This is a very interesting property for shape recognition because it permits to distinguish between transformed objects, but we can also know if two objects are related by an affine transformation by analyzing their Radon transforms. Let's see these properties:

Periodicity

$$R_f(\rho,\theta) = R_f(\rho,\theta + 2k\pi) \quad (2)$$

For any integer k the period is 2π.

Symmetry

$$R_f(\rho,\theta) = R_f(-\rho,\theta \pm \pi) \quad (3)$$

Translation of a vector $u_0 = (x_0, y_0)$: $R_f(\rho - x_0\cos\theta - y_0\sin\theta, \theta)$. A translation of f results in the shift of its transform in the variable ρ by a distance equal to the projection of the translation vector on the line $\rho = x\cos(\theta) + y\sin(\theta)$.

Rotation by angle θ_0 **:** $R_f(\rho,\theta+\theta_0)$. Implies a shift of the radon transform in the variable θ .

Scaling of α : $\frac{1}{|\alpha|}R_f(\alpha\rho,\theta)$. A scaling of f results in a scaling of both the ρ coordinates and the amplitude of the transform.

3. R-transform

R-transform has been developed by Tabbone in [15], [16], [18]. Its principle is simple: It consists to do for a given value of theta (i.e. within of the same of column the Radon matrix), the sum of squared elements. Can be written as:

$$R_T = \int_{-\infty}^{+\infty} R_f{}^2(\rho,\theta)d\rho \quad (4)$$

Where R_f is the Radon transform of the function f .

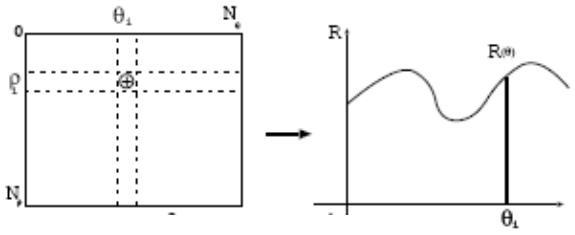


Fig. 7 definition of the R-transform

We applied here a Radon transform used an array of sources. The study of this R-transform allows us to define the following properties:

- **Periodicity:**

$$R_T(\theta \pm \pi) = \int_{-\infty}^{+\infty} R_f{}^2(\rho, \theta \pm \pi) d\rho = R_T(\theta) \qquad (5)$$

The R-transform is periodic and the period is π.

- **Rotation:**

For a rotation of the shape by an angle θ_0, the R-transform is:

$$\int_{-\infty}^{+\infty} R_f{}^2(\rho, \theta + \theta_0) d\rho = R_T(\theta + \theta_0) \qquad (6)$$

So a rotation of the shape by an angle θ_0 implies a translation of the R-transform of θ_0.

- **Translation:**

For a translation of vector $u_0 = (x_0, y_0)$ the R-transform is:

$$\int_{-\infty}^{+\infty} R_f{}^2((\rho - x_0\cos(\theta) - y_0\sin(\theta)), \theta) d\rho = \int_{-\infty}^{+\infty} R_f{}^2(\nu, \theta) d\nu \qquad (7)$$
$$= R_T(\theta)$$

We set $\nu = \rho - x_0\cos(\theta) - y_0\sin(\theta)$

The R-transform is invariant under a translation of f by a vector $u_0 = (x_0, y_0)$

- **Scaling:**

For a scaling $\alpha > 0$ the R-transform is :

$$\frac{1}{\alpha^2}\int_{-\infty}^{+\infty} R_f{}^2(\alpha\rho, \theta) d\rho = \frac{1}{\alpha^3}\int_{-\infty}^{+\infty} R_f{}^2(\nu, \theta) d\rho \qquad (8)$$
$$= \frac{1}{\alpha^3} R_T(\theta)$$

we set $\nu = \alpha\rho$

A zoom (before or back) of α generates a zoom of α^3 for R-transform.

4. R_{3D} - transform

The principle of 2D R-transform is applied to the binary image a distance transforms (After binarization of the grayscale image). A distance transform of a binary image specifies the distance from each pixel to the nearest non-zero pixel. The result of the distance transform is a gray level image and we capture both the internal structure and the boundaries of the shape. There are different families of distance transformation see [19], [20] for more information.

Given the distance transform of a shape, the distance image is segmented into n equidistant levels to keep the segmentation isotropic. For each distance level, pixels having a distance value superior to that level are selected and at each level of segmentation, an R-transform is computed [15]. The Figure (9) present the result obtained The X-axis present the angle θ in the Radon transform and the Y-axis reports the number of cuts in the distance transform.

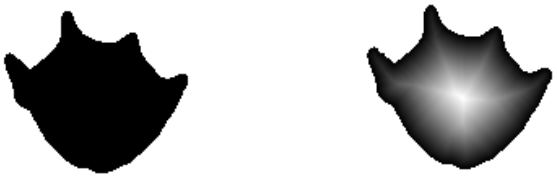


Fig. 8 binary shape and his Chamfer distance transform

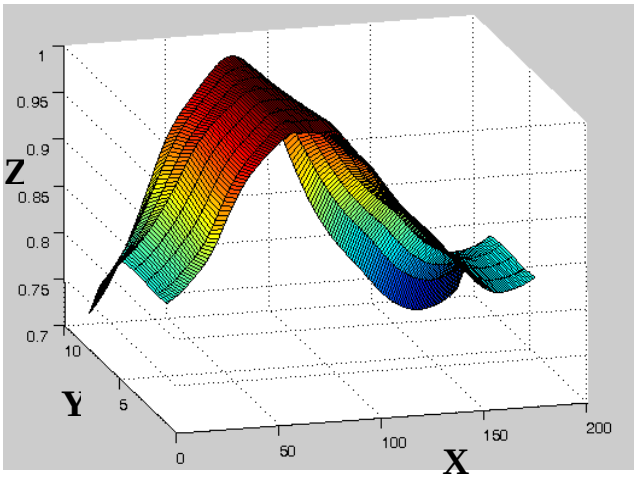


Fig. 9 surface visualisation of the 2D R-transform

Generally 2D R-transform is sufficient to obtain accurate results by considering only the shape of the object. It is poorly adapted for greyscale objects recognition.

To overcome this problem we used a transform called R_{3D} transform. We considered the grayscale image as a set of cuts obtained from successive binarization for each gray level in image.
Set O an object composed of k gray level and I_i is a binary threshold i cut applied to the object O. We have $R_{3D} = \bigcup \{R_{I_i}\}_{i=1,....K}$ with R_{I_i} is the R- transform computed on the binary image I_i. By definition an R_{3D}-transform is a set of R-transform checking separately the properties of R-transform described previously
and as the photometry is invariant to rotation, translation and scale, R_{3D} transform preserve the properties of a conventional R-transform ie the invariance to rotation, translation and scale. We normalized by the volume of each R_{3D} transform.

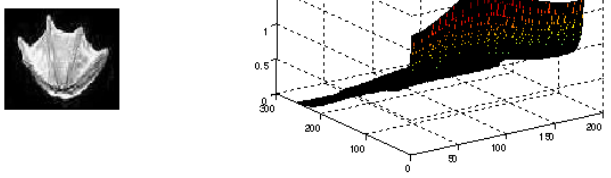


Fig. 10 Cheirothyris fleuriausa image and his R_{3D} transform

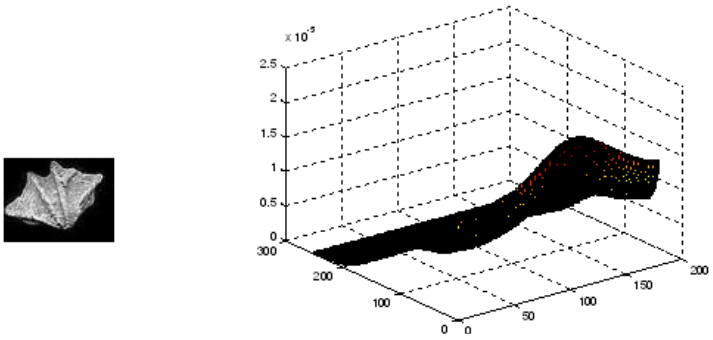


Fig. 11 Anathyris ezquerai image and his R_{3D} transform

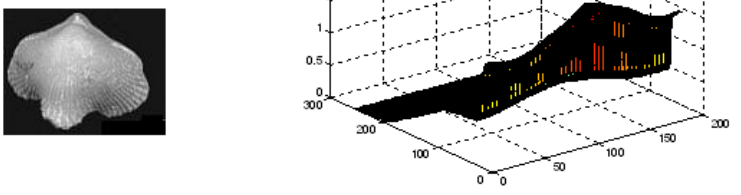


Fig. 12 Cyclothyris vespertilio image and his R_{3D} transform

5. Classification of the Brachiopod

Each shape (Brachiopod) is described by 3D R-transform. Consider two shapes A et B, R_{3D}^{A} and R_{3D}^{B} are the normalized R_{3D}. To measure the similarity between two shapes we used the ratio of similarity RS, the Shift of θ cyclyque are applied to R_{3D}^{B}, and RS is obtained by maximizing the index (min divided by max). that is insensitive to object rotation and scaling. This ratio is expressed as a percentage calculated between R_{3D}^{A} and R_{3D}^{B} is defined as:

$$RS = 100 \max_{x \in [0,p[} \left\{ \frac{\sum_{\theta=0}^{p} m^{AB}(\theta, x)}{\sum_{\theta=0}^{p} M^{AB}(\theta, x)} \right\}$$

With p is the nomber of orientations

and $m^{AB}(\theta, x) = \sum_{i=1}^{k} \min(R_{I_i}^{A}(\theta), R_{I_i}^{B}(\theta + x))$

$$M^{AB}(\theta, x) = \sum_{i=1}^{k} \max(R_{I_i}^{A}(\theta), R_{I_i}^{B}(\theta + x))$$

6. Conclusion

In this paper we presented a new method for classification of the Brachiopods based on the Radon transform from their grayscale image. We used an adaptation of Radon

transform called R-transform, we considered the grayscale image as a set of cuts obtained from successive binarization for each gray level in image. The proposed method is robustness to noise, and invariant to common geometrical transformations scale, translation and rotation. The proofs of transformation invariance including translation, scaling and rotation are provided. Our method gives satisfactory and encouraging results.

References

[1] F. L. Bookstein, 'The study of shape transformation after D'arcy Thompson'. Mathematical Biosciences, 1977, pp. 177- 219.
[2] F. L. Bookstein, B. Chernoff, R. L. Elder, J.M. Humphries, G.R. Smith and R.E. Strauss, 'Morphometrics in Evolutionary Biology' 15. The Academy of Natural Sciences of Philadelphia Special Publication, 1985, pp. 277.
[3] F. L. Bookstein, ' Morphometric tools for landmark data' Geometry and Biology. Cambridge University Press, 1991, pp. 435.
[4] G. p. Lohmann, 'Eigenshape analysis of microfossils: a General morphometric procedure for describing changes in shape', Mathematical Geology vol. 15, 1983, pp. 659-672.
[5] M. Foote, 'Perimeter-based Fourier analysis: a new morphometric method applied to the Trolobite cranidium', Journal of Paleontology, vol. 63, 1989, pp. 880–885.
[6] J. S. Crampton, 'Elliptic Fourier shape analysis of fossil bivalves: some practical considerations' Lethaia 28, 1995, pp. 179–186, Oslo.
[7] A. Bachnou, 'Modélisation des contours fermés: «An-Morph» outil mis au point pour maîtriser les composantes des profils latéraux des ostracodes. Perspective d'application systématique', Géobios vol. 32, 1999, pp. 733–742.
[8] K. El Hariri, P. Neige and J.L. Dommergues, 'Morphométrie des côtes chez des Harpoceratinae (Ammonitina) pliensbachiens. Comparaison des formes du Haut-Atlas (Maroc) avec celles de l'Apennin central (Italie)', Comptes Rendus Acade´mie Sciences Paris vol. 322, 1996, pp. 693–700.
[9] K. El Hariri, 'Analyse morphométrique des côtes chez des Graphoceratidae (Ammonitina) du Maroc'. Revue Paléobiologie, Genève vol. 20, 2001, pp. 367–376.
[10] K. El Hariri and A. Bachnou, 'Describing Ammonite shape using Fourier analysis', Journal of African Earth Sciences vol. 39, 2004, pp. 347–352.
[11] P. Fränti, A. Mednonogov , V. Kyrki. and H. Kälviäinen, 'Content-based matching of line-drawing images using the Hough transform', International Journal on Document Analysis and Recognition, vol. 3, 2000, pp. 117–124.
[12] V. F. Leavers, 'Use of the Radon transform as a method of extracting information about shape in two dimensions', Image Vision and Computing, vol. 10, 1992, pp. 99–107.
[13] V. F. Leavers, 'Use of the Two-Dimensional Radon Transform to Generate a Taxonomy of Shape for the Characterization of Abrasive Powder Particles', IEEE Transactions on PAMI, vol. 22, 2000, pp. 1411–1423.
[14] V. F Leavers and J.F.Boyce, 'The Radon transform and its application to shape parametrization in machine vision', Image Vision and Computing, vol. 5, 1987, pp. 161–166.
[15] S. Tabbone, L. Wendling and J.P. Salmon, 'A new shape descriptor defined on the Radon transform', Computer Vision and Image Understanding, vol. 102, 2006, pp. 42–51.
[16] S. Tabbone, L. Wendling and K. Tombre, 'Matching of Graphical Symbols in Line- Drawing Images Using Angular Signature Information', International Journal on Document Analysis and Recognition, vol. 6, 2003, pp. 115-125.
[17] S.R. Deans, 'Applications of the Radon Transform', Wiley Interscience, 1983, Publications, New York.
[18] S. Tabbone and L. Wendling, 'Technical Symbols Recognition Using the Two-dimensional Radon Transform', Proceedings of the 16th International Conference on Pattern Recognition, Québec (Canada), vol. 2, 2002, pp. 200-203.
[19] G. Borgefors, 'Distance transformations in arbitrary dimensions', Comput. Vis. Image Process. Vol. 27, 1984, pp. 321–345.
[20] G. Sanniti diBaja and E. Thiel 'Skeletonization algorithm running on path-based distance maps', Image Vis. Comput. Vol. 14, 1996, pp. 47–57.

Youssef AIT KHOUYA received in 2008 his Master in telecommunications and Computer Networks from the University of Cadi Ayyad, Morocco. He is currently a Ph.D. Student at the Faculty of Sciences and Technology Marrakech. His research interest is Computer Science.

Pr. NourEddine ALAA received his Master of Science and his Ph.D. degrees from the University of NancyI France respectively in 1986 and 1989. In 2006, he received the HDR in Applied Mathematics from the University of Cadi Ayyad, Morocco. He is currently Professor of modeling and scientific computing at the Faculty of Sciences and Technology of Marrakech. His research is geared towards non-linear mathematical models and their analysis and digital processing applications.

A Detailed Study on Energy Efficient Techniques for Mobile Adhoc Networks

Mrs. Suganya Senthil[1], Dr.Palaniammal Senniappan[2]

[1] Sr.Lecturer , Dept. of Computer Applications,
TamilNadu College of Engineering,
Coimbatore-641 659, India

[2]Prof. & Head, Dept. of Science and Humanities,
VLB Janakiammal College of Engineering & Tech.,
Coimbatore-641 042, India

***Abstract*--The performance of wireless networks in the applications like transferring video files is subjected to dual constraints. Both minimization of power and other QoS requirements like delay, throughputs are have to be take care properly. Mobile Ad Hoc Networks are more perceptive to these issues where each mobile device is active like a router and consequently, routing delay adds considerably to overall end-to-end delay. This paper presents a survey on power efficient routing protocols for Mobile Ad-Hoc Networks. This survey centered on recent progress on power saving algorithms. In addition we suggest one energy efficient technique which will reduce energy consumption as well as increase the lifetime of node and network.**

Keywords- Mobile adhoc networks; Quality of Services; Minimum Energy consumption; Network life time;

1. Introduction

The Mobile ad hoc network [MANET] is a distributed network where mobile nodes are connected together by wireless links without any fixed infrastructure like base stations, fixed links, routers, centralized servers. In such a network the data can be transmitted or routed by intermediate nodes which are not in the fixed location. A large scale of independence and self organizing capability formulate it completely different from other networks. The topology of mobile Ad Hoc network is not static and depends upon the mobility of the nodes so it can adjust rapidly and suddenly. Mobile Ad Hoc networks are useful in many areas such as, vehicular network, Communication in front line, Disaster recovery areas, agro sensing, Institutions and Colleges, Space and astronomy related projects, pollution monitoring and Medical Field[1].

Mobile Ad Hoc networks have few challenges like Limited wireless transmission range, broadcast nature of the wireless medium, hidden terminal and exposed terminal problems, packet losses due to transmission errors and mobility, stimulated change of route, Battery constraints and security problem [2,3]. The power level basically affects many features of the operation in the network including the throughput of the network. Power control also affects the conflict for the medium and the number of hops in turn it will affect the delay time. Transmission power also influences the important metric of energy consumptions. Therefore the energy efficient protocol is must to increase the lifetime of node as well as the lifetime of network [4]. So the designed Ad Hoc routing protocol must meet all these challenges to give the average performance in every case.

Routing is the process of path establishment and packet forwarding from source node to sink node. It carried out in two steps, first selecting the route for different pair of source-sink and delivers the data packets to the target node. Various protocols and data structures are available to maintain the routes and to execute this process. This survey paper is paying attention on how these protocols are selecting energy efficient routes. Routing in ad-hoc networks has some distinct characteristics such as, Energy of node which depends on the limited power supply battery, Mobility of the nodes which may cause frequent route failures and Wireless channels required variable bandwidth compare to wired network. The key solution for the above requirements is energy efficient routing protocols [5]. In the protocols the energy efficiency can be achieved by using efficient metric for selection of route such as cost, node energy, and battery level. The energy efficiency is not intended only on the less power consumption, it also focuses on increasing the life time of node where network maintains certain performance level [6]. Recently it is reported in the literature that energy efficiency can be made at all layer of the network protocol stack. Various study recommended different techniques for handling the energy issue. In this paper we discussed the features of few protocols which increase the network lifetime and performance and propose a technique to minimize the consumption of energy as well as increase the lifetime of network. The technique recommended pertain power control

at node level to condense the transmission power of a node and energy-inefficient nodes are detached to increase network lifetime. The rest of the paper structured as follows. Section II presents related work. Section III addressed on the classification of routing techniques based on different approaches and conventional properties of various routing. Proposed energy efficient technique is discussed in section IV and finally Section V Concludes the paper.

2. Related Work

This section consists of complete study on conventional protocols and energy efficient protocols published in different journals which has proposed so much innovation and new ideas in this field. Since energy preservation is an open issue to all layer of the network protocols stack, and power is main anxiety in mobile ad-hoc wireless networks different techniques were recommended by different study and focus has been given on different layer design to preserve energy more efficiently. None of the energy efficient protocol can perform well in every condition[7, 8]. It has some advantages and inadequacy which depends on the network parameters. Energy preservation on the mobile nodes should maintain not only during active communication but also when they are inactive[9]. The standard protocols proposed for wireless networks have two types of power managements [10]. First type is power save (PS) mode for infrastructure based wireless network and the second type is named as independent basic service set power save (IBSS PS) mode, which is for infrastructure-less networks. In the first method, power consumption of the nodes in PS mode is less than the power consumption of nodes which are in active mode. The power saving mechanism is implemented using the access points in the network. But this is not suitable for ad hoc network environment since there is no central coordinator like access point. Conversely, IBSS PS mode is applicable to entirely connected single hop network where all the nodes are within the radio range of each other. Coordinated beacon interval is reputable by the node which initiates the IBSS and is maintained in a distributed approach.

Dynamic power saving mechanism [DPSM] is a conflict of the above protocols we discussed by using the concept of adhoc traffic indication message (ATIM) window and beacon interval. During ATIM window all nodes are conscious and those nodes have no traffic to receive or send are goes to sleep mode after the end of ATIM window. In the paper author Freeny [11] concluded that if ATIM window is fixed then energy saving cannot be efficient. DPSM improves this performance by using the variable ATIM window. It allows the sender and receiver node to change the ATIM window dynamically. The ATIM window size can be increased while a few packets are still in waiting stage after the current window has expired. The data packets carry the current length of the ATIM window to help the nodes to adjust their ATIM window length. The energy saving performance of DPSM is better as compare to IEEE 802.11 distributed coordination function (DCF) in term of power saving but it is more complex in computations. The author Sahoo [12] proposed a distributed transmission power control protocol for wireless network to achieve energy conservation at the level of node. The protocol uses distributed algorithm to construct the power saving hierarchy topologies without taking the local information of the nodes and provide a simple way to keep the network on account of changing the transmission power.

3. Classification of Routing Techniques

Transmission power control, load distribution and Power Management are the approaches to minimize the energy on active communication and sleep/power-down approach is used to minimize energy during inactivity. The protocols are designed based on the energy related metrics like energy consumed per packet to provide the minimum power path which is used to minimize the overall energy consumption for delivering packet. The next important metric is inconsistency in node power levels which is a simple indication of energy balance and in turn it can be used to extend network lifetime.

Table 1: Techniques of energy efficient routing protocols.

circumstances	Name of approach	objective
Minimize Active Communication Energy	Transmission Power Control	The total transmission energy is minimized by avoiding low energy nodes.
	Load Distribution	Distribute load to energy comfortable nodes
	Power Management	Minimize the energy consumption by using separate channels for data and control
Minimize Inactivity Energy	Sleep/Power-Down Mode	Minimize energy consumption when node in an

		idle state

Here the transmission power is to be fine-tuning to transmit packets using intermediate nodes [13]. It is like a finding shortest path in a graph problem, where each edge is weighted with the distance corresponding to the required transmission power as shown in the fig.1(e.g., p(SA) for the edge S->A). Finding the most energy efficient route from *S* to *D* is equivalent to finding the shortest path in the weighted graph.

The following Fig.1 illustrates the technique of transmission power control using two models. In the constant link model the routing path S-> D is direct path without fine tuning the transmission power. But in the adjustable model *S->B->D* is more energy efficient than the route *S->D* since $p(SD) > p(SB) + p(BD)$. Node *S* preserve energy by lowering its radio power just enough to reach node *B*, but not enough to reach node *D*.

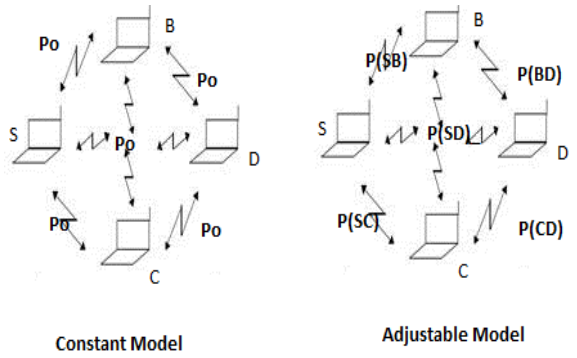


Fig.1 Transmission power control

The routing protocols available under the technique transmission power optimization[16] is uphold additional information at each node other than that acquired during operation such as link costs of all edges , costs of all nodes and data generation rate at all nodes . With the help of the information available the protocol select the max-min path among a number of best min-power paths and few protocols regulate the transmission power just enough to reach the next hop node in the given routing path.

The objective of the load distribution approach is to balance the energy usage of all nodes by selecting a route with nodes which are not used frequently instead of the shortest route [15]. The result of this approach may involve more nodes in a route but packets are routed only through energy comfortable intermediary nodes. Protocols based on this approach are not necessarily offer the lowest energy route, but prevent certain nodes from being overloaded, and guarantees for longer network lifetime. One of such protocol is named as Localized Energy-Aware Routing (LEAR). The LEAR routing protocol is conflict from DSR in the process of route discovery procedure for balanced energy consumption. In DSR, when a node receives a route-request message, it attaches its identity in the header of message and forwards it in the direction of destination [17]. Therefore, an intermediate node always relay messages if the corresponding route is selected. On the other hand, in LEAR, a node has to decide whether to forward the route-request message or not depends on its residual energy. If the residual energy is higher than a threshold value, then the node forwards the route-request message. Otherwise, it abandons the message and decline to participate in transmitting packets. Consequently, the destination node will receive a route-request message only when all intermediate nodes in the route have good energy levels, and nodes with low energy levels can preserve their battery power.

The Power Management Based Protocols are focused to achieve the energy efficiency goal by using two separate channels, one channel for control and another for data. RTS/CTS signals are transmitted through the control channel while data are transmitted over data channel. The protocol named power aware multi-access protocol (PAMAS)[18,20] in which the nodes sends a RTS message over the control channel when it ready to transmit and waits for CTS, if the CTS message not receives within a precise time then node enters to a power off state. In the receiving end, the node transmits a busy tone over the control channel to its neighbors when its data channel is busy. The control channel is used to determine when and how long the node to be in power off state. After turn to active state, a node can transmit data over the data channel. Conversely, once CTS is received, then the node transmits the data packet over the data channel.

Contrasting the previous techniques discussed, the sleep/power-down mode approach focused on inactive time of communication [17]. In MANET when all the nodes in a sleep mode packets cannot be delivered to a destination node. To overcome this problem, choose a special node named as *master* which can manage the communication on behalf of its neighboring slave nodes. At this moment, slave nodes may be in sleep mode for saving battery energy. Each slave node once in a while wakes up and communicates with the master node to detect if any data it has to receive or not. If no packed for the slave it may back to previous mode to save energy. In a multihop MANET, more than one master node can identified to handle the entire MANET. Fig.2 shows the master-slave network architecture, where nodes except master nodes can save energy by setting their power hardware into low state.

Geographic Adaptive Fidelity (GAF)[19] is the protocol fall under this category which uses location information to determine node equivalence with the help of GPS. The algorithm divides the entire network area into small virtual

framework. The nodes present in one virtual framework can communicate to the nodes present in its neighboring framework. Here the power management technique applies to place some of the node in to sleep state to conserve energy. The nodes can be in any of the states like, discovery, active or sleep.

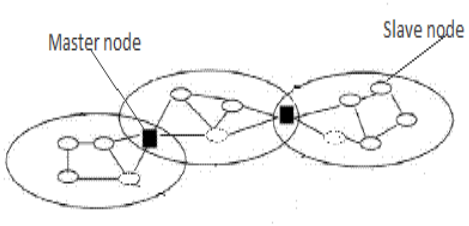


Fig.2 Master-slave Architecture

It applies load balancing approach to balance the lingering energy in a distributed manner. Any node with maximum lingering energy became the active node while its neighboring node goes to sleep state. This approach initiates more computation delay, extra messaging overhead, more energy consumption at each node.

4. Proposed Energy Efficient Technique

In this section we present the outline of our proposed technique. We regard as a network which consists of *N* nodes organized at randomly in the given area. We assume that all nodes may transmit at any power level P which is ≤ Pmax. All nodes that wish for transmission in the current session should have the minimum residual energy that is 15% of maximal battery capacity. We also assume that all nodes maintain their residual capacity all the time and have maximum bandwidth resources. When the node has capacity which is less than 15% of initial capacity, we push the node become in the sleep mode and marked it as rationally dead. It cannot forward packets to any further extent, but still it has enough energy to send packets. The node which marked as rationally dead can forward the high priority packet when this node is the only node that can forward the packet to destination node. After propel few packets in this emergency stage the node to become referred as actually dead. The algorithms proposed so far are minimize energy consumption per packet, consequently it minimize the total power needed to transmit a packet in a established route, or the algorithms focus on load distribution where the objective is to extend the minimum lifetime for the node. On the other hand, minimizing energy consumption is not taking care of the residual capacity of nodes, which decreases the life time of node when the traffic through the node is higher. Thus using power aware algorithm may exhaust all their energy very fast and die within a short period of time. On the other hand, when load distribution algorithms are used with the main consideration of power by each node, not taking into account the cost inspired during transmissions. It may lead to involve more number of nodes in the route. The proposed solution consists of using the algorithm which combines both energy consumption and shortest path for route algorithms and it also consider the node's residual capacity. As a result, we suggested that always using the path that consists of nodes having enough residual capacity which is larger than some predefined threshold. The objective of applying both techniques is to minimize the total power consumption by avoiding nodes with minimum battery lifetimes as well as increase the lifetime of network.

5. Conclusion

A Mobile Ad Hoc network (MANET) is a collection of nodes that can communicate with one another without any fixed networking infrastructure. Energy efficiency is one of the main problem in a MANET, especially in designing a routing protocol. In this paper, we surveyed and classified a number of energy aware routing techniques. Each technique has its own assumptions and different objectives and different methodologies in the implementation. For instance, in the transmission power control approach the power level is essential but the cost is not considered. The load distribution approach is efficient to improve the energy imbalance problem. There are different channels for sending data and control packets to reduce the energy consumption in power management approach but it increase the network traffic. The sleep/power-down mode approach is different from the other approaches as it focuses on inactivity energy. The proposed power aware algorithm combines the features of existing techniques to decrease the energy consumption and increase the lifetime of node & network.

References

[1] C.E.Perkins, "Ad Hoc Networking", Addison Wesley, 2001.

[2] S.Misra,I.Woungang and S.C. Misra, "Guide to Wireless Ad Hoc Networks", Springer science, 2009.

[3] Ashwani Kush, Sunil Taneja and Divya Sharma, "Energy Efficient Routing for MANET", IEEE, 978-1-4244-9703-4/101, 2010.

[4] D.Zhou and.T.H.Lai, " A scalable and adaptive clock synchronization protocol for IEEE 802.11-based multihop ad hoc Networks" IEEE International Conference on Mobile Ad hoc and Sensor Systems Conference, 2005 , Nov 2005.

[5] Li Q, AslamJ, Rus D. "Online Power-aware Routing in Wireless Ad-hoc Networks", Proceedings of International Conf. on Mobile Computing and Networking (MobiCom'2001) 2001.

[6] Chang J-H, Tassiulas L. "Energy Conserving Routing in Wireless Ad-hoc Networks", Proceedings of the Conf. on Computer Communications (IEEE Infocom 2000) 2000; 22-31.

[7] B. H. Liu, Y. Gao, C. T. Chou and S. Jha, "An Energy Efficient Select Optimal Neighbor Protocol for wireless Ad Hoc Networks," Technical Report, UNSW-CSE-TR-0431, Network Research Laboratory, University of New South Wales, Sydney, Australia, October 2004.

[8] Chen Huang, "On Demand Location Aided QoS Routing in Adhoc Networks, " 33rd International Conference on Parallel Processing (ICPP 2004), 15-18 August 2004, Montreal, Quebec, Canada. IEEE Computer Society 2004, pp 502-509.

[9] Nikolaos A. Pantazis, and Dimitrios D. Vergados, " A Survey on Power Control Issues In Wireless Sensor Networks", IEEE Communications Surveys & Tutorials, VOLUME 9, NO.4, 2007, pp.86-107

[10] I.W.Ho and S.C.Liew, "Impact of Power Control on performance of IEEE 02.11Wireless network", IEEE Transaction on Mobile Computing, vol. 6(11), pp. 1245-1258, November 2007.

[11] L.M. Freeny, "Energy efficient communication in ad hoc networks", Mobile Ad Hoc Networking, Wiley-IEEE press, pp. 301-328, 2004.

[12] P.K.Sahoo, J.P.Sheu and K.Y.Hsieh, "Power control based topology construction for the distributed wireless sensor Networks", Science Direct, Computer Communications, vol. 30, pp. 2774-2785, June 2007.

[13] Stojmenovic I, Lin X. "Power-Aware Localized Routing in Wireless Networks", IEEE Trans.Parallel and Distributed Systems 2001; 12(11):1122-1133.

[14] Jangsu Lee, Seunghwan Yoo, and Sungchun Kim, "Energy aware Routing in Location based Ad-hoc Networks", in communications control and signal processing (ISCCSP), 2010 4th international Symposium on 3-5 march 2010.

[15] Woo K, Yu C, Youn HY, Lee B. "Non-Blocking, Localized Routing Algorithm for Balanced Energy Consumption in Mobile Ad Hoc Networks", Proceedings of International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS 2001) 2001;117-124.

[16] Yu Wang, Wen-Zhan Song, Weizhao Wang, Xiang-Yang Li and Teresa A. Dahlberg, "LEARN: Localized Energy Aware Restricted Neighborhood Routing for Ad-hoc Networks", in Third Annual IEEE Communications Society Conference on Sensor, Mesh and Ad-hoc Communications (IEEE SECON 2006).

[17] Floriano, De Rango, Marco Fotino and Salvatore Marano, "EE-OLSR: Energy Efficient OLSR Routing Protocol For Mobile Ad-hoc Networks", in Proceedings of Military communications(MILCOM'08), San Diego, CA, USA, November 17-19, 2008.

[18] Benamar KADRI, Mohammed FEHAM and Abdallah M'HAMED,
"Weight based DSR for Mobile Ad Hoc Networks," in 3rd International Conference on Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. pp. 1- 6, 7-11 April 2008

[19] J.Gomez and A.T.Campbell, " Variable-Range Transmission Power Control in Wireless Ad Hoc Networks", IEEE Transactions on Mobile Computing, vol. 6(1), pp.87-99, January 2007.

[20] Rong Zheng and Robin Kravats, "On Demand Power Management for Adhoc Networks, " Journal of Adhoc Networks, Elsevier, Vol. 3, pp 51-68, 2005.

Author's Biography

Ms. S.Suganya received B.Sc Degree in Computer Science in 1994, MCA Degree in Computer Applications in 1997. She is currently working as a Sr.Lecurer in the Department of MCA, Tamilnadu College of Engineering, Coimbatore, Tamilnadu, India. She is currently pursuing Ph.D. Her research interest includes Resource allocation in Mobile Computing and Simulation. She has published 4 technical papers in International journals and National, International conferences.

Dr S.Palaniammal is working as Professor and Head, Department of Science & Humanities,VLB Janakiammal College of Engineering and Technology, Coimbatore,Tamil Nadu, India. She has 25 years of teaching experience at the Under graduate and Post graduate Levels for the Engineering students.She is Board of studies Member and Doctoral Committee member of various universities and Advisory Committee Member for several National/ International Conferences.She has organised many Seminars / Conferences /Workshop. 10 Research Scholars are pursuing their Ph.D under her guidance.Her area of interest includes DataMining, Queuing Theory and Computer Networks. She has published more than 56 papers in the National and International Journals and Conferences. She has authored 7 Books in Mathematics for various branches of Engineering Students.

A Study on Cyber Crimes and protection

M.Loganathan[1], Dr.E.Kirubakaran[2]
[1] Research Scholar
Department of Computer Science
Vinayaka Missions University, Salem
TamilNadu, India,636308

[2] Additional General Manager
Outsourcing Department
Bharat Heavy Electricals Limited,
Tiruchirappalli, TamilNadu, India, 620014

Abstract – **Information technology has widened itself over the last two decades and has become the axis of today's global development. The world of internet provides every user all the required information and latest information making it the most valuable source of information. With the advancement of internet, the crime has also widened its roots in all possible directions which claim to be the biggest threat in the near future. The cyber crimes pose a threat to the under developed, developing and the developed nations as a whole. One such major cyber crime is Phishing. It targets not just big organization but also individual users. In this paper we explore the Cyber crimes, the online security vulnerabilities and the available strategies and techniques for protection**

Index terms **– Security threats, Online Security, Cyber crime, Phishing**

I. INTRODUCTION

Crimes are as old as man himself and computer crimes are as old as computers themselves. The more advanced computers and technologies become, the more rise in computer crimes especially with the widespread of networks. People are very reliant on information systems and the Internet making them easy targets for cyber criminals. According to a report from McAfee based on a survey conducted globally on more than 800 IT company CEO's in 2009, data hacking and related cyber crimes have cost multinational companies one trillion US dollars. Cyber crimes take different forms and shapes and could be carried out, not only by using personal computers, but also through cell phones and PDA's.[1] To understand cyber crimes it is necessary to take a detailed view into the crimes

II. PHISHING

Phishing is the criminally fraudulent process of attempting to acquire sensitive information such as usernames, passwords and credit card details. Phishing often directs users to enter details in a fake website who's URL, look and feel are almost identical to the legitimate one. Even when using SSL with strong cryptography for server authentication it is practically difficult to detect that the website is fake. Phishing is an example of social engineering techniques used to fool users, and exploit the poor usability of current web security systems.[2]

Once the attacker has established a realistic and convincing fake web site that mimics a trusted brand, their main challenge is how to divert users of a legitimate web site to the fake web site instead. Unless the Phisher has the ability to alter the DNS for a target web site (DNS poisoning) or somehow otherwise redirect network traffic. A technique sometimes referred to as Pharming, they must instead rely on some form of content level trickery to lure unfortunate users to the fake web site The better the quality of the lure, and the wider the net that can be thrown, the greater the chance of an innocent user mistakenly accessing the fake website and in the process potentially providing the Phisher with the victim's credentials or other personal data)

III. URL OBSIFUCATION

Using URL obfuscation techniques, the attacker tricks the customer into connecting to their proxy server instead of the real server.[2] For example, the customer may follow a link to http://www.mybank.com.ch/ instead of the original link http://www.mybank.com/

IV. PHARMING

Pharming is a hacker's attack aiming to redirect a website's traffic to another bogus website. Pharming can be conducted either by changing the hosts file on a victim's computer or by exploitation of a vulnerability in Domain Name System's (DNS) server software. DNS servers are computers responsible for resolving the Internet names into their real addresses. Compromised DNS servers are sometimes referred to as "Poisoned". DNS cache poisoning is a maliciously created or unintended situation that provides data to a Domain Name Server that did not originate from authoritative DNS sources. Once a DNS server has received such non-authentic data and caches it for future performance, it is considered poisoned, supplying the non-authentic data to the clients of the server.

In recent years both have been used to steal the end user's identity information. Sophisticated measures known as anti-Pharming are required to protect against this serious threat. Antivirus software and spy ware removal software cannot guarantee to protect against Pharming.

V. ANATOMY OF PHISHING

A raw phishing message can be split into two components: the content and the headers. These components are commonly accepted as being the major components of a message.

1) Content: The content is the part of the message that the user sees and is used by phishing message producers to deceive users. It can be subdivided into two parts.

(i) The cover is the content which is made to look likea message from the legitimate organization, and usually informs the user of a problem with their account. Early phishing messages could be identified based only on their cover, due to imperfect grammar or spelling mistakes (which are uncommon in legitimate messages). Over time, the covers used in phishing messages have become more sophisticated, to the point where they even warn the users about protecting their password and avoiding fraud.

(ii) The sting is the part of the content that directs the victim to take remedial actions. It usually takes the form of a clickable URL that directs the victim to a fake website to log into their account or enter other personal details. We call this the sting, as this is the part of the content that inflicts pain, by means of financial loss or other undesirable action after the victim enters their details on the website. Typically the sting is hidden by using HTML to display a legitimate looking address, instead of the address of the fake website.

2) Headers: The headers are the part of the message which is primarily used by the mail servers and the mail client to determine where the message is going and how to unpack the message. Most users do not see these headers, but in terms of determining if a message is phishing or not, this part of the message can be quite useful. Headers can be subdivided into three parts based on the entities which add them to the message:

(i) Mail clients typically add headers such as "To:", "From:", "Subject:" and some client specific headers. Examples of mail client headers are X-MSMail-Priority, X-Mailer, and X-MimeOLE, Phishing messages may try to fake a particular header and in doing so, give away that the message is fake. For example, if the X-Mailer header indicates that a HTML message has been composed using MS Outlook but the message only contains HTML (without plaintext), this is an indication that the message is fake, as MS Outlook cannot send HTML only messages.

(ii) Mail relays will add headers along the path of the message. These are usually "Received" headers, which can be used to determine the originating IP of the message and the path taken by the message.

(iii) Spam-filters or virus-scanners will usually add headers to the message to indicate results of the tests run over the message. These headers can then be used by the receiving client to determine (based on a user-set threshold) what to do with the message.[3]

VI. MAN IN THE MIDDLE ATTACK

One of the most successful vectors for gaining control of customer information and resources is through man-in-the-middle attacks. In this class of attack, the attacker situates themselves between the customer and the real web-based application, and proxies all communications between the systems. From this vantage point, the attacker can observe and record all transactions. This form of attack is successful for both HTTP and HTTPS communications. The customer connects to the attackers' server as if it was the real site, while the attackers' server makes a simultaneous connection to the real site. In the case of secure HTTPS communications, an SSL connection is established between the customer and the attackers proxy (hence the attackers system can record all traffic in an unencrypted state), while the attackers proxy creates its own SSL connection between itself and the real server. For man-in-the-middle attacks to be successful, the hacker must re-direct the user to his proxy server instead of the real server.[2] This may be carried out through a

- DNS Cache Poisoning
- URL Obfuscation

VII. IDENTITY THEFT

Identity theft is undertaken by an individual or numerous individuals to facilitate criminal activity. [4] Specifically, it involves stealing another person's "identity"—personal and financial information—for the purpose of committing other

crimes constituting fraud. More often than not, these fraudulent acts are perpetrated by someone known by the victim such as a relative, friend, employee, or coworker, etc. Further, the success of these criminal acts directly depends upon the victim not knowing about it and the perpetrator of the act having an authentic address (one, however, that is actually bogus for the criminal).

(i) Constructions of Identity

According to Finch (2003) "identity theft spans a wide spectrum of conduct and covers varying degrees of fraudulent behavior." p.86 She states that in considering the nature of identity theft, it is important and necessary to distinguish between individual, social, and legal constructions of identity, terms she developed based on the work of Goffman (1963). Clearly, her intent in making these distinctions is to establish a clear delineation between identity and identifiability. In her taxonomy, "Individual identity is concerned with the question of 'who am I'." It is "what most of us think of when we think of the deepest and most enduring features of our unique selves that constitute who we believe ourselves to be.". "It can be seen as the sense of self that is that is based upon the internalization of all that is known about oneself...Hence individual identity is more than simply self-perception; rather, it is a subjective construction of the self that is modified by reflections on the views of others and the individual's interactions in the social world. As such individual identity is not a static construction but one that is constantly evolving and readjusting in line with the individual's life experience." Social identity, on the other hand, is concerned with the question of 'what is the nature of this person.' [4]

While individual identity "can be influenced by the way an individual is received in society," social identity "is contingent upon the way in which individuals present themselves." "For Goffman, social identity is based upon the categorisation of an individual to determine the acceptability of the membership of certain social groups." The key point to consider here, according to Finch, is that while both individual identity and social identity may be affected by identity theft, neither can be stolen. Legal identity is of concern in discussions of identity theft, because given its "fixed" and "immutable" nature; it has the greatest potential of being abused. Legal identity is concerned with the question of 'who is this person.' and "is more concerned with identifiability rather than identity as it seeks to make the link between a collection of facts and the person to whom they relate. . Therefore, it is clear that "the legal construction of identity gives primacy to factual information regarding an individual; information that is largely unalterable." [4]

(ii)Traditional Versus Online Identity Theft

According to the better business bureau, "identity theft is more prevalent offline with paper than on-line." On-line channels are blamed in only 9% of cases. Traditional means of obtaining information fraudulently include:

(1) dumpster diving (going through trash bins for checks, credit card numbers, identification numbers, pins, passwords, social security numbers, mail, receipts, or other sensitive information);

(2) shoulder surfing (involving watching someone enter personal information or eavesdropping on personal conversation/information);

(3) insider abuse (stealing on the job, bribing employees, etc);

(4) and lost wallets or purses (providing access to credit cards, checks, etc).

Online Identity theft happens in a number of ways including:

(1) Social Engineering—where users are manipulated into giving sensitive information (also used in f-t-f);

(2) Phishing—As explained in detail in the previous section of this paper where a spurious site imitates a well-known site;

(3) Pharming— As explained in detail in the previous section of this paper where malware redirects traffic destined for a legitimate website to one which looks like the original site;

(4) and Hacking—where the perpetrator intrudes into the system illegally and steals files (can be a method that is part of phishing or pharming).[4]

VIII.SOCIAL ENGINEERING

Social engineering is the practice of manipulating users to obtain confidential or sensitive information. Rather than exploiting the security of the technology, the social engineer exploits the weaknesses of the human user to trust the manipulator. It can apply to either to face-to-face, telephone, or internet manipulation to gain access to the physical computer itself or the information on it. Advance-fee scams (i.e. 419 scams) are an example of social engineering. The scam artist, pretending to be anyone from a government official to a surviving spouse, uses fee solicitation to acquire personal information with the promise of sharing inheritances, lottery winnings, and other sums of money. They play on the goodness and compassion of the victim with poignant stories, polite rhetoric, and the "guarantee" of financial gain for all involved. It usually involves the victim first being persuaded to open an e-mail attachment, followed by a malicious attack on the victim's computer, and the victim's computer or information then being used for criminal purposes, be it sending spam or stealing identities.[4]

IX. PROTECTION FROM PHISHING ATTACKS

There are several technical and non-technical ways to prevent Phishing attacks:[2]

1) Educate users to understand how Phishing attacks work and to be alert when Phishing-alike e-mails are received.
2) Use technical methods to stop Phishing attackers.

In this paper, we only focus on the technical aspect. Technically, if we can cut off one or several of these steps that are needed by a Phishing attack, we then successfully prevent that attack. In what follows, we briefly review these approaches:

A. Detect and block Phishing in time

If we can detect the Phishing Web sites in time, we can block the sites and prevent Phishing attacks. It's relatively easy to (manually) determine whether a site is a Phishing site or not, but it's difficult to find those Phishing sites out in time. Here we list three methods for Phishing site detection:

B. DNS Scan

The web master of a legal web site periodically scans the root DNS for suspicious sites. (e.g. www.icci.com vs. www.icici.com)

C. Enhance the security of the web sites

The business websites such as the web sites of banks can take new methods to guarantee the security of users' personal information. One method to enhance the security is to use hardware devices. For example, the Barclays bank provides a hand-held card reader to the users. Before shopping in the net, users need to insert their credit card into the card reader, and input their (personal identification number) PIN code, then the card reader will produce a onetime security password, users can perform transactions only after the right password is input. Another method is to use the biometrics characteristic (e.g. voice, fingerprint, iris, etc.) for user authentication. For example, Pay pal had tried to replace the single password verification by voice recognition to enhance the security of the Web site. With these methods, the Phishers cannot accomplish their tasks even after they have gotten part of the victims' information. However, all these techniques need additional hardware to realize the authentication between the users and the websites, which hence will increase the cost and bring certain inconvenience. Therefore, time is needed for these techniques to be widely adopted.

D Install online anti-Phishing software in user's computers

Despite all the above efforts, it is still possible for the users to visit the spoofed web sites. As a last defense, users can install anti-Phishing tools in their computers. The anti-Phishing tools in use today are categorized as: blacklist/White list based.

• Blacklist/White list: When a user visits a Web site, the anti-Phishing tool searches the address of that site in a blacklist stored in the database. If the visited site is on the list, the anti-Phishing tool then warns the users. Though the developers of these tools announced that they can update the blacklist in time, they cannot prevent the attacks from the newly emerged (unknown) Phishing sites which pop up in the internet frequently.[2]

X. BIOMETRICS

In addition to using antivirus software, firewalls, digital signature ... etc. to protect from cyber crimes, biometrics is used. Usually there are three different security measures used to authenticate or identify a person. [1]

1) What a person remembers: like password, personal identification number or any other keyword.

2) What a person can carry: smartcard, card key, token ... etc.

3) The person himself: the biological aspects like finger print, face, iris, or sound and that is called biometrics.

Biometrics is authentication techniques used in computer security in trying to stop cyber crimes. There are different methods of identification used. We will detail two of them:

1) Fingerprint: examination of unique fingerprints.

2) Palm Geometry: examination of the shape of the hand and fingers.

3) Voice: examination of the tone, pitch and frequency of voice.

4) Signature: examination how a person signs his name.

5) Retina: examination of the capillary vessels located at the back of the eye.

6) Iris: examination of the colored ring around the eye's pupil and this is done not by using infrared or any other type of rays but a simple camera and a person can even stand away from it. Each person of us has a unique iris that never changes as we grow old unlike the retina that does change.
The iris scan is not affected by contact lenses, eye glasses, refractive surgery, cataract surgery or cornea transplant.. . etc.

Some airports around the world have already started implementing the iris biometrics technology. The first airport to implement iris scanning technology was CharlotteDouglas International airport in North Carolina USA. By using a normal camera that shoots 30 frames per second in black and white, the images are digitized and stored in a database. The iris code with the person's

name and journey details are stored on a hard drive in a 512 bytes file and a resolution of 640x480.

7) Face: examination of facial characteristics. The distance between the eyes, width of the nose, depth of the eyes, and shape of the cheeks ... etc, more than 80 nodal points are reported to a database. All these nodal points are represented by a special number for each face and stored in the database. A total of 14 to 2 2 nodal points are enough to identify a person.

This works in five different stages:

a) Detection: the software will identify the face within the range of a video camera. 2010 International Conference on Networking and Information Technology

b) Alignment: it automatically adjusts the alignment to store the details of the position, size, type ... etc of the detected face.

c) Normalization: the software will try to normalize and to fix the image by correcting the size, or to rotate the image of the detected head with the proper background.

d) Representation: all the nodal points of the face are represented as a unique number.

e) Matching: the new collected-detected data are compared with the database for matching.

The image is stored as an 84 byte face print file and could be compared to other face prints in a huge database. The software can compare six million face prints/minute from the memory or 1.5 million/minute from the hard disk. Faces are used as passwords to enter into restricted areas or any site where a password is required.[1]

XI. FUSION CENTERS

Fusion centers were created in order to provide the capability to examine seemingly disparate pieces of information and to draw from them a picture of a pending or future attack. Fusion is the key term and, according to the DHS/DOJ Guidelines, means "turning information and intelligence into actionable knowledge." "Actionable" is the key term in this description. For fusion centers to be successful they need to not just produce vast quantities of information and reports but should instead produce knowledge that is actionable – knowledge and information that leaders can use to take actions that could prevent an attack from occurring. Again from Proceedings of the 41st Hawaii International Conference on System Sciences - 2008 2the DHS/DOJ guidelines we read that "For purposes of this initiative, fusion refers to the overarching process of managing the flow of information and intelligence across all levels and sectors of government and private industry. It goes beyond establishing an information/intelligence center or creating a computer network.

The fusion process supports the implementation of risk-based, information-driven prevention, response, and consequence management programs." This passage introduces the idea that fusion centers do not just rely on government organizations but have a private industry component as well. Building upon this idea the guidelines continue and state "data fusion involves the exchange of information from different sources—including law enforcement, public safety, and the private sector—and, with analysis, can result in meaningful and actionable intelligence and information. The fusion process turns this information and intelligence into actionable knowledge." What this fundamentally means is that for fusion centers to function, they need to be gathering information not just from law enforcement and intelligence agencies but from industry and the private sector as well. To this effect, the guidelines later state "ideally, the fusion center involves every level and discipline of government, private sector entities, and the public—though the level of involvement of some of these participants will vary based on specific circumstances." This is an important concept that becomes even more critical when considering cyber issues later.

The guidelines refer to the "fusion process" several times. This process is, quite simply, the steps necessary to turn information into actionable knowledge. The fusion process will:

• Allow local and state entities to better forecast and identify emerging crime and public health trends.

• Support multidisciplinary, proactive, riskbased, and community-focused problem solving.

• Provide a continuous flow of intelligence to officials to assist in developing a depiction of evolving threats.

• Improve the delivery of emergency and nonemergency services.

Building a fusion center capability is a phased process. This was true for the development of the initial creation of the fusion center concept and is equally true as entities develop their own fusion capability. The first phase is the introduction of the law enforcement and intelligence component. This is the backbone of every fusion center. The center will rely on individuals who have the training to perform an analysis of disparate data in order to form clear pictures of what might be indicated.

The second phase of building a fusion capability is the incorporation of public safety elements. This primarily means incorporating inputs from traditional first responders

within communities. It also includes individuals from the transportation, agriculture, and environmental protection sectors as well.

The third phase in constructing a fusion center capability is the inclusion of the private sector component. This last phase is critical to the success of fusion centers. Nearly 85 percent of the critical infrastructures needed by the nation on a daily basis are found in the private sector. But it is not just the critical infrastructures found in the private sector that are included in the final phase of building a fusion capability, it also includes private citizens and their inputs. Similar to the concept of the neighborhood watch program found in communities around the nation, private citizens can aid fusion centers by maintaining a certain level of vigilance in observing when abnormal activities occur within their communities. Law enforcement personnel and first responders can't be everywhere; citizens need to shoulder some responsibility for maintaining the security of the communities in which they live.

What citizens, the private sector, and the first responder community bring to the fusion center process is the gathering of data that will be examined by the intelligence analysts who will transform the various pieces of information into the actionable knowledge that we keep referring to. For conventional attacks and weapons of mass destruction, what information is needed is a fairly well understood process. (That is not to say that it is an easy process, just that we can describe what needs to occur for the process to be successful.) In intelligence terminology, what is being searched for are the various "indicators" of a pending attack. In the cyber realm this is a different matter. Very little has been done in terms of incorporating cyber into fusion centers and little is understood about what constitutes an indication of a potential cyber attack. Put simply, what is needed is a list of the things that people should be looking for and reporting on that would serve to indicate an attack may be in the planning or early stages of the execution process.[5]

XII. CONCLUSION

Cyber Crime becoming a serious security threat which causes loss of sensitive data like passwords, credit card information etc. which in turn causes loss in billions of dollars to both consumers and e-commerce companies. In this paper a detailed study has been made on the existing Cyber crimes and the available mechanisms which are used to counter attack the crimes. On a complete study, it is fair to say a new revolutionary technique is the need for the hour which will incorporate cyber laws into the technological ream to counter attack the cyber crimes to a greater extent.

ACKNOWLEDGEMENT

The authors would like to thank the Lord God Almighty for pouring his grace and strength upon us. Next, we would like to thank Vinayaka Misson University. We would also like to thank Bharat Heavy Electricals Limited for its valuable contribution and continuous motivation.

REFERENCES

[1] Alex Roney MathewDepartment of Information Technology,College of Applied Sciences, NizwaSultanate of Oman, "Cyber Crimes: Threats and Protection"; 2010 International Conference on Networking and Information Technology

[2] K.Nirmal, S.E Ewards, K Geetha, "Maximizing online security by implementing a three factor authentication to counter attach Phishing"; INTERACT Conference

[3] Danesh Irani, Steve Webb, Jonathon Giffin and Calton PuCollege of Computing Georgia Institute of TechnologyAtlanta, "Evolutionary Study of Phishing"; eCrime Researchers Summit, 2008

[4] Rae Carrington SchipkeDepartment of EnglishCentral Connecticut State University, "The Language of Phishing, Pharming, and OtherInternet Fraud—Metaphorically Speaking"; Technology and Society, 2006. ISTAS 2006

[5] Natalie Granado, Gregory WhiteCenter for Infrastructure Assurance and SecurityThe University of Texas at San Antonio, "Cyber Security and Government fusion Centers"

Corporate Customers Usage of Internet Banking in East Africa

Nelson Jagero and Silvance O Abeka

Abstract

The purpose of this paper was to identify the factors that influence corporate customers' adoption of Internet banking services in Kenya, Uganda, Tanzania and Rwanda. The hypotheses are empirically evaluated by using Trade Finance customers of an East African bank as the target sample. The study involved 137 respondents from Kenya, Uganda, Tanzania and Rwanda.

Due to the quantitative nature of the study, the results are analysed with statistical measures. The analysis reveals that corporate users are not motivated by the same factors as private users. In order to become Internet banking customers, it is extremely important for corporate users to have a system that is easy to use and operate with full support from the bank.

Keywords: Internet Banking, East Africa, Cooperate customers, Quantitative methodology

1. Introduction

Originally information technology was utilized in back offices for batch data processing, which was something not that obvious to the customers. Consumer oriented innovations became more important during 1980-1995. This time period is called the "diffusion period of the information revolution in commercial banking" (Bátiz-Lazo and Wood, 2002). Mainly this was possible due to Personal Computers (PC's), which enabled new contacts between banks and customers. But as expected, it didn't end there. After PC's invaded homes and workplaces, customers themselves could start communicating with the bank electronically from their own PC's. The information between customers' PC's and bank's systems did not transfer on-line at that time. Only after emergence of the Internet, banks have been able to provide real-time banking services electronically to a larger audience without a need to install anything on the customer's PC. (Bátiz-Lazo and Wood, 2002)

Historically branches and physical distribution channels have been the very cornerstones to most banks' market success. However, the emerging electronic channels have forced banks to change their entire management approach. Much of this is thanks to the fact that geographical and time restrictions do not limit the use of banking services anymore (Karjaluoto et al. 2002). As long as customers are connected to the Internet, they should be able to use the services when and where ever. The whole banking strategy has changed as a result of this; people are not dependent on the bank having branch closest to them physically, as it used to be. They can choose whichever bank offering its services online - or even several banks to serve different banking needs. This kind of development has shifted banks' attention more from marketing and selling of services and products towards building and managing customer relations.

Hypotheses

i. Perceived Usefulness positively influences use of Trade Finance Internet Services in East Africa.
ii. Perceived Ease of Use positively influences use of Trade Finance Internet Services in East Africa.
iii. Organizational Support positively influences use of Trade Finance Internet Services in East Africa.
iv. Bank Support positively influences use of Trade Finance Internet Services in East Africa.

2. Literature Review

Technology Acceptance Model

Technology Acceptance Model (TAM) was initially suggested by Fred Davis in 1989. It is one of the most studied and used models in the investigations of user acceptance of information technology. The model is adapted from Theory of Reasoned Action (TRA), which was originally proposed by Fishbein and Ajzen in 1975. Technology Acceptance Model is an information system theory, which purpose is simply to predict and explain the user acceptance of information technology. The model addresses the reasons why users either accept or reject particular piece of information technology. The revised model by Davis et al. (1989) is constructed from external variables (external stimulus), perceived usefulness and perceived ease of use (cognitive response), behavioral intention, and actual usage (behaviour). (Davis et al. 1996a)

Quite a few researchers have applied TAM when studying acceptance of Internet banking. Liao et al. (2002) even made an invariance analysis concluding that TAM is a well suitable instrument for evaluating Internet banking acceptance, but also that the suitability is independent of the respondent characteristics such as gender, age and information technology competence. The current research done about Internet banking and Technology Acceptance Model are reviewed next, presenting the major findings of them and the empirical environment.

Sudarraj et al. (2005) used deconstructed TAM to measure the importance of usefulness and ease of use in online and telephone banking. They successfully validated the model with Canadian university students. Karjaluoto et al. (2002) built a model based on TRA and TAM, which was empirically tested with private Finnish retail bank customers. Their conclusion is, that “prior computer experience, prior technology experience, personal banking experience, reference group, and computer attitudes strongly affect attitude and behaviour towards online banking.” (Karjaluoto et al. 2002).

Supporting findings were those of Lassar et al. (2005) who studied online banking adoption in the United States in the light of TAM. They concluded that the intensity of Internet usage is significantly influencing individuals’ adoption of Internet banking. These findings suggest that the more experienced the consumers are in using the computers and the Internet, the more likely it is that they will start using Internet banking.

Another Finnish study investigated consumers’ acceptance of online banking: Pikkarainen et al. (2004) added perceived enjoyment, information on online banking, security and privacy and quality of Internet connection to the model. Surprisingly, they found only PU and information of online banking significantly affecting use of Internet banking services in Finland. Hong Kong students were used to empirically test another modification of TAM; in this study Chau and Lai (2004) also discovered that PU could be the only major factor directly influencing the attitude towards online banking. PEOU influenced also directly, but mainly via PU. Other measured factors like alliance services, personalization and task familiarity

influenced through PU, and accessibility through PEOU.

Suh and Han (2002) added trust to the original TAM model. They studied their model by empirically evaluating responses from personal customers of five major banks in South Korea and discovered trust to be a very significant determinant of user acceptance of Internet banking. Trust had a significant positive effect on both PEOU and PU, out of which PU appeared to be stronger in predicting the intention to use Internet banking.

Trust was handled also by Wang et al. (2003). Their research aimed on recognizing the determinants of user acceptance of Internet banking. In this research they introduced perceived credibility as a new factor to TAM, in addition to self-efficacy, perceived usefulness and perceived ease of use. The model was empirically tested by phone interviews
with Taiwanese consumers. Surprising results were found: perceived ease of use and perceived credibility were more significant than perceived usefulness in predicting the behavioural intention to use Internet banking. The surprising factor in this was, that majority of TAM related research has concluded that PU is the ruling factor over PEOU. Self-efficacy

3. Methodology

Research population

The researcher targeted population was business process in the organization, corporate Customers, Current system if any, Capability of the organization's technology infrastructure and the management of the organization

Research Instruments

Quantitative analysis was chosen to test the research model, as it is good for measuring how many and in what proportion. In addition, with statistically reliable quantitative research it is possible to generalize the results: if the same questions are asked from different people with the same characteristics, the answers should support the outcome of the study.

The method for collecting empirical data for the statistical analysis was customer survey. Questionnaires were sent out to randomly selected Trade Finance customers of the case bank; (Kenya Commercial Bank) in Kenya, Uganda, Rwanda and Tanzania. The questionnaires were developed together with this banks best Trade Finance specialists. With the help of the expertise of these specialists, the questionnaire content and validity of the questions were confirmed to facilitate achieving the goal of the study in the best possible way. In addition the questionnaires contained questions outside of this research, mainly related to customer service and open-ended comments. The responses to those questions are used for further analysis only for the case bank's purposes.
The survey questions and their relation to the hypotheses are presented in the table below.

Table 1. Questionnaire questions for hypothesis testing

FACTOR	VARIABLE	HYPOTHESIS	SURVEY QUESTION
Perceived usefulness	PU	H1	I find / I think I would find TFIS useful in conducting Trade Finance banking transactions
Perceived ease of use	PEOU_1	H2	a) I find / I think I would find it easy to do what I want to in TFIS
	PEOU_2	H2	b) I find / I think I would find TFIS easy to use
Organizational support	OSU_1	H3	a) It is / would be important for me to have someone else in my organization to help out in case of non-technical* problems with TFIS
	OSU_2	H3	b) It is / would be important for me to have someone else in my organization to help out in case of technical** problems with TFIS
Bank support	BSU_1	H4	a) It is / would be important for me to have someone to help out in the bank in case of nontechnical* problems with TFIS
	BSU_2	H4	b) It is / would be important for me to have someone to help out in the bank in case of technical** problems with TFIS

* Non-Technical problem could be for example creating a template, finding a deal via Inquiry, etc)
**Technical problem c

Data Analysis Method

The analysis was done with a system designed for statistical analyses (SPSS). Descriptive statistics and regression analysis, completed with Pearson product-moment correlation analysis, were selected as the methods for interpreting and analysing the empirical data. With the help of these statistical measures, the validity of the theoretical model and hypothesis are tested.

Regression analysis was chosen, for it fits well for hypotheses testing and analysing how independent variables can be used to predict a dependent variable. Linear regression is based on correlation between the variables, in this case Pearson product-moment correlation, but it enables more detailed and sophisticated examination of the interrelationship of the variables.

Analysis called ANOVA is conducted in order to determine the statistical significance of the correlations between the selected variables. The p-value of the F-test indicates the level of association between the dependent and independent variables in the model. When the significance p-value is less than 0.05, it

means there is a statistically significant association between the dependent and independent variables. P-value 0.10 refers to weakly significant association. If the p-value is more than 0.10, then the model chosen is not statistically significant.

4. Analysis of Findings

Use of the system

Most of the responses came from users of the system (90%). Customers, who do not currently use the system, but reported that they will in the future cover 7%. Only 3% of all responses came from customers who do not use the system, and do not intend to. All Kenyan customers were users of the system, and also in Uganda only one of the responses came from a non-user. The most non-users were registered from Rwanda (36%). This can probably be explained by the distinct difference in TFIS between Rwanda and the other countries.

Table 2 User statistics

Usage	Frequency (%)
Don't Use and Won't	**3**
Don't Use but will	**7**
Users	**90**

T-tests

An independent T-test was conducted to compare the scores for each of the variables between users and non-users, females and males, older and younger, and between those with higher and lower education.

Differences between users and non-users

A t-test was conducted to compare the outcomes for each of the variables between users and non-users. Table 3 contains the outcome for this test.

Table 3. T-tests between users and non-users

	Mean		Levene's Test for Equality of Variances		t-test for Equality of Means
	Non-user	**User**	**F**	**Sig.**	**Sig. (2-tailed)**
PU	4.10	4.27	1.445	0.232	0.555
PEOU_1	3.56	3.97	0.197	0.658	0.185
PEOU_2	3.22	4.01	0.074	0.786	0.011
OSU_1	2.80	2.63	0.080	0.777	0.688
OSU_2	3.20	2.96	0.171	0.680	0.582
BSU_1	3.83	4.56	0.987	0.323	0.001
BSU_2	3.55	4.54	1.143	0.287	0.000

** F-value for Equal variances assumed was lower than 0.05. Therefore values for equal variances not assumed are used.*

As can be seen from the table above, both users and non-users find the system useful. Non-users seem to be more aware of using the system, and have more confidence on them when it comes to using it. Non-users also have more experience on using other bank services provided in the Internet. For non-users organisational support is more important. The only variables that are statistically significant between users and non users are PEOU_2 and BSU_1 and BSU_2 ($p < 0.05$). These three are all scored higher among the users. The finding about bank support is also in line with the regression analysis results for the adjusted model.

Differences between females and males

A t-test was conducted to compare the outcomes for each of the variables between females and males. The results of this comparison can be seen in table 4.

Table 4. T-tests between males and females

	Mean		Levene's Test for Equality of Variances		t-test for Equality of Means
	Female	**Male**	**F**	**Sig.**	**Sig. (2- tailed)**
PU	4.29	4.18	1.036	0.311	0.561
PEOU_1	4.10	3.53	0.896	0.346	0.002
PEOU_2	4.06	3.65	0.866	0.354	0.027
OSU_1	2.70	2.52	2.396	0.124	0.503
OSU_2	2.99	2.97	0.491	0.485	0.942
BSU_1	4.56	4.29	2.643	0.107	0.077
BSU_2	4.54	4.21	3.263	0.074	0.049

Based on the T-test results, there is statistically significant difference between the scores of males and females in Perceived Ease of Use and Bank Support. Both PEOU_1 and PEOU_2 have received higher scores by the females. Both BSU_1 and BSU_2 are statistically significant: BSU_2 somewhat more strongly (p-values<0.01 and $P<0.05$ respectively. Hence, females think the system is easier to use than males, but to them the importance of support received by the bank is bigger than for males – especially technical support. That is not very surprising if traditional roles and areas of interest are considered; men in general tend to be more self-assured about technical matters.

Differences between age groups

A t-test was conducted to compare the outcomes for each of the variables between respondents of different ages. They were divided into two categories: respondents between 24-45 years and 46-65 years. The results of this comparison can be seen in table 5.

Table 5. T-tests between Age Scales

	Mean		Levene's Test for Equality of Variances		t-test for Equality of Means
	24-45 years	46-65 years	F	Sig.	Sig. (2- tailed)
PU	4.28	4.20	0.188	0.665	0.622
PEOU_1	3.97	3.91	1.417	0.236	0.721
PEOU_2	3.94	3.98	1.270	0.262	0.843
OSU_1	2.67	2.56	0.145	0.705	0.656
OSU_2	2.94	3.00	0.436	0.510	0.819
BSU_1	4.51	4.43	0.066	0.798	0.599
BSU_2	4.60	4.18	5.181	0.025	0.009

According to the T-test between respondents of age 24-45 and 46-65, there is statistical significant difference in variables EXP_1 and BSU_2. The results indicate that the older the users are, the more experience they have in other Internet bank services and the less technical support they need from the bank. This is surprising when considering the common impression that younger are more familiar with electronic banking services, which also has been supported by empirical results in few of the studies (Karjaluoto et al. 2002).

Differences between education levels

A t-test was conducted also for comparison of scores for each of the variables between respondents with different education levels. They were divided into two categories: respondents with elementary school, high school education, and those with university bachelor's or master's degree.

Table 6. T-tests between Low and High educated

	Mean		Levene's Test for Equality of Variances		t-test for Equality of Means
	Lower education	Higher education	F	Sig.	Sig. (2- tailed)
PU	4.37	4.19	1.347	0.248	0.315
PEOU_1	4.00	3.96	0.938	0.335	0.806
PEOU_2	3.95	4.00	0.586	0.446	0.782
OSU_1	2.65	2.62	0.393	0.532	0.897
OSU_2	3.10	2.91	0.667	0.416	0.469
BSU_1	4.61	4.41	1.068	0.304	0.186
BSU_2	4.54	4.390	0.083	0.774	0.390

The T-test results indicate that there is a big difference in previous experience. Similarly surprising results can be seen with the education level of the respondents, as with the age and use of Internet banking of females: Clearly the higher the level of education, the less experience the respondent has with both Internet banking and other Internet services.

Again the common understanding and empirical evidence from studies done before do not support this notion.

Differences between nationalities

In order to distinguish the differences between Kenyan, Ugandan, Tanzanian and Rwandan respondents, a t-test was also made to compare the scores of each of the variables. The analysis of the countries and the differences of scores were done by pairing the countries for the analysis. This approach was chosen to see the differences in more detailed.

Table 7. Mean values for Kenya, Uganda, Rwanda and Tanzania

	Mean			
	RWANDA	**UGANDA**	**TANZANIA**	**KENYA**
PU	4.54	4.20	4.05	4.42
PEOU_1	4.00	3.84	3.78	4.29
PEOU_2	3.77	4.02	3.70	4.13
OSU_1	2.75	2.62	2.96	2.33
OSU_2	3.17	2.81	3.17	3.08
BSU_1	4.14	4.62	4.00	4.88
BSU_2	4.36	4.57	3.92	4.75

Table 8. T-tests between Kenya, Uganda, Rwanda and Tanzania

	Levene's Test for Equality of Variances and t-test for Equality of Means											
	RWANDA-UGANDA		**RWANDA - KENYA**		**RWANDA-TANZANIA**		**UGANDA - KENYA**		**UGANDA-TANZANIA**		**TANZANIA - KENYA**	
	F	**Sig. (2-tailed)**	**F**	**Sig. (2-tailed)**	**F**	**Sig. (2-tailed)**	**F**	**Sig. (2-tailed)**	**F**	**Sig. (2-tailed)**	**F**	**Sig. (2-tailed)**
PU	1.056	0.262	0.255	0.616	0.347	0.074	3.443	0.337	3.567	0.531	0.115	0.083
PEOU_1	0.102	0.558	0.863	0.341	0.269	0.494	0.521	0.039	0.097	0.810	0.143	0.054
PEOU_2	0.874	0.354	0.011	0.247	0.091	0.831	1.289	0.609	2.329	0.154	0.259	0.123
OSU_1	2.462	0.749	2.998	0.387	0.014	0.587	0.663	0.381	4.394	0.267	4.696	0.100
OSU_2	1.345	0.376	8.810	0.875	0.004	1.000	7.151	0.431	2.051	0.240	12.841	0.837
BSU_1	5.312	0.018	14.732	0.001	0.823	0.624	12.620	0.068	0.197	0.000	4.657	0.000
BSU_2	0.204	0.342	1.438	0.123	0.611	0.167	1.735	0.306	2.674	0.002	3.924	0.002

When looking at the table 8, several small statistically significant differences can be distinguished from the T-test results: Variable Awareness is statistically significantly lower in Uganda than in Rwanda and Kenya. Experience in other Internet services is significantly different among Rwandans and Tanzanians and Rwandans and Kenyans.

Although all the countries value support from the bank in both non-technical (BSU_1) and technical (BSU_2) issues, there is clear statistical difference between Kenyans, Ugandans and the other

countries. Especially Tanzanians give the lowest scores to both: M=4.00 and M=3.92 respectively. Especially score for the technical support is significantly lower than it is for Kenyans and Ugandans. Rwandans do not see non-technical support as important as Kenyans and Ugandans either.

Tanzanians and Rwandan corporate customers obviously do not value support from the bank as much as Kenyans and Ugandan customers do. Therefore it is good to keep in mind that most of the respondents of this research were from Uganda and Kenya. However, none of the nations seem to demand much of support from their own organisations. In general, Rwandans are the least experienced, and Ugandans have the least confident and lowest level of awareness of the system usage.

References

Bbátiz-Lazo, B. & Wood, D. An Historical Appraisal of Information Technology in Commercial Banking, Electronic Markets, Vol. 12, No. 3, pp.192-205, 2002

Chau, P.Y.K. & Lai, V.A.K. An Empirical Investigation of the Determinants of User Acceptance of Internet Banking, Journal of Organizational Computing and Electronic Commerce, Vol. 13 No. 2 pp. 123-145, 2003.

Davis, F. D., Bagozzi, R. P & Warshaw, P. R. User Acceptance of Computer Technology: a Comparison of Two Theoretical Models, Management Science, Vol. 35, No. 8, 1989.

Davis, F. D & Venkatesh. V. A critical assessment of potential measurement biases in the technology acceptance model: three experiments, Int . J . Human – Computer Studies, Vol. 45 pp. 19-45, 1996

Karjaluoto, H., Mattila, M. & Pento, T. Factors underlying attitude formation towards online banking in Finland, International Journal of Bank Marketing Vol. 20, No. 6 pp. 261-272, 2002

Lassar, W. M., Manolis, C. & Lassar, S. S, The relationship between consumer Innovativeness, personal characteristics, and online banking adoption, International Journal of Bank Marketing, Vol 23 No. 2 pp. 176-199, 2005.

Liao, Z. & Cheung, M. T. Internet based e-banking and customer attitudes: an empirical study, Information & Management, Vol. 39 pp. 283-295, 2002.

Pikkarainen, T., Pikkarainen, K. Karjaluoto, H. & Pahnila, S.. Consumer acceptance of onlinebanking: an extension of the technology acceptance model, Internet Research, Vol. 14 No. 3, pp. 224–235 2004.

Suh, B. & Han, I. , Effect of trust on customer acceptance of Internet banking, Electronic Commerce research and applications, Vol. 1, pp. 247-263,2002.

Sudarraj, R.P., & Wu, J. Using information-systems constructs to study online and telephonebanking technologies, Electronic Commerce research and applications, Article in press, 17 pages. 2005.

Wang, Y.S., Wang, Y-M., Lin, H-H. & Tang, T.I. Determinants of user acceptance of Internet banking: an empirical study, International Journal of Service Industry Management, Vol. 14 No. 5, pp. 501-519, 2003.

About the authors:

Dr Nelson Jagero is a senior lecturer in the school of postgraduate studies and research at Kampala International University Dar es Salaam Collage. His areas of specialization include Research methods, statistical methods, quantitative analysis in business and public administration and Operation Research. He has been a lecturer at Maseno University in Kisumu Kenya. He has a PhD (2009) from Maseno University, Masters in 1999 from Moi University and bachelors 1990 from Kenyatta University

Silvance O Abeka is the director of marketing at Kampala International University Dar es Salaam Collage. He holds a masters degree in Business Administration (Information Technology) (2009) and currently he is undertaking a PhD in Management Information Science (MIS), At Kampala International University Dar es Salaam Collage. His interests include Management Information Systems, Principles of Statistics and E- Commerce. H is also a lecturer in the school of computer studies.

An Alternative Process Documentation for Data Warehouse Projects

Jyothi Prasad K S S [1], Smt. G Hima Bindu[1], Smt. G Lakshmeeswari[1]

[1]Department of Computer Science
GITAM University, Visakhapatnam, Andhra Pradesh, India.

ABSTRACT

It is a well-known fact that software documentation is, in practice, poor, incomplete and flexible. Projects may wish to add, change, remove or ignore any part of any document. Some may also believe that aspects of one document would sit better in another. If this is the case then users of this document and these templates are encouraged to change them to fit their needs.

The paper describes a process for the documentation that describes the template in data warehouse projects. We focus our attention to develop a series of guides and checklists. This ensures that small teams of relatively skilled resources developing the system can cover all aspects of the project whilst being free to deal with the specific issues of their environment to deliver exceptional solutions, rather than a rigid methodology that ensures that large teams of relatively unskilled staff can meet a minimum standard.

Keywords: *Software Documentation, Template, Document Standards.*

1. INTRODUCTION

Software engineering projects, as defined by the IEEE/EIA, consist of a number of development activities [1]. Each activity is characterized by a set of deliverables, normally in the form of code or documentation. Providing a structured template for software documentation assists the software engineering project. These templates provide a guide to the expected format and content of the documentation deliverables based on international standards. They also provide a framework for the evaluation of the project.

For large software projects, it is usually the case that documentation starts being generated well before the development process begins. For some types of systems, a comprehensive requirements document may be produced which defines the features required and expected behavior of the system. During the development process itself, all sorts of different documents may be produced – project plans, design specifications, test plans etc.

It is not possible to define a specific document set that is required – this depends on the contract with the client for the system, the type of system being developed and its expected lifetime, the culture and size of the company developing the system and the development schedule that it expected [2].

However, we can generally say that the documentation produced falls into two classes:

1. *Process documentation:*

These documents record the process of development and maintenance. Plans, schedules, process quality documents and organizational and project standards are process documentation.

2. *Product documentation:*

This documentation describes the product that is being developed. System documentation describes the product from the point of view of the engineers developing and maintaining the system; user documentation provides a product description that is oriented towards system users.

Process documentation is produced so that the development of the system can be managed. Product documentation is used after the system is operational but is also essential for management of the system development. The creation of a document, such as a system specification, may represent an important milestone in the software development process.

Further more documentation standards act as a basis for document quality assurance[3]. Documents produced according to appropriate standards have a consistent appearance, structure and quality. The standards that may be used in the documentation process are:

a. Process standards

These standards define the process which should be followed for high-quality document production.

b. Product standards

These are standards which govern the documents themselves.

c. Interchange standards

It is increasingly important to exchange copies of documents via electronic mail and to store documents in databases. Interchange standards

ensure that all electronic copies of documents are compatible.

Standards are, by their nature, designed to cover all cases and, consequently, can sometimes seem unnecessarily restrictive. It is therefore important that, for each project, the appropriate standards are chosen and modified to suit that particular project. Small projects developing systems with a relatively short expected lifetime need different standards from large software projects where the software may have to be maintained for 10 or more years.

This paper has looked at a consistent set of documents developed at **process documentation phase** that reflects a desire to develop the right amount of documentation at the right time in the project lifecycle and stored in the right place. It is essential to the success of a data warehouse project that a culture of open access is fostered and that the documentation is seen as the entry point to the data warehouse projects.

Here we have identified three aspects to essential documentation:

• A roadmap that describes what documentation is required and how it fits together.
• Team members within the project to use the templates, create quality documents and store them to the project repositories.
• Easy access for people outside the project team to the documentation including publication or notification of changes, updates and new releases.

2. Process documentation

Effective management requires the process being managed to be visible. Because software is intangible and the software process involves apparently similar cognitive tasks rather than obviously different physical tasks, the only way this visibility can be achieved is through the use of process documentation.

Here the process documentation in data ware housing project proposes a six phase approach that maintains focus on the critical success factors along the development path:

i) Committed user and technical staff involvement from the beginning
ii) Clear definition of scope to prevent paralyzing scope-creep,
iii) Early executive review and buy-in to ensure priorities are met,
iv) Careful attention to configuring a platform that will enable rapid response time to queries,
v) Intense scrutiny of the data loading and cleansing process to ensure data integrity from source to data warehouse, and
vi) Documentation and training of technical and production staff and end users to guarantee active use, refinement, and custodianship of the data warehouse.

2.1 Data Warehouse Development approach

Data Warehouse Development in 6 phases		
Phase	Tasks	Results
1	Workshops & develop Prototype data warehouse	Data Warehouse Prototype
2	Procure the data warehouse Equipment & consulting Services	Hardware, Software and Implementation Plan
3	Develop the data Ware house software & convert the Initial Data	Operational software, Initial Queries, Reports & Data
4	Install the Data Warehouse Hardware, Software & Converted Data	Data Warehouse Goes live
5	Train the Data Warehouse Users & Operational Staff	Software Docs and User Manuals
6	Refine the Data Warehouse Data, Queries & Report	Revised Queries & Reports

As per the phases mentioned in the above development approach we define a model for process documentation in below figure (Fig1 Process of Software Documentation). Here phases 2 and 3 define the SDD ,phases 4 and 5 define the technical document and further more phases 5 and 6 defines the testing document and the product document .

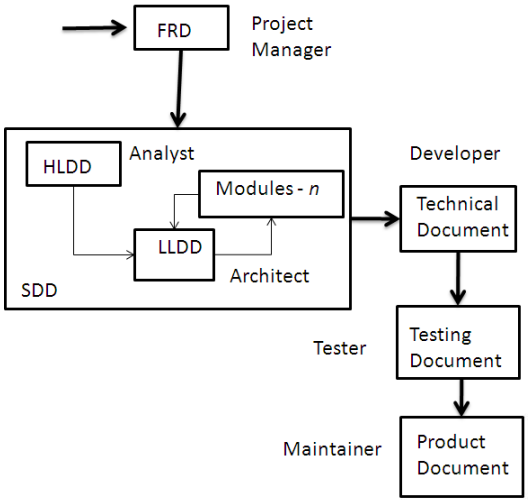


Fig 1: Process of Software Documentation

The process starts with collecting FRD (Functional Requirement Document) which contains Business Requirements and System Requirement This collection takes place at client location. These are submitted to the company or organization which should produce end product and this is handled by the Project Manger (PM)

The documentation contains the basic need of project i.e. scope, out of scope ,the purpose, resource, risk factor and other considerations of project with cover page attached which contains project title, project code number, Team Members (Roles, experience), deadline, cost and etc. This FRD forms the input for next phase of SDLC for Analysis and Design phase. The analyst converts the FRD into Software Description Document which is turn divided into two ways[5].

First is the High Level design Document (HLDD), this document is produced by the analyst after a long instant of the analysis of the FRD.

Second is the Low Level Design Document (LLDD) developed by the architect who contains the information of the modules that exist in the project. Here the documentation contains the templates that represent the information of module-*n* with module identity, name of module, status of module etc.

Further more the above information about the Modules helps in easy development (coding) of project by the developer. The developer starts coding based on these modules. After the completion of coding the unit testing is performed by the developer and in resultant produces a Technical document which contains start and ending time schedules of module tested and the testing report with input provided and expected outputs resulted etc.

This Technical document serves as basis input for the testing document phase and makes tester task easy by not dealing with unnecessary errors i.e. while developing the project by the help of tools, there are some packages that show errors or upload errors at some constraints. The tester will skip such errors and these error are supported neither with hardware or software

Finally the Tester produces Testing document by performing various forms of testing the following various types testing carried out in testing environment.

1) Development Integration Testing
2) System Integration Testing
3) User Acceptance Testing

Further if there exists inconsistency with in the testing report, retesting take place for error free project satisfying all the needs in the Software Requirement Specification. This testing document produced is finally submitted to PM. Based on the testing report the maintainer start producing product documentation which is used by end user.

*Note this paper provides specific assessment criteria: it describes the development process of software documentation and it does not cover the product documentation which contains the following documentation (user manual, reference manual, installation manual etc.

3. Procedure of process documentation in Data ware house projects:

This process of documentation is produced at the end of phases. Know let us examine how this process is used in Data Warehouse (DWH) project. As the client approach the organization or company in bidding process, the Project manager (PM) of that selected organization is solely responsible for returning error free, quality product to client. DWH project neither be an existing or new one. The client do explain the requirements that gives a projection of the aspects of the project Based on these requirements FRD is produced by the PM which leads the process into the next phase of process documentation i.e. SDD .Here with in the SDD phase firstly, the DWH project deals with databases design, data sources design, data marts design, which is developed with the analysis of the design models such as E-R diagrams, dataflow diagrams, use cases etc.

Next the Low Level Design Document (LLDD) developed by the architect which contains the information of the modules that exist in the project which further helps the Based on design models the architect divided into set modules and developers for writing the code easily with the help of respective

To avoid risks in developing, the meetings are conducted once in a week to share Knowledge Transfer. This meeting follows a mesh wise procedure i.e. one of the team member starts the meeting by questioning or explaining the risks aspects and in turn any one of the team member can give solution or suggestions for that aspects. Concluded multiple views can be shared.

Knowledge Transfer and Testing are the main documents to be produced error free [4][6].

Each document phase follows a template basing upon the requirement here basing upon the

requirement some of the information within the templates can be ignored.

4.TEMPLATES

COVER PAGE (contents & Layout)

Name of Document

Project Title

Document Version Number

Printing Date

Location of Electronic version of file

Domain

Fig 2: Template 1

REVISION PAGE (Contests)

Overview
Target Audience
Project Team Members
Version Control History:

Version	Primary Author(s)	Description of Version	Date Completed
Draft/final			

Signatures of Approval

Fig 3: Template 2

FUNCTIONAL REQUIREMENT DOCUMENTATION (FRD)

Cover page
Revision page
Table of contents
1. Introduction
   1.1 Purpose
   1.2 Scope
   1.3 Out of Scope
   1.4 Reference
   1.5 Assumptions and constraints
2. Project Manager and Methodology Selection
3. Functional Requirements
   3.1 Context
   3.2 User Requirements
   3.3 Data Flow Diagrams
   3.4 Logical Data Model /Data Dictionary
4. Other Requirements
   4.1 Interface Requirements
   4.2 Data Conversion Requirements
   4.3 Hardware/software Requirements
   4.4 Operational Requirements

Fig 4: Template 2

HIGH LEVEL DESIGN DOCUMENT (HLDD)

Cover page
Revision Page
1. Introduction
   1.1 Scope
   1.2 Assumptions
2. Database Architecture
   2.1 Metadata Design
   2.2 Fundamental Entities Design
   2.3 Transaction Entities Design
   2.4 Data marts Design
   2.5 Classification model design
   2.6 Transformation from logical to physical design
3. Mapping to the Data Warehouse Model
   3.1 Mapping requirements to DW model
   3.2 Mapping Fundamental Entities
   3.3 Mapping of Transaction Entities
   3.4 Mapping of classifications
4. Design Tools

Fig 5: Template 3

LOW LEVEL DESIGN DOCUMENT (LLDD)

Cover page
Revision page
1. Environment
2. Tools
3. Construction process
   3.1 Logic Design (Coding)
   3.2 Unit Testing and report
4. Meeting / Review (Internal)
   Knowledge Transfer
5. Review Report

Fig 6: Template 4

MODULE – *n*

Cover Page
1. Module –*n*
   1.1 Module id
   1.2 Module Name
   1.3 Module Status
   1.4 Module Error
2. Report

Fig 7: Template 5

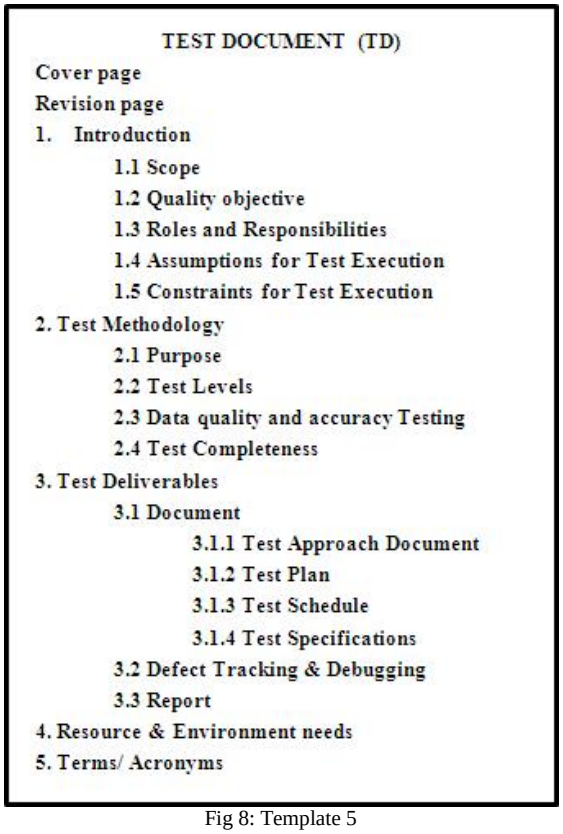


Fig 8: Template 5

5. TABLES

Knowledge Transfer :						
Project Id:						
Project name:						
S.No	Module ID	Member name	Role	Risk Found	Suggestions/ idea	Remarks

Table1: Knowledge Transfer

Module Report :						
Project Id:						
Project name:						
S.No	Module ID	Member name	Status	Start Date	End Date	Remarks

Table 2: Module Report

Testing Report :						
Project Id:						
Project name:						
Test Case ID	Test Case	Input	Excepted Output	No. of Error/Defects	Comments	Status

Table 3: Test Report

6. Conclusion:

Many data warehousing projects are both long running and poorly documented. It does'nt mean that there is a lack of documentation, but just a lack of the right documentation in the right place. It is the quality and availability of the documentation that leads to an understanding of what is available and hence to the value and reputation of the data warehouse itself.

7. References

[1]. Garg, P. K. and Scacchi, W. 1990. 'A hypertext system to maintain software life-cycle document' IEEE Software, 7 (3), 90–8.

[2]. IEEE, 1987. IEEE Standard for Software User Documentation, IEEE-Std1063-1987

[3]. Chapter 30 from book Software Engineering, 4th edition, published by Addison Wesley in 1992.

[4]. IEEE Std. 829-1998 IEEE Standard for Software Test Documentation

[5]. IEEE Std. 830-1998 IEEE Recommended Practice for Software Requirements Specifications

[6]. IEEE Std. 1008-1997 IEEE Standard for Software Unit Testing

Jyothi Prasad K S S, student of M.Tech(CST) in the Dept. of Computer Science and Engineering , GIT, GITAM University Visakhapatnam, AP, India.

G.HimaBindu,Asst.Professor, Dept. of Computer Science and Engineering , GIT, GITAM University Visakhapatnam, AP, India. Her research interests include Software Engineering ,image processing, Security and Computer Networks. She has 6 years of teaching experience.

G Lakshmeeswari Asst.Professor, Dept. of Computer Science and Engineering , GIT, GITAM University Visakhapatnam, AP, India. Her research interests include Software Engineering ,image processing, Security and Computer Networks. She has 12 years of experience and is now doing her Ph.D in Steganography.

Analysis and Improvement of DSDV Protocol

Nayan Ranjan Paul[1], Laxminath Tripathy[2] and Pradipta Kumar Mishra[3]

[1] Department of Computer Science and Engineering, KMBB College of Engineering and Technology, Khurda, Odisha, India

[2] Department of Information Technology, Eastern Academy of Science and Technology, Bhubaneswar, Odisha, India

[3] Department of Computer Science and Engineering ,Hi-Tech Institute of Technology, Bhubaneswar, Odisha, India

Abstract

An ad-hoc network is a group of mobile wireless nodes that cooperatively form a network among themselves without any fixed infrastructure. Each node in ad-hoc network forward packets for other nodes to allow nodes not within direct wireless transmission range to communicate. There has been considerable research on conserving power in the routing protocol. Although most of these researches focused on controlling the transmission power of the sender network interface. Increasing power consumption and packet storming within ad-hoc network is becoming a core issue for these low power mobile devices. This work focuses on an approach for energy conservation as well as reducing packet storming within the routing protocol of the ad-hoc network. A wireless network interface in sleep mode consumes less power than idle mode. In this work we propose an improvement on DSDV protocol to allow sleep mode to take part in communication of the ad-hoc network.

Keywords: Ad-hoc network, DSDV, Sleep mode

1. Introduction

An ad-hoc network is a group of mobile wireless nodes that cooperatively form a network among themselves without any fixed infrastructure. Each node in ad-hoc network forward packets for other nodes to allow nodes

not within direct wireless transmission range to communicate. Energy is a limiting factor in the successful deployment of ad hoc networks since nodes are expected to have little potential for recharging their batteries. In this chapter, we investigate the energy costs of wireless communication and discuss the mechanisms used to reduce these costs for communication in ad hoc networks. We then focus to reduce energy consumption during both active communication and idle periods in communication. There has been considerable research on conserving power in the routing protocol. Although most of these researches focused on controlling the transmission power of the sender network interface.

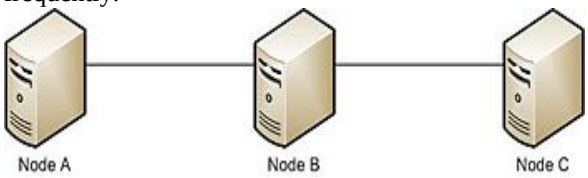
We address different problems in DSDV[1] protocols. In DSDV protocol nodes consumes more power, we use the concept of sleeping mode to reduce power consumption in this thesis.

Rest of the paper organized as follows. Section 2 describe the DSDV protocol , Section 3 describes the literature survey, different problems of DSDV is outlined in section 4, proposed protocol is given in Section 5 and section 6 presents conclusion and future work.

2. Destination-Sequenced Distance-Vector Routing (DSDV)

It is a table-driven routing scheme for ad hoc mobile networks based on the Bellman-Ford algorithm. It was developed by C. Perkinsand P.Bhagwat in 1994. The main contribution of the algorithm was to solve the routing loop problem. Each entry in the routing table contains a sequence number, the sequence numbers are generally even if a link is present; else, an odd number is used. The number is generated by the destination, and the emitter needs to send out the next update with this number. Routing information is distributed between nodes by sending *full dumps* infrequently and smaller incremental updates more

frequently.



For example the routing table of Node A in this network is

Destination	Next Hop	Number of Hops	Sequence Number	Install Time
A	A	0	A46	001000
B	B	1	B36	001200
C	B	2	C28	001500

Naturally the table contains description of all possible paths reachable by node A, along with the next hop, number of hops and sequence number.

3. Literature Survey

DSDV protocol[1] is silent about energy conservation at nodes that is every node in the network should be active all the time in the communication process even if some of them are not currently taking part in the data forwarding process. So that unnecessarily energy is wasted at the nodes.

Another problem in DSDV[1] protocol is packet storming that is even if there is no data communication taking place still control packets are transmitted among nodes consuming much of the bandwidth.

Increasing power consumption and packet storming within ad-hoc network is becoming a core issue for these low power mobile devices. This work focuses on an approach for energy conservation as well as reducing packet storming within the routing protocol of the ad-hoc network. A wireless network interface in sleep mode consumes less power than idle mode. In this work we propose an improvement on DSDV protocol to allow sleep mode to take part in communication of the ad-hoc network.

Each node in ad-hoc network forward packets for other nodes to allow nodes not within direct wireless transmission range to communicate. There has been considerable research on conserving power in the routing protocol. Although most of these researches focused on controlling the transmission power of the sender network interface. The network interface hardware at receiver node can operate in any of four different modes [3].

1. Transmit mode
2. Receive mode
3. Idle mode
4. Sleep mode

Transmit mode- a node is in transmitting mode when it goes to transmit packet.

Receive mode- When a node receive a packet.

Idle mode- When a node neither sends nor receive packet. This mode consumes power because the node has to listen network continuously in order to detect a packet that it should receive, so that node can then switch to receive mode.

Sleep mode- This mode has very low power consumption. The network interface at a node in sleep mode can neither transmit nor receive packet. It must be wake up to idle mode first by explicit information from the node.

Feeny[2] shows the specification and actual measured current drawn by one popular wireless network interface card in the four possible modes. Receive and idle mode require similar power, and transmit mode requires slightly greater power. Sleep mode requires more than an order of magnitude less power than idle mode. These measurements show that the network interface consumes similar energy, whether it is just listening or actually receiving data. Hence, intelligently switching to sleep mode whenever possible will generally create significant energy savings.

Our proposed protocol will solve above problem.

4 Problems of DSDV routing protocol-

- No sleeping nodes are used
- Overhead: most routing information never used
- It only considers hop count as metric but is not considering efficiency (processing speed) of nodes.
- It is also not considering the status (free/busy) of internal nodes.
- It is also silent about the convergence.

The protocol is unable to detect significant change in the network

5 Proposed protocol

Every node in network can interleave between sleep mode and idle mode. Sleeping condition of a node is the condition that every node in the network knows that the node is in sleep mode but that node will interleave between sleep mode and idle mode, during that sleeping condition without revealing to the network. A node can go to sleep mode when it will only receive control packet for some fixed amount of time. The time may not same for each node in the network that is every node will take a random amount of time.

When a node ready to go to sleep node it will transmit a control message indicating its address. When all other nodes receive that message will update their routing table by setting a flag for that node. After a node goingto sleep mode it will periodically wake up to idle mode but it will not revel this information to the network.

When a node is in sleeping condition and receives a sleep mode message of another node it will just update the table for that node but will not wake up.

When a node gets a request to wake up message (RW) then it will reveal that it is wake up by sending a wake up message containing its routing table information to its neighbors. It will remain in wake up state during data packet forwarding or receiving.

Sending and receiving

When a node is in sleeping condition and wants to transmit data to another node which is in wake up state then first it will wake up and broadcast wake up message along with current routing table information. When its neighbors get wake up message they will also wake up and also update its table and then according to current table information sender will send data packet.

When a node wants to send data to another node that is in sleeping condition then it will first broadcast RW message by Flooding. When any sleeping node receive that RW message will wake up and communicate as usually.

6 Conclusions and Future work

In this work we discussed on the problem of energy consumption and packet storming in DSDV and used the concept of sleeping mode, idle mode, transmit mode and receive mode to reduce energy consumption and packet storming unnecessarily when there is no data communication takes place. It also conserve energy at the nodes. These proposed improvements can be simulated for performance analysis by using ns-2 simulator.

References

[1] DSDV (Highly Dynamic Destination-Sequenced Distance Vector routing protocol) - C. E. PERKINS, P. BHAGWAT Highly Dynamic Destination- Sequenced Distance Vector (DSDV) for Mobile Computers Proc. of the SIGCOMM 1994 Conference on Communications Architectures, Protocols and Applications, Aug 1994, pp 234-244.

[2] L. Feeney and M. Nilsson. Investigating the Energy Consumption of a Wireless Network Interface in an Ad Hoc Networking Environment. In Proceedings of INFOCOM 2001, volume 3, pages 1548–1557, Anchorage, Alaska, Apr. 2001.

[3] S. PalChaudhuri. Power Mode Scheduling for Ad Hoc Network Routing. Masters Thesis, Computer Science, Rice University, May 2002.

Nayan Ranjan Paul received his MTech. Degree in Computer Science and Engineering from IIIT-BH.He is currently an Assistant Professor in Department of Computer Science and Engineering of KMBB College of Engineering and Technology, Khurda .His research interest include Ad-Hoc network, Cryptology and Network security.

Laxminath Tripathyl received his MTech. Degree in Computer Science and Engineering from IIIT-BH. He is currently an Assistant Professor in Department of Information Technology of Eastern Academy of Science and Technology, Bhubaneswar .His research interest include Ad-Hoc network, Cryptology and Network security.

Pradipta Kumar Mishra received his MTech. Degree in Computer Science and Engineering from IIIT-BH. He is currently an Assistant Professor in Department of Computer Science and Engineering of Hi-tech Institute of Technology, Bhubaneswar .His research interest include Ad-Hoc network, Wireless sensor network and Network security.

Classification of Load Balancing Conditions for parallel and distributed systems

Zubair Khan[1] Ravendra Singh[2] Jahangir Alam[3] Shailesh Saxena[4]

[1,4]Department Of Computer Science and engineering Invertis University Bareilly India

[2]Department of CS&IT MJP Rohilkhand University Bareilly, India

[3]Woman Polytechnic Department of CSE AMU Aligargh India

Abstract

Although intensive work has been done in the area of load balancing, the measure of success of load balancing is the net execution time achieved by applying the load balancing algorithms. This paper deals with the problem of load balancing conditions of parallel and distributed applications. Parallel and distributed computers have multiple-CPU architecture, and in parallel system they have shared memory. While in distributed system each processing element has its own private memory and connected through networks. Parallel and distributed systems communicate to each other by Message-passing mechanism. Based on the study of recent work in the area, we propose a general classification for describing and classifying the growing number of different load balancing conditions. This gives an overview of different algorithms, helping designers to compare and choose the most suitable strategy for a given application .To illustrate the applicability of the classification, different well-known load balancing algorithms are described and classified according to it. Also, the paper discusses the use of the classification to construct the most suitable load balancing algorithms for different parallel algorithmic paradigms.

Keywords: *Load Balancing, Load Matching, Under load, Over load, processor communication, Network(Topology)*

1. Introduction

Load balancing is one of the central problems which have to be solved to achieve a high performance from a parallel computer. For parallel applications load balancing attempts to distribute the computation load across multiple processors or machines as evenly as possible to improve performance. Since load imbalance leads directly to processor idle times, high efficiency can only be achieved if the computational load is evenly balanced among the processors. Generally a load balancing scheme consists of three phases-1.Information collection, 2.Decision-Making and 3.Data migration.

1.1 Information collection: During this phase the load balancer gathers the information of workload distribution and the state of computing environment and detects whether there is a load imbalance.

1.2 Decision-Making: This phase focuses on calculating an optimal data distribution.

1.3 Data migration: This phase transfer the excess amount of workload from overloaded processor to under loaded ones.

Three kinds of load balancing schemes have been proposed and reviewed in the literature [2], and they can be distinguished depending on the knowledge about the application behavior. The first one, **static load balancing**, is used when the computational and communication requirements of a problem are known a priori. In this case, the problem is partitioned into tasks and the assignment of the task-processor is performed once before the parallel application initiates its execution.

The second approach, **dynamic load balancing** schemes, is applied in situations where no priori estimations of load distribution are possible. It is only during the actual program execution that it becomes apparent how much work is being assigned to the individual processor. In order to retain efficiency, the imbalance must be detected and an appropriate

dynamic load balancing strategy must be devised. Some dynamic strategies that use local information in a distributed architecture, have been proposed in the literature. These strategies describe rules for migrating tasks on overloaded processors to under loaded processors in the network of a given topology. In this survey dynamic load balancing techniques (also referred as Resource-Sharing, Resource scheduling, job scheduling, task- migration etc.) in large MIMD multiprocessor systems are also studied. Dynamic load balancing strategies have been shown to be the most critical part of an efficient implementation of various algorithms on large distributed computing systems. A lot of dynamic load balancing strategies have been proposed in the last few years. With this large number of algorithms, it becomes harder for designers to compare and select the most suitable strategy. A load-balancing algorithm must deal with different unbalancing factors, according to the application and to the environment in which it is executed. Unbalancing factors may be static, as in the case of processor heterogeneity, or dynamic. Examples of dynamic unbalancing factors include the unknown computational cost of each task, dynamic task creation, task migration, and variation of available computational resources due to external loads.

The third one is **hybrid load balancing** condition when dynamic and static are merge together and perform to take the advantages of both conditions.

2. The Classification

The proposed classification is represented in Fig. 1. In order to define a Load-balancing algorithm completely, the main four sub-strategies (initiation, location, exchange, and selection) have to be defined. The goal of this Classification is to understand load balancing algorithms. This Classification provides a terminology and a framework for describing and classifying different existing load balancing algorithms, facilitating the task of identifying a suitable load balancing strategy.

A detailed discussion of the Classification is presented in the following sections:

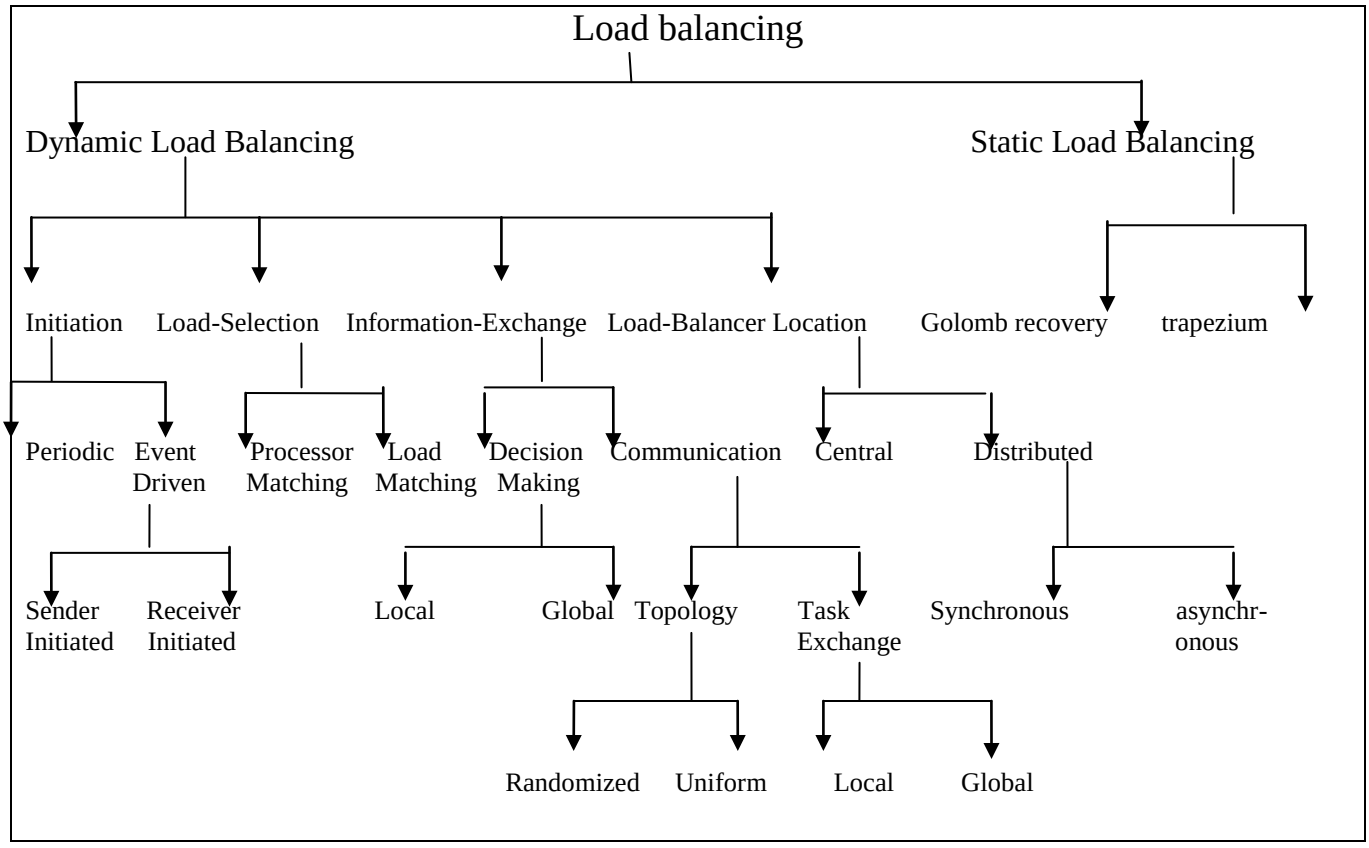


Figure 1 Grouping of Load Balancing Algorithms.

2.1Initiation

The initiation approach specifies the system, which invokes the load balancing behavior. This may be a *episodic* or *event-driven* initiation. Episodic initiation is a timer based initiation in which load information is exchanged every preset time interval. The event-driven is a usually a load dependent policy based on the observation of the local load. Event-driven strategies are more reactive to load imbalances, while episodic policies are easier to implement. However, episodic policies may result in extra overheads when the loads are balanced.

2.2 Load-balancer location

This approach specifies the location at which the algorithm itself is executed. The load balancing algorithm is said to be *vital* if it is executed at a single processor, determining the necessary load transfers and informing the involved processors. Distributed algorithms are further classified as *synchronous* and *asynchronous*. A synchronous load-balancing algorithm must be executed simultaneously at all the participating processors. For asynchronous algorithms, it can be executed at any moment in a given processor, with no dependency on what is being executed at the other processors.

2.3 Information exchange

This specifies the information and load flow through the network. The information used by the dynamic load-balancing algorithm for *decision-making* can be *local* information on the processor or gathered from the surrounding neighborhood. The communication policy specifies the connection ***topology (network)*** of the processors in the system, by sending the messages to its neighboring processing elements. This network doesn't have to represent the actual physical topology of the processors. **A *uniform* network** indicates a fixed set of neighbors to communicate with, while in **a *randomized* network** the processor randomly chooses another processor to exchange information with.

2.4 Load selection

The load selection is very vital part of system in which the processing elements decide from which node to exchange load . Apart from that, it specifies the appropriate load items(tasks) to be exchanged. Local averaging represents one of the common techniques. The overloaded processor sends load-packets to its neighbors until its own load drops to a specific threshold or the average load.

Table 1: Classification of dynamic load balancing algorithms

Algorithm	**Inform-ation exchange**	**Processor Matching**	**Load Matching**	**Communi-cation**	**Applica-tions**	**LB Locations**	**Initia-tion**
SASH	Global	Processor that will produce the fastest turn around	Cost function	Randomized Global	independent tasks	Central (dedicated)	Event driven (Shortest execution time)
Dynamic Load Balancing (DLB)	Local/ Global	According to the load balancer	cost function which uses past to predict future	Randomized Global	Independent loops	Central/ Distributed	Receiver initiated
Automatic Heteregene-ous Supercomputing	Global	According to the load balancer	N/A	Randomized Global	Whole programs	Central	Event driven (User)
Direct Neighbor-hood Repeated (DNR)	Local	Least loaded processor in the neighborhood of the receiver	load is sent. If the load difference exceeds a threshold, a percentage of the	Uniform Local	Independent loops	Central/ Distributed	Receiver initiated
Neighbor	Local	Adjacent processors	Load is distributed over the nodes of the island.	Uniform Local	Independent tasks	Distributed Asynchronous	Receiver initiated

Central Algorithm	Global	Match overloaded with idle	Divides the load among the loaded and idle peers	Randomized Global	Independent tasks	Distributed Asynchronous	Periodic
Pre -Computed Sliding	Local	Adjacent processors	An extra step, which calculates the required total number of transfers required, is done before transfer	Uniform Local	Independent tasks	Distributed Asynchronous	Periodic
Rendez-Vous	Global	Matches most load	Divides the most loaded with least loaded	Randomized Global	Independent tasks	Distributed Asynchronous	Periodic
Random	Local	Random	Each new task is redistributed randomly	Randomized Global	Independent tasks	Distributed Asynchronous	Periodic
Rake	Local	Adjacent processors	Load above the average workload is transferred to the adjacent processor	Uniform Local	Independent Tasks	Distributed Synchronous	Periodic
Tilling (DN)	Local	Balances processors within same window	The load is distributed among processors in the window.	Uniform Local	Independent tasks	Distributed Synchronous	Periodic
X-Tilling	Local	Balances processors connected in the hypercube	The load is distributed among processors in the hypercube	Uniform Global	Independent tasks	Distributed Synchronous	Periodic

3. Categorization of Different Load Balancing Algorithms

In this paper we will illustrate how the proposed Classification is capable of classifying diverse load-balancing algorithms. A number of Dynamic load balancing algorithms are existing for different systems; a small description is presented for each algorithm, followed by a detailed classification.

3.I. Decision and Migration based algorithms[25],[26],[8]

These algorithms are classified as fallows.

3.1.1. Local Decision and Local Migration (LDLM): In this strategy a processor distributes some of his load to its neighbors after each fixed time interval. This is a LDLM because the decision to migrate a load unit is done purely local. The receiver of a load is also a direct neighbor. The balancing is initiated by a processing unit which sends a load unit. We implemented this strategy after x iterations of the simulator y load units are sent to random neighbors.

3. I.2. Direct neighborhood (d-N) : if the local load increased by more than Up percent or decrease by more than Down percent, actual value is broadcasted to direct neighbors. if the load of a

processing element exceeds that of its least neighbor load by more than d percent, then it sends one unit to that neighbor.

3. I.3.Local Decision and Global Migration (LDGM): in this strategy the load units are migrated over the whole network to a randomly chosen processor.

3. I.4.Global Decision and Local Migration (GDLM) : The Gradient Model method discussed above in section 2,was introduces by Lin & Keller [10].it belongs to the group of GDLMr-strategies ,because decisions are based on Gradient information. Gradients are vectors consisting of load respectively, distance information of all processing elements. Which means that each processing element wants to achieve well approximated global state information on network?

3. I.5.Global decision and Global Migration (GDGM) : This method is classified as fallows

a) Bidding algorithm: it is also a state controlled algorithm. The number of processing elements able to take load from a processor in state H depends on the distance between these processors.

b) Drafting Algorithm: in this a processor can be one of the three states L(low),n(normal),H(high) which represent actual load situation. Each processor maintains a load table which contains the most recent information of the so called"candidate processors". A candidate processors is a processor from which load may be received.

Since the workload of system changes dynamically, X-gradient surface can only be approximated. This is done by the protocol that is used in original gradient model For this we recursively define a pressure function

p: V-> {0,……D(G)+1 And the suction function S: V ->{0,……D(G)+1}

3.2 The Random Algorithm [16]

Each time an element (task) is created on a processor, it is sent on a randomly selected node anywhere in the system. For each node, the expectation to receive a part of the load is the same regardless of its place in the system.

3.3 The Tiling (Direct Neighborhood DN) Algorithm [13]

It divides the system in small and disjointed sub-domains of processors called windows. A perfect load balancing is realized in each window using regular communications. In order to propagate the work over the entire system, the window is shifted (slightly moved so that they overlap only a part of the old domain) for the next balancing phase.

3.4 The X-Tiling Algorithm [13]

Similar to Tiling algorithm but extra links are added to the current topology of the processor to form a hypercube topology.

3.5 The Rake Algorithm [13]

It uses only regular communications with processors in the neighborhood set. Firstly, the average load is processed and broadcasted. In the first transfer phase, during p iterations, each processor sends to its right neighbor the data over the average workload It uses only regular communications with processors in the neighborhood set. Firstly, the average load is processed . In the second transfer phase, during the extra workloads, each processor sends to its right the work over the average workload + 1.

3.6 The Pre-Computed Sliding Algorithm [13]

It is an improvement of the Rake algorithm. Instead of transferring data over the average workload of the system, it computes the minimal number of data exchanges needed to balance the load of the system. Unlike the Rake algorithm, it may send data in two directions.

3.7 The Average Neighbor Algorithm [17], [20]

The architecture is made of islands. An island is made of a center processor and all the processors in its neighborhood. It works on the load balancing every node in the island. The partial overlapping allows the load to propagate.

3.8 The Direct Neighborhood Repeated Algorithm [21]

Once a sender-receiver couple is established, the migrating load can move from the sender to the receiver. In its turn, the receiver can have an even less neighborhood. The receiver is allowed to directly forward the migrating load to the less loaded nodes. Load migration stops when there are no more useful transfers.

3.9 The Central Algorithm [11], [28]

Firstly, the average workload is computed and broadcasted to every processor in the system. Then, the processors are classified into 3 classes: idle, overloaded, and the others. The algorithm tries to match each overloaded node with an idle peer.

4. Dynamic Load Balancing (DLB) [18]

Synchronization is triggered by the first processor that finishes its portion of the work. This processor then sends an interrupt to all the other active slaves, who then send their performance profiles to the load balancer. Once the load balancer has all the profile information, it calculates a new distribution. If the amount of work to be moved is below a threshold, then work is not moved else a profitability analysis routine is performed. This makes a trade-off between the benefits of moving work to balance load.

According to the first run, the application adjusts itself for one of the load balancing strategies: global-centralized (GCDLB), global-distributed (GDDLB), and local-centralized (LCDLB) and, local-distributed (LDDLB). The compilation phase is used to collect information about the system and generate cost functions, and prepare the suitable libraries to be used after the first run.

4.1. Automatic Heterogeneous Supercomputing (AHS) [19]

Uses a quasi-dynamic scheduling strategy for minimizing the response time observed by a user when submitting an application program for execution. This system maintains an information file for each program that contains an estimate of the execution time of the program on each of the available machines. When a program is invoked by a user, AHS examines the load on each of the networked machines and executes the program on the machine that it estimates will produce the fastest turn-around time.

4.2. Self-Adjusting Scheduling for Heterogeneous Systems (SASH)[20]

It utilizes a maximally overlapped scheduling and execution paradigm to schedule a set of independent tasks on to a set of heterogeneous processors. Overlapped scheduling and execution in SASH is accomplished by dedicating a processor to execute the scheduling algorithm. SASH performs repeated scheduling phases in which it generates partial schedules. At the end of each scheduling phase, the scheduling processor places the tasks scheduled in that phase on to the working processors' local queues.

The SASH algorithm is a variation of the family of branch-and-bound algorithms. It searches through a space of all possible partial and complete schedules. The cost function used to estimate the total execution time produced by a given partial schedule consists of cost of executing a task on a processor and the additional communication delay required to transfer any data values needed by this task to the processor. As observed from Table 1, that any dynamic load-balancing algorithm may be classified according to the Classification. Accordingly, this makes it simpler for designers to compare and select the proper algorithm for their application to be executed on a certain computing environment. The next section will illustrate how to select the most suitable dynamic load-balancing category for the different parallel programming paradigms.

5. Dynamic Load-Balancing Conditions

A number of parallel algorithmic paradigms have emerged for parallel computing like:

1. Gradient Model ,
2. Sender Initiated Diffusion (SID),
3. Receiver Initiated Diffusion (RID) ,
4. Hierarchical Balancing Method (HBM) ,
5. The Dimension Exchange Method (DEM)
6. Phase Parallel,
7. Divide and Conquer,
8. Pipeline,
9. Process Farm ,

Each paradigm has its own characteristics. A brief description is given for each paradigm and a suitable load-balancing algorithm is suggested for each based on the Classification. It should be noted that scalability and low communication cost are the main considerations affecting the choice of the following suggested strategies.

Gradient Model [9]

The gradient model is a demand driven approach [8]. The basic concept is that underloaded processors inform other processors in the system of their state, and overloaded processors respond by sending a portion of their load to the nearest lightly loaded processor in the system.

Sender Initiated Diffusion (SID) [15]

The SID strategy is a, local, near-neighbor diffusion approach which employs overlapping balancing domains to achieve global balancing. A similar strategy, called Neighborhood Averaging, is proposed in [12]. The scheme is purely distributed and asynchronous. for an N processor system with a total system load L unevenly distributed across the system, a diffusion approach, such as the SID strategy, will eventually cause each processor's load to converge to L/N.

Receiver Initiated Diffusion (RID) [16]

The RID strategy can be thought of as the converse of the SID strategy in that it is a receiver initiated approach as opposed to a sender initiated approach. However, besides the fact that in the RID strategy underloaded processors request load from overloaded neighbors, certain subtle differences exist between the strategies. First, the balancing process is initiated by any processor whose load drops below a prespecified threshold (Llow). Second, upon receipt of a load request, a processor will fulfill the request only up to an amount equal to half of its current load. When a processor's load drops below the prespecified threshold Llow, the profitability of load balancing is determined by first computing the average load in the domain, Lpavg [(8)]. If a processor's load is below the average load by more than a prespecified amount, L threshold &, it proceeds to implement the third phase of the load balancing process.

Hierarchical Balancing Method (HBM) [15],[16]

The HBM strategy organizes the multicomputer system into a hierarchy of balancing domains, thereby decentralizing the balancing process. Specific processors are designated to control the balancing operations at different levels of the hierarchy.

The Dimension Exchange Method (DEM) [17], [19]

The DEM strategy [17], [19] is similar to the HBM scheme in that small domains are balanced first and these then combine to form larger domains until ultimately the entire system is balanced. This differs from the HBM scheme in that it is a synchronized approach, designed for a hypercube system but may be applied to other topologies with some modifications. In the case of an N processor hypercube configuration, balancing is performed iteratively in each of the log N dimensions.

All processor pairs in the first dimension, those processors whose addresses differ in only the least significant bit, balance the load between themselves

Phase parallel [12]: The parallel program consists of a number of super steps, and each super step has two phases. A computational phase, in which, multiple processes, each perform an independent computation C. In the subsequent interaction phase, the processors perform one or more synchronous interaction operations, such as a barrier or blocking communication.

This paradigm is also known as the loosely synchronous paradigm and the a general paradigm. It facilitates debugging and performance analysis, but interaction is not overlapped with computation, and it is difficult to maintain balanced workloads among the processors. Suggested load balancing algorithm:

***Initiation*:** event driven, with every synchronization step.

***Load balancer location*:** central or distributed synchronous.

***Information exchange*:**

 ***Decision-making*:** would be global to observe the different loads.

 ***Communication*:** Global Randomized, as this is the nature of the paradigm.

***Load selection*:** processor matching and selection is application dependent.

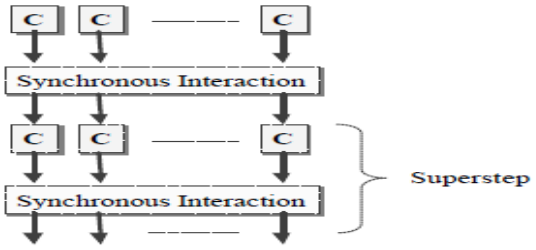


Fig. 2. Phase Parallel

Divide and conquer [12]: In this a problem id divided into small parts and then we start to conquer it. In dynamic load balancing a parent process divides its workload into several smaller pieces and assigns them to a number of child processes. The child processes then compute their workload in parallel and the results are merged by the parent. This paradigm is difficult to maintain balanced workloads among the processors.

The parent is the one, which distributes the load among its children, and accordingly it should be the one to balance the load between them. Suggested load balancing algorithm:

 ***Initiation*:** event driven, sender/receiver (child) initiated.

***Load balancer location*:** distributed asynchronous. Each parent is responsible to load balance its children.

***Information exchange*:**

 ***Decision-making*:** would be local based on the children only.

 ***Communication*:** Local Uniform, as the children can only communicate to parents and their children.

 ***Load selection*:** load is exchanged among the children and selection of load is flexible according to the application.

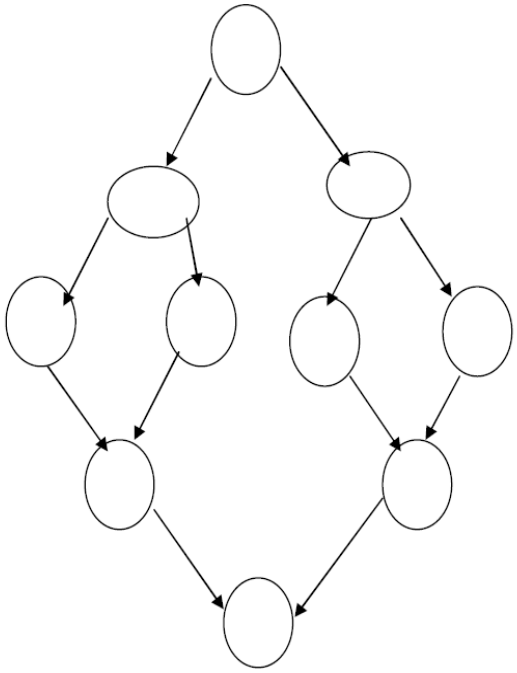


Fig. 3. Divide and conquer

Pipeline [12]: In this the output of one stage works as an input for next stage, hence a pipe is seem to be created called Virtual pipe. A number of processors

form a virtual pipe. A continuous data stream is fed into the pipeline and the processes execute at different pipeline stages simultaneously in an overlapped fashion. The pipeline paradigm is the basis for SPMD. Each processor runs the same code with different data. Interface data is exchanged between adjacent processors. Suggested load balancing algorithm:

• ***Initiation*:** event driven, sender/receiver initiated.

• ***Load balancer location*:** Central/distributed. Depends on the number of the processors involved in the synchronization. For scalability reasons, a distributed asynchronous strategy is suggested.

☐***Information exchange*:**

☐ ***Decision-making*:** Global/Local. Local is recommended for scalability.

☐***Communication***: Local Uniform, as the processors only communicate with their neighbors.

☐ ***Load selection:*** load is exchanged among the neighbors and selection of load is application dependent.

Fig.4. Pipe Line

Process Farm: This is a very common paradigm shown in fig.5. A master process executes the sequential part of the parallel part of the program and spawns a number of slave processes to execute the parallel workload. When a slave finishes its workload, it informs the master which assigns a new workload to the slave. This is a very simple paradigm, but the master could become the bottleneck.

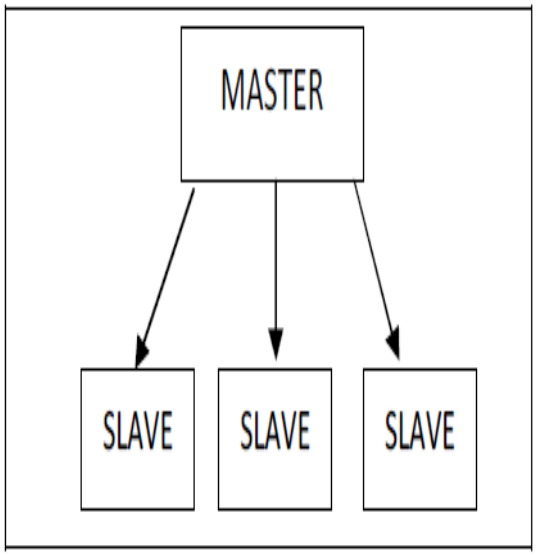


Fig.5. Process farm

6. Conclusion and Future Work

In the paper it has been illustrated how to suggest new algorithms for different application paradigms. The Classification is considered helpful for designers to compare different load-balancing algorithms and design new ones tailored for their needs.

In the future, we intend to develop a framework for applications with load balancing that utilizes this Classification and helps the designer tailor his own algorithm. The framework would generate the required libraries needed and the corresponding coding that will facilitate the development of parallel applications.

References

[1]M. Willeheek-LeMair and A. P. Reeves, "Region growing on a hyper- cube multiprocessor," in Proc. 3rd Conf Hypercube Concurrent Comput. and Appl., 1988, pp. 1033- 1042.

[2] M. WiIIebeek-LeMair and A. P. Reeves, "A general dynamic load balancing model for parallel computers," Tech: Rep. EE-CEG-89-I) Cornell School of Electrical Engineering, 1989.

[3]T. L. Casavant and .I. G. Kuhl, "A taxonomy of scheduling in general purpose distributed computing systems," IEEE' Trans. Software Eng., vol. 14, no. 2, pp. 141-154, Feb. 1988.

[4]Y.-T. Wang and R.I. T. Morris, "Load sharing in distributed systems," IEEE Trans. Cornput., vol. C-34, pp.

204-211, Mar. 1985. M. J. Berger and S. H. Bokhari, "A partitioning strategy for nonuniform problems on multiprocessors," IEEE Trans. Cornput., vol. C-36, pp. 570-580, May 1987.
[5] G. C. Fox, "A review of automatic load balancing and decomposition methods for the hypercube," California Institute of Technology, C3P- 385, Nov. 1986.
[6] K. Ramamritham, I. A. Stankovic, and W. Zhao, "Distributed scheduling of tasks with deadlines and resource requirements," IEEE Trans. Comput. pp. 1110-1123, Aug. 1989.
[7] K. M. Baumgartner, R. M. Kling, and B. W. Wah, "Implementation of GAMMON: An efficient load balancing strategy for a local computer system," in Proc. 1989 Int. Conf Parallel Processing, vol. 2, Aug. 1989, pp. 77-80.
[8] F. C. H. Lin and R. M. Keller, "The gradient model load balancing method," IEEE Tran. Software Eng., vol. 13, no. 1, pp. 32-38, Jan. 1987.
[9]. Casavant, T. L. and Kuhl, J. G.: A Taxonomy of Scheduling in General-Purpose Distributed Computing Systems. IEEE Trans. on Soft. Eng. 14 (1998) 141-154
[10]. Plastino, A., Ribeiro, C. C. and Rodriguez, N. R.: Load Balancing Algorithms for SPMD Applications. Submitted for publication (2001)
[11]. Hillis, W.D.: The Connection Machine. MIT press, 1985.
[12]. Fonlupt, C., Marquet, P. and Dekeyser, J.: Data-parallel load-balancing strategies. Parallel Computing 24 (1998) 1665-1684.
[13]. Dekeyser, J. L., Fonlupt, C. and Marquet, P.: Analysis of Synchronous Dynamic Load Balancing algorithms", Parallel Computing: State-of-the Art Perspective (ParCo'95), volume 11 of Advances in Parallel Computing, pages 455--462, Gent, Belgium (September 1995)
[14] V. A. Saletore, "A distrubuted and adaptive dynamic load balancing scheme for parallel processing of medium-grain tasks," in Proc. Fifth Distributed kemorycomput. ?onf, Apr. 1990, pp. 995-999.
[15] K. G. Shin and Y.-C. Chang, "Load sharing in distributed real time systems with state-change broadcasts," IEEE Trans. Cornput., pp . 1124-1142, Aug. 1989 .V. A. Saletore, "A distr
[16]. Subramanian, R. and Scherson, I.: An Analysis of Diffusive Load Balancing. Proceedings of Sixt h Annual ACM Symposium on Parallel Algorithms and Architectures, (June 1994) 220-225
[17]. Saletore, V. A.: A distributive and adaptive dynamic load balancing scheme for parallel processing of medium-grain tasks. Proceedings of the 5th Distributed Memory Conference (April 1990) 995-999
[18] G. Cybenko, "Dynamic load balancing for distributed memory multiprocessors," J. Parallel and Distributed Comput., vol. 7:279-301, October, 1989.
[19]D. P. Bertsekas and J. N. Tsitsiklis, Parallel and Distributed Computation: Numerical Methods. Englewood Cliffs, NJ: Prentice-Hall,
[20]. Willebeek- LeMair, M.H. and Reeves, A.P.: Strategies for dynamic load balancing on highly parallel computers. IEEE Trans. on parallel and distributed systems, vol. 4, No. 9 (Sept. 1993)
[21]. Corradi, A., Leonardi, L. and Zambonelli, F.: Diffusive load-balancing policies for dynamic applications. IEEE Concurrency Parallel, Distributed and Mobile Computing (January-March 1999) 22-31
[22]. Zaki, M. J., Li, W. and Parthasarathy, S.: Customized dynamic load balancing for a network of workstations. Proceedings of the 5th IEEE Int. Symp., HPDC (1996) 282-291
[23]. Dietz, H. G., Cohen, W.E. and Grant, B. K.: Would You Run it Here... or There? (AHS: Automatic Heterogeneous Supercomputing. International Conference on Parallel Processing, Volume II: Software (1993) 217-221
[24]. Hamidzadeh, B., Lilja, D. J. and Atif, Y. : Dynamic scheduling techniques for heterogeneous computing systems. Concurrency: Practice and Experience, vol. 7 (1995) 633-652.
[25]S.Zhou, A Trace Driven Study Of Load balancing , IEEE Trans.On Software Engineering,Vol.14,no.9,1988,pp,1327-1341.
[26]F. Ramme, Lastausgleichsverfahren in Vertileilten Systemen, Master Thesis,University of Paderborn.1990.
[27]] F. C. H. Lin and R. M. Keller, "The gradient model load balancing method," IEEE Tran. Software Engineering 13, 1987,pp.32-38
[28] Powley, C., Ferguson, C. and Korf, R. E.: Depth-First Heuristic Search on a SIMD Machine. Artificial Intelligence, vol. 60 (1993) 199-242

A study for Issues of the Location Management In Mobile Networks

Sami M. Halawani[1] **Dr. Ab Rahman bin Ahmad[2]** **M. Z.ubair Khan[3]**

[1,2]**Faculty of Computing and Information Technology, Rabigh Campus**
King Abdulaziz University Saudi Arabia

[3]**Department Of Computer Science and engineering Invertis University Bareilly India**

Abstract

This paper presents a hierarchical model for location management of mobile agents in global networks, Location Management in PCS systems, important issues in Location Management, Performance of Location Management, What will happen in future.
Location management is a key issue in personal communication service networks to guarantee the mobile terminals to continuously receive services when moving from one place to another. In this paper we study about the location management and its Issues in mobility networks. We also study about different system model and its components of location management. Mobility management is a necessity in highly dynamic and large-scale mobile agents' network, especially in a multi-region environment in order to control and communicate with agents after launching. Existing mechanisms for locating mobile agents are not efficient as these do not consider the effect of location updates on migration time and produce network overload. A location management protocol consists of location updates, searches and search updates. An update occurs when a mobile agent changes location. A search occurs when a mobile agent needs to be located.

Keywords: *Location Management, Personal communication services (PCSs), Visitor Location Register, mobile network.*

1. Introduction

Location management Location Management is the process to determine the current location of a mobile terminal [1]. A Mobile agent (MA) is a software process, which can move autonomously from one physical network location to another. The agent performs its job wherever and whenever it is found appropriate and is not restricted to be co-located with its client. Thus, there is an inherent sense of autonomy in the mobility and execution of the agent. Agents can be seen as automated errand boys who work for users. MA research evolved over the past years from the creation of many different monolithic mobile agent systems (MASs), often with similar characteristics and built by research groups spread all over the world, for optimization and better understanding of specific agent issues.

MOBILE computing and wireless communications are perhaps the fastest growing areas in recent years. Not only have we seen a variety of emerging wireless networking technologies (such as GSM, GPRS, WCDMA, cdma2000,IEEE 802.11 WLAN, and Bluetooth), but also are there numerous portable computing devices widely available (such as laptops, tablet PCs, PDAs, and handsets). The marriage of these two fields has made ubiquitous computing and communications possible [2]. Most MASs has the following common features:
(a) MAs are launched to complete some tasks. They may roam around the network automatically from host to host. They normally end at the launching point with their results or submit results at the last host in the itinerary.
(b) The agent management centre keeps locating these roaming agents so that it can set up

communication with them at their current locations whenever necessary. The second feature given above is actually the function of MAS mobility management. The basic operations associated with mobility management are:

1. A roaming agent updates its location frequently to the central management server (e.g. a directory server).
2. The agent management server refreshes the current location record of the agent in its location database.
3. When there is a request asking for the location of the agent, the management server searches the database and replies with the current location of the MA. Beside these three basic steps, the management server may also process issues such as out-of-date location records. Most existing MASs have provided partial mobility management, by defining different naming and locating mechanisms.

The ability of mobile hosts (MHs) to autonomously move from one part of the network to another part in a mobile computing system, sets it apart from static networks. Unlike static networks, the network configuration and topology keep changing in mobile computing systems [6]. The mobility of some nodes in the network raises interesting issues in the queried. Hence, a location management strategy should address issues (iii) and (iv) so as to ensure

1.1 Components in the location management-

i. **Base Station:-** A tower or antenna transmitting and receiving radio signals over a cell in a wireless network.

ii. **Base Station Controller (BSC):-** An agent performing functions on behalf of a group of base stations. The BSC handles the allocation of radio channels, controls handovers, performs paging and interfaces with the central network and HLR.

iii. **Cell:-**A geographical area serviced by a base station in a wireless network, also used to refer to one or more collocated base stations. Cells are the 'building blocks' of a cellular network, with overlapping cells defining the coverage area of a particular network.

iv. **Global System for Mobile Communication (GSM):-** The dominant standard for second generation mobile phone communication, defining the protocols for communication between mobile devices and network cells [12].

v. **Handoff:-** The process of transferring an in-progress call from one cell or base station to a neighboring cell without interruption.

vi. **Home Location Register (HLR):-** The central database in a cellular network, containing information on all subscribers to a particular carrier. This database also contains a record of each user's location, used to route calls to the correct cell.

vii. **Location Area (LA):-** A group of neighboring cells combined to form a larger *meta-cell*. Devices are free to move within this Location Area without performing a Location Update. Location Areas may be fixed, as in current static schemes, or allocated dynamically on a Location Update.

viii. **Location Management (LM):-** The maintenance of a record of cell locations for devices in a mobile network. The study of Location management of location information of these nodes. Creating a xed location directory of all the nodes a priori is not a solution. The location directory has to be dynamically updated to account for the mobility of the MHs. The design of a location directory whose contents change dynamically raises important issues. Some of them are as follows: (i) When should the location directory be updated? If the updates are done each time an MH's location changes, the directory will always have the latest location information, reducing the time and effort in locating an MH. However, such a policy imposes a heavy burden on the communication network and the location servers, i:e:, nodes that maintain the directory. (ii) Should the location directory be maintained at a centralized site, or should it be distributed? A central location server has problems with regard to robustness and scalability. Hence, a distributed directory server is preferable. This leads us to the next questions. (iii) How should the location information be distributed among the location servers? and (IV) should location information about an MH be replicated across multiple location servers? It is not possible to a priori determine the variations in spatial distribution of MHs in the network, and the frequency with which node location will be updated or
fair distribution of responsibility among all the location servers, and be scalable.

Management aims to reduce the net cost involved in maintaining this information.

ix. **Location Update (LU):-** Performed by a device in a wireless network to inform the network of the cell in which it resides. This Location Update is usually performed only when leaving the Location Area previously assigned to the device.

x. **Paging:-** Under a Location Area scheme, the network does not know the precise location of a device, only its general area. Paging is performed on an incoming call and involves sending a message to all cells in the Location Area to determine which one contains the destination device.

xi. **Spectrum:-** A portion of the electromagnetic spectrum containing a limited frequency range within which a mobile device may communicate. It is vital that multiple signals transmitted on the same frequency do not interfere and hence the allocation of sections of this spectrum is governed by regulatory bodies. A communications provider must purchase a license for a particular frequency band within this spectrum to Broadcast cellular data.

xii. **Subscriber Identity Module (SIM):-** A small *smart card* used in mobile phones operating under the GSM standard. This SIM card contains user identification information, as well providing storage space for phone numbers and associated data.

xiii. Third Generation (3G):- A new wireless communication specification replacing second generation technologies such as GSM. Third generation cellular networks provide for high-speed data access in addition to audio communication, with goals of high-quality multimedia and advanced global roaming[15].

2. SYSTEM MODEL

We assume a cellular communication system that divides the geographical region served by it into smaller regions, called cells. Each cell has a base station, also referred to as the mobile service station (MSS). Figure 1 shows a logical view of a mobile computing system. The mobile service stations are connected to each other by a xed wire network. A mobile service station can be in wireless communication with the mobile hosts in its cell. The location of a mobile host can change with time. It may move from its present cell to a neighboring cell while participating in a communication session or it may stop communicating with all nodes for a period of time and then pop-up in another part of the network. A mobile host can communicate with other units, mobile or static, only through the mobile service station of the cell in which it is present. If a node (static or mobile) wishes to communicate with a mobile host, rst it has to determine the location of the MH (the cell in which the MH is currently residing). This location information is stored at location servers. Depending on the frequency of location updates, this location information may be current, or out-of-date. Once the location of the MH has been determined, the information is routed through the xed wire network to the MSS of the cell in which the MH is present[4][5].

Then the MSS relays the information to the destination MH over a wireless channel. We assume that MSSs act as location servers. Hence, all the MSSs collectively maintain the location directory.

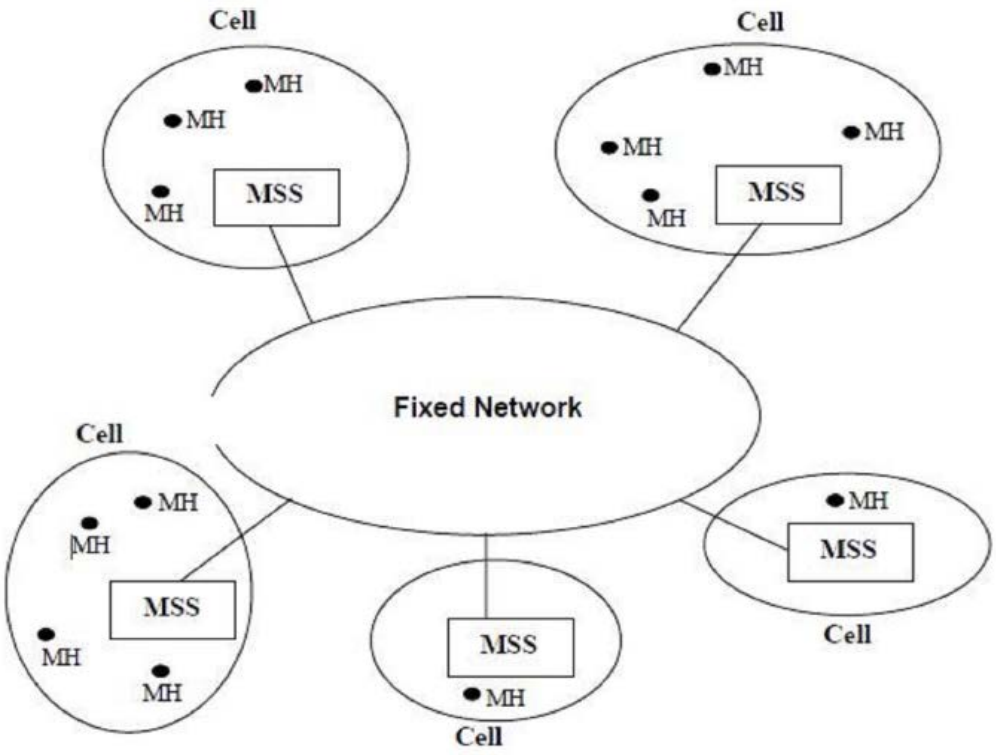


Fig.1 Logical view of mobile computer syste

2.1 Mechanism for location management

The Base Transceiver Station (BTS) of every cell continuously transmits the location area identity on the control channel (BCCH). When the mobile station detects that the broadcast location area identity is different from the one stored in the SIM card, it performs a location update. If the mobile subscriber is unknown to the Mobile Services Switching Center/Visitor Location Register (MSC/VLR) (that is, the broadcast location area belongs to a new MSC/VLR serving area), then the new MSC/VLR must be updated with subscriber information. This subscriber information comes from the Home Location Register (HLR). This location updating procedure is described in the steps below and in Figure2.

i. The mobile station requests a location update to be carried out in the new MSC/VLR. The IMSI is used to identify the mobile station. An International Mobile Equipment Identity (IMEI) check is also performed.

ii. In the new MSC/VLR, an analysis of the IMSI number is carried out. The result of this analysis is a modification of the IMSI to a mobile global title which is used to address the HLR.

iii. The new MSC/VLR requests the subscriber information for the mobile station from the HLR.

iv. The HLR stores the address of the new MSC/VLR.

v. The HLR sends the subscriber data to the new MSC/VLR.

vi. The HLR also orders the old serving MSC/VLR to cancel all information for the subscriber because the mobile subscriber is now served by another MSC/VLR.

vii. When the new MSC/VLR receives the information from the HLR, it sends a location updating confirmation message to the mobile station.

viii. Note: The HLR is not informed if the mobile subscriber moves from one location area to another within the same MSC/VLR serving area.

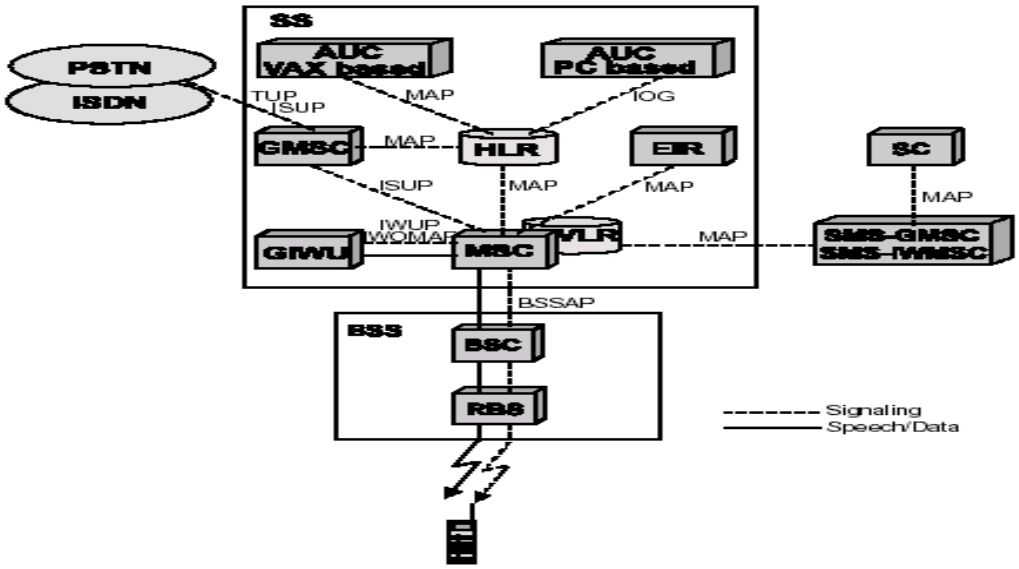


Fig 2- Architechure view of a mobile computing system

2.2 LOCATING USER:

Location management deals with how to keep track of an active mobile station within the cellular network. In this paper there are two basic operations involved in location management is discussed. These are location update and paging. The cellular network performs the paging operation. When an incoming call arrives for a mobile station, the cellular network will page the mobile station in all possible cells to find out the cell in which the mobile station is located so the incoming call can be routed to the corresponding base station. The number of all possible cells to be paged is dependent on how the location update operation is performed. An active mobile station performs the location update operation.

A location update scheme can be classified as either global or local, and given in figure 3. A location update scheme is global if all subscribers update their location at the same set of cells, and a scheme is local if an individual subscriber is allowed to decide when and where to perform the location update. A local scheme is also called individualized or per-user-based. A location update scheme is static if there is a predetermined set of cells at which a mobile station regardless of its mobility must generate location updates. A scheme is dynamic if a mobile station in any cell depending on its mobility can generate a location update. A global scheme is based on aggregate statistics and traffic patterns, and it is usually static too. Location management involves signaling in both the wire line portion and the wireless portion of the cellular network [8]. However, most researches only consider signaling in the wireless portion due to the fact that the radio frequency bandwidth is limited, whereas the bandwidth of the wire line network is always expandable. Location update involves reverse control channels whereas paging involves forward control channels. The total location management cost is the sum of the location update cost and paging cost. There is a tradeoff between the location update cost and the paging cost. If a mobile station updates its location more frequently the network knows the location of the mobile station better. Then the paging cost will be lower when an incoming call arrives for the mobile station. Therefore, both location update and paging costs cannot be minimized at he same time. However, the total cost can be minimized or putting a bound on the other cost can minimize one cost. Locating users who are on the move and often to locations, which are remote from home, is a challenging task [11]. In general, it is unnecessary to track locations of all users all the time. Hence, a database, which stores locations of users, will often be imprecise in terms of the exact user's location. For instance, a user's location may only be updated when the user crosses the border between two different areas or zones as opposed to updates on crossing a small cell. This, in general, will save on the number of location updates that the moving user will have to perform but will put an

additional burden on the search process if the exact location of the user is sought.

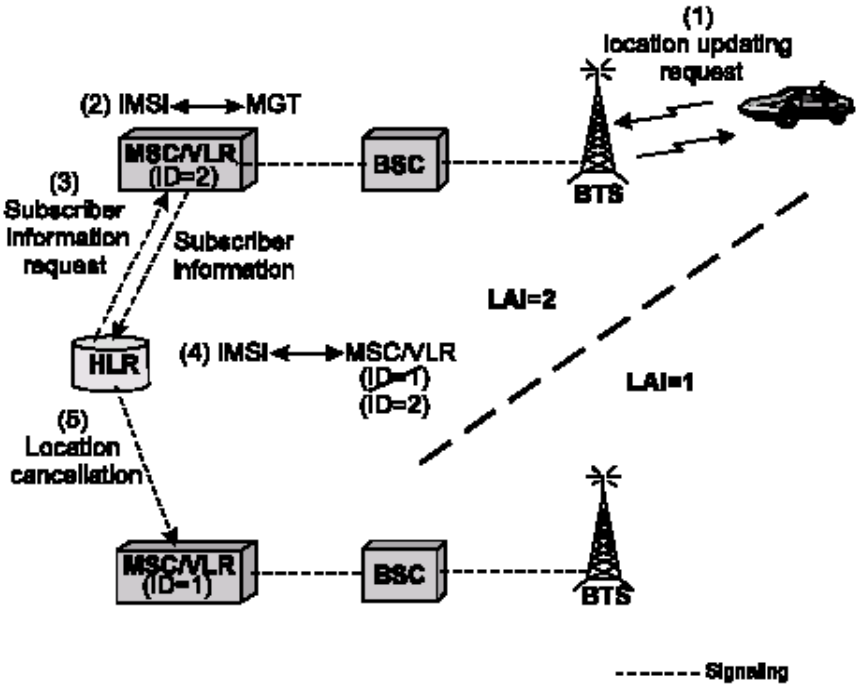


Figure 3- Location updating

2.2.1 Location management in PCS:-

Personal communication services (PCSs) include many wireless access and mobile communication services. Its goal is to provide communications for users at any time and any place [2] [3][10] as given in figure 4. The network architecture of a PCS network is shown in Fig. 1. It consists of several components which we briefly describe them as follows:

i. Mobile stations (MSs) are devices used to send and receive calls.

ii. Base stations (BSs) use radio protocol to communicate with mobile stations and wired protocol to connect with the mobile switching center (MSC). The range of a base station is called a cell. Several BSs are connected to anMSC. The coverage area of those BSs is called a Registration Area (RA).

iii. The MSC supports switching function for wireless communications and serves as an interface between MSs and the public switched telephone network (PSTN). The functions of mobile switching center (MSC) include location registration, call delivery, paging, handoff, etc.

iv. The home location register (HLR) is a centralized database containing user profiles. The user profiles record information such as the types of services subscribed, the quality of service (QoS) requirements, the billing information, and the current locations of the MSs.

v. The visitor location registers (VLRs) are dynamic and distributed databases. A PCS network usually implements hierarchical database with one HLR and several VLRs below the HLR. The VLR stores the information of MSs that are currently in its RA. The user profiles of an MS may be replicated from HLR to its current serving VLR. In general, one or more MSCs can share one VLR.We consider only one MSC is served by each VLR in this paper and the VLR and the MSC are located in the same place . This kind of implementation is more widely adopted in today's PCS systems [7].

vi. The public switched telephone network (PSTN) represents a wired backbone network. Wireless networks are not like wired networks in which every node has a fixed address and location that can be used to find the node. Hence in PCSs,

we cannot use the node's identification number to find its current location. Thus, an efficient location management scheme is required. In PCS systems today, the popular standards are GSM and IS-41. IS-41 is generally used in North America and GSM is common in Europe. The two standards both use hierarchical databases.

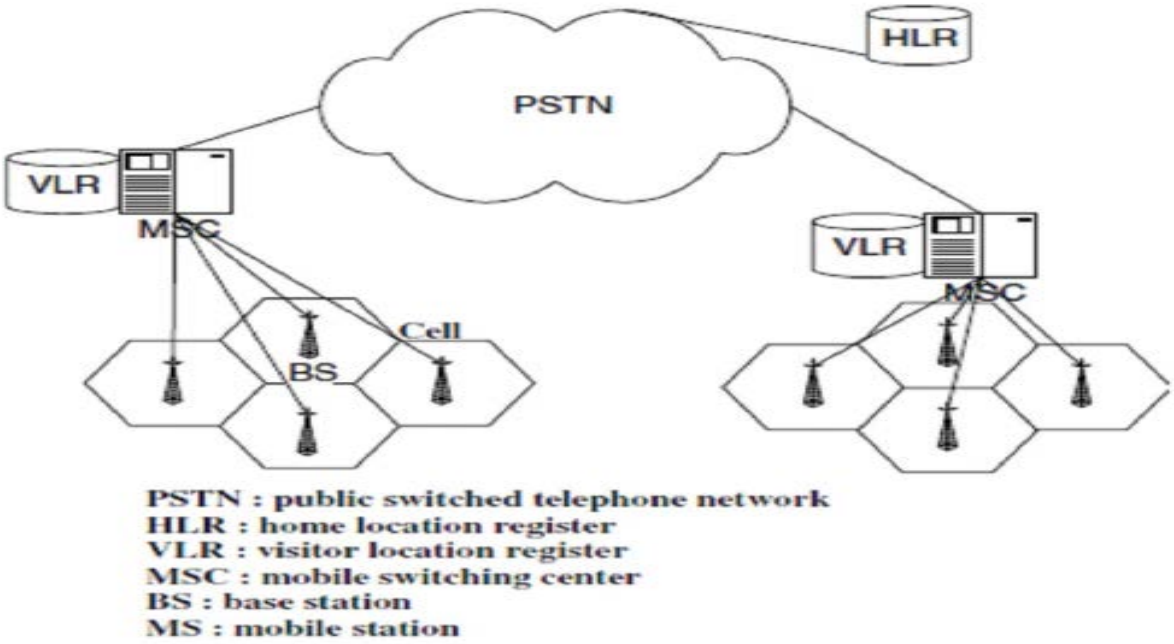


Fig. 4 PCS Network Architecture

3. Issues in Location Management:-

The study shows that Location management has three issues:

3.1 Location registration. When an MS moves to a new MSC, messages are exchanged among the HLR and the new and the old VLRs. Exchanging messages are for recording and updating the new location information of the MS in databases. The process is called location registration. There are several location registration schemes. (A) Geography: A user updates the system only when it moves to a new RA from an old RA. (B) Timer: The user updates the location only periodically with a timer. (C) Stimulus: The user performs the location update only when there is a request. (D) ON/OFF: A location update occurs only between the time that the MS is powered on and the time that the MS is powered down. In this paper, we will emphasize on geography based location registration, which is used in most of second-generation cellular systems.

3.2 Call delivery. The purpose is to find the called MS (callee) from the calling MS (caller) when the caller makes a call. Messages are exchanged among VLRs of the caller and callee and the HLR of the callee. In PCS systems, when the callee is an MS, the caller must query the callee's HLR via its VLR to know the VLR location of the callee. Then it queries the VLR of the callee to learn the current location of the callee. The MSC of the callee may assign a temporary location directory number (TLDN) for the caller and sends the TLDN back to the HLR [8]. The TLDN is forwarded to the MSC of the caller. If the location query is successful, the MSC of the caller establishes a connection to the MSC of the callee.

3.3 Paging. When an MSC receives a call for an MS, the MSC knows whether the MSis in its RA. However, it does not know which cell the MS is in. The MSC will send a paging message to all BSs belonged to the MSC. If the MS is within the cell of a BS, the BS will send an

acknowledgement back to the MSC. The paging is usually not included in counting the costs of a location strategy because every call delivery needs the paging and we can assume the paging costs are the same for all schemes. Hence, paging costs will also not be counted in this paper.

TABLE I
LOCATION MANAGEMENT ISSUES IN DIFFERENT MOBILE NETWORKS.

	identity	location	tracking strategies
GSM/ cdma2000	MSISDN	location area (LA)/ enhanced Cell ID/ <latitude, longitude>	time/movement/distance-based, BS-initiated, centralized servers (HLR/VLR), A-GPS/TDOA/E-OTD centralized servers
IP networks	IP address	subnet	time-and-movement-based, terminal-initiated, centralized servers (home agent)
ad hoc/sensor networks	IP/MAC address	2D/3D coordinates	time/distance-based, host-initiated, distributed servers (virtual home zone, grid, etc.)

Managing location information of mobile nodes is an important issue in mobile computing systems. Location management is one of the fundamental issues in cellular networks. It deals with how to track subscribers on the move and how to update his or her movements. In mobile communication environment, they are going to accommodate more subscribers; the size of the cell must be reduced to make more efficient use of the limited frequency spectrum allocation. This will add to the challenge of some fundamental issues in cellular networks. Location management consists of updating the location of the user, searching the location and performing search-updates. Various strategies can be discussed in this paper for the efficient performance of updating, searching and search-updating strategies throughout the execution.

In a cellular network, a service coverage area is divided into smaller hexagonal areas referred to as cells. A base station serves each cell. The base station is fixed. It is able to communicate with mobile stations such as cellular telephones using its radio transceiver. The base station is connected to the mobile switching centre (MSC), which is, in turn, connected to the public switched telephone network (PSTN). The frequency spectrum allocated to wireless Communication is very limited, so the cellular concept was introduced to reuse the frequency. Each cell is assigned a certain number of channels. To avoid radio interference, the channels assigned to one cell must different from the channels assigned to its neighboring cells. The radio interference between them is tolerable. By reducing the size of the cells, the cellular network is able to increase its capacity, and therefore to serve more subscribers. A mobile station communicates with another station, either mobile or land, via a base station. A mobile station cannot communicate with another mobile station directly. To make a call from a mobile station, the mobile station first needs to make a request using a reverse control channel of the current cell. If the request is granted by the MSC, a pair of voice channels will be assigned for the call. To route a call to a mobile station is more complicated. The network first needs to know the MSC and the cell in which the mobile station is currently located. How to find out the current residing cell of mobile station is an issue of location management. Once the MSC the cell of the mobile station, it can assign a pair of voice channels in that cell for the call. If a call is in progress when the mobile station moves into a neighboring cell, the mobile station needs to get a new pair of voice channels in the neighboring cell from the MSC so that the call can continue. This process is called as ‘handoff ‘or ‘handover ‘. The MSC usually adopts a channel assignment strategy that prioritizes handoff calls over new calls. Providing connection-oriented services to the mobile host requires that the host be always connected to the rest of the network in such a manner that its movements are transparent to the users. This would require efficient location management in order to minimize the time taken for updates and searches, so that there is no loss of connection. The ability of mobile hosts (MHs) to autonomously move from one part of the network to another part in a mobile computing system sets it apart from static networks. Unlike static networks, the network configuration and topology keep changing in mobile computing systems. The mobility of some nodes in the network raises interesting issues in the

management of location information of these nodes.

Location server is maintaining the details about mobile user, it consist separate location directory for each MH. Creating a fixed location directory of all the nodes a priori is not a solution. The location directory has to be dynamically updated to account for the mobility of the MHs.

Another Issue in the location management is Data management as given in figure 5. It is necessary to manage users' location information as efficient as possible, since users move around the network and their current locations should be updated in the databases. Especially when there are many users in a network, the mobile system suffers from the scalability problem. By *scalability*, we mean "the ability of a network to adjust or maintain its performance as the size of the network increases (and the demands made upon it increases), yet the performance of a network tends to degrade as the number of mobile users increases".

To resolve the scalability problem in a mobile computing system, Pitoura et al. proposed a hierarchical system with a tree topology[6]. This system relieves the scalability problem by updating the databases in the system locally. In a hierarchical database system, clustering the

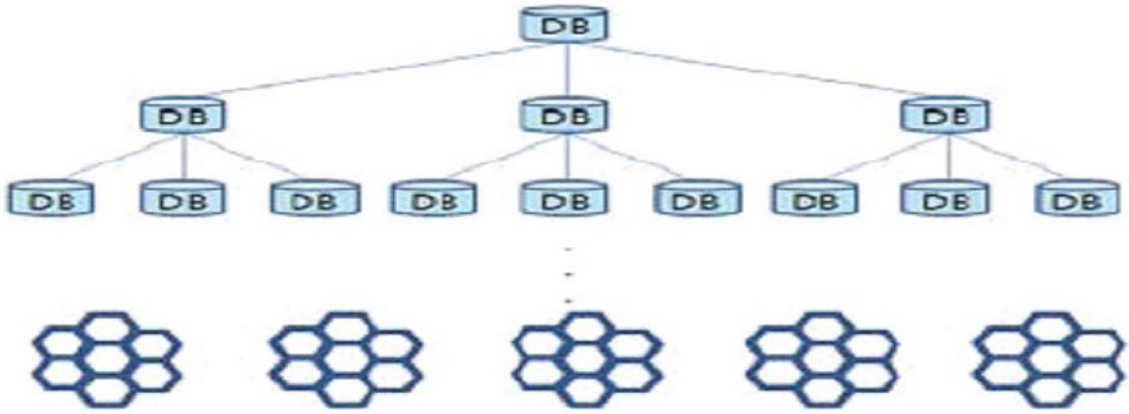


Fig. 5 A hierarchical location database system

Databases are a very important issue to reduce the update cost. But the optimal clustering can only be obtained exhaustively because the user moving patterns are dynamic in their nature.

Jixiong et al. developed a location database clustering algorithm and later called it *the set-cover algorithm*. Their algorithm utilizes the "greedy" approximation set-cover algorithm for clustering with a bottom-up approach. However, once some of the databases in cells are grouped into a cluster at the bottommost level, it is difficult that the movement information among the cells is used properly for clustering in the upper levels toward the root. In this paper, we propose a top-down clustering algorithm for the location databases [9]. In our clustering algorithm, we consider the number of visits to each cell by users, called *the visit count* of a cell, as well as that is, our algorithm takes into account both the node (cell) and edge information, while the set-cover algorithm utilizes only the edge information for clustering [13].

4. Future Work

The location management scheme proposed here is a significant departure from current research and signals a wide range of future work. Much of this is scheduled for completion in the near future with continual developments to progress for a significant period of time. It is expected that this research will form the basis of a series of publications, ideally signaling a new direction for research into location management techniques, in 3rd generation mobile networks that is DYNAMIC-3G and STATIC- 3G location management schemes for 3G wireless cellular systems (particularly UMTS), Distributed computing and object oriented [15].

5. Conclusion.

In this paper several static location management strategies for identification of user, update the user location in location server based on a hierarchical tree structure database are discussed.

Static location management uses one combination of search, update and search-update strategies throughout the execution. It was noticed that performing search-updates significantly reduced aggregate costs. Dynamic location management and tracking scheme are also discussed. Location management about mobile host is replicated, so, not all MSSs need to store the location of every mobile host. Mobile hosts that are query more often than others have their location information stored at a greater number of MSSs. The set of MSSs that store a mobile host's location change dynamically as the host moves from one part of the network to another. Also, MSSs that store location information of frequently queried mobile hosts store information about fewer hosts than the MSSs that only store location information of infrequently queried mobile hosts. As a result, the location directory is fairly distributed throughout the network, and no single MSS is overburdened with the responsibility of responding to location queries. This paper also discussed about the location in PCS (Personal Communication System) and Issues in location management. Also covering the database issues in location management. After discussing issues of location management, the performance of location management and future work also discussed.

6. REFERENCES

[1] A. Rahaman, J. Abawajy, and M. Hobbs :Taxonomy and Survey of Location Management Systems. *IEEE International Workshop on Component-Based Software Engineering*, pp. 369--374, (2007)

[2] M. Mouly and M.-B. Pautet. *The GSM System for Mobile Communications*. Telecom Publishing, 1992.

[3] Akyildiz, I. and Ho, J, 1996. On Location Management for Personal Communication Networks. IEEE Communications 34(9), 138-145.

[4] Das, Sajak K. and Sen, Sanjoy K, 1999. Adaptive Location Prediction Strategies Based On a Hierarchical Network Model in a Cellular Mobile Environment. The Computer Journal 42(6) 473-486.

[5] Jannink, J., Lam, D., Shivakumar, N., Widom, J. and Cox, 1997. Efficient and Flexible Location Management Techniques for Wireless Communication Systems. Wireless Network 3(5), 361-374.

[6] Krishna, P., Vaidya, H., and Pradhan, K, 1996. Static and adaptive location management in mobile wireless networks. Computer Communications 19(4) , 321-334.

[7] Levy, H. and Naor, Z., 1999. Active tracking: Locating Mobile Users in Personal Communication Service Networks. Wireless Network 5(6), 467-477.

[8] Plassmann, D., 1994. Location Management Strategies for Mobile Cellular Networks of 3rd Generation. In: Proceedings of 44th IEEE Vehicular Technology Conference, pp. 649-653.

[9] C. Jixiong, L. Guohui, X. Huajie, C. Xia, and Y. Bing :Location Database Clustering to Achieve Location Management Time Cost Reduction in a Mobile Computing System. *Wireless Communications, Networking and Mobile Computing*, vol. 2, 23--26, pp. 1328--1332 , (2005)

[10] Levy, H. and Naor, Z., 1999. Active tracking: Locating Mobile Users in Personal Communication Service Networks. Wireless Network 5(6), 467-477.

[11] Xie, H., Tabbane, S. and Goodman, D., 1993. Dynamic Location Area Management and Performance analysis. In: Proceeding of 43rd IEEE Vehicular Technology Conference, pp. 536-539.

[12] Mobile Location Protocol (MLP), LIF TS 101 Specification v3.0.0. *Location Inter-operability Forum (LIF)*, Jun. 2002.

[13] Guo-Hui Li, Kam-Yiu Lam, Tei-Wei Kuo, and Shi-Wu Lo on Location Management in Cellular Mobile Computing Systems with Dynamic Hierarchical Location Databases.

[14] James Cowling on Dynamic Location Management in Heterogeneous Cellular Networks Kwang-Jo Lee, Jin-Woo Song, Sung-Bong Yang on Effective Top-down Clustering Algorithms for Location Database Systems

[15] Yang Xiao, Member, IEEE, Yi Pan, Senior Member, IEEE, and Jie Li, Member, IEEE on Design and Analysis of Location Management for 3G Cellular Networks

Comparative Performance Analysis of Different Radio Channel Modelling For Bluetooth Localization System

Idigo Victor.[1] , Okezie C.C[2]Akpado kenneth [3], Ohaneme C.O[4]

[1]Department of Electronics and Computer Engineering, Nnamdi Azikiwe University, Awka Anambra State(234),Nigeria

[2]Department of Electronics and Computer Engineering, Nnamdi Azikiwe University, Awka Anambra State(234),Nigeria

[3]Department of Electronics and Computer Engineering, Nnamdi Azikiwe University Awka AnambraState(234),Nigeria

[4]Department of Electronics and Computer Engineering, Nnamdi Azikiwe University, Awka Anambra State(234),Nigeria

Abstract

This paper studied the possibility in a Bluetooth network of using channel simulated results as alternatives to on-site measurements and compares the average localization error as a function of a radio map resolution for the free-space path loss wall Attenuation Factor and Ray-Tracer (RT) models. Three reference radio maps were generated; one for each model. The Nearest Neighbour (NN) RSS based localization algorithm was used in this work. This algorithm was applied to the three models with different number of reference points. Results obtained show that RT proved to be more robust technique, especially for grid resolution greater than 10 meters. The simulation results using RT software in these situations is very close to the on-site results. On the other hand, the WAF model produced results that are very close to the on-site results for grid resolution less than 8 meters.

Keywords : path loss, localization, resolution, Bluetooth, Access points

1. Introduction

Position information is essential in many indoor applications. A hospital or health-care facility may wish to be informed of the location of all its patient at all times, a military unit may wish to extend its reconnaissance capabilities beyond the line of sight, a search and a rescue team can locate and provide help quickly if the locations of the individual in distress may be known accurately in advance[1]. Recent indoor applications include mobile e-commerce (m-commerce), e-museums, locating objects in warehouses and big shops(mall), locating books in libraries, etc.

Accurately predicting the location of an individual or an object can be a difficult task producing ambiguous results because of the harsh wireless environment[2]. The harsh site- specific multipath environment in indoor areas introduces difficulties in accurately tracking the position of objects and people. The behaviour of the channel changes from building to building and even within a single floor of a building. The

channel may vary with added objects and people moving in the vicinity. As a result, considerable work is needed for modelling the indoor channel for position applications.

Basically, the indoor localization procedure begins with collecting metrics related to the positions of the mobile terminal relative to the reference point. Almost all sort of metrics which are used in telecommunications system can also be used in position estimation systems. Angle of Arrival (AOA) and Received Signal Strength (RSS) are the most popular ones but Time of Arrival (TOA) and Phase of Arrival(POA) can be used as well[3]. The second step is to process the gathered metrics and estimate the location of the desired person or objects. This step usually requires signal processing knowledge unless the finger printing method is used. In using the finger printing method, it is required that a grid-network be built prior to any location estimation. After building the database for a new location, the new metric is measured irrespective of the viewed location and compares it with the database to find the best node, which could be referred to the desired point.

Bluetooth specification [4,5] provides no specific support for positioning service. In the absence of such support, various research efforts have been made in this area with alternating conclusions. The Bluetooth signal strength information has been used to create a system for locating and tracking users inside building[2]. Again, the concept of reference tags and readers that work with both the possibilities of Bluetooth supporting and not supporting the signal strength parameter has also been introduced[6]. However, other works suggest an unreliable relationship between the positioning and the signal strength and hence avoids this parameter for positioning with Bluetooth[7].

The objective of this work is to compare the performance of a Bluetooth indoor localization systems that is modelled using on-site measurements of RSSI values with other channel modelling approaches such as wall attenuation factor and Ray-tracing models.

2. Recent Emerging Indoor Models

In general, indoor channels may be classified either as line-of-sight(LOS) or obstructed (OBS), with varying degree of clutter. Some of the key models which have recently emerged are presented below.

- Partition Losses(Same Floor).
  Buildings have a wide variety of partitions and obstructions which form the internal or external structure. Partitions that are formed as part of building structure are called “ hard partitions”, and partitions that may be moved and which do not span to the ceiling are called “ soft partitions”. Partitions vary widely in their physical and electrical characteristics making it difficult to apply general models to specific indoor installations. Average signal loss measurement obstructed by common building material is given in [3].
- Partition Losses between floors

The losses between floors of a building are determined by the external dimensions and materials of the building, as well as the type of construction used to create the floors and the external surroundings[3]. Even the number of windows in a building and the presence of tinting(which attenuates the radio energy) can impact the loss between floors.

- Log distance path loss model
  Indoor path has been shown by many researchers to obey the distance power law as shown in equation (1) [3]:

  $PL(dB) = PL(d_0) + 10\alpha\log(d/d_0) + X\sigma$ (1)

  Where α = path loss exponent which indicates the rate at which the path loss increase with distance and depends on the surroundings and building type, and Xσ represents a normal random variable in dB having a standard deviation of σ dB. d_0 is the close- in reference distance which is determined from measurements close to the transmitter while d is the transmitter-receiver distance.
- Attenuation Factor Model
  This model provides flexibility and was shown to reduce the standard deviation between measured and predicted path loss to around 4dB, as compared to 13dB when only a log-distance model is used in two buildings[2]. The attenuation factor model is given by[3,8]

  $PL(d)\ [dB] = PL(d_0)\ [dB] + 10\alpha_{nf}\log(d/d_0) + FAF[dB] + \sum PAF[dB]$ (2)

  Where α_{nf} = the exponent value for "same floor" measurement, FAF = floor attenuation factor for a specific number of building floors and PAF = Partition attenuation factor for a specified obstruction encountered by a ray drawn between the transmitter and receiver.
- Wall Attenuation Factor (WAF) Model
  This model was proposed in [9], which included attenuation factor for building floors to disregard the effects of floors and instead consider the effects of obstacles(walls) between the transmitter and receiver. The wall Attenuation factor (WAF) model is described by[2,3]:

  $P(d)\ [dBm] = P(d_0)[dB] + 10\alpha\log(d/d_0) - \begin{pmatrix} nW.WAF, nW < c \\ c.WAF, nW \geq c \end{pmatrix}$

  Where α = path loss exponent that indicates the rate at which the path loss increases with distance, $P(d_0)$ = the signal power at some reference distance d_0 , d = transmitter- receiver separation, C = the maximum number of walls up to which the attenuation factor makes a difference, nW = the number of walls between the transmitter and the receiver, WAF = the wall attenuation factor.
- Channel Modelling using Ray Tracer

Ray – Tracing (RT) is a simulation tool encompassing the geometrical information of a floor plan in addition to the reflection and transmission coefficients of building materials that models the radio channel behaviour in different areas [10]. For a pair of transmitter-receiver at some known locations, RT determines the necessary information of a channel such as arrival angle, departure angle, phase, number of reflections and number of transmissions by sending a set of rays from the transmitter and tracing them until they either reach the receiver or largely attenuated that they can not be detected by the receiver. The TOA, magnitude, and phase of each path are recorded for each ray.

The predictions from ray tracing software are partition largely accurate for propagation of radio signals at frequencies greater than 900MHZ where electromagnetic waves can be described as travelling along localized ray paths. This method is shown to be accurate for indoor environments[3,10]. RT can be used to produce large database of channel impulse responses for statistical analysis of the channel. Therefore, RT is a viable alternative to physical measurements.

3. Methodology

Three reference radio maps were generated for on-site measurement, WAF and RT channel models. The nearest neighbour (NN) algorithm was used as the localization algorithm. The experimental test bed is the Electronic and communications laboratory of Nnamdi Azikiwe University, Awka, Nigeria. The test bed was segmented by a square of 1x1 meters as shown in figure 1. A Pentium based laptop equipped with Bluetooth device was placed at positions shown in the test bed. The mobile host carried by the user being tracked, was another Pentium based laptop equipped with a Bluetooth devices.

The protocol stack used in this equipment was provided by the Bluetooth simulator(BlueHoc). BlueHoc is IBM's new Bluetooth simulator released under IBM public license. It allows one to evaluate how Bluetooth performs under various ad-hoc scenarios.

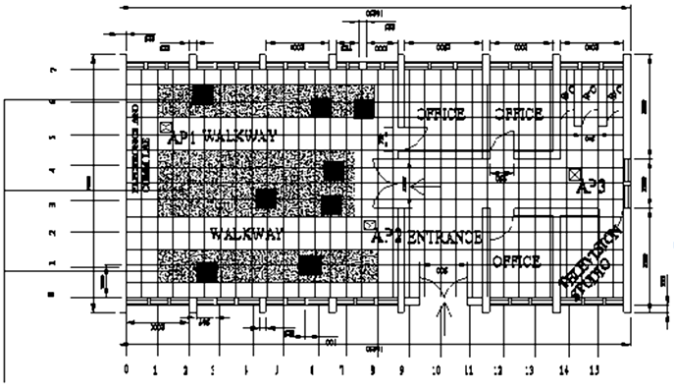


Figure 1 : Experimental Test bed

3.1 Channel characterization using on-site measurements.

A total of 75 RSSI data sets were collected from a square grid of 25 positions in the test bed. First, the

measurements were carried out using one transmitting Bluetooth, then two and finally three. At any given location, the RSSI measuring laptop(MS) can receive signals from transmitting Bluetooth. Each set of measurements consists of five minutes at approximately 0.25 second sampling interval. That is approximately 1200 data points were collected for each set. A sample measurement of RSSI data collected from transmitting Bluetooth is shown in figure 2

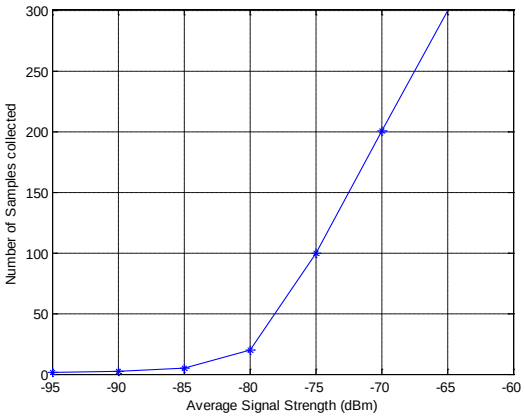


Fig. 2 Analysis of RSSI measurements over 5 minutes

3.2 Channel Characterization using WAF models

Two experiments that leads to the determination of WAF were conducted. First several measurements were taken to show the distribution of signal strength as a function of T-R separation derived from the empirical data collected in the off-line or data collecting phase. Secondly, experiments were conducted to determine the WAF. Finally, the effect of applying correction for intervening walls between the station and the mobile user was shown

Figure 3 illustrate how the average signal strength at each point varies with distance between the transmitter and also the receiver. The wide difference in signal strengths between points at similar distances is explained as follows: the layout of the rooms in the building, the placement of the base station and the location of the mobile user all have an effect on the received signal. Signals transmitted may be attenuated by different amounts due to the difference in the number and types of obstructions they encountered. For instance, in Fig. 3 , it was observed that the strength of the signal from two locations approximately 1.5meters from the base stations were approximately 8dBm apart. This is because there were several wall between one of the locations and the base stations, while the other locations had line-of-sight to the station.

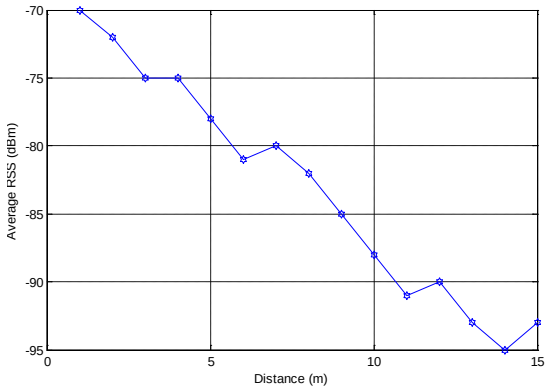


Figure 3: Average RSSI (dBm) versus Distance(m)

Furthermore, the following experiment was conducted to determine the wall Attenuation factor (WAF): first the measurement of the signal strength at the receiver when the receiver and the transmitter had line of sight was taken . Then the signal strength measurement with varying but known number of walls between the receiver and the transmitter (n=1,2,3,) were taken. Computation of the average of the difference between the signal strength values was used to determine the WAF. Fig. 4 shows the WAF graph obtained from this experiment for n = 1(WAF1), n = 2 (WAF2), and n = 3 (WAF3).

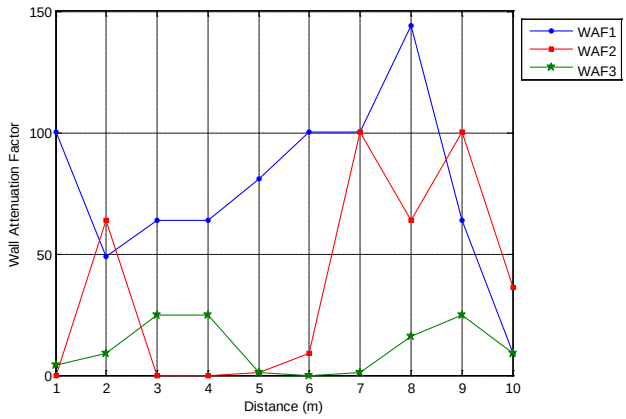


Figure 4 :Wall Attenuation Factor Graph

It is observed that the amount of additional attenuation dropped off as the number of walls separating the transmitting Bluetooth device and MS increased. Based on Measurements, we choose WAF to be 3.0dBm and C to be 4 (where C represents the number of walls that are factored into the model).

Figure 5 shows the results after the measured signal strength has been compensated for signal loss due to the intervening walls between the transmitting Bluetooth device and MS. This figure also shows a trend similar to that of the free-space loss(fig.3). This demonstrates that WAF propagation model compensates effectively attenuation due to obstructions.

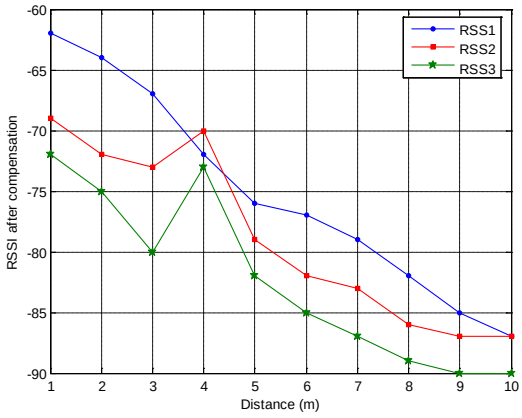


Figure 5 :Average RSS values after compensation

3.3 Channel Characterization using RT model

Figure 6 shows a typical channel impulse response generated by RT in the left side of the test bed . This model contains all the geometrical information of the floor plan such as doors and windows .

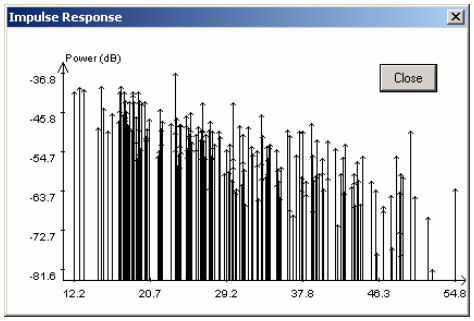


Figure 6: A channel impulse response generated by RT

4. Comparative Analysis of On-Site measurement(free space loss), WAF and RT models

4.1 Experimental Setup

The performance of the empirical free space loss model was compared with the results from WAF and RT models. The objectives are to :

- To study the possibility of using channel simulated results as alternative to on – site measurements.
- To compare the average localization error as a function of a radio map resolution for the three models

Major components of the experiment are shown in figure 7. Three reference radio maps were generated for on-site measurement, WAF and RT channel models. The nearest neighbour (NN) was used as localization algorithm.

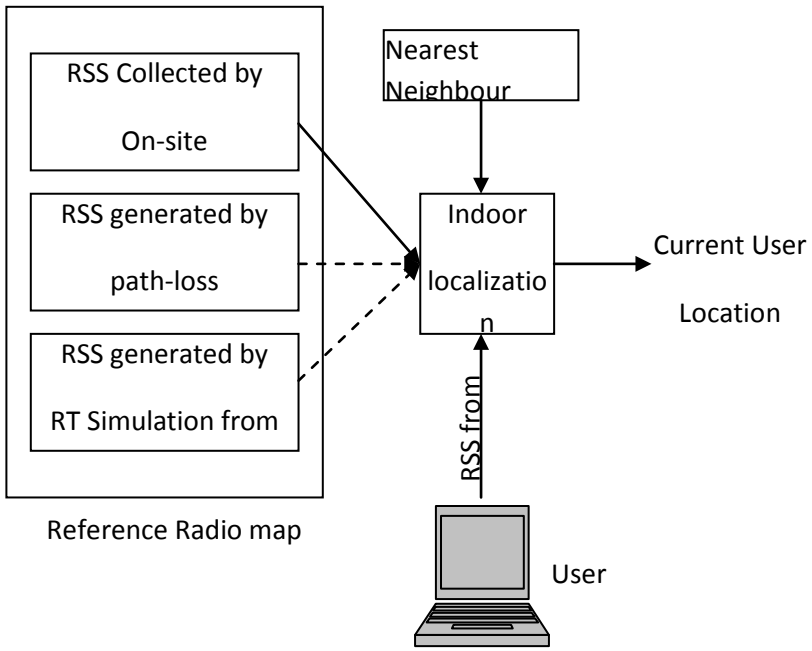


Figure 7 : Major Components of the comparison Test bed

4.2 Results and Discussion

An important performance metric in a positioning system called localization error was used. This depends on the resolution of the reference radio map. As shown in [6], the performance of a typical localization algorithm can be improved by using a finer grid of reference points. Figure 9 shows the impact of the number reference points in the distribution of localization accuracy in a system using the nearest neighbour algorithm on the three models: Increasing the number of reference point from 4 to 250 reduces the ninety percentile (P_{90}) of the localization error from 20 meters to 5 meters.

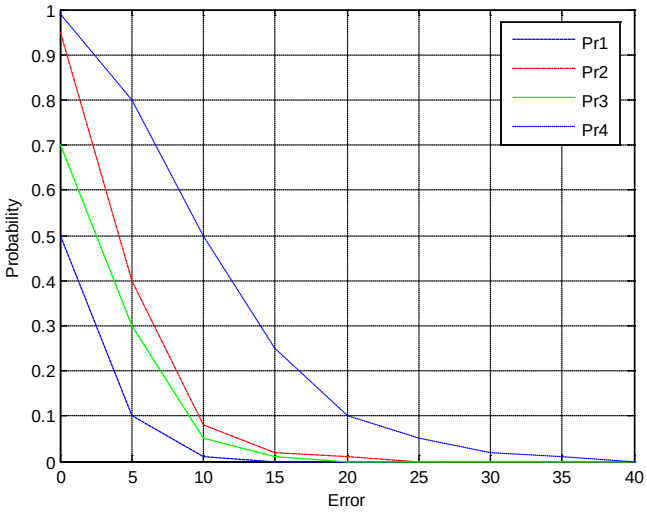


Figure 8: Impact of number of reference points on localization error using NN algorithms

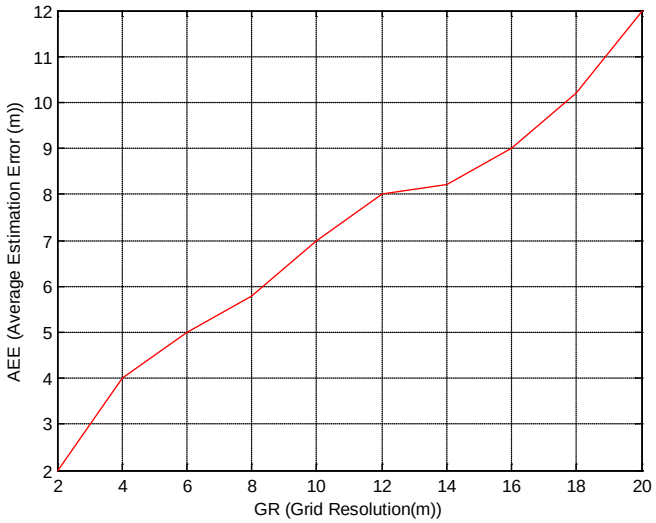


Figure 9 (a)

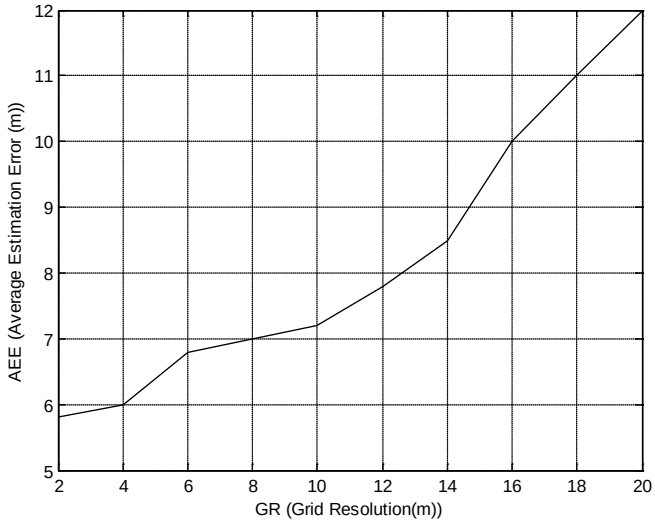


Figure 9 (b)

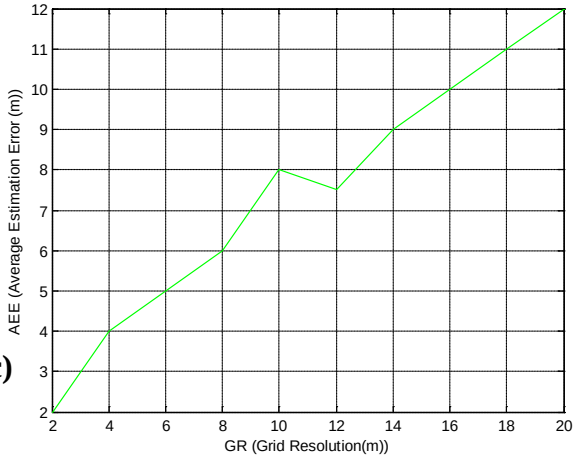


Figure 9 (c)

Figure 9: Average localization for the three models

(a) for a system trained with on-site measurement

(b) for a system trained with RT Channel Simulator

(c) for system trained with WAF model

Figure 9 (a-c) shows the average distance error estimation with NN algorithm as a function of grid resolution for the three radio maps. Using these curves, a system designer can determine the optimum number of reference points to achieve a desired localization accuracy which is dictated by the application requirements. From these plots it can be seen that the performance of an RT trained RSS based localization system is comparable to a system which is trained by an on-site measurements. For example, NN algorithm in a system which uses a RT generated reference radio map with 10 meter grid resolution is 7.2 meters which is a good approximation for the corresponding value of 7.0 meters for the case of on-site mode for any practical deployments.

Between the WAF and RT modelling approaches, RT proved to be more robust technique(especially for grid resolution> 10 meters). The simulation results using RT software in these situations is very close to the on-site results. On the other hand, the WAF model produced results that are very close to the on-site results for grid resolutions < 8 meters.

Using the RT modelling technique, we can easily increase the grid resolution to achieve higher localization while increasing the number of reference points using on-site measurements is a real challenge. In both WAF and RT modelling technique, the system needs to know the exact locations of the transmitting Bluetooth device in order to create the reference radio map. However, a positioning system trained with on-site measurement does not require knowing the location of the transmitting Bluetooth device.

5. Conclusion

This work compares the performance of a Bluetooth indoor localization systems that models the radio channel using RSSI values obtained from on-site measurement, WAF, and RT models. Three reference radio maps were generated for on-site measurement using free-space pathloss, WAF, and RT models. The NN algorithm was used as the localization algorithm.

An important performance metric for accessing the performance of positioning system called the localization error was used in accessing the models. This metric depends on the resolution of the reference map. The performance of the localization algorithm was found to improve by using finer grid of reference points. Finally, the study showed that using channel simulated results can be an alternative to on-site measurements.

References

[1]. Kiran thapa, and steven case , (2006) , "An Indoor positioning service for Bluetooth Ad-Hoc Networks", Department of computer and Information Science, Minnesota state university Mankato< www.mnsu.edu>

[2]. Kamol K. And Prastiant K. (2004), " Modelling of indoor positioning system based on location fingerprinting", School of information science, university of Pittsburgh publications. pp 15-23.

[3]. Rappaport, T.S. (2003), Wireless Communications, Pearson Education Inc, India

[4]. Bluetooth special Interest Group (2001), " Specification volume 1, specification by Bluetooth system, core" version 1.1, SIG. 18th Feb. 2006. <www.bluetooth.com>

[5]. Bluetooth special Interest Group (2001)," Specification volume 2, specification by Bluetooth system, profiles" version 1.1, SIG. 18th Feb. 2006. <www.bluetooth.com>

[6]. Bahl, P. And Padmanabham, V. (2000), "RADAR: An In-Building RF-Based user location and tracking system" Proceedings of IEEE infocom 2000, Tel-Aviv, Israel, Vol 2, pp 755-784

[7]. Patil, A. (2002), " Performance of Bluetooth Technologies and their applications to location sensing", Michigan state university publications, 4th edition , pp 15-25

[8]. Moustafa, P. 22nd Sept. 2007. <www.moustafa.usv.edu>

[9]. Orino. 10th Jan. 2008. < www.orinocowireless.com>

[10]. Spiros D. Mantis (1999), " Localization of Wireless Emitters Using Time Difference of Arrival (TDOA) methods in Noisy Channels" Masters thesis Naval PG school Monterey, California.

Data Visualization Technique Framework for Intrusion detection

Alaa El - Din Riad [1] , Ibrahim Elhenawy[2], Ahmed Hassan[3] and Nancy Awadallah[4]

[1]Head of Information Systems Dep, Faculty of Computer Science and Information Systems, Mansoura University, Egypt
Mansoura,,DK, 35513, Egypt

[2]Faculty of Computer Science and Information Systems, Zagazig University, Egypt
Zagazig, Egypt

[3]Faculty of Engineering , Mansoura University , Egypt
Mansoura,,DK, 35513, Egypt

[4]Faculty of Computer Science and Information Systems, Mansoura University, Egypt
Mansoura,,DK, 35513, Egypt

Abstract

Network attacks have become the fundamental threat to today's largely interconnected computer system. Intrusion detection system (IDS) is indispensable to defend the system in the face of increasing vulnerabilities.
While a number of information visualization software frameworks exist, creating new visualizations, especially those that involve novel visualization metaphors, interaction techniques, data analysis strategies, and specialized rendering algorithms, is still often a difficult process. To facilitate the creation of novel visualizations this paper presents a new framework that is designed with using data visualization technique for analysis and visualizes snort result data for user. The framework suggests PHP and CSS as data visualization technique and snort as intrusion detection system (IDS).

Keywords: *Intrusion Detection System, Visualization techniques, Snort, PHP and CSS.*

1. Introduction

Intrusion Detection Systems (IDS) look for attack signatures, which are specific patterns that usually indicate suspicious or malicious intent. Computer network administrators use IDS as a security management tool to monitor networks [1].

2. Related Research

J. Blustein, C. Fu and D. L. Silver presents proposed system that utilizes spatial hypertext workspace as the user interface could reduce the impact of high false alarm from IDS. This system may improvement the user's willingness to continuously monitor the system [1].

Network security visualization is a new research field as a result of introducing information visualization into network security area. Taking advantage of the ability of human vision perception to model structure, this technique turns abstract network and system data into graphical displays to help analysts explore network status and identify network anomalies or intrusion and even forecast the trend of security events [2].

Using data visualization technique to support the result of snort (IDS) , we consider that PHP and CSS as data visualization technique , we will deal with data of snort database to detect which data will be useful for network administrator to be visualized .

The framework introduced here is powerful because it is general, it can be applied to a wide domain of visualization problems. This research will assist users of visualization to explore, communicate, and understand their results.
The organization of this paper: next section discusses related research, section 3 presents proposed framework by using data visualization techniques for intrusion detection.

R.F.Erbacher discuss how user behavior can be exhibited within the visualization techniques, the capabilities provided by the environment, typical characteristics users should look out for (i.e., how unusual behavior exhibits itself), and exploration paradigms effective for identifying the meaning behind the user's behavior [3] .

H.Koike and K.Ohno propose a visualization system of a NIDS log named SnortView, which supports administrators in analyzing NIDS alerts much faster and much more easily. Instead of customizing the signature DB, they propose to utilize visualization to recognize not only each alert but also false detections [4].

N.Rangaraju and M.Terk describe a framework that is designed to simplify the process of building immersive visualization of structural analysis of building structures. They describe the components of the framework and describe two applications that were created to test their functionality [5].

J.Peng, C.Feng and J.W.Rozenblit propose a hybrid intrusion detection and visualization system that leverages the advantages of current signature-based and anomaly detection methods. The hybrid instruction detection system deploys these two methods in a two staged manner to identify both known and novel attacks.

When intrusion is detected, autonomous agents that reside on the system will automatically take actions against misuse and abuse of computer system, thus protecting the system from internal and external attacks [6].

Y.Park and J.Park presents Web Application Intrusion Detection System (WAIDS); an intrusion detection method based on an Anomaly Intrusion Detection model for detecting input validation attacks against web applications. Their approach is based on web application parameters which has identical structures and values. WAIDS derives a new intrusion detection method using generated profile from web request data in normal situation. By doing this, it is possible to reduce analysis time and false positives rate [7].

R.U. Rehman consider snort as an open source packet sniffer and logger that can be used as a lightweight Intrusion Detection System (IDS) to detect a variety of attacks and probes such as buffer overflows, stealth port scans, CGI attacks, and more. The Basic Analysis and Security Engine (BASE) displays and reports intrusions and attacks logged in the Snort database in a web browser for convenient analysis [8].

A.Komlodi, J. R. Goodall and W.G. Lutters report a framework for designing information visualization (IV) tools for monitoring and analysis activities.They studied ID analysts' daily activities in order to understand their routine work practices and the need for designing IV tools [9].

K.Abdullah presents new techniques to aid in network security using information visualization. Research contributions have been made in network data scaling and processing, port activity visualization, useful visualization showing a larger amount of information than textual methods, scaling port numbers and IP address for maximum use of screen space without occlusion, performing and using user study results to design an IDS alarm visualization tool [10].

From previous studies we present our framework which be overcome on the problem of how to describe intrusion detection system results for network administrator.

3. Proposed Framework

This research aims to design a system for visualize intrusion detection by using PHP & CSS as data visualization technique .The system introduces four components which are described in detail on next sections .

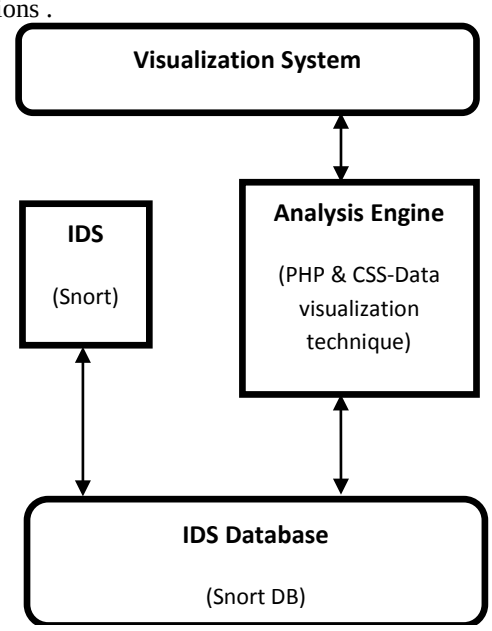


Fig. 1: Proposed System Structure

3.1 Snort Component

Snort is logically divided into multiple components. These components work together to detect particular attacks and to generate output in a required format from the detection system. A Snort-based IDS consists of the following major components:

- Packet Decoder
- Preprocessors
- Detection Engine
- Logging and Alerting System
- Output Modules [8]

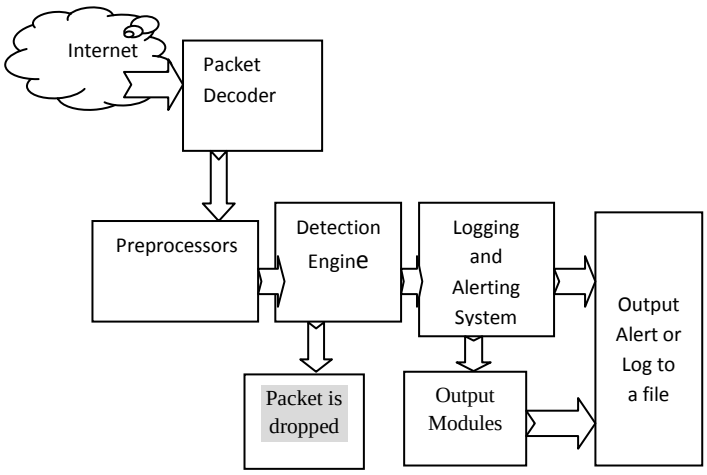


Fig. 2: Components of Snort [8]

Fig. 2 shows how these components are arranged. Any data packet coming from the Internet enters the packet decoder. On its way towards the output modules, it is either dropped, logged or an alert is generated. [9]

3.2 IDS (Snort) Database

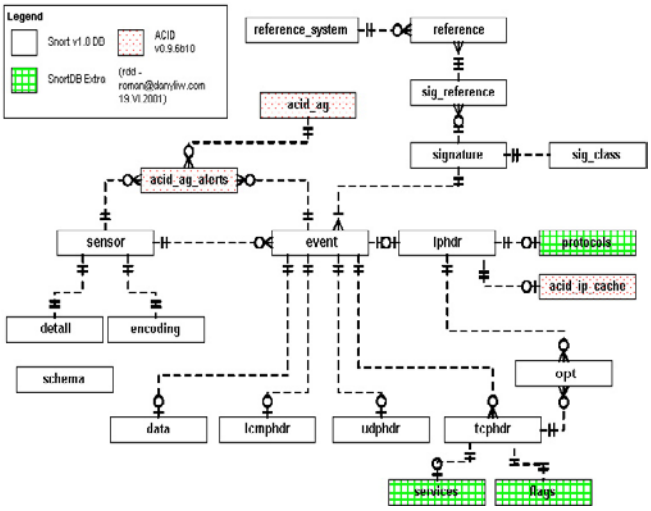


Fig. 3: Snort database schema [11]

The next table illustrates each table of Snort database from which component related to (Snort or ACID) and its description.

Table 1: snort tables [11]

Table	Component	Description
schema	Snort	Self-documented information about the database
sensor	Snort	Sensor name
event	Snort	Meta-data about the detected alert
signature	Snort	Normalized listing of alert/signature names, priorities, and revision IDs
sig_reference	Snort	Reference information for a signature
reference	Snort	Reference IDs for a signature
reference_system	Snort	(lookup table) Reference system list
sig_class	Snort	Normalized listing of alert/signature classifications
data	Snort	Contents of packet payload
iphdr	Snort	IP protocol fields
tcphdr	Snort	TCP protocol fields
udphdr	Snort	UDP protocol fields
icmphdr	Snort	ICMP protocol fields

opt	Snort	IP and TCP options
detail	Snort	(lookup table) Level of detail with which a sensor is logging
encoding	Snort	(lookup table) Type of encoding used for the packet payload
protocols	SnortDB extra	(lookup table) Layer-4 (IP encoded) protocol list
services	SnortDB extra	(lookup table) TCP and UDP service list
flags	SnortDB extra	(lookup table) TCP flag list
acid_ag	ACID	Meta-data for alert groups
acid_ag_alert	ACID	Alerts in each alert group
acid_ip_cache	ACID	Cached DNS and who is information

3.3 Analysis Engine

This component responsible for retrieving data from snort database which be detected from snort (IDS) to be analyzed and processed it by CSS & PHP.

3.4 Visualization System

This component will be user interface for snort intrusion detection system result implemented by CSS & PHP (Data Visualization Technique).

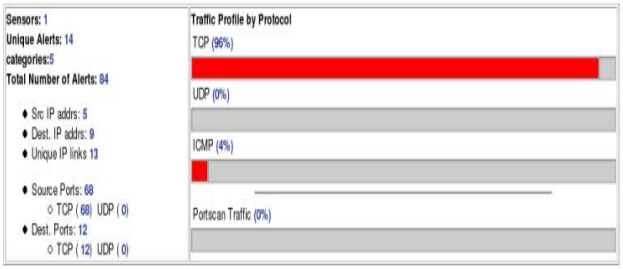


Fig. 4: Visualization System

4. Conclusion and Future Work

Intrusion detection is an information intensive and deeply analytic process that cannot be undertaken without the assistance of a computer.
Intrusion detection systems must handle masses of information (often in real-time) so as to report the abnormal use of networks and computer systems.
Our proposed system has proven to be effective for visually the intrusion which be detected by snort system.
In the future work we will use JQUERY as data visualization technique to visually intrusion detection and comparing between using it, CSS & PHP technique.

References

[1] J.Blustein , C.Fu , D.L.Silver ," Information Visualization for an Intrusion Detection System",ACM , 2005.
[2] Nurbol, H. Xu, H.Yang, F.Meng, L. Hu,"A real-time intrusion detection security visualization framework based on planner-scheduler",IEEE ,2009 .
[3] R.F.Erbacher," Intrusion Behavior Detection Through Visualization", IEEE , 2003 .
[4] H.Koike and K.Ohno ,"SnortView: Visualization System of Snort Logs" IEEE , 2004 .
[5] N.Rangaraju and M.Terk ," Framework for Immersive Visualization of Building Analysis Data " , IEEE , 2001 .
[6] J.Peng, C.Feng and J.W.Rozenblit,"A Hybrid Intrusion Detection and Visualization System", IEEE , 2006 .
[7] Y.Park and J.Park ," Web Application Intrusion Detection System for Input alidation Attack" , IEEE , 2008 .
[8] R.U. Rehman " Intrusion Detection Systems with Snort Advanced IDS Techniques Using Snort, Apache, MySQL, PHP, and ACID" , Publishing as Prentice Hall PTR Upper Saddle River, New Jersey, 2003.
[9] A.Komlodi, J.R. Goodall, W. G. Lutters , "An Information Visualization Framework for Intrusion Detection , 2004 , IEEE
[10] K.Abdullah , " Scaling and Visualizing Network Data to Facilitate in Intrusion Detection Tasks " ,Phd., School of Electrical and Computer Engineering ,Georgia Institute of Technology ,May 2006 .
[11] http://www.andrew.cmu.edu/user/rdanyliw/snort/acid_db_er_v102.html - last visit 22/07/2011

Authors

Alaa El - Din Riad, Head of Information Systems Department, Faculty of Computer and Information SciencesMansoura University

Ibrahim Elhenawy , Faculty of Computer and Information Sciences, Zagazig University
Ahmed Hassan ,Department of Electrical Engineering ,Faculty of Engineering , Mansoura University

Nancy Awadallah Researcher Assistant , Master in E-commerce Security 2008 , Faculty of computer science & Information System - Mansoura University - Egypt

Generation of Random Fields for Image Enhancement and Reconstruction

1. B. Srinivasa Rao

2. K. Srinivas

3. Dr. LSS Reddy

ABSTRACT

Noise Suppression from images is one of the most important concerns in digital image processing. Impulsive noise is one such noise, which may corrupt images during their acquisition or transmission or storage etc. A variety of techniques are reported to remove this type of noise. It is observed that techniques which follow the two stage process of detection of noise and filtering of noisy pixels achieve better performance than others. In this work such schemes of impulsive noise detection and filtering thereof are proposed. Two models of impulsive noise are considered in this work. The first one is *Salt & Pepper Noise* (SPN) model, where the noise value may be either the minimum or maximum of the dynamic gray scale range of the image. And, the second one is *Random Valued Impulsive Noise* (RVIN) model, where the noise pixel value is bounded by the range of the dynamic gray scale of the image. This work deal with SPN model and deal with RVIN model of noise. The first scheme is based on second order difference of pixels in order to identify noisy pixels. The second scheme for SPN model uses fuzzy technique to locate contaminated pixels. The contaminated pixels are then subject to median filtering. This detection–filtration is done recursively so that filtered pixels take part in the detection of noise in the next pixel. In the propose schemes for adaptive threshold selection is emphasizing. Incorporation of adaptive threshold into the noise detection process may be leads to more reliable and more efficient detection of noise. Based on the noisy image characteristics and their statistics, threshold values are selected. It may be observed, in general, that the proposing schemes are better in suppressing impulsive noise at different noise ratios than their counterparts.

Introduction

Noise removal from a contaminated image signal is a prominent field of research and many researchers have suggested a large number of algorithms and compared their results. The main thrust on all such algorithms is to remove impulsive noise while preserving image details. These schemes differ in their basic methodologies applied to suppress noise. Some schemes utilize detection of impulsive noise followed by filtering whereas others filter all the pixels irrespective of corruption. In this section an attempt has been made for a detail literature review on the reported articles and studies their performances through computer simulation. We have classified the schemes based on the characteristics of the filtering schemes. First one is Filtering without Detection, in this type of filtering a window mask is moved across the observed image. The mask is usually of size $(2N+1)^2$ where N is a positive integer. Generally the center element is the pixel of interest. When the mask is moved starting from the left-top corner of the image to the right-bottom corner, it performs some arithmetical operations without discriminating any pixel. Second one is Detection followed by Filtering; this type of filtering involves two steps. In first step it identifies noisy pixels and in second step it filters those pixels. Here also a mask is moved across the image and some arithmetical operations are carried out to detect the noisy pixels. Then filtering operation is performed

only on those pixels which are found to be noisy in the previous step, keeping the non-noisy intact. And third one is Hybrid Filtering, in such hybrid schemes; two or more filters are suggested to filter a corrupted location. The decision to apply a particular filter is based on the noise level at the test pixel location or performance of the filter on a filtering mask.

Motivation

Most of the traditional reported schemes work well under SPN but fails under RVIN, which is more realistic when it comes to real world applications. Even though some of the reported methods claim to be adaptive, they are not truly adaptive for the simple reason of not considering the image and noise characteristics. These schemes generally use a threshold value for the identification of noise. A predefined parameter is compared with this threshold value. If it exceeds, the pixel is marked as contaminated otherwise not. Usually the threshold value used is either a constant or a set of four/five values. A threshold, which is optimal in one environment, may not be good at all in a different environment. By environment we mean, the type of image, characteristic and density of noise. Further, there has been little or no usage of soft computing techniques in the reported schemes. Soft computing methodologies mimic the remarkable human Capability of making decision in ambiguous environment. It embraces approximate reasoning, imprecision, uncertainty and partial truth. There exists scope for improving the detector's performance using soft computing techniques. These facts motivated us

• To work towards improved and efficient detectors for identifying contaminated pixels.
• To devise adaptive thresholding techniques so that noise detection would be more reliable.
• To exploit the computational power of soft computing techniques in predicting the threshold value by adapting to the environment with a greater ease.

In this work all the existing filters will be analyzed and based on the limitations innovative methods will be suggested for Images Enhancement and restorations which will produce better results.

Families of Common Nonlinear Filters

Boolean, stack, OS and morphological filters. Many nonlinear signal processing methods have their origin in statistics. In fact, the median filter was first introduced in statistics for smoothing economical time series [24]. It soon became evident that the median filter performs very well especially in image processing applications where sharp transitions are common. Especially in urban or other "man-made" scenes we almost always have sharp edges and these edges usually are the most important information in the image. Attempts to retain sharp edges in linear filtering lead to "ringing" effects that are often more disturbing than noise. In applications involving images, image sequences and color images, order statistics and their close kin morphological filters have by far been the most prominent and successful classes of nonlinear techniques, see [1], [3], [5], [6], [17], and [21].One of the greatest limitations of order statistics filters is the fact that they are "smoothers". Without additional processing or combinations, their use remains limited to restoration applications, in which they excel especially in the presence of heavy-tailed noise (to be removed) and important signal details (to be preserved). General Boolean filters and morphological filters with non-flat structuring elements do not suffer from such a shortcomings; however, they do not benefit from the stacking property which unifies all subclasses of stack filters; ranked order, median and weighted median and weighted order statistics filters. The stacking property says that the Boolean function that defines the filter is positive (or increasing as is the standard term in mathematical morphology).

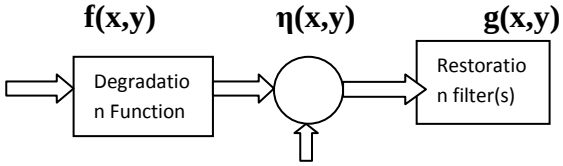
There is usually no underlying physical model that would demand the filter to be increasing. The power of increasing
filters comes from the fact that concept narrows the filter class in a way that fits well to design processes. For instance, if we have information of the possible desired signal form expressed in Boolean vectors in such a form that it does not conflict the positivity of the defining Boolean function, designing optimal increasing filter becomes straightforward, see [4], [11], [13], [26], and [28]. In some problems, notably in document image processing, noise is loosely speaking signal dependent binary union and intersection noise and increasing filters turn out to perform quite badly [14]. Here one must give up positivity and the penalty is that the large number of parameters and non-robust behavior of an unconstrained Boolean function makes the design of filters with large window sizes impossible. Recently there have emerged new ways to constrain the function leading to much better performance for large window sizes [20].

Challenges in Filter Design

A unified and efficient framework for nonlinear filter design remains one of the most challenging tasks in this field. Even though we can not hope to obtain a framework as powerful as the techniques
for designing linear filters we should be able to build a methodology that would tie together the conditions and assumptions of the problem, the major nonlinear filter classes, relevant cost functions and accessible optimization algorithms. It is clear that the methodology must be able to deal with both statistical and deterministic aspects of the problem and filters. This framework cannot be obtained by one step (leap) but it will emerge as the result of incremental steps from the joint
efforts of the signal processing community. However it is good to keep the ultimate goal in mind while solving problems for more immediate demands. Few attempts have been made to this end, see for instance [8] and [29]. Here we consider some problems whose solutions will clearly take us forward on this path. We all agree that it would be important to be able to devise a feasible optimization procedure with a suitable cost function, even for a specific application, e.g. image restoration. The unification of two or more existing filter classes will un-doubtfully increase the modeling power of the framework. Therefore, it would of great interest to determine the class of problems (signals) that can be solved (represented) by the new framework.

A related problem is that of the filter structure, or more specifically, the filter size. An often asked question is how large should the filter size be. Most of the answers have been “try and see” type. In [22], we proposed a solution to this problem, in which we combined both the optimization and the filter structure in a recursive manner. Another equally important challenge to the nonlinear signal and image processing community is to develop new and attracting applications. Next to a mature theory (still developing), interesting applications would be the driving force to open up new frontiers in the field. Most of the current applications remain in the areas of signal (1- and M-D) restoration, enhancement, edge detection and interpolation. Recently, stack and Boolean filters were successfully used as predictors in a DPCM lossless image compression scheme, [16]. More such endeavors are needed in other areas such as speech analysis and processing, telecommunications and data analysis and communication.

Model for image degradation/restoration process noise



We will assume that a degradation function exists, which, together with additive noise, operates on the input image $f(x,y)$ to produce a degraded image $g(x,y)$.

The objective of restoration is to obtain an estimate for the original image from its degraded version $g(x,y)$ while having some knowledge about the degradation function H and the noise $\eta(x,y)$.

Mean Filter

The Mean Filter is a linear filter which uses a mask over each pixel in the signal. Each of the components of the pixels which fall under the mask are averaged together to form a single pixel. This new pixel is then used to replace the pixel in the signal studied. The Mean Filter is poor at maintaining edges within the image.

Mean Filter(x1…, xN) =1/N ∑ xi (1)
For i=1…….N

The use of the median in signal processing was first introduced by J. W. Tukey [1]. The Median Filter is performed by taking the magnitude of all of the vectors within a mask and sorting the magnitudes, as defined in (2). The pixel with the median magnitude is then used to replace the pixel studied. The Simple Median Filter has an advantage over the Mean filter in that it relies on median of the data instead of the mean. A single noisy pixel present in the image can significantly skew the mean of a set. The median of a set is more robust with respect to the presence of noise.

Median filter(x1…..xN) = Median (‖x1‖^2,….., ‖xN‖^2)

When filtering using the Simple Median Filter, an original pixel and the resulting filtered pixel of the sample studied are sometimes the same pixel. A pixel that does not change due to filtering is known as the root of the mask. It can be shown that after sufficient iterations of median filtering, every signal converges to a root signal [2].

The Component Median Filter, defined in (3), also relies on the statistical median concept. In the Simple Median Filter, each point in the signal is converted to a single magnitude. In the Component Median Filter each scalar component is treated independently. A filter mask is placed over a point in the signal. For each component of each point under the mask, a single median component is determined. These components are then combined to form a new point, which is then used to represent the point in the signal studied. When working with color images, however, this filter regularly outperforms the Simple Median Filter. When noise affects a point in a grayscale image, the result is called “salt and pepper” noise. In color images, this property of “salt and pepper” noise is typical of noise models where only one scalar value of a point is affected. For this noise model, the Component Median Filter is more accurate than the Simple Median Filter. The disadvantage of this filter is that it will create a new signal point that did not exist in the original signal, which may be undesirable in some applications.

$$\mathbf{CMF(x1,\ldots,xN)} = \begin{bmatrix} \mathbf{Median(x1r,\ldots,xNr)} \\ \mathbf{Median(x1g,\ldots,xNg)} \\ \mathbf{Median(x1b,\ldots,xNb)} \end{bmatrix} \quad \ldots\ldots\ldots \quad \mathbf{.(3)}$$

The Vector Median Filter (VMF) was developed by Astola, Haavisto, and Neuvo in 1990 [3]. In the VMF (4), a filter mask is placed over a single point. The sum of the vector magnitude differences using the *L*2 norm from each point to each other point within the mask is computed. The point with the minimum sum of vector differences is used to represent the point in the signal studied. The VMF is a well-researched filter and popular due to the extensive modifications that can be performed in conjunction with it.

$$\mathbf{VMF(x1,\ldots,xN)= Min\ (\sum \|x1-xi\| \ldots \sum \|xN-xi\|)} \quad (4)$$
$$\mathbf{For\ i=1\ldots N}$$

For each of the four algorithms discussed, experimental results will be shown that indicate which algorithm is best suited for the purpose of impulse noise removal in digital color images. Those results will be compared to two new algorithms for noise removal: the Spatial Median Filter and the Modified Spatial Median Filter.

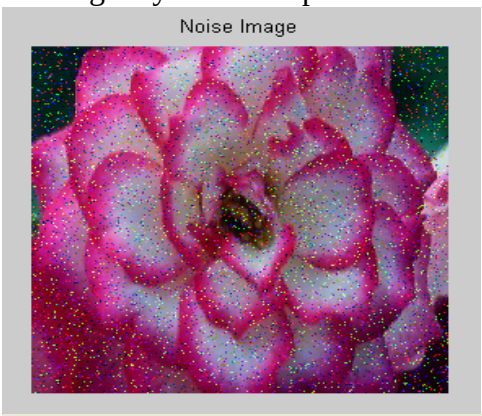
In testing, we consider sets of images containing various amounts of artificial noise. Impulse noise represents random spikes of energy that happen during the data transfer of an image. To generate noise, a percentage of the image is damaged by changing a randomly selected point channel to a random value from 0 to 255. The noise model, *In*, is given by

$$\mathbf{In(i,j)} = \begin{cases} \mathbf{I(i,j)} & \mathbf{x \geq p} \\ \mathbf{(Ir(i,j),Ig(i,j),z)} & \mathbf{y<1/3} \quad \mathbf{x<p} \\ \mathbf{(Ir(i,j),z,Ib(i,j))} & \mathbf{1/3 \leq y<2/3} \quad \mathbf{x<p} \\ (z,Ig(i,j),Ib(i,j)) & 2/3 \leq y \quad x<p \end{cases}$$

where *I* is the original image, *Ir*, *Ig*, and *Ib* represent the original red, green, and blue component intensities of the original image, *x,y 2* [0,1] are continuous uniform random numbers, *z 2* [0,255] is a discrete uniform random number, and *p 2* [0,1] is a parameter which represents the probability of noise in the image.

Restoration in the presence of noise only – Spatial filtering

Selecting an incorrect sign in contra harmonic filtering may lead to unpleasant results...



	MSE	PSNR
Mean	347.4479	22.7219
Median	107749458.	32.1933
CMF	208.0758	24.9486
VMF	24.023	38.2154

Adaptive local noise reduction filter:
since the mean gives a measure for average intensity in the region and variance characterizes contrast in that region, they both are reasonable parameters to base an adaptive filter.

Considering a local region *Sxy* of a noisy version *g*(*x,y*) of the image *f*(*x,y*) with the overall noise variance ση*2*, local mean of pixels in the region *mL* and their local variance σ*L2*, the following rules are to be implemented:

Conclusion

We have introduced two new filters for removing impulse noise from images and shown how they com- pare to four well-known techniques for noise removal. The Spatial Median Filter is proposed based on the Vector Median Filter and the Spatial Median quantile order statistic. Seeing that the order statistic can be utilized to make a judgment as to whether a point in the signal is considered noise or not, a Modified Spatial Median Filter has been proposed. This filter accepts a threshold parameter *T* indicating the estimated number of uncorrupted pixels under the mask.

References

[1] J. Astola and P. Kuosmanen, December 1999.
[11] P. Kuosmanen, *Statistical analysis and optimization of stack filters*, Acta Polytechnica Scandinavica, Electrical Engineering Series, 77, Helsinki, 1994.
[12] H. Leung and S. Haykin, "Detection and estimation using an adaptive rational function filter," *IEEE Transactions on Signal Processing*, vol. 42, no. 12, pp. 3366-3376, December 1994.
[13] J.H. Lin, T.M. Sellke and E. Coyle, "Adaptive stack filtering under the mean absolute error criterion," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 38, pp. 938-954, June 1990.
[14] R.P. Loce and E.R. Dougherty, *Enhancement and restoration of digital documents*, SPIE Optical Engineering Press, Bellingham, Washington, 1997
[15] V.J. Mathews and G.L. Sicuranza, *Polynomial Signal Processing*, Wiley, 2000.
[16] D. Petrescu, I. Tabus, and M. Gabbouj, "Prediction capabilities of Boolean and stack filters for lossless image compression," *Multidimensional Systems and Signal Processing*, Volume 10, No. 2, April 1999, pp. 161-187.
[17] I. Pitas and A.N. Venetsanopoulos, *Nonlinear Digital Filters, Principles and Applications*, Kluwer Academic Publishers, 1990.
[18] G. Ramponi, "The rational filter for image smoothing," *IEEE Signal Processing Letters*, vol. 3, no. 3, pp. 63-65, 1996.
[19] G. Ramponi, "Image processing using rational functions," *Proceedings of Cost 254 Workshop*, Budapest, Hungary, 6-7. February 1997.

[20] O.V. Sarca, E.R. Dougherty and J. T. Astola, "Secondarily constrained Boolean filters," *Signal Processing*, vol 71, No 3, pp. 247-263, December 1998.
[21] *Image Analysis and Mathematical Morphology*, J. Serra, Ed., Academic Press, 1988.
[22] I. Tabus and M. Gabbouj, "Fast Order-Recursive Algorithms for Optimal Stack Filter Design," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Detroit, MI,

Fundamentals of Nonlinear Digital

Filtering, CRC Press, 1997.

[2] F.A. Cheikh, L. Khriji, M. Gabbouj and G. Ramponi, "Color image interpolation using vector rational filters," *Proceedings SPIE Conference on Nonlinear Image Processing IX*, San Jose, CA, 24-30 January 1998.
[3] E.R. Dougherty and J. Astola, *An Introduction to Nonlinear Image Processing*, SPIE Press, Vol. TT 16, 1994.
[4] M. Gabbouj and E. Coyle, "Minimum mean absolute error stack filtering with structural constraints and goals," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 38, pp. 955-968, June 1990.
[5] M. Gabbouj, E.J. Coyle and N.C. Gallagher, Jr., "An Overview of Median and Stack Filtering," *Circuits, Systems, and Signal Processing*, Special issue on Median and Morphological Filtering, vol. 11, no. 1, pp. 7-45, 1992.
[6] M. Gabbouj, "Weighted Median Filtering -- Striking Analogies to FIR Filters," in *Circuits and Systems Tutorials*, C. Toumazou, ED., Oxford, UK, IEEE ISCAS'94, 1994, pp. 5-21.
[7] P. Heinonen, and Y. Neuvo, "FIRmedian hybrid filters with predictive FIR substructures," IEEE Transactions on Acoustic Speech and Signal Processing, vol. 36, pp. 892-899, June 1988.
[8] V.G. Kamat, E.R. Dougherty and J. Barrera, "Bayesian multiresolution filter design," *Proc. SPIE Nonlinear Image Processing XI*, San Jose, California, 24-25 January 2000, vol. 3961, pp. 22-33.
[9] L. Khriji and M. Gabbouj, "Medianrational hybrid filters for image restoration," *Electronic Letters*, vol. 34, no. 10, pp. 977-979, May 1998.
[10] L. Khriji and M. Gabbouj, "A new class of multichannel image processing filters: vector median-rational hybrid filters," *IEICE Transactions on Information and Systems*, vol. E82-D, no. 12, pp. 1589-1596, 8-12 May 1995, pp. 2391-2394.
[23] I. Tabus, D. Petrescu and M. Gabbouj, "A training framework for stack and Boolean filtering –Fast optimal design procedures and robustness case study," *IEEE Transactions on Image Processing, Special issue on nonlinear image processing*, vol. 5, no. 6, pp. 809-826, June 1996.
[24] J.W. Tukey, "Nonlinear (nonsuperposable) methods for smoothing data," in *Congr. Tec. EASCON-74*, p. 673, 1974 (abstract only).
[25] J.L. Walsh, "The existence of rational functions of best approximation," *Transactions Am. Math. Soc.*, vol. 33, pp. 668-689, 1931.
[26] R. Yang, L. Yin, M. Gabbouj, J. Astola and Y. Neuvo, "Optimal weighted median filters under structural constraints," *IEEE Transactions on Signal Processing*, vol. 43, pp. 591-604, March 1995.
[27] L. Yaroslavsky, *Seminar Notes*, Tampere International Center for Signal Processing, Tampere, Finland, 1999.
[28] L. Yin, R. Yang, M. Gabbouj and Y. Neuvo, "Weighted median filters: a tutorial," *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, vol. 43, pp. 157-192, March 1996.

Bibliography:

1. Author: B. Srinivasarao, did his B.Tech from Nagarjuna University, M. Tech from JNTU Anantapur, and pursuing his Ph. D under the guidance of Dr. LSS Reddy. Currently he is working as Assoc. Professor in the department of IT.

2. Author: K. Srinivas, did his MCA from Nagarjuna University, M. Tech from Nagarjuna University. Currently he is working as Assoc. Professor in the department of MCA.

3. Author: Dr. LSS Reddy, did his B.Tech Electronics & Communication Engineering from J.N.T. University Hyderabad. M. Phil Computer Science Engineering from Central University Hyderabad. Ph.D. Computer Science Engineering from BITS, PILANI. Currently working as Director of Lakireddy Balireddy college of engineering.

Rahul Fixed Priority Enhance Classes Bandwidth Exploitation in TDM EPON

Muhammad Bilal
Faisalabad, 38000, Pakistan

Abstract

A Passive Optical Network (PON) is a single, collective optical fiber that used low-cost optical splitters to divide the single fiber into split strands feed individual subscribers. PON'S are called passive because, other than at the CO (Central Office) and subscriber endpoints, there are no active electronics inside the access network. With the development of services offered by the Internet, the "last mile" restricted access problems keep it up to increase step by step.
Many algorithms were developed for making TDM EPON efficient similar to Scheduling (No class Solution) and Priority Swapping, IPACT etr. These all algorithms have problems like delay, QoS and channel under- utilization. We focused the well-organized bandwidth utilization in TDM EPON by managing time slots within ONUs and reducing latency and increasing quality of service.

Our Rahul Fixed Priority Enhance Classes Bandwidth (RFPECB) algorithm is an intra-ONU bandwidth allocation algorithm, which is used to enhance the network performance by evaluating the parameters like channel underutilization, delay and Quality of Service. The issues which are lacking in the already made algorithms are being resolved with our RFPECB Algorithm. The main problem time slots management issue solved in RFPECB algorithm.

Keywords: Last Mile, QoS (Quality of Services), Rahul Fixed Priority Enhance Classes Bandwidth (RFPECB) Algorithm, ONU (Optical Network Unit) etr.

1. Introduction

Passive Optical Networks (PON's) are point-to-multipoint optical networks. There are no active elements such as amplifier, router switch in the signals path from source to destination. The elements used in such networks are passive combiners, couplers, and splitters.

PON technology is receiving additional furthermore more interest by the telecommunication production as the "last Mile" solution. The "Last Mile" solution is also called "First Mile" solution.

1.2 PON Components

There are two types of PON components.

I. Active Network Elements
II. Passive Network Elements

1.2.1 Active Network Elements

Vendors of the Network elements mainly focus on active network elements for instance CO chassis and ONU, because these elements can reduce the cost of laying network. The CO chassis is located at service provider's CO, head end. [1]

Optical Line Terminal (OLT):

Optical Line Terminal is placed in CO (Central Office). Its functional unit is dependent upon which type of multiplexing used a TDM, WDM or hybrid, but main functional unit is transponder. [1]

The OLT generates time stamped messages to be used as global time reference. It also assigns bandwidth and performs ranging operations. [4]

Optical Network Unit (ONU):

Optical Network Unit provides interface between the purchaser's data, video and telephony networks and the PON. Its main function is to receive traffic in optical format and then convert it to the user desired format (Ethernet, IP multicast etc.). [1]

The ONU performs an auto-discovery process which Includes ranging and the assignment of both Logical Link IDs and bandwidth. [4]

1.2.2 Passive Network Element

These elements are placed between OLT and ONUs.

i. Optical Coupler/Splitter.
ii. Combiner

1.3 EPON Protocol:-

For controlling the P2MP fiber network, EPON uses the Multi-Point Control Protocol (MPCP).
MPCP perform bandwidth assignment, bandwidth polling, auto-discovery, and ranging. It is implemented in the MAC Layer, introducing new 64-byte control messages: [4]

- GATE and REPORT are used to assign and request bandwidth
- REGISTER is used to control the auto discovery process

1.4 PON Topologies:

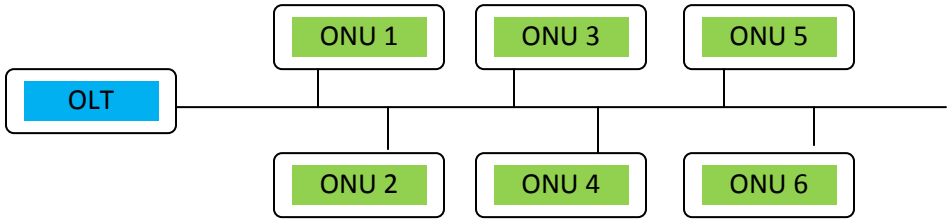


Fig. 1 Bus Topology [2]

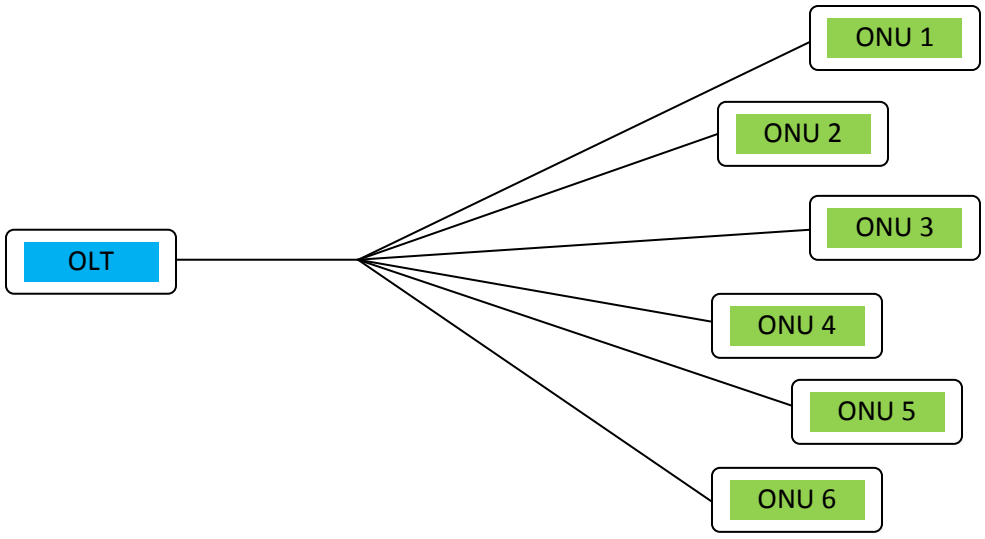


Fig. 2 Tree Topology [2]

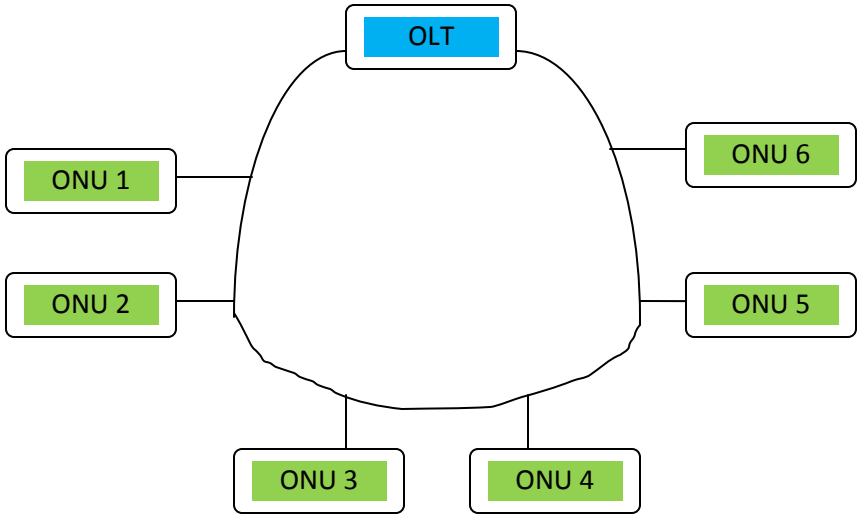


Fig. 3 Ring Topology [2]

PON Topologies

Ring Topology is better than others but mostly Tree

Topology used.

1.5 Transmission in EPON:-

There are two types of transmission in EPON are used:-

1. Downstream (Broadcast from OLT to ONU's). Point to multipoint network.
2. Upstream (Joint from ONU's to OLT). Multipoint to point network.

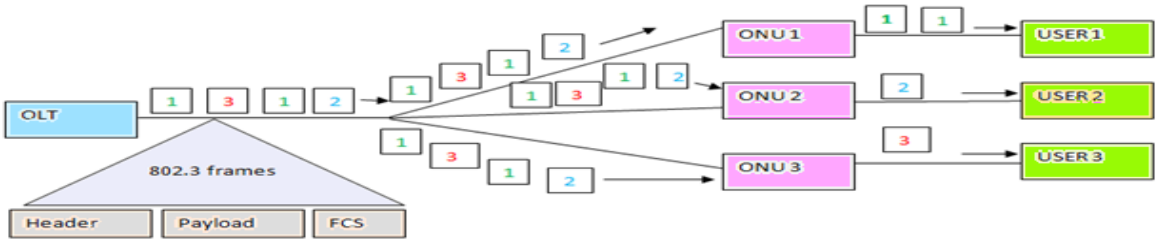


Fig. 4 Downstream Traffic in EPON [2]

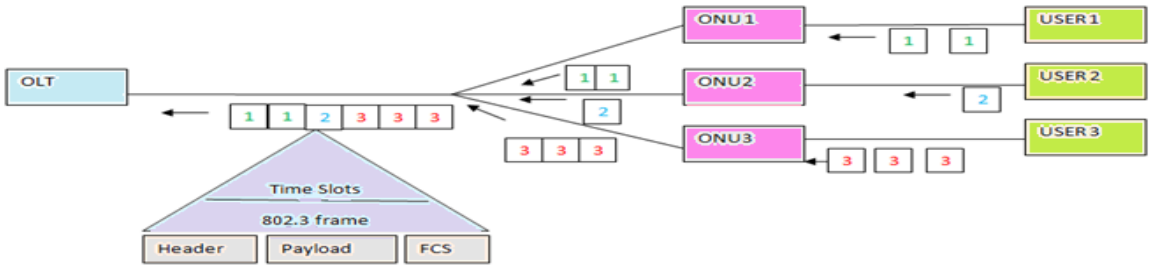


Fig. 5 Upstream Traffic in EPON [2]

2. Available Solution:-

In Scheduling algorithm (No class) except delays there is also one another part for the load administration and bandwidth deployment by the ONUs, as we have to broadcast the packets on the basis of the timeslot, if the size of the packet is greater than the timeslot being obtainable by the OLT, for transmission then that trace has to wait for the next time slot, this may cause channel underutilization, we can avoid it by implementing scheduling, at the ONUs. [3]

In Scheduling algorithm (With Class) We will implement the scheduling in the way that assume there are five packets in the buffer, if the size of the first three and fifth one is up to that is offered by the timeslot, then we will not wait for the fourth packet that is not fitting in the timeslot, we will allow fifth one to move first in the timeslot, by doing so channel will not be underutilized, and less timeslots will be required for packet broadcast. [3]

Two Major Problems in available Solution are:-

- Quality of Service
- Delays

3. Stuff and Methods

Our Fixed Priority Enhance Classes Bandwidth (RFPECB) Algorithm is an intra-ONU bandwidth allocation algorithm focusing to handle in better way parameters, like channel underutilization, delay and Quality of Service.

We have compared our solution with simple "scheduling" algorithm in which no classes were implanted for the purpose of quality of service.

In RFPECB algorithm, I implemented the four classes.

- Expedited Forwarding (EF)
- Assured Forwarding (AF)
- Best Effort (BE)
- Text Forwarding (TF)

EF, AF and BE are IEEE classes. But RT is an additional class (Not an IEEE class). According to RFPECB algorithm EF deals with Video type data. AF deals with voice type data. BE deals with audio type data. TF deals with text type data.

In our RFPECB algorithm EF class bandwidth is fixed. EF bandwidth fixed as 40 %. Simply, it means 0nly 40 % data can be sent at a time in EF class (In every starting time slot). If the data is more than 40 % than second time slot T2 sent. Every time slot starts with EF data if exits. AF and BE and RT classes bandwidth are not fixed but priority phenomenon used. In three remaining classes which priority less than others move first and if tie condition exit among these three classes AF, BE, TF than AF first 2nd BE than TF move last. Because AF data is important than BE and BE data important than TF.

Network Traffic Schemes

Scheme	Class EF	Class AF	Class BE	Class TF
	300 / 30%	400 / 40%	200 / 20%	200 / 20%
	200 / 20%	300 / 30%	300 / 30%	200 / 20%
	400 / 40%	200 / 20%	100 / 10%	300 / 30%

Table 1

3.1 Relative Analysis through Gantt Charts

Network Traffic Schemes for scheduling (no class Solution) and RFPECB algorithm:-

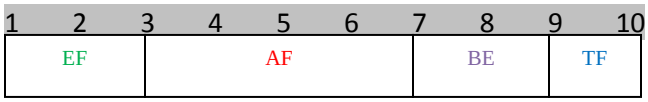
In no class solution no queue is implemented so bits have to move on the basis of their appearance, while in FPECB algorithm always EF 1st and AF, BE and TS base own priority which is less move 2nd.
In RFPECB delays are less than others.
In No class solution, I checked the delay through any given order assigns:-

Order No. 1 AF > EF > BE > TF
Order No. 2 AF < EF < BE < TF
Order No. 3 EF > AF > BE > TF
Order No. 4 EF < AF < BE < TF
Order No. 5 BE > EF > AF < TF
Order No. 6 BE < EF < AF < TF
Order No. 7 TF > EF > AF > BE
Order No. 8 TF < EF < AF < BE
Order No. 9 AF > BE > EF > TF
Order No. 10 AF < BE < EF < TF
Order No. 11 EF > BE > AF > TF
Order No. 12 EF < BE < AF < TF
Order No. 13 BE > AF > EF < TF
Order No. 14 BE < AF < EF < TF
Order No. 15 TF > AF > EF > BE
Order No. 16 TF < AF < EF < BE

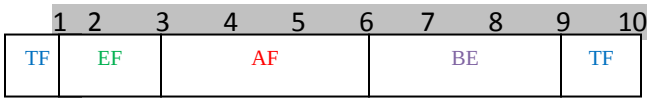
I checked the delay for scheduling algorithm through order No. 3. Because EF data is important than AF and AF data is important than BE and BE data is important than TF. Well, order No. 3 is better for comparison of RFPECB algorithm.

For No Class Solution

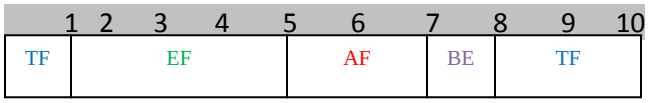
Order No. 3 enter



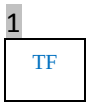
T1



T2



T3



T4

EF Delays = 2 µs

AF Delays = 11 µs

BE Delays = 20 µs

TF Delays = 25 µs

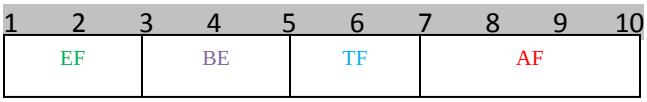
Gantt chart 1

Summary of Delays

Instances	EF (µs)	AF (µs)	BE (µs)	TF (µs)
1st	0	3	7	9
2nd	1	3	6	8
3rd	1	5	7	8

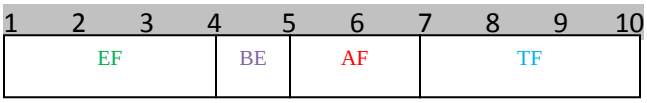
Table 1

For RFPECB Algorithm



T1

T2



T3

EF Delays = 0 µs

AF Delays = 16 µs

BE Delays = 14 µs

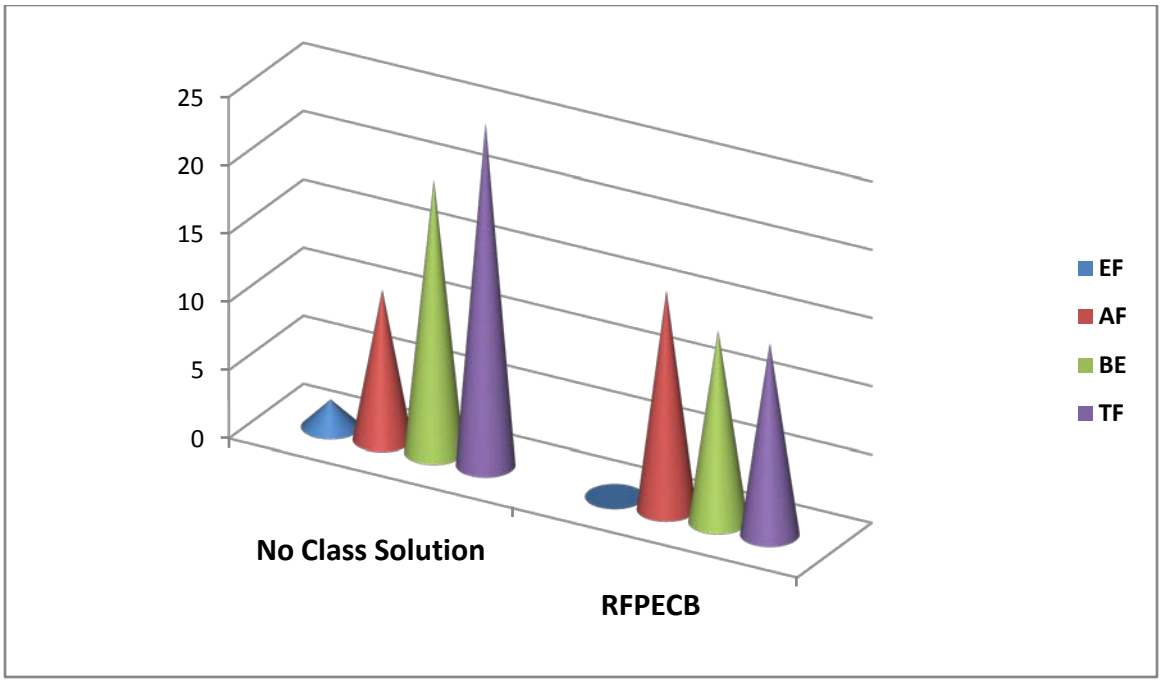
TF Delays = 14 µs

Gantt chart 2

Summary of Delays

Instances	EF (µs)	AF (µs)	BE (µs)	TF (µs)
1st	0	7	3	5
2nd	0	4	7	2
3rd	0	5	4	7

Table 2



Graph 1

4. Results and Argument

It is observable from the graphs that in **No class Solution,** data that arrives first, occupies the timeslot. So at rush hours our important data may get very high delays and our communication is disturbed very much, such as in case of voice and video conferencing in daily life.

RFPECB Algorithm is better than No Class Solution and because RFPECB algorithm eliminate the drawbacks such as Delay and QoS be eliminated allocating more bandwidth to the insistent data class.

Scheduling (No Class Solution) algorithm is compared with RFPECB algorithm

Delay of EF was high for "no class solution", because in this solution no priority is given to any class, delay for "FPECB" delay for EF is zero because data of EF moves always on first turn (Bandwidth fixed).

But AF, BE and TF based own priority phenomenon. Which priority is lesser than others move first.

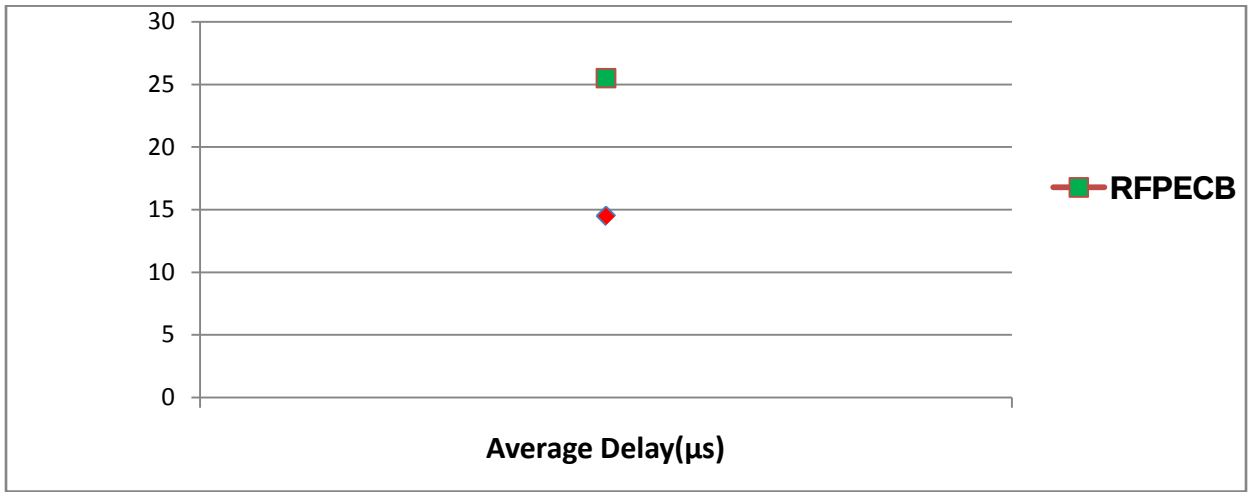
4.1 Total Delay Table & Graph:

Total Delay

Algorithms	EF Delays(µs)	AF Delays(µs)	BE Delays(µs)	TF Delays(µs)	Average Delays(µs)
No Class Solution	2	11	20	25	**14.5**
RFPECB	0	16	14	14	**11**

Table 3

Average Delays (µs) Graph



Graph 2

6. Conclusion

It is accomplished that, TDM EPON transmission in point to multi point networks is the productive technique because all components are passive. It will be better if it is changed according to RFPECB algorithm because the parameters which are affecting its QoS are handled in much better way in our solution. Hence, it is consummate that TDM EPON is better technology till now if it is used with a better scheduler such as RFPECB algorithm. Delay less than Scheduling algorithms.

7. Future Directions

In future, other professionals work on starvation to progress the better Quality of Services (QoS).

References

[1] "Interleaved Polling with Adaptive Cycle Time (IPACT): A Dynamic Bandwidth Distribution Scheme in an Optical Access Network" Ipact.pdf
By:- Glen Kramer, Biswanath Mukherje & Gerry Pesavento

[2]*"Fixed Advance Priority based Bandwidth Utilization in TDM EPON" in IJCSI May 2011 (Volume 8 Issue 3)* by **Muhammad Bilal**
By: Glen Kramer and Gerry Pesavento, Alloptic February 2002
[3] Ethernet PON (ePON): Design and Analysis of an Optical Access Network.

By: - Glen Kramer, Biswanath Mukherjee and Gerry Pesavento

[4] Ethernet Passive Optical Network (EPON) A Tutorial

By: The Metro Ethernet Forum 2005

http://www.metroethernetforum.org

About Author

I Muhammad Bilal was born on November 14, 1987 at Faisalabad, Pakistan. I belong to a family of educationists, who are offering the best of their services for the last four decades. My research interest is in Computer Networks. I recently got the B.Sc (Hons) Computer Science degree from UET, Lahore. I want to pay tribute to my parents who always encouraged, motivated me and are a constant source of inspiration throughout my academics. I would also like to thank all my teachers who always guide and help me to explore new world. Apart from all above playing and watching cricket boosted up my morale high up till sky. My favorite batsman's are Adam Gilchrist and Misbah-Ul-Haq.

A Routing Algorithm based on Cellular Automata for Mobile Ad-hoc Networks

Azadeh Ghalavand[1], Ahmad khademzadeh[2], Arash Dana[3] and Golnoosh Ghalavand[4]

[1] Dept. of Computer. Eng, South Tehran Branch, Islamic Azad University Tehran, Iran

[2] Education and International Scientific Cooperation Dept, Iran Telecommunication Research Center

[3] Dept. of Elect. Eng. ,Central Tehran Branch, Islamic Azad University, Tehran, Iran

[4] Dept. of Computer. Eng, Science And Research Branch, Islamic Azad University, Tehran, Iran

Abstract

Mobile ad hoc networks (MANETs) are self organizing, adaptive and infrastructure less networks. Analyzing these networks is a complex task due to the high mobility and rapid topology change. Routing in MANETs is one of the challenging tasks. Although a lot of researches have done in this area, the usage of cellular automata in routing for MANETs has not been explored. The proposed routing algorithm is a new cellular automaton based routing algorithm for mobile ad hoc networks, which finds a route that not only has least number of hops but also supports Quality of service**.** The solution presented here, is based on selecting a delay constrained shortest path between source and destination as a best route by using cellular automata. A simulator has been developed to evaluate a routing protocol and the obtained results indicate that the efficiency of the proposed protocol especially in satisfying QoS requirements.

Keywords: *Mobile Ad-hoc Networks, Routing, Quality of Service, Cellular Automata.*

1. Introduction

Mobile ad hoc networks (MANETs)[1] are networks composed of a set of mobile nodes able to spontaneously communicate with each other over a wireless medium without any fixed infrastructure. The topology of such networks changes dynamically as the nodes are free to move about arbitrarily. MANETs are useful in many applications as virtual classrooms, military communications, emergency search, rescue operations, at airport terminals for workers to share files, and etc [2, 3] .

One of the challenging aspects in the study of these networks is the routing algorithm for data transmission between the nodes. Any routing algorithm has to take into account the highly dynamic nature of the mobile nodes in the network and the limited resources of such networks. Many routing protocols for MANETs has been proposed, such as: Ad hoc On Demand Vector (AODV) [4], Associativity-Based Routing (ABR) [5], Dynamic Source Routing (DSR)[6], Temporally Ordered Routing Algorithm (TORA)[7], and etc. They are mostly designed for best effort transmission without any guarantee of quality of Service. Quality of Service (QoS) models in mobile ad hoc networks become more and more required because more and more real time applications are implemented on the network. For considering QoS, extensions, often given in terms of bandwidth or delay, can be added to the messages used during the route discovery process. Several works have been done to provide QoS routing algorithm such as Quality of Service for Ad-hoc on Demand Distance Vector (QoS-AODV) [8] which is based on AODV and creates routes according to the QoS requirements of the applications, a brief survey about this issue can be found in[9,10].

Nowadays, many areas of research have been benefited with the theories related to the cellular automata (CA)[11], mainly the areas that need to deal with systems that are constantly changing or have a random behavior as physics (mainly thermodynamics), ecology, medicine ,computer science, and chemistry. Cellular Automata (CA) consist of n-dimensional lattice of cells with states that evolve according to a set of rules. These rules are defined by the current state of the cell and the possible states of it's adjacent cells . These discrete time decentralized systems are useful and efficient as a computation tool for describing complex systems like physical systems.[12, 13].

Real ad-hoc networks infrastructures are still very expensive. Therefore, most of the evaluations of new protocols are being made through simulation tools and Different modeling, [14,15].Recently, different authors [16,19,17,18] decided to solve the problems in ad hoc networks using CA. Subrata and Zomaya [16] address the location tracking/management problem. They use CA combined with a genetic algorithm to create an evolving parallel reporting cells planning algorithm. The genetic algorithm is used to discover the 'best' CA transition rules for the application. Their results showed that the

discovered CA rules describe near-optimal solutions to the reporting cells problem. In a related application, Kirkpatrick and Van Scoy [17] investigate the use of CA to model message broadcasting in highly mobile ad hoc networks. They point out some strengths of the CA approach (discrete time steps useful in defining rapidly changing systems, and the approach is conducive to the design of a GUI), and derive an upper bound on the time required by a broadcast algorithm to distribute a message across a network. A paper by Cunha et al. [18] shows the applicability of CA to simulate the topology control problem in sensor networks. They concluded that CA can be a worthwhile simulation tool for large WSNs, because it allows for the fast and objective verification of network properties such as degree of coverage and degree of connectivity. The approach presented in[19]is modeling some features of ad-hoc networks such as mobility, network coverage and routing in wireless ad-hoc networks using Cell-DEVS which is an extension to Cellular Automata in which each cell in the system is considered a DEVS model. Their research shows that routing protocols, such as AODV can be successfully mapped onto Cell-DEVS.

Although several researchers have attempted to use CA for modeling some aspects of ad hoc networks, they don't consider the problem of routing with adequate QoS in MANET. The objective of this work is to verify the applicability of using cellular automata to solve this problem in mobile ad-hoc networks. The solution presented is based on a modified version of classical Lee's Algorithm [13] to find out the shortest path that satisfies the delay constraint as a QoS metric between two communicating nodes on a network plane. In response to this need, we add an extension to the routing request message in AODV protocol to define delay constraints which must be met by nodes during the discovery procedure of route, so the route with the least number of hops that satisfied delay constraint is chosen as a best route. We define our protocol and also mobility model with the concept of the CA, then we evaluate the performance of it in comparison with AODV and QoS for AODV routing protocols (QoS-AODV) in terms of average delay, and packet delivery fraction. Simulations are presented in different delay constraints and mobility speeds .

The remainder of the paper is structured as follows. The main concept of cellular automata is presented in section 2, including their formal definition and some practical applications. Details of our routing protocol are given in Section 3. Section 4 describes the simulation environment and results. Finally, our concluding remarks and future work are presented in Section 7.

2. Cellular Automata Concept

Cellular automata (CA)[11] are interesting computation systems to study because of their simplicity and inherently parallel operation. Such systems have the potential to model the behavior of complex systems in nature. For this reason CA and related architectures have been studied extensively in the natural sciences, mathematics, and in computer science. They have been used as models of physical and biological phenomena, such as fluid flow, galaxy formation, earthquakes, and biological pattern formation. They have been considered as mathematical objects about which formal properties can be proved. They have been used as parallel computing devices, both for the high-speed simulation of scientific models and for computational tasks such as image processing.[21]

CA can be formally defined as a 4-tuple (L, S, N, f) where[18]:

- L is a regular grid. The elements that compose this grid are called cells.
- S is a finite set of states, S={0,1….,s-1}.
- N is a finite set (of size $|N| = n$) of neighborhood index, such that $\forall c \in N$ and $\forall k \in L : k + c \in L$
- $f : S^n \longrightarrow S$ is a transition function *where* n is the size of the neighborhood.

The grid can be in any finite number of dimensions (e.g. a line or a cube of cells) and the time basis of the system is synchronous (t = 0, 1, 2,…). A configuration $C_t : L \rightarrow S$ is a function that associates one state with one cell of the grid. The task of this function is to change the configuration C_t to a new configuration C_{t+1} (see Equation1).

$$C_{t+1}(k) = f(C_t(i) \mid i \in N(k)) \quad (1)$$

where $N(k)$ is the set of all neighbors of the cell k.

$$N(k) = i \in L \mid k - i \in N \quad (2)$$

Generally, the CA starts at time t = 0 with each of the cells in one of their n possible states. This global state is known as the initial configuration. All cells have their own state transition function, also called the local rule, whose input is the states of their neighbors, and output is one of the possible states. At time step t, each cell uses the states of its neighbors as input to the state transition function, and the output of the function is the cell's new state at time t + 1.The rule for updating the state of cells is the same for each cell and does not change over time, and the entire CA updates synchronously in this fashion, though exceptions are known. Fig.1 shows the evolution of a one-dimensional, binary-state cellular automaton with infinite cells. As you see the neighborhood of cells is radious-1 (r

= 1)which refers the neighbors of a cell are one cell from it . The two fundamental types of neighborhood in CA systems are *von Neumann* neighborhood and *Moor* neighborhood, these two standard radius-1 neighborhoods in a two-dimensional cellular grid are depicted in Figure 2. The Moore neighborhood consists of the 8 surrounding cells of the cell depending on whether or not the central cell is counted, and the von Neumann neighborhood consisting of the 4 cell array - defined as north, south, east and west; that are strictly adjacent to the central cell.

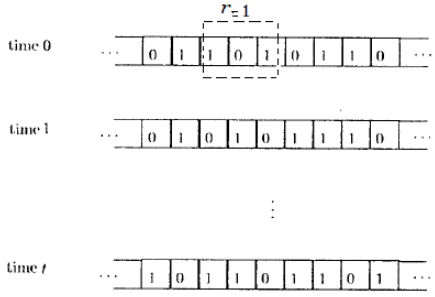


Fig.1. Evolution of 1-d, binary-state, nearest-neighbor (r =1) CA

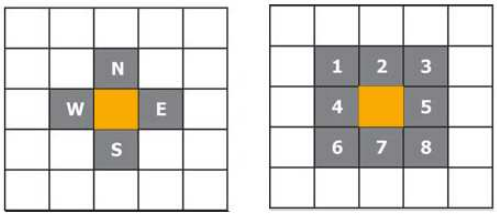


Fig. 2. Moor neighborhood (right) & von Neumann neighborhood (left) in Two- Dimensional CA.

2.1. Lee Algorithm

A very well known approach to routing problems is the Lee algorithm [20]. Its purpose is to find an optimal path between two points on a regular grid. The algorithm finds the path on the grid with the lowest sum of weights. By adjusting the weights of the grid points the user has some control over what is supposed to be an optimal path.

At a first glance this algorithm looks like it perfectly fits onto a cellular automaton. Unfortunately the number of states required to perform the algorithm is related to the longest path or more precisely to the largest accumulated weight that may occur. Thus in [13] the writers decided to develop a version of the algorithm which has a constant number of states. Hochberger and Hoffman [13] use a cellular processing model to simulate the Lee algorithm for the routing of connections on printed circuit boards. A modification was made to the Lee algorithm so that in a CA implementation it only requires 14 states. They postulate that mapping classical algorithmic problems onto the CA model may lead to new ideas in their theory and implementation.

The brief description of their algorithm is as follows:

The algorithm works in two phases and with fourteen states. At the beginning of the first phase all cells, except the start and the end point, are in the **free** state. During the first phase, all cells check whether there is any cell in the neighborhood that already has a wave mark. If a wave mark is found, the cell itself becomes a wave mark towards the already marked cell for forming a reverse path to the starting point. Fig.3 shows a sample grid at the end of phase one where S and D represent respectively, a sender and destination. The wave marks are symbolized by small arrows and the black squares are obstacles.

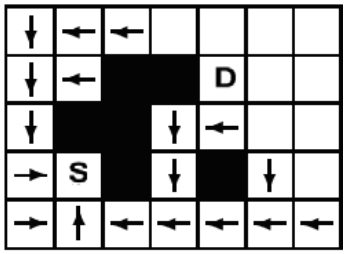


Fig.3. Sample grid at the end of phase 1

Phase one ends when the end point is reached by the wave. Now the path is built backward towards the start point along the wave marks**.** If a cell is one of the **wave** states and it sees a neighbor cell that is a path towards this cell, then this cell becomes a path in the direction of its previous mark (Fig.4).

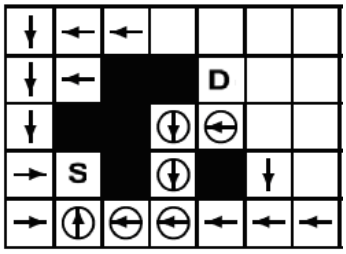


Fig.4 The path along the former marks

All wave marks from phase one have to be cleared in order to allow subsequent routing passes**.** For this purpose all cells that see a neighbor cell which is a path not pointing towards this cell are cleared. Such a cell will never become part of the path. Also all cells are cleared, that see a neighbor in the clear state. A cell that is in the **clear** state will become free in the next generation**.** Since the building of the path moves along the shortest path, it is impossible that a cell in **clear** state can reach a cell which will be in the path but is not yet part of it.

The algorithm terminates when the **start** cell sees one if its neighbors become part of the path. The **start** cell changes its state to ready and thereby signals that the path is complete**.**

3. MODELING the Routing problem using CA

In this section, we are going to discuss the methodology used to model the problem of finding route which has minimum hop counts and satisfies end to end delay requirement in cellular automata. We use the idea of QoS-AODV, a QoS version of AODV, which was proposed in 2000 by Perkins et al[8]. AODV is one of the first ad-hoc on demand routing algorithms chosen by IETF, that just find the shortest path without supporting any QoS attributes, but QoS-AODV establishes a route between source and destination by specifying Quality of Service parameters, namely maximum delay or minimum bandwidth .To give a more accurate definition of our objective, we first describe , an abstract description of this famous routing protocol.

3.1. Quality of Service for AODV

The AODV [4], a well known ad-hoc routing protocol, creates routes on demand. Whenever a node needs to communicate with another one, it broadcasts a Route Request message (RREQ) to its neighbors. They re-broadcast the message and set up a reverse path pointing towards the source. When the intended destination receives a RREQ message, it replies by sending a Route Reply (RREP) that travels along the reverse path set up by RREQ. AODV is based on the shortest path between two nodes in an ad hoc network plane and does not support QoS features. in contrast, QoS-AODV by considering quality-of-service (QoS), finds the best route that satisfies the end to-end QoS requirement. Infact, QoS-AODV modifies the route discovery and maintenance mechanisms of AODV to provide QoS assurance.

A QoS extension for AODV routing packets was proposed in [8]**.** This QoS object extension includes the Minimum bandwidth or Maximum delay parameters of each application, and it also has a "session *ID"* which is used to identify each QoS flow that is established according to the application. The extensions are added to the route messages (RREQ, RREP) during the phase of route discovery. A node which receives a RREQ with a quality of service extension must be able to meet the service requirement in order to either rebroadcast the RREQ, or unicast a RREP to the source. The maximum delay used in RREQ indicates the maximum time allowed to be used for a transmission from the current node to the destination and the minimum bandwidth extension specifies the minimum amount of bandwidth that must be made available along an acceptable path from the source to the destination.. When a node receives RREQ, the maximum delay will be compared with the node traversal time (should include queuing delays, interrupt processing times and transfer times) The NODE TRAVERSAL TIME is by default set to 40 milliseconds although could have a different value. If the maximum allowable delay is smaller than the node traversal time, the RREQ will not be broadcasted further, otherwise, the RREQ will be broadcasted and before the broadcasting, traversal time of this node must be subtracted from the maximum delay field in RREQ message. In RREP message, the maximum delay extension field will be filled with zeroes, and delays are cumulative from the destination along the route that the RREP come from to the source node, and the source node will choose the one satisfying the delay requirement. Similarly, the minimum bandwidth extension can be also appended to a RREQ by a requesting node. In this case, before forwarding the RREQ, an intermediate node must compare its available link capacity to the bandwidth field in the extension. If the requested amount of bandwidth is not available, the node must discard the RREQ and not process it any further. Otherwise, the node continues processing the RREQ as specified in [8].When the destination generates a RREP in response to a RREQ with a minimum bandwidth extension; the bandwidth field in the RREP is initially infinity. Each node forwarding the RREP compares the bandwidth field in the RREP and its own link capacity and maintains the minimum of the two in the bandwidth field of the RREP before forwarding the RREP. This value also goes in the route table entry for the corresponding destination (as per the destination IP address in the RREP) and indicates the available minimum bandwidth to the destination. It enables the intermediate node to respond to a later RREQ with a minimum bandwidth extension by comparing the requested minimum bandwidth and the available bandwidth recorded in the route table entry.

3.2. 1. Delay constraint Assumption

As a RREQ may include a maximum delay extension or a minimum bandwidth extension, for simplicity in this work, we only consider the delay constraint. So by giving the delay requirement as *Dmax*, the goal is to detect a path from source node *s* to destination node *d* such that the delay on the path does not exceed *Dmax*.

To control dissemination of RREQ with this QoS requirement (delay) the following assumptions are considered:

First,we assume 40 milisecond for NODE_TRAVERSAL_TIME .

Second, If the NODE_TRAVERSAL_TIME is greater than or equal to the maximum delay in delay field of RREQ, the intermediate node must discard the RREQ.

Third If the NODE_TRAVERSAL_TIME is less than the maximum (remaining) delay in message by subtracting from its value of the NODE_ TRAVERSAL_TIME, the

intermediate node should send a RREP to the originator or must continue broadcasting the RREQ .

4.3. Proposed Model

In order to find a path with adequate QoS between two communicating nodes in the network we consider only delay as the QoS constraint. we assume a two dimensional network plane consisting of a number of mobile nodes spread randomly . There is no hierarchy between nodes, and The network plane is homogeneous (all nodes have the same properties). Data movement between two cells represents one hop. Every node represents a cell of the cellular automaton. Each cell has a state which corresponds to the status of the node that occupies it. The neighborhood of each cell for finding the route, is the range of communication of each node defined as a Von-Neumann neighborhood governed by a circle of radius 1 around the cell .The network plane also contains *dead cells* through which communication cannot take place. These cells represent physical obstacles (such as a high-rise building) or simply the absence of a communication link. Two nodes with a dead cell between them cannot communicate directly to each other .As network nodes can move randomly to any of the neighboring cells the Moor neighborhood will be used for mobility. By using the modified Lee's Algorithm [13] with some variances, we define our routing method which satisfies the assumptions discussed in section 3.2.

In our method, for discovering a QoS path that satisfies the delay-constraint as the QoS parameter, the source node broadcasts QRREQ message with the delay requirement of the connection request (maximum delay(*Dmax*)), to its communicating neighbors which includes, all the nodes on its top, down, left and right side. According to lee algorithm,at first, the nodes which are in neighborhood of the source node become wave . But for providing a delay-constraint Dmax must be first subtract from NODE-TRAVERSAL-Time(NTT) in every intermediate node before becoming as a wave node. Therefore two conditions maybe happen:

1. If the NTT is greater than the maximum(remaining) delay in delay field of QRREQ, the intermediate node must drop the QRREQ, and don't become a wave node.
2. If the NTT is less than the maximum delay in message, by subtracting from its value of the NODE_ TRAVERSAL_TIME, the intermediate node will become wave node and continue broadcasting the RREQ.

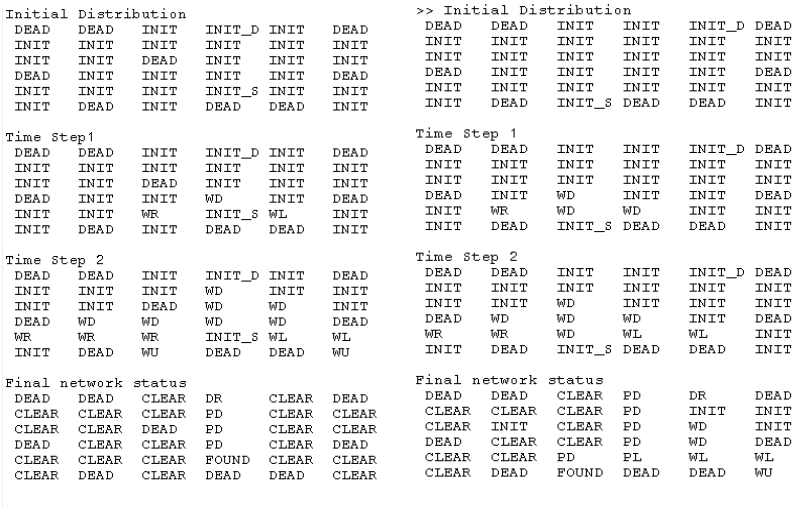
The wave nodes re-broadcast the message to their neighbors, and set up a reverse path to the sender, which is represented with pointing arrows in the Fig.4. These nodes further if provide delay constraint and have a wave node in their neighbors, become wave and re-broadcast this message and set up a reverse path to the nodes from which they received the message.

This process continues until the message reaches the destination node or the delay experienced by the packet exceeds the limit *Dmax*. Since there are more than one path from the sender to the destination, the destination may receive multiple QRREQ message for the same sender. However, the route through which the destination node receives the QRREQ message first is the shortest path between the sender and the destination which guarantees the Quality assurance, thus the destination replies to the first QRREQ message by sending a QRREP message using the reverse path set up when the QRREQ messages are forwarded. All the wave nodes that lie on this route between the source and the destination become the *path* nodes (represented with circles containing arrows in the Fig.4). All communications between the source and the destination from this point onwards takes place using this path until the topology of the network changes. All other wave nodes that don't see a path node in their neighbor pointing towards them, are sent a clear state message to move them from the *wave* state to a *clear* state.

The local states of each node are defined by adding one state as mentioned in [19] to the 14 states of the lee algorithm. These states are as follows:

Init (initial state of each node except source and destination), **Dead** (Dead Cell), **Init_S** (Initial State of the Source node), **Init_D** (initial state of the destination), **DR** (Destination Ready; state of the destination node after it has received a RREQ message), **WU** (Wave Up), **WD** (Down), **WR** (Right), **WL** (Left), **PU** (Path Up), **PD** (Path Down), **PR** (Path Right), **PL** (Path Left), **Clear** (final state of the node that received a wave message but is not going to become a path node)and **Found** (route found; final state of the Source node).With these set of states we consider Von-Neumann neighborhood (see Fig.2) in which each cell can communicate to all the cells on its top, down, left and right. The local transition rules for each cell based on these neighborhoods are defined according to the routing mechanism.

For implementing mobility we assume that nodes can move only to neighboring cells. A Moore neighborhood is chosen because it allows movement in more different directions than von Neumann neighborhood ,and as we are using a two-dimensional grid, each node can have no



```
Initial Distribution
 DEAD   DEAD   INIT   INIT_D INIT   DEAD
 INIT   INIT   INIT   INIT   INIT   INIT
 INIT   INIT   DEAD   INIT   INIT   INIT
 DEAD   INIT   INIT   INIT   INIT   DEAD
 INIT   INIT   INIT   INIT_S INIT   INIT
 INIT   DEAD   INIT   DEAD   DEAD   INIT

Time Step1
 DEAD   DEAD   INIT   INIT_D INIT   DEAD
 INIT   INIT   INIT   INIT   INIT   INIT
 INIT   INIT   DEAD   INIT   INIT   INIT
 DEAD   INIT   INIT   WD     INIT   DEAD
 INIT   INIT   WR     INIT_S WL     INIT
 INIT   DEAD   INIT   DEAD   DEAD   INIT

Time Step 2
 DEAD   DEAD   INIT   INIT_D INIT   DEAD
 INIT   INIT   INIT   WD     INIT   INIT
 INIT   INIT   DEAD   WD     WD     INIT
 DEAD   WD     WD     WD     WD     DEAD
 WR     WR     WR     INIT_S WL     WL
 INIT   DEAD   WU     DEAD   DEAD   WU

Final network status
 DEAD   DEAD   CLEAR  DR     CLEAR  DEAD
 CLEAR  CLEAR  CLEAR  PD     CLEAR  CLEAR
 CLEAR  CLEAR  DEAD   PD     CLEAR  CLEAR
 DEAD   CLEAR  CLEAR  PD     CLEAR  DEAD
 CLEAR  CLEAR  CLEAR  FOUND  CLEAR  CLEAR
 CLEAR  DEAD   CLEAR  DEAD   DEAD   CLEAR
```

```
>> Initial Distribution
 DEAD   DEAD   INIT   INIT   INIT_D DEAD
 INIT   INIT   INIT   INIT   INIT   INIT
 INIT   INIT   INIT   INIT   INIT   INIT
 DEAD   INIT   INIT   INIT   INIT   DEAD
 INIT   INIT   INIT   INIT   INIT   INIT
 INIT   DEAD   INIT_S DEAD   DEAD   INIT

Time Step 1
 DEAD   DEAD   INIT   INIT   INIT_D DEAD
 INIT   INIT   INIT   INIT   INIT   INIT
 INIT   INIT   INIT   INIT   INIT   INIT
 DEAD   INIT   WD     INIT   INIT   DEAD
 INIT   WR     WD     WD     INIT   INIT
 INIT   DEAD   INIT_S DEAD   DEAD   INIT

Time Step 2
 DEAD   DEAD   INIT   INIT   INIT_D DEAD
 INIT   INIT   INIT   INIT   INIT   INIT
 INIT   INIT   WD     INIT   INIT   INIT
 DEAD   WD     WD     WD     INIT   DEAD
 WR     WR     WD     WL     WL     INIT
 INIT   DEAD   INIT_S DEAD   DEAD   INIT

Final network status
 DEAD   DEAD   CLEAR  PD     DR     DEAD
 CLEAR  CLEAR  CLEAR  PD     INIT   INIT
 CLEAR  INIT   CLEAR  PD     WD     INIT
 DEAD   CLEAR  CLEAR  PD     WD     DEAD
 CLEAR  CLEAR  PD     PL     WL     WL
 CLEAR  DEAD   FOUND  DEAD   DEAD   WU
```

(a) *(b)*

Fig.5*: Illustration of the 6*6 grid before mobility changes (a) and after movement of nodes (b).*

more than 8 nodes as neighbors (Fig.2). So in this experiment, a node can with deterministic speed, move around in random fashion to one of each surrounding cells, by selecting any direction of North, North-East, North-West, East, West, South-West, South-East and South. As cellular automata are discrete event time systems, in each time step, nodes move to one of neighboring cells and then routing algorithm is being performed. The execution of mobility rules is considered to be higher than routing rules, it means that first topology is changed then the process of finding the route between source and destination is performed.

Using this specification as a base, the model was implemented in mat lab. Fig.5, shows the procedure of our method after 4 steps before the topology changes(a), and after that (b). As you see in the figure, the destination node (INIT_D) and the source node (INIT_S) will move to the east and south-east, respectively. It's obvious that dead nodes which are represented as DEAD notations, are static and didn't change their location after the topology have changed. The figure also shows that the routing mechanism is performed well as the path with adequate route is discovered.

Our proposed routing protocol with the help of cellular automata doesn't need to store accumulated delay for calculating the delay in phase of unicasting RREP instead of QoS-AODV, by defining only constant set of states .This leads to simply model and evaluate the behavior of routing algorithms in MANET.

4. Simulation and Results

To evaluate the performance of our method, we simulated our proposed method via mat lab software[22].The model will be tested by having different initial distributions of the node and checking whether the algorithm successfully determines the delay-constraint path between two nodes (if one exists).

We consider a rectangular network of 500m×300m of mobile nodes placed randomly in the area. Our simulation has been conducted at different pause times and speeds. We varied the pause time from 0 to 600 seconds which is chosen to be between high mobility rate (pause time 0 seconds) and no mobility (pause time 600seconds), and studied its effect on the performance of the routing protocols. The average node speed in this group of simulations is chosen to be 5 m/s. we also studied the effect of mobility. Node speeds are 5, 10, 15, 20 m/s. The pause time used for these scenarios was 60seconds. For both scenarios our protocol is introduced with and without QoS delay constraint. Delay constraints were chosen to be 0.1 and 0.3 seconds.. our method's performance is compared with AODV and The original QoS-AODV protocol .

5.1. Performance metrics

Two performance metrics, which are used for evaluating the performance of the routing protocols, are listed below.

- **Average end-to-end Delay**

This is the average end-to-end delay of talking parties in the simulation and it includes all possible delays caused by buffering during route discovery latency, queuing at the interface queue, retransmission delays at the MAC, propagation and transfer times.

$$D = \frac{1}{S}\sum_{i=1}^{S} r_i - s_i$$

D: Average end-to-end delay S : Number of successfully received packets. i : Unique packet identifier. r_i : Time at which a packet with unique identifier i is received. Si: Time at which a packet with unique identifier i requests a route to be send.

- **Packet Delivery Fraction**

The fraction of successfully received packets, which will be survived while finding their destination. Successful packet delivery is calculated such that all data packets with unique identifier leaving the source MAC are counted and defined as originating packets. Received packet identifiers are compared to collected transmission data and each unique packet is counted once to ensure prevention of counting excess receptions, which are mainly caused by multiple paths as a result of mobility. The result is the average of the ratio of uniquely received and all uniquely transmitted packets as seen in the following equation.

$$F = \frac{1}{C}\sum_{f=1}^{C} \frac{R_f}{T_f}$$

F: Fraction of successfully delivered packets. ***C*** : Total number of flows, connection. ***f*** : Unique flow id. R_f : Count of unique packets received from flow ***f***. T_f : Count of packets transmitted to flow ***f***.

Fig.6 shows the average end to end delay under various pause times. According to the graph, the average delay of all the protocols decreases as the time increases. When there is no delay constraint, the performance of our proposed method and QoS-AODV protocol is better than AODV with little bit differences. By Using delay constraints both QoS protocols have always better delay than AODV because they force the network to satisfy certain delay constraints , so the delay achieved is always less than or equal to the delay required even for high delay constraints (low delay bound 0.1 seconds) .For both protocols on the average, the delay achieved is half that required, but the delay of our method at higher pause times is much lower than the QoS-AODV routing protocol. The low end to end delay of packets ensures the on time transmissions required by real time traffic transmissions. As is shown in Fig.7 the packet delivery fractions are almost the same for AODV protocol, QoS-AODV, and the proposed routing protocol without delay constraints and with high delay bounds (0.3 seconds).QoS-AODV routing protocol and our method with low delay bound, have low performance by having a low packet delivery fraction, however, our method has better ratio (about 1.2%). Indeed, with delay constraints especially in delay bound 0.1 seconds, more packets are being dropped because the routes available for them do not satisfy the QoS requirements.

Simulation results also show that when mobile nodes moved at higher mobility speeds, our method with no delay constraint has average 95% packet delivery fraction, which is almost similar to the QoS-AODV (with no delay constraint) and AODV (Fig.8).The performance of our method and QoS-AODV with delay constraints also drops to provide QoS assurance. The proposed method outperforms slightly QoS AODV with 0.1 and 0.3 delay bound.

Fig.9 shows the average delay in different speeds. From Fig.9 we notice that, higher mobility leads to a higher delay for all protocols. Although the average delay increases as the mobility changes, the delay achieved is still better than required for both our protocol and QoS-AODV with different delay bounds.

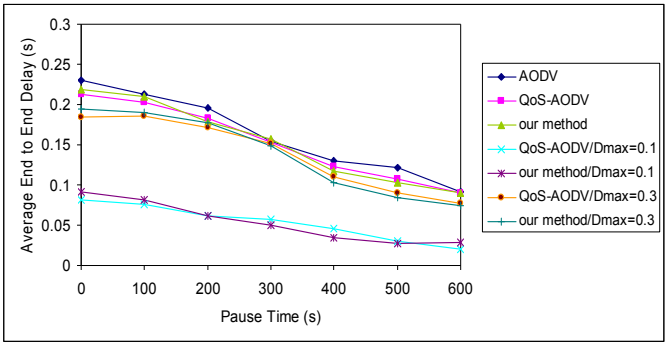


Fig.6.Average end to end delay versus pause time

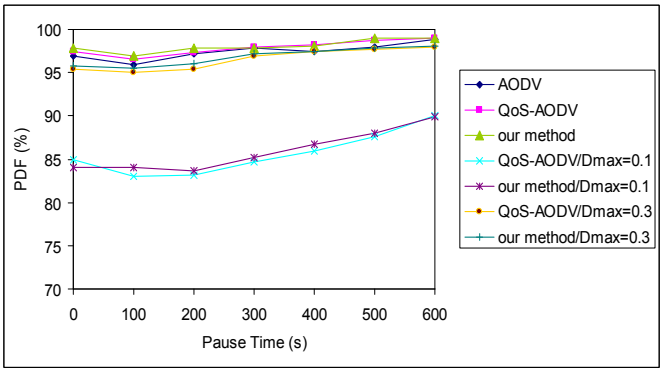


Fig.7. Packet delivery fractions versus pause time

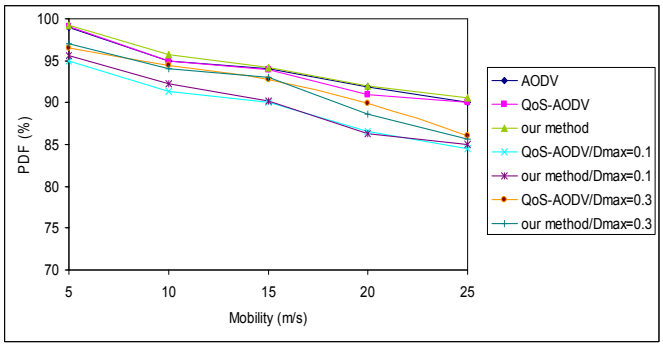


Fig. 8. Packet delivery fractions versus mobility

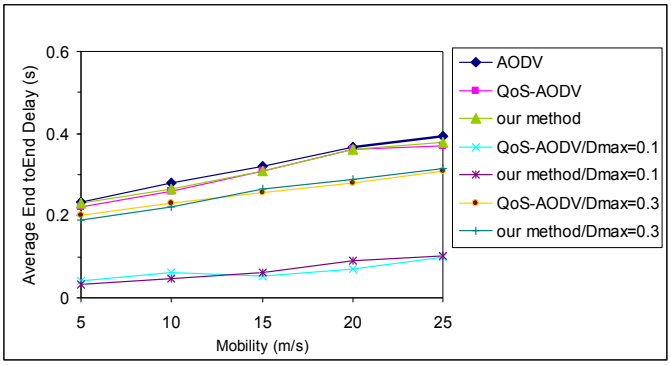


Fig.9. Average data packet delay versus mobility

7. Conclusion

Number of problems can be solved if they are described as cellular automata, one of the NP-complete problems is the problem of providing a shortest route with high quality of service among two communicating mobile nodes in MANET. In this paper we presented a new routing algorithm in mobile ad-hoc networks using cellular automata by taking delay into account as the QoS parameter. The variant of Lee routing algorithm for finding the shortest path between two points was modified into a more efficient form, which could find the on demand delay constraint path like QoS AODV but with lower consuming resources only by defining some infinite states. The Two following performance metrics are used to compare the performance of the protocols: (a) average end-to-end delay, (b) packet delivery fraction, in various pause times and mobility. The obtained results indicate that cellular automata can be used with success to simulate mobile ad-hoc networks and the proposed protocol performs well in supporting the QoS feature.

For further work , another extension like a bandwidth extension can be used to satisfy minimum bandwidth requirements and other optimization can be taken into consideration like reliability of the route.

References

[1] T.S. Rappaport. “Wireless Communications: Principles and Practice”, Second Ed., Prentice Hall, 2002

[2] Chen Niansheng,Li Layuan. Research on the Basis of QoS Routing Protocol of Ad Hoc Network. DCABES 2004Proceedings. Wuhan:Hubei Science and Technology Press,2004:206-210.

[3] Niansheng Chen, Layuan Li, Zongwu Ke. A Multicast Routing Algorithm of Multiple QoS for Mobile Ad Hoc Networks. International Symposium on Distributed Computing and Applications to Business, Engineering and Science, DCABES 2007, PROCEEDING: 301-305.

[4] C. Perkins, E. Belding-Royer, S. Das, "Ad-hoc On Demand Distance Vector (AODV) Routing", IETF Network Working Group, RFC 3561, July 2003.

[5] C.K. Toh, A novel distributed routing protocol to support ad-hoc mobile computing, Proceedings of the fifteenth IEEE Annual International Phoenix Conference on Computers and Communications, March, 1996 pp. 480–486.

[6] D.B. Johnson, D.A. Maltz, Dynamic Source Routing in Adhoc Wireless Networks, Kluwer, 1996.

[7] V. Park, M.S. Corson, A highly adaptive distributed routing algorithm for mobile wireless networks, Proceedings of the 1997 IEEE INFOCOM, Kobe, Japan, April, 1997 pp. 1405–1413.

[8] C. Perkins and E. Royer, "Quality of service for ad hoc on-demand distance vector routing," in Zntemet draft, draft-perkins-manet-aodvqos-02. txt, Oct. 2003. [Online]. Available: http://people.nokia.net/ charliep/txt/aodvid/qos.txt

[9] A.M. Abbas, O. Kure, "A Quality of Service in Mobile Ad-hoc Networks: A Survey", International Journal of Ad hoc and Ubiquitous Computing (IJAHUC), vol. 6, no. 2, pp.75-98, June 2010.

[10] G.Snthi and A.Nachiappan, "A Survey of QoS Routing Protocols for Mobile Ad hoc Networks",", International Journal of Computer Science and Information Technology (IJCSIT),vol.2,No.4, pp.125-135, August 2010.

[11] S. Wolfram "A new kind of science". Wolfram Media,Inc. 2002.

[12] B. Chopard and M. Droz. Cellular Automata Modeling of Physical Systems. New York: Cambridge Univ. Press 1998.

[13] C. Hochberger; R. Hoffmann. "Solving routing problems with cellular automata", in Proceedings of the Second conference on Cellular Automata for Research and Industry, Milan, Italy. 1996.

[14] L. Bajaj, M. Takai, R. Ahuja, K. Tang, R. Bagrodia, and M. Gerla. "GloMoSim: A Scalable Network Simulation Environment". UCLA Computer Science Department Technical Report 990027, May 1999.

[15] Kevin Fall, Kannan Varadhan, Eds. "The ns Manual (formerly ns Notes and Documentation)", http://www.isi.edu/nsnam/ns/. Accessed February 2004.

[16]R. Subrata and A.Y. Zomaya. "Evolving Cellular Automata for Location Management in Mobile Computing Networks". IEEE Trans. on Parallel and Distributed Systems 14:1 (Jan 2003), 13- 26.

[17] Michael Kirkpatrick and Frances L. Van Scoy. Using cellular automata to determine bounds for measuring the eficiency of broadcast algorithms in highly mobile ad hoc networks. In ACRI, pages 316-324, 2004.

[18] R.O. Cunha, A.P. Silva, A.A.F. Loreiro, and L.B. Ruiz. Simulating large wireless sensor networks using cellular automata. Simulation Symposium, 2005. Proceedings.38th Annual, pages 323{330, 4-6 April 2005.

[19] U. Farooq, Gabriel Wainer, and B. Balya, "DEVSmodeling of mobile wireless ad hoc networks", In Simulation Modelling Practice and Theory 15 of Elsevier,pp.285–314, December2007.

[20] C. Y. Lee, "An algorithm for path connections and its applications", in IRE Transaction on Electronic Computers, pp. 345-365, Sep. 1961.

[21] P. SARKAR, "A Brief History of Cellular Automata",ACM ACM Computing Surveys,Vol.32, No.1, pp.80-107,March 2000.

[22] The Matlab Simulator, http://www.mathworks.com.

IJCSI CALL FOR PAPERS JANUARY 2012 ISSUE

Volume 9, Issue 1

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: 30th November 2011
Notification: 04th January 2012
Revision: 12th January 2012
Online Publication: 31st January 2012

- 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