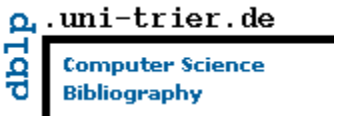
IJCSI

International Journal of Computer Science Issues

Volume 10, Issue 1, No 3, January 2013
ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

www.IJCSI.org

IJCSI proceedings are currently indexed by:



IJCSI Publicity Board 2013

IJCSI Editorial Board 2013

EDITORIAL

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

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

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

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

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

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

We are pleased to present IJCSI Volume 10, Issue 1, No 3, January 2013 (IJCSI Vol. 10, Issue 1, No 3). The acceptance rate for this issue is 33.09%.

IJCSI Editorial Board
January 2013 Issue
ISSN (Online): 1694-0814

www.IJCSI.org

IJCSI Reviewers Committee 2013

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

Mr. Prabu Mohandas, Anna University/Adhiyamaan College of Engineering, india

Dr. Manish Kumar Jindal, Panjab University Regional Centre, Muktsar, India

Prof Mydhili K Nair, M S Ramaiah Institute of Technnology, Bangalore, India

Dr. C. Suresh Gnana Dhas, VelTech MultiTech Dr.Rangarajan Dr.Sagunthala Engineering College,Chennai,Tamilnadu, India

Prof Akash Rajak, Krishna Institute of Engineering and Technology, Ghaziabad, India

Mr. Ajay Kumar Shrivastava, Krishna Institute of Engineering & Technology, Ghaziabad, India

Dr. Vu Thanh Nguyen, University of Information Technology HoChiMinh City, VietNam

Prof Deo Prakash, SMVD University (A Technical University open on I.I.T. Pattern) Kakryal (J&K), India

Dr. Navneet Agrawal, Dept. of ECE, College of Technology & Engineering, MPUAT, Udaipur 313001 Rajasthan, India

Mr. Sufal Das, Sikkim Manipal Institute of Technology, India

Mr. Anil Kumar, Sikkim Manipal Institute of Technology, India

Dr. B. Prasanalakshmi, King Saud University, Saudi Arabia.

Dr. K D Verma, S.V. (P.G.) College, Aligarh, India

Mr. Mohd Nazri Ismail, System and Networking Department, University of Kuala Lumpur (UniKL), Malaysia

Dr. Nguyen Tuan Dang, University of Information Technology, Vietnam National University Ho Chi Minh city, Vietnam

Dr. Abdul Aziz, University of Central Punjab, Pakistan

Dr. P. Vasudeva Reddy, Andhra University, India

Mrs. Savvas A. Chatzichristofis, Democritus University of Thrace, Greece

Mr. Marcio Dorn, Federal University of Rio Grande do Sul - UFRGS Institute of Informatics, Brazil

Mr. Luca Mazzola, University of Lugano, Switzerland

Mr. Hafeez Ullah Amin, Kohat University of Science & Technology, Pakistan

Dr. Professor Vikram Singh, Ch. Devi Lal University, Sirsa (Haryana), India

Dr. Shahanawaj Ahamad, Department of Computer Science, King Saud University, Saudi Arabia

Dr. K. Duraiswamy, K. S. Rangasamy College of Technology, India

Prof. Dr Mazlina Esa, Universiti Teknologi Malaysia, Malaysia

Dr. P. Vasant, Power Control Optimization (Global), Malaysia

Dr. Taner Tuncer, Firat University, Turkey

Dr. Norrozila Sulaiman, University Malaysia Pahang, Malaysia

Prof. S K Gupta, BCET, Guradspur, India

Dr. Latha Parameswaran, Amrita Vishwa Vidyapeetham, India

Ms. Habib Izadkhah, Tabriz University, Iran

Dr. Lokesh Kumar Sharma, Chhattisgarh Swami Vivekanand Technical University Bhilai, India

Mr. Kuldeep Yadav, IIIT Delhi, India

Dr. Naoufel Kraiem, Institut Superieur d'Informatique, Tunisia

Prof. Frank Ortmeier, Otto-von-Guericke-Universitaet Magdeburg, Germany

Mr. Ashraf Aljammal, USM, Malaysia

Mrs. Amandeep Kaur, Department of Computer Science, Punjabi University, Patiala, Punjab, India

Mr. Babak Basharirad, University Technology of Malaysia, Malaysia

Mr. Avinash singh, Kiet Ghaziabad, India

Dr. Miguel Vargas-Lombardo, Technological University of Panama, Panama

Dr. Tuncay Sevindik, Firat University, Turkey

Ms. Pavai Kandavelu, Anna University Chennai, India

Mr. Ravish Khichar, Global Institute of Technology, India

Mr Aos Alaa Zaidan Ansaef, Multimedia University, Cyberjaya, Malaysia

Dr. Awadhesh Kumar Sharma, Dept. of CSE, MMM Engg College, Gorakhpur-273010, UP, India

Mr. Qasim Siddique, FUIEMS, Pakistan

Dr. Le Hoang Thai, University of Science, Vietnam National University - Ho Chi Minh City, Vietnam

Dr. Saravanan C, NIT, Durgapur, India

Dr. Vijay Kumar Mago, DAV College, Jalandhar, India

Dr. Do Van Nhon, University of Information Technology, Vietnam

Dr. Georgios Kioumourtzis, Researcher, University of Patras, Greece

Mr. Amol D.Potgantwar, SITRC Nasik, India

Mr. Lesedi Melton Masisi, Council for Scientific and Industrial Research, South Africa

Dr. Karthik.S, Department of Computer Science & Engineering, SNS College of Technology, India

Mr. Nafiz Imtiaz Bin Hamid, Department of Electrical and Electronic Engineering, Islamic University of Technology (IUT), Bangladesh

Mr. Muhammad Imran Khan, Universiti Teknologi PETRONAS, Malaysia

Dr. Abdul Kareem M. Radhi, Information Engineering - Nahrin University, Iraq

Dr. Manuj Darbari, BBDNITM, Institute of Technology, A-649, Indira Nagar, Lucknow 226016, India

Ms. Izerrouken, INP-IRIT, France

Mr. Nitin Ashokrao Naik, Dept. of Computer Science, Yeshwant Mahavidyalaya, Nanded, India

Mr. Nikhil Raj, National Institute of Technology, Kurukshetra, India

Mr. Sunil Kashibarao Nayak, Bahirji Smarak Mahavidyalaya, Basmathnagar Dist-Hingoli., India

Prof. Nikhil gondaliya, G H Patel College of Engg. & Technology, India

Mr. Nisheeth Joshi, Apaji Institute, Banasthali University, India

Mr. Nizar, National Ingineering School of Monastir, Tunisia

Prof. R. Jagadeesh Kannan, RMK Engineering College, India

Prof. Rakesh.L, Vijetha Institute of Technology, Bangalore, India

Mr B. M. Patil, Indian Institute of Technology, Roorkee, Uttarakhand, India

Dr. Intisar A. M. Al Sayed, Associate prof./College of Science and IT/Al Isra University, Jordan

Mr. Thipendra Pal Singh, Sharda University, K.P. III, Greater Noida, Uttar Pradesh, India

Mrs. Rajalakshmi, JIITU, India

Mr. Shrikant Ardhapurkar, Indian Institute of Information Techonology, India

Ms. Hemalatha R, Osmania University, India

Mr. Hadi Saboohi, University of Malaya - Faculty of Computer Science and Information Technology, Malaysia

Mr. Sunil Kumar Grandhi, Maris Stella College, India

Prof. Shishir K. Shandilya, NRI Institute of Science & Technology, INDIA

Dr. Umesh Kumar Singh, Vikram University, Ujjain, India

Prof. Prasun Ghosal, Bengal Engineering and Science University, India

Dr. Nagarajan Velmurugan, SMVEC/Pondicherry University, India

Dr. R. Baskaran, Anna University, India

Dr. Wichian Sittiprapaporn, Mahasarakham University College of Music, Thailand

Mr. Lai Khin Wee, Universiti Teknologi Malaysia, Malaysia

Dr. Kamaljit I. Lakhtaria, Atmiya Institute of Technology, India

Mrs. Inderpreet Kaur, PTU, Jalandhar, India

Mr. Palaniyappan, K7 Virus Research Laboratory, India

Mr. Guanbo Zheng, University of Houston, main campus, USA

Mr. Arun Kumar Tripathi, Krishna Institute of Engg. and Tech-Ghaziabad, Affilated to UPTU, India

Mr. Iqbaldeep Kaur, PTU / RBIEBT, India

Mr. Amit Choudhary, Maharaja Surajmal Institute, New Delhi, India

Mrs. Vasudha Bahl, Maharaja Agrasen Institute of Technology, Delhi, India

Dr. Ashish Avasthi, Uttar Pradesh Technical University, India

Dr. Manish Kumar, Uttar Pradesh Technical University, India

Prof. Vinay Uttamrao Kale, P.R.M. Institute of Technology & Research, Badnera, Amravati, Maharashtra, India

Mr. Suhas J Manangi, Microsoft, India

Mr. Shyamalendu Kandar, Haldia Institute of Technology, India

Ms. Anna Kuzio, Adam Mickiewicz University, School of English, Poland

Mr. Vikas Singla, Malout Institute of Management & Information Technology, Malout, Punjab, India, India

Dr. Dalbir Singh, Faculty of Information Science And Technology, National University of Malaysia, Malaysia

Dr. Saurabh Mukherjee, PIM, Jiwaji University, Gwalior, M.P, India

Mr. Senthilnathan T, Sri Krishna College of Engineering and Technology, India

Dr. Debojyoti Mitra, Sir Padampat Singhania University, India

Prof. Rachit Garg, Department of Computer Science, L K College, India

Dr. Arun Kumar Gupta, M.S. College, Saharanpur, India

Dr. Todor Todorov, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria

Mrs. Manjula K A, Kannur University, India

Mrs. Sasikala R., K S R College of Technology, India

Prof. M. Saleem Babu, Department of Computer Science and Engineering, Vel Tech University, Chennai, India

Dr. Rajesh Kumar Tiwari, GLA Institute of Technology, India

Mr. Rakesh Kumar, Indian Institute of Technology Roorkee, India

Prof. Amit Verma, PTU/RBIEBT, India

Mr. Sohan Purohit, University of Massachusetts Lowell, USA

Mr. Anand Kumar, AMC Engineering College, Bangalore, India

Dr. Samir Abdelrahman, Computer Science Department, Cairo University, Egypt

Dr. Rama Prasad V Vaddella, Sree Vidyanikethan Engineering College, India

Dr. Manoj Wadhwa, Echelon Institute of Technology Faridabad, India

Mr. Zeashan Hameed Khan, UniversitÃ© de Grenoble, France

Mr. Arup Kumar Pal, Indian SChool of Mines, Dhanbad, India

Dr. Pouya, Islamic Azad University,Naein Branch, iran

Prof. Jyoti Prakash Singh, Academy of Technology, India

Mr. Muraleedharan CV, Sree Chitra Tirunal Institute for Medical Sciences & Technology, India

Dr. E U Okike, University of Ibadan, Nigeria Kampala Int Univ Uganda, Nigeria

Dr. D. S. Rao, Chitkara University, India

Mr. Peyman Taher, Oklahoma State University, USA

Dr. S Srinivasan, PDM College of Engineering, India

Dr. Rafiqul Zaman Khan, Department of Computer Science, AMU, Aligarh, India

Ms. Meenakshi Kalia, Shobhit University, India

Mr. Muhammad Zakarya, Abdul Wali Khan University, Mardan, Pakistan, Pakistan

Dr. M Gobi, PSG college, India

Mr. Williamjeet Singh, Chitkara Institute of Engineering and Technology, India

Mr. G.Jeyakumar, Amrita School of Engineering, India

Mr. Osama Sohaib, University of Balochistan, Pakistan

Mr. Jude Hemanth, Karunya University, India

Mr. Nitin Rakesh, Jaypee University of Information Technology, India

Mr. Harmunish Taneja, Maharishi Markandeshwar University, Mullana, Ambala, Haryana, India

Dr. Sin-Ban Ho, Faculty of IT, Multimedia University, Malaysia

Dr. Mashiur Rahman, Institute for Molecular Science, Japan

Mrs. Doreen Hephzibah Miriam, Anna University, Chennai, India

Mr. Kosala Yapa Bandara, Dublin City University, Ireland.

Mrs. Mitu Dhull, GNKITMS Yamuna Nagar Haryana, India

Dr. Chitra A.Dhawale, Professor, Symbiosis Institute of Computer Studies and Research, Pune (MS), India

Dr. Arun Sharma, GB Technical University, Noida, India

Mr. Naoufel Machta, Faculty of Science of Tunis, Tunisia

Dr. Utpal Biswas, University of Kalyani, India

Prof. Parma Nand, IIT Roorkee, India

Prof. Mahesh P K, Jnana Vikas Institute of Tevhnology, Bangalore, India

Dr. D.I. George Amalarethinam, Jamal Mohamed College, Bharathidasan University, India

Mr. Ishtiaq ahmad, University of Engineering & Technology, Taxila, Pakistan

Mrs. B.Sharmila, Sri Ramakrishna Engineering College, Coimbatore Anna University Coimbatore, India

Dr. Muhammad Wasif Nisar, COMSATS Institue of Information Technology, Pakistan

Mr. Prabu Dorairaj, EMC Corporation, India/USA

Mr. Neetesh Gupta, Technocrats Inst. of Technology, Bhopal, India

Dr. Ola Osunkoya, PRGX, USA

Ms. A. Lavanya, Manipal University, Karnataka, India

Dr. Jalal Laassiri, MIA-Laboratory, Faculty of Sciences Rabat, Morocco

Mr. Ganesan, Sri Venkateswara college of Engineering and Technology, Thiruvallur, India

Mr. V.Ramakrishnan, Sri Venkateswara college of Engineering and Technology, Thiruvallur, India

Prof. Vuda Sreenivasarao, St. Mary's college of Engg & Tech, India

Prof. Ashutosh Kumar Dubey, Assistant Professor, India

Dr. R.Ramesh, Anna University, India

Mr. Ali Khadair HMood, University of Malaya, Malaysia

Dr. Vimal Mishra, U.P. Technical Education, India

Mr. Ranjit Singh, Apeejay Institute of Management, Jalandhar, India

Mrs. D.Suganyadevi, SNR SONS College (Autonomous), India

Mr. Prasad S.Halgaonkar, MIT, Pune University, India

Mr. Vijay Kumar, College of Engg. and Technology, IFTM, Moradabad(U.P), India

Mr. Mehran Parchebafieh, Douran, Iran

Mr. Anand Sharma, MITS, Lakshmangarh, Sikar (Rajasthan), India

Mr. Amit Kumar, Jaypee University of Engineering and Technology, India

Prof. B.L.Shivakumar, SNR Sons College, Coimbatore, India

Mr. Mohammed Imran, JMI, India

Dr. R Bremananth, School of EEE, Information Engineering (Div.), Nanyang Technological University, Singapore

Prof. Vasavi Bande, Computer Science and Engneering, Hyderabad Institute of Technology and Management, India

Dr. S.R.Balasundaram, National Institute of Technology, India

Dr. Prasart Nuangchalerm, Mahasarakham University, Thailand

Dr. M Ayoub Khan, C-DAC, Ministry of Communications & IT., 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

Mr. Amin Daneshmand Malayeri, Young Researchers Club, Islamic AZAD University, Malayer Branch, Iran

Dr. Arti Arya, PES School of Engineering, Bangalore (under VTU, Belgaum, Karnataka), India

Mr. Pawan Jindal, J.U.E.T. Guna, M.P., India

Dr. Soumen Mukherjee, RCC Institute of Information Technology, India

Dr. Hamid Mcheick, University of Qubec at Chicoutimi, Canada

Dr. Mokhled AlTarawneh, PhD computer engineering/ Faulty of engineerin/ mutah university, jordan

Prof. Santhosh.P.Mathew, Saintgits College of Engineering, Kottayam, India

Ms. Suman Lata, Rayat Bahara institue of engg. & Nanotechnology,Hoshiarpur, India

Dr. Shaikh Abdul Hannan, Vivekanand College, Aurangabad, India

Prof. PN Kumar, Amrita Vishwa Vidyapeetham, India

Dr. P. K. Suri, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, India

Dr. Syed Akhter Hossain, Daffodil International University, Bangladesh

Mr. Sunil, Vignan College, India

Mr. Ajit Singh, TIT&S Bhiwani, Haryana, India

Mr. Nasim Qaisar, Federal Urdu Univetrsity of Arts , Science and Technology, Pakistan

Ms. Rshma, Maharishi Markandeshwar University, India

Mr. Gaurav Kumar Leekha, M.M.University, Solan (Himachal Pardesh), India

Mr. Ordinor Tucker, Ministry of Finance Jamaica, Jamaica

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

Dr. G R Sinha, SSCET, India

Mr. Chetan Sharma, TechMahindra India Ltd., India

Dr. Nabil Mohammed Ali Munassar, University of Science and Technology, Yemen

Prof. T Venkat Narayana Rao, Department of CSE, Hyderabad Institute of Technology and Management , India

Prof. Vasavi Bande, HITAM, Engineering College, India

Prof. S.P.Setty, Andhra University, India

Dr. C. Kiran Mai, J.N.T.University,Hyderabad/VNR Vignana Jyothi Institute of Engineering & Technology/, India

Ms. Bindiya Ahuja, Manav Rachna International University, India

Mrs. Deepa Bura, Manav Rachna International University, 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

Mr. Abhishek Shukla, RKGIT, India

Prof. Manpreet Singh, M. M. Engg. College, M. M. University, Haryana, India

Mr. Syed Imran, University College Cork, Ireland

Dr. Intisar Al Said, Associate Prof/Al Isra University, Jordan

Dr. Namfon Assawamekin, University of the Thai Chamber of Commerce, Thailand

Dr. Shiv KUmar, Technocrat Institute of Technology-Bhopal (M.P.), India

Dr. Shahaboddin Shamshirband, Islamic Azad University, Iran

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

Prof. Jakimi, Faculty of Science and technology my ismail University, Morocco

Dr. R. Manicka Chezian, N G M College, Pollachi - 642 001, Tamilnadu, India

Dr. P.Dananjayan, Pondicherry Engineering College, India

Mr. Manik Sharma, Sewa Devi SD College Tarn Taran, 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

Mr. Mueen Uddin, Universiti Teknologi Malaysia, Malaysia

Mr. Manoj Gupta, Apex Institute of Engineering & Technology,Jaipur (Affiliated to Rajasthan Technical University,Rajasthan), Indian

Mr. S. Albert Alexander, Kongu Engineering College, India

Dr. Shaidah Jusoh, Zarqa Private University, Jordan

Dr. Dushmanta Mallick, KMBB College of Engineering and Technology, India

Mr. Santhosh Krishna B.V, Hindustan University, India

Dr. Tariq Ahamad Ahanger, Kausar College Of Computer Sciences, India

Dr. Chi Lin, Dalian University of Technology, China

Prof. VIJENDRA BABU.D, ECE Department, Aarupadai Veedu Institute of Technology, Vinayaka Missions University, India

Mr. Merugu Gopichand, JNTU/BVRIT, India

Mr. Rakesh Kumar, M.M. University, Mullana,Ambala, India

Mr. M. Sundar, IBM, India

Prof. Mayank Singh, J.P. Institute of Engineering & Technology, India

Dr. Saurabh Pal, VBS Purvanchal University, Jaunpur, India

Mr. Khaleel Ahmad, S.V.S. University, India

Mr. Amin Zehtabian, Babol Noshirvani University of Technology / Tetta Electronic Company, Iran

Mr. Rahul Katarya, Department of Information Technology , Delhi Technological University, India

Dr. Vincent Ele Asor, University of Port Harcourt, Nigeria

Ms. Prayas Kad, Capgemini Australia Ltd, Australia

Mr. Alireza Jolfaei, Faculty and Research Center of Communication and Information Technology, IHU, Iran

Mr. Nitish Gupta, GGSIPU, India

Dr. Mohd Lazim Abdullah, University of Malaysia Terengganu, Malaysia

Ms. Suneet Kumar, Uttarakhand Technical University/Dehradun Institute of Technology, Dehradun, Uttarakhand, India

Mr. Rupesh Nasre., Indian Institute of Science, Bangalore., India.

Mrs. Dimpi Srivastava, Dept of Computer science, Information Technology and Computer Application, MIET, Meerut, India

Dr. Eva Volna, University of Ostrava, Czech Republic

Prof. Santosh Balkrishna Patil, S.S.G.M. College of Engineering, Shegaon, India

Mr. Mohd Dilshad Ansari, Jaypee University of Information Technology Solan (HP), India

Mr. Ashwani Kumar, Jaypee University of Information Technology Solan(HP), India

Dr. Abbas Karimi, Faculty of Engineering, I.A.U. Arak Branch, Iran

Mr. Fahimuddin.Shaik, AITS, Rajampet, India

Mr. Vahid Majid Nezhad, Islamic Azad University, Iran

Ms. C. Divya, Dr G R Damodaran College of Science, Coimbatore-641014, Tamilnadu, India

Prof. D. P. Sharma, AMU, Ethiopia

Dr. Sukumar Senthilkumar, School of Mathematical Sciences, Universiti Sains Malaysia, Malaysia

Mr. Sanjay Bhargava, Banasthali University, Jaipur, Rajasthan, India

Prof. Rajesh Deshmukh, Shri Shankaracharya Institute of Professional Management & Technology, India

Mr. Shervan Fekri Ershad, shiraz international university, Iran

Dr. Vladimir Urosevic, Ministry of Interior, Republic of Serbia

Mr. Ajit Singh, MDU Rohtak, India

Ms. Shiva Prakash, Madan Mohan Malaviya Engg. College, Gorakhpur, India

Dr. Kristijan Kuk, School of Electrical Engineering and Computer Science Applied Studies, Belgrade, Serbia

Prof. Dinesh Vitthalrao Rojatkar, Govt. College of Engineering, Chandrapur, India

Prof. Lalji Prasad, RGTU/TCET, Indore, india

Dr. A. John Sanjeev Kumar, Thiagarajar College of Engineering, Madurai, Tamilnadu, India

Mr. Harishbabu Kalidasu, Priyadarshini Institute of Technology and Science, Tenali, Guntur(DT), Andhra Pradesh, India

Prof. Vaitheeshwaran, Priyadharshini Indira Gandhi College of Engineering, India

Mrs. P.Salini, Pondicherry Engineering College, India

Mr. Vivek Bhambri, Desh Bhagat Institute of Management and Computer Sciences,Mandi Gobindgarh(Punjab), India

Mr. Slavko Zitnik, Faculty of Computer and Information Science Ljubljana, Slovenia

Ms. Sreenivasa Rao, CMJ University/Yodlee Infotech, India

Mr. Shihabudheen P M, TATA ELXSI LTD, India

Dr. Ahmed Moustafa Elmahalawy, Faculty of Electronics Engineering, Computer Science and Engineering, Egypt

Mr. Kamlesh Kumar, Kumaun University, Nainital, India

TABLE OF CONTENTS

Personality Types Classification for Indonesian Text in Partners Searching Website Using Naïve Bayes Methods

Ni Made Ari Lestari[1], I Ketut Gede Darma Putra[2] and AA Ketut Agung Cahyawan[3]

[1]Department of Information Technology, Udayana University
Bali, 80119, Indonesia

[2]Department of Information Technology, Udayana University
Bali, 80119, Indonesia

[3]Department of Information Technology, Udayana University
Bali, 80119, Indonesia

Abstract
The development of digital text information has been growing fast, but most of digital text is in unstructured form. Text mining analysis is needed in dealing with such unstructured text. One of the activities important in text mining is text classification or categorization. Text categorization itself currently has a variety of approaches such as probabilistic approaches, support vector machines, and artificial neural network or decision tree classification. Naive Bayes probabilistic method has several advantages of simplicity in computing. Naïve Bayes method is a good method in machine learning based on training data using conditional probability as the basic. This experiments use text mining with Naïve Bayes method to classify the personality type of user and use the type to find their couples based on the compatibility of their personality type.
Keywords: *text mining, classification, personality, naive bayes*

1. Introduction

Development of science and computer technology has given an enormous influence in Information technology's world, thereby encouraging the appearance of various types applications, such as desktop, web, or mobile. Among the three applications, web is the most rapidly progressing now, that's make internet has become a primary requirement. Percentage of internet users today is very large. Almost all people know and use the internet for daily needs. Starting from simple things such as communication, social networking to business. About 85% of the data available on the internet has an unstructured format, so it needs to be developed a system that is able to automatically categorize and classify the data is not structured [1]. Automatic text categorization is one of the solutions to the problem because they can significantly reduce the cost and time manual categorization. The abundance of information unstructured text has encouraged the appearance of a new discipline in text analysis, namely text mining that tries to find patterns of information that can be extracted from a text that is not structured. By that understanding the text mining term refers also to the text data mining (Hearst, 1997) or knowledge discovery from text databases (Friedman and Dagan, 1995). Text mining can provide a solution to the problem of processing, organizing, and analyzing the unstructured data in large numbers. According to Saraswati (2011), the current text mining has gained attention in many areas, such as security application, biomedical applications, software and applications, online media, marketing applications, and academic applications. [2]

Documents classification based on similarities features or content of the document. Classification is done by entering documents into categories predetermined. That classification method is called supervised learning. Generally, the method of classification divided into two, are supervised learning and unsupervised learning. First, supervised learning is a method of grouping documents, which class or category of documents predefined; whereas unsupervised learning is clustering documents automatically without define a category or class first. [3]

From numerical based approach group, Naïve Bayes has several advantages such as simple, fast and high accuracy. Naïve Bayes for classification or categorization of text using word attributes that appear in a document as a basis for classification. Some research showed that although the assumption independence between words in a document is not fully met, but performance in the NBC classification is relatively very good. Previous experiments results showed the accuracy of Naive Bayes is to reach 90%.[4]

Allport (1937) defined Personality as the dynamic organization within the individual of those psychophyscal systems that determine his unique adjustment to his environment. Temperament appears from our genetic endowment and influences or is influenced by the experience of each individual, and one of its outcomes is the adult personality [5]. There are many theories about personality. The most commonly known personality theory is the theory of the four temperaments from Hippocrates. Hippocrates divided the human temperaments into 4 big categories. Each category can be mixed and have a dominant trait in the

human body and form a mixed personality. It also has a match temperament between temperaments that can be used to determine a match between human beings who have different temperaments. [6]

2. Previous Research

Research related to text mining using Naïve Bayes method with several research objects as follows:

The study entitled Application of Naive Bayes for classification SMS Customer's Voice (Case Study PT. Pertamina UPMS V Surabaya). The raised issued is how to implement Naive Bayes in classifying SMS customers voice into categories determined by PT. Pertamina UPMS V Surabaya and classify SMS customers voice based on department which is determined by PT. Pertamina UPMS V Surabaya. In the Naive Bayes algorithm, SMS data voice subscribers in the past, will be entered for the training process that will result in probabilistic models. This research use 40 learning document and 40 classification document. And the result for accuration rate is 97,5%. [7]

The study entitled text Mining with Naïve Bayes Method Classifier and Support Vector Machines for Sentiment Analysis. Test performed to compare the use of Naive Bayes and SVM for Sentiment Analysis. Sentiment Analysis is a computational studies of the opinions of people, appraisal and emotion through entities, events and attributes owned (Biu, L. 2010). In this research is used the Indonesian and English documents. Each data has positive and negative values, each of which will be tested by the method of NBC and SVM. From the test results that the SVM can provide good results for the positive test data and NBC gave good results for the negative test data. [8]

The study entitled Text Document Keywords Extraction Using Naïve Bayes Method. Tests conducted to determine the influence the use of two features (TFxIDF and PD) and 4 features (TFxIDF, PD, PT, and PS) on the accuracy of the system generated keywords. The first test conducted on 10 documents at 20, 30, and 40 documents training with stopwords elimination. The second test performed on 10 documents of different tests at 20 and 30 training documents the elimination of stopwords. Then see the results by preccicion values, recall, and f-measure. The result is training documents provide the value of certain tendencies how should the value of the features of the keyword, with more and more features that use the word (which is the keyword) the value of the probability becomes greater keywords and words (which rather than keywords) decreases the probability values. [9]

The study entitled Spam Email Classification with Naïve Bayes Classifier Method use Java Programming. This study tested the validity of a document whether or not including spam. The accuracy of the test results obtained the error rate when categorizing spam use NBC. The biggest error rate is when the training data used reaches 40. That is because the difference the number of keywords in the second category is too much. So that lead to a greater level of error than others. [10]

3. Methodology

The overview diagram of this research is shown in Figure 1.

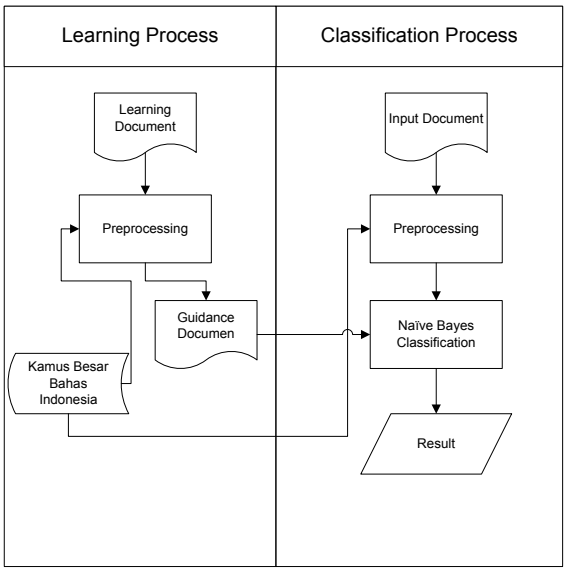


Figure 1 Research Overview Diagram

3.1 Preprocessing Text

Text preprocessing phase has been showed in figure 2.

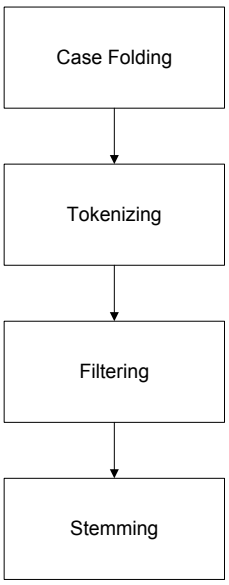


Figure 2 Text Processing Step

1. Case folding is the phase of changing uppercase to lowercase in the document then the elimination of

punctuation other than the "a" to "z" letter which is considered as the delimiter character.

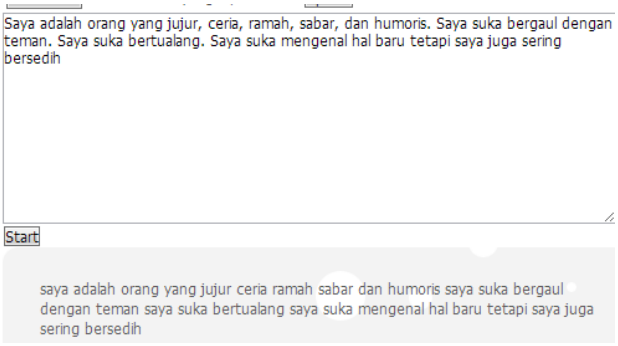


Figure 3 Case Folding Process

2. Tokenizing is the phase of splitting sentence to words. With the word's splitting first, the string that has been input will be simpler because showed in each words according to space which split it, so with that form, will make easier the changing process to be a word stem.

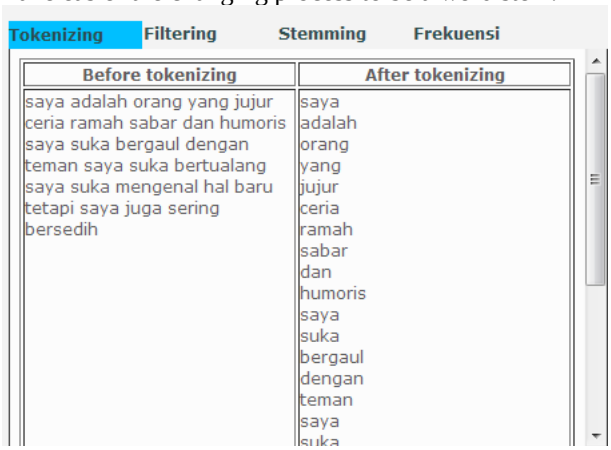


Figure 4 Tokenizing Process

3. Filtering is the phase of removal the words is not considered contain any meaning or thought there should be exist (Stopwords). Words in the stopwords list must be removed.

Figure 5 Filtering Process

4. Stemming is the phase of disposal affixes the words, either a prefix or a suffix. The flowchart for stemming process as seen in figure 6.

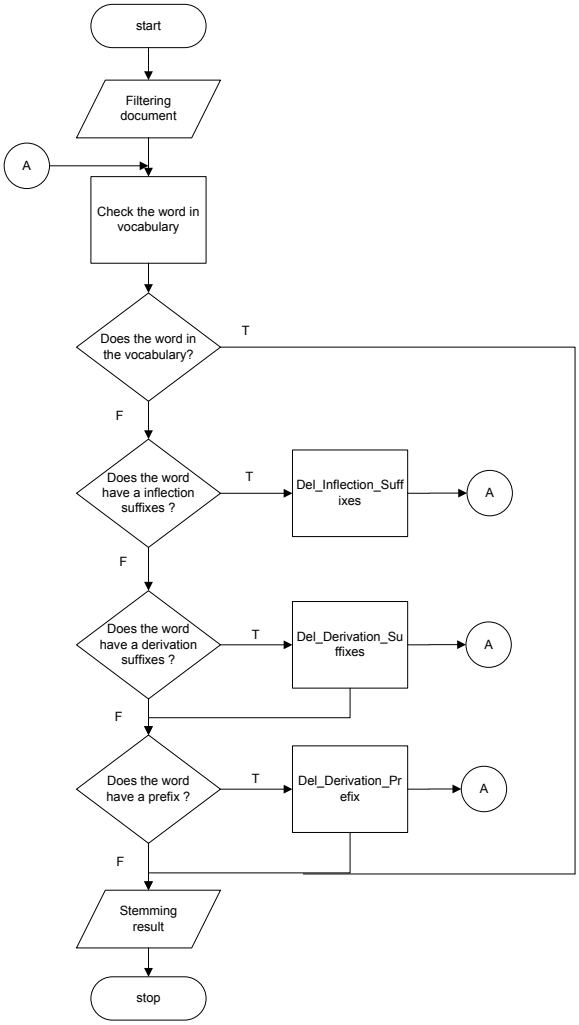


Figure 6 Stemming Algorithm

Process begins with the entry of input filtering results before. Then go into the process of checking the vocabulary. If the word entered is already contained in the vocabulary of the word is to be output directly to the process of stemming, whereas if not, the words is going through the process of checking further. In the program, words that do not qualify in checking vocabulary will undergo three processes, namely:

1. **Delete inflection suffixes** process is words removal process that have the suffix "-lah", "-kah", "-ku", "-mu", or "-nya". for example if there is a word "sebelumnya", in this process the suffix "-nya" in the word "sebelumnya" is removed, so that the results is "sebelum".

2. **Delete derivation suffixes** process is words removal process that have the suffix "-i", "-an" or "-kan". for example if there is the word "pukuli" in this process, the suffix "-i" in the word "pukuli" will be removed, so that the results is "pukul".

3. **Delete prefix derivation** process is words removal process that have the prefix "di-", "ke-", "se-", "te-",

“ber-”, “me-”, or "pe-". for example if there is the word "dibaca", in this process the prefix "di-" in the word "dibaca" will be removed so that the result is "baca". In some words, prefixes can change the form. For example, for the prefix "me-" could turn out to be "mem-", "meng-", "menge-", "menye-", "mempe-", "men-", "meny-", and prefix “pe-“ could turn out to be “per-“, “pem-“, “pen-“, peng-“, “penge-“, “peny-“, pel-“,and else depending on the first letter from the word.

After all words through the process above, the output is stemming results in the form of word stem. For the real example, the input of the filtering process is the word "menyesali". First, system checks whether the word "menyesali" already exists in the database vocabulary. If it is true it will be output directly, but in this case, the word "menyesali", not in the vocabulary database, then the next process is delete inflection suffixes. System check, if the word "menyesali" having the suffix "-lah”, “-kah”, “-ku”, “-mu”, or "-nya". If true, then the word "menyesali" will have the suffix deletion. Yet in the word "meyesali" is no inflection suffix, then process further to delete the derivation suffixes. System checks whether the word "menyesali" having the suffix "-i", "-an" or "-kan". If it is false, then the system will go directly to the next process. Yet in this case, the word "menyesali" there is the suffix "-i", the suffix will undergo a process of elimination. Results obtained from this process in the form of the word "menyesal". Furthermore, the system checks whether the word "menyesal" was in the database, if it is true then the system will go directly to the output. Because it is false, then the process continues to delete prefix derivation. The next process is the delete derivation prefix. System checks whether the word "menyesal" has a prefix. if it is false, the system will immediately to output, but in this case, the word "menyesal" has a prefix, the "me" that change form to "meny-“ (me + sesal = menyesal, according to the Indonesian dictionary), the word “menyesal” having replacement prefix . The prefix "meny-" replaced with vocal alphabets (aiueo) or the letter "s-' that one by one matched to the database vocabulary. Because the word that existing in database is “sesal”, then the output that comes out is the word “sesal”. After that, the process stops.

The example of stemming process in this program can be seen in figure 7.


Tokenizing Filtering Stemming Frekuensi

Filtering	Word Stem
jujur	jujur
ceria	ceria
ramah	ramah
sabar	sabar
humoris	humoris
bergaul	gaul
bertualang	tualang
mengenal	kenal
bersedih	sedih

Figure 7 Stemming Process

3.2 Classification with Naïve Bayes

Naïve bayes method consist of two phases, they are learning phase and classification phase.

1. Learning phase is the phase where the document preprocessing result through the learning process to get a learning data. This process is used to get probabilistic value from P(Vj) and P(Wk|Vj). Flowchart of learning process can be seen in figure 6.

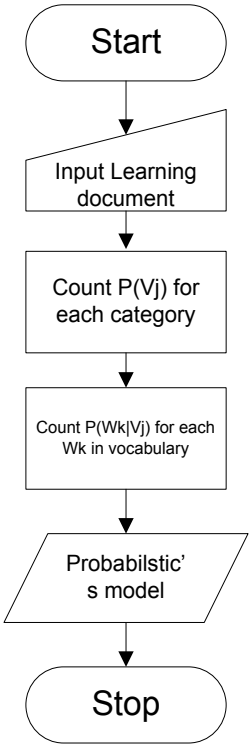


Figure 8 Naïve Bayes Learning Process

The process of learning begins with the input is the learning document then start the forming of vocabulary. Vocabulary is the set of all the unique words of the data training which then the amount being calculate. Furthermore, calculating P (Vj) for each category using the formula:

$$P(V_j) = \frac{|fd(Vj)|}{|D|} \quad (1)$$

Which is fd (Vj) is the number of words in the category j and D is the number of documents used in training. Furthermore, calculating P (Wk | Vj) for each Wk in the vocabulary with formula:

$$P(W_k \mid V_j) = \frac{f(W_k \mid V_j) + 1}{N + |W|} \quad (2)$$

Where P (Wk | Vj) is the amount of occurrences of word wk in the category Vj, N is the amount of all words in

the category Vj and |W| is the number of unique words (distinct) on all training data.

2. The classification phase is the phase where the new document will undergo a process of classification based on data previously coached there. Flowchart for the classification phase can be seen in the Figure 7.

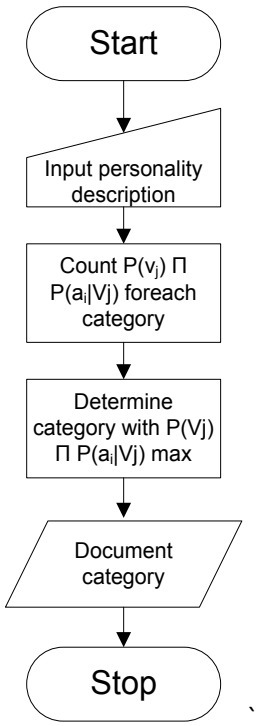


Figure 9 Naïve Bayes Classification Process

In the classification process, the input is personality documents and probabilistic model that has generated in the learning phase. The next stage VMap calculated by the formula:

$$V_{MAP} = \underset{V_j \in V}{\arg\max} P(V_j) \prod_i P(Wk|V_j) \quad (3)$$

After obtained the calculation for each category, then selected categories with maximum VMap that used to classify the personality document. Personality document will be classified according to the categories that have the maximum VMap value.

3.3 Personality Types

According to a book written by Florence Littauer called Personality Plus, more than 400 years before Christ, Hippocrates, a physician and philosopher from Greece, suggested a theory of personality that says that there are basically four types of temperament, they are Sanguine, Choleric, Melancholic and Phlegmatic. Each personality based on Hippocrates theory formed by the bile. Then Galenus refine this theory by stating that the four liquid is present in the body in a certain proportion, whereby if one fluid is more dominant than the other liquids, the liquid can form a personality. Here are the personality types and their characteristics:

1. Sanguine has a cheery and light hearted personality traits, friendly, talkative, likes to smile, outgoing, personality type who would rather party.
2. Choleric personality characterized by a life of passion, hard, heart-flammable, great fighting spirit, optimistic, tough, irritable, regulators, authorities, vengeful, and serious.
3. Personality traits of melancholy have easily disappointed, small guts, grim, pessimistic, fearful, and stiff.
4. Personality Phlegmatic characterized dislike to rush, calm, not easily influenced, loyal, cool, peaceful, relaxed and patient.

In addition there are four mix personalities where there are two dominant types of the same personality. The personality mixture is:

1. Natural mixed personality is the mixed personality that has similar properties. Included are sanguine-Choleric and melancholy-Phlegmatic
2. Complementary mixed personality is the mixed personalities who blend the two are complementary. Included are Choleric-melancholic and sanguine-Phlegmatic
3. Opposite mixed personality is the mixed personality which is the two personality are contradictory. Included are sanguine-melancholic and Choleric-Phlegmatic.

3.4 Couple Compatibility by Type Personality

Everything will attract the opposite. In the personality's type, when there are two types of personalities met will find a match with one another. The cheerful sanguine will improve the life's spirit of melancholy as well as melancholy will make sanguine life more scheduled.

The peaceful phlegmatic dislike to be pressed, but if not, they never find what they want. Meanwhile, choleric is the people who quick to make a decisions, having a goal and diligent, so both of them will match each other.

4. Experiments and Results

Naïve Bayes Method is a supervised learning, so they need require prior knowledge to be able to taking a decision. The success rate of this method depending on initial knowledge that given.

For example, user input the data of personality, as follow: "**Saya adalah orang yang jujur, ceria, ramah, sabar, dan humoris. Saya suka bergaul dengan teman. Saya suka bertualang. Saya suka mengenal hal baru tetapi saya juga sering bersedih**".
That document will through the text mining process, the result will calculate with Naïve Bayes method as seen as table below.

Table 1. Result of Text Mining Process (1)

Category	$P(V_j)$	$P(W_k \mid V_j)$				
		jujur	ceria	ramah	sabar	humoris
Sanguine	1/4	1/200	2/200	2/200	1/200	2/200
Choleric	1/4	1/200	1/200	1/200	1/200	1/200
Melancholic	1/4	1/200	1/200	1/200	1/200	1/200
Phlegmatic	1/4	1/200	1/200	2/200	2/200	1/200

Table 2. Result of Text Mining Process (2)

Category	$P(V_j)$	$P(W_k \mid V_j)$			
		gaul	tualang	kenal	sedih
Sanguine	1/4	1/200	1/200	1/200	1/200
Choleric	1/4	1/200	2/200	1/200	1/200
Melancholic	1/4	1/200	1/200	1/200	1/200
Phlegmatic	1/4	2/200	1/200	1/200	1/200

After knowing the P(Vj) and P(Wk|Vj) then count the VMap for each category.

P(sanguine|document) = $1/4 \times 1/200 \times 2/200 \times 2/200 \times 1/200 \times 1/200 \times 1/200 \times 1/200$

$= 8 / (2{,}048 \, x \, 10^{21})$

$= \mathbf{3{,}09 \times 10^{-21}}$

P(choleric|document) = $1/4 \times 1/200 \times 1/200 \times 1/200 \times 1/200 \times 1/200 \times 2/200 \times 1/200 \times 1/200$

$= 2 / (2{,}048 \, x \, 10^{21})$

$= 0{,}977 \times 10^{-21}$

P(melancholic|document) = $1/4 \times 1/200 \times 1/200 \times 1/200 \times 1/200 \times 1/200 \times 1/200 \times 1/200$

$= 1 / (2{,}048 \, x \, 10^{21})$

$= 0{,}488 \times 10^{-21}$

P(phlegmatic|document) = $1/4 \times 1/200 \times 2/200 \times 2/200 \times 1/200 \times 2/200 \times 1/200 \times 1/200 \times 1/200$

$= 8 / (2{,}048 \, x \, 10^{21})$

$= \mathbf{3{,}09 \times 10^{-21}}$

After see the formula above, the category which has maximum VMap are sanguine and phlegmatic. That's mean the result of text mining with Naïve Bayes Method for the document above is Sanguin and Phlegmatic.

In system, the output of this program are personality types and their potential partner. First, before start using the method to classify the personality, the non registered user must register them. After that, they can login, and use this program.

Figure 10 shows the personality's paragraph which is written in text box area. Besides written the input, user can also input it through the file with .txt extension.

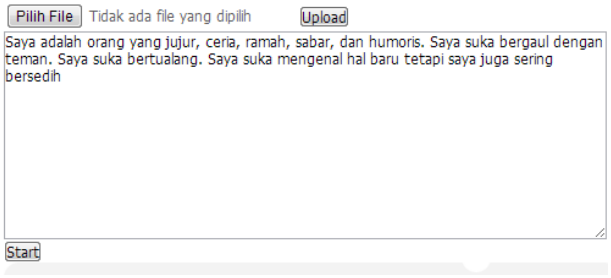


Figure 10 Input personality's data process

After the finish written or upload the data, click the start button. Then the result will appear as shown in figure 10.

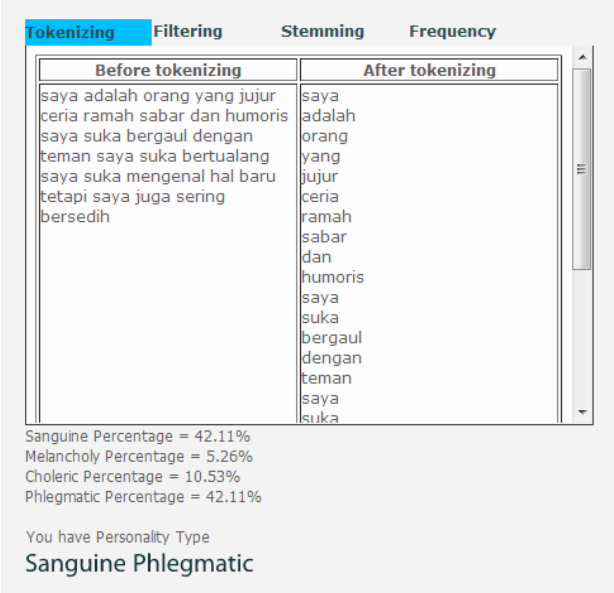


Figure 11 Result of Personality Type

After knowing the personality types, users can find their potential mates. As example above, user has a complementary mixed personality, which is sanguine and phlegmatic. As the theory of couple compatibility, the sanguine is a mate of melancholy and phlegmatic is a mate of choleric. So their mate must be a person who have melancholy, choleric, or two of them. Figure 11 will show the result of the matching couples.

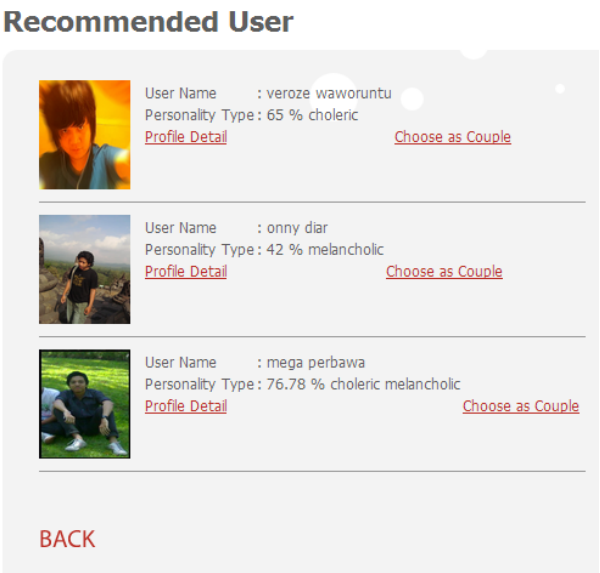


Figure 12 Result of matching couples

This experiment use 40 documents training and has 160 learning documents. In table 3, is the result of the details of personality classification for 40 training data which has been classified. There are 3 errors (error) which has produced the three data are unidentified category. So the percentage error reached,

$$\text{Accuracy percentage} = \frac{sum\ of\ correct\ \ classifica\ tion}{training\ \ documents} \text{ x } 100\% \quad (4)$$

$$= \frac{37}{40} \text{ x } 100\% = 92{,}5\%$$

Table 3. Result of Classification Training Document

Document Number	Classification Result	True/Flase
1	Phlegmatic	True
2	Melancholy	True
3	Phlegmatic	True
4	Phlegmatic	True
5	Sanguine Choleric	True
6	Melancholy	True
7	Phlegmatic	True
8	Phlegmatic	True
9	Sanguine	True
10	Choleric	True
11	Sanguine	True
12	Choleric	True
13	Melancholy	True
14	Melancholy	True
15	Unidentified category	False
16	Unidentified category	False
17	Sanguin Phlegmatic	True
18	Choleric Melancholy	True
19	Sanguine Phlegmatic	True
20	Choleric Phlegmatic	True
21	Sanguine	True
22	Melancholy	True
23	Choleric	True
24	Choleric Phlegmatic	True
25	Sanguine Melancholy	True
26	Melancholy	True
27	Sanguine	True
28	Sanguine	True
29	Phlegmatic	True
30	Sanguine Melancholy	True
31	Choleric melancholy	True
32	Sanguine choleric	True
33	Unidentified category	False
34	Choleric	True
35	Choleric	True
36	Choleric Phlegmatic	True
37	Melancholy	True
38	Sanguine Phlegmatic	True
39	Melancholy	True
40	Sanguine Phlegmatic	True

With the number of training data with error percentage as such, the 40 training data will use as learning data in the database for classify the training data in subsequent experiments and is expected to shrink error percentage in selecting or classifying personality types.

5. Conclusion

This experiment has successfully obtained the type of personality and finds a mate based on personality types by using the text mining with Naïve Bayes method for personality classification. In this experiment, some of the user data personality is used as learning document in the learning process of Naive Bayes methods. The success rate of the classification depends on the amount of learning document used. Personality classification process is done by the determination of the biggest VMap from each category. For matching couple output, the programs use Personality compatibilities theory, where the matching couples are the couples who have opposite personalities.

Acknowledgments

Our thank goes to Department of Information Technology Udayana University, Bali, who has helped organize this research's in Indonesia.

References

[1] Reddy V, Siva RamaKrishna, dkk. Classification of Movie Reviews Using Complemented Naïve Bayesian Classifier: Prithyi Information Solutions Limited: India

[2] Hamzah, Amir. Text Classification with Naive Bayes classifier (NBC) for Abstract Grouping Text and Academic News. Prosidign Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III. Yogyakarta. 3 November 2012.

[3] Abdur Rozaq, Abdur., Agus Zainal Arifin., Diana Purwitasari. Arabic Language Text Document Classification using Naive Bayes Algorithm: Surabaya

[4] Kim, Jangwoo., Daniel X. Le, and George R. Thoma. Naïve Bayes Classifier for Extracting Bibliographic Information from Biomedical Online Articles: National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894: USA

[5] Rothbart, Mary.K., Stephan A. Ahadi., David E. Evans. Temperament and Personality: Origins and Outcomes. Journal of Personality dan Social Psychology 2000, Vol. 78. No 1. 122-135

[6] Littauer, Florence. 1992. Personality Plus. Jakarta Barat: Binarupa Aksara

[7] Aprilia, Krisma Dini. 2008 Application of Naive Bayes for classification SMS Customer's Voice (Case Study PT. Pertamina UPMS V Surabaya): Stikom Digilib : Surabaya

[8] Saraswati, Ni Wayan Sumartini. Text Mining dengan Metode Naïve Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis: Denpasar. 2011

[9] Maharsi, Lisa. Text Document Keywords Extraction Using Naïve Bayes Method: Bandung. 2009

[10] Anugroho, Prasetyo., Idris Winarno., S.ST M.Kom., Nur Rosyid M., S.Kom. Spam Email Classification with Naïve Bayes Classifier Method use Java Programming: Surabaya.

[11] Indranandita, Amalia., Budi Susanto, and Antonius Rachmat C. Classification System and Journal Search using Naive Bayes Methods and Vector Space Model. Jurnal Informatika, Volume 4 Nomor 2, November 2008.

[12] Trisedya, Bayu Distiawa and Hardinal Jais. Document Classification using Naive Bayes algorithm with the addition of Parameter Probability Parent Category: Jakarta.2009

[13] Nurhayati, Sri. Implementation of Text Mining for Classification of Traditional Arts with NBC method (Naive Bayes Classifier): Bandung

[14] Destuardi.I dan Surya Sumpeno. 2009 Emotion Classification for Indonesian Language Text Using Naïve Bayes Method: Jurnal Teknik Elektro ITS : Surabaya

[15] Feldman, Ronen., James Sanger. 2007. The Text Mining Handbook. United Kingdom: Cambridge University Press

[16] Khodra, Masayu Leylia. Text Mining Text Categorization Naïve Bayes : Informatika ITB: Bandung

AA Ketut Agung Cahyawan, ST., MT received his ST degree and MT degree in Electrical Engineering from Institut Teknologi Bandung. He is lecturer at Electrical Engineering Department (major in Computer System and Informatics) of Udayana University, lecturer at Information Technology Department of Udayana University

Ni Made Ari Lestari study in Information Technology, Department of Information Technology Udayana University since August 2008, and now working her research for S.Ti. degree in Information Technology.

Dr. I Ketut Gede Darma Putra, S.Kom., MT received his S.Kom degree in Informatics Engineering from Institut Teknologi Sepuluh Nopember University, his MT. degree in Electrical Engineering from Gajah Mada University and his Dr. degree in Electrical Engineering from Gajah Mada University. He is lecturer at Electrical Engineering Department (major in Computer System and Informatics) of Udayana University, lecturer at Information Technology Department of Udayana University.

A Novel Feature Extraction Technique for Facial Expression Recognition

***Mohammad Shahidul Islam[1], Surapong Auwatanamongkol[2]**

**[1] Department of Computer Science, School of Applied Statistics,
National Institute of Development Administration,
Bangkok, 10240, Thailand**

**[2] Department of Computer Science, School of Applied Statistics,
National Institute of Development Administration,
Bangkok, 10240, Thailand**

Abstract

This paper presents a new technique to extract the light invariant local feature for facial expression recognition. It is not only robust to monotonic gray-scale changes caused by light variations but also very simple to perform which makes it possible for analyzing images in challenging real-time settings. The local feature for a pixel is computed by finding the direction of the neighboring of the pixel with the particular rank in term of its gray scale value among all the neighboring pixels. When eight neighboring pixels are considered, the direction of the neighboring pixel with the second minima of the gray scale intensity can yield the best performance for the facial expression recognition in our experiment. The facial expression classification in the experiment was performed using a support vector machine on CK+ dataset The average recognition rate achieved is 90.1 ± 3.8%, which is better than other previous local feature based methods for facial expression analysis. The experimental results do show that the proposed feature extraction technique is fast, accurate and efficient for facial expression recognition.

Keywords: *Emotion Recognition, Facial Expression Recognition, Image Processing, Local Descriptor, Pattern Recognition.*

1. Introduction

Facial Expression plays an important role in human-to-human interaction, allowing people to express themselves beyond the verbal world and understand each other from various modes. Some expressions incite human actions, and others fertilize the meaning of human communication. Human-centered interfaces must have the ability to detect shades of and changes in the user's behavior and to start interactions based on this information rather than simply responding to the user's commands. Facial expression recognition is a challenging problem in computer vision. Due to its potential important applications, it attracts much attention of the researchers in the past few years (Z. Zeng *et al*, 2009). Appearance-based methods have been heavily employed in this domain with great success. Popular methods are Gabor filters, local binary patterns (LBP) descriptors, Haar wavelets and subspace learning methods. Facial expression recognition process is a part of facial image analysis. A. Mehrabian (1968) mentioned in his paper that the verbal part of a message contributes only 7% of its meaning as a whole, the vocal part contributes 38% while facial movement and the expression gives 55% to the effect of that message, see Fig. 1. This means that the facial part does the major contribution in human communication. There are seven basic types of facial expressions. They are contempt, fear, sadness, disgust, anger, surprise and happiness. From the review of papers on facial expression, it is clear that most of the facial expression recognition systems (FERS) were based on the Facial Action Coding System (FACS), Y.L. Tian *et al.* (2001), Y. Tong *et al.* (2007), M. Pantic *et al.* (2000). In this system, the changes in the facial expression are described with FACS in terms of 44 different

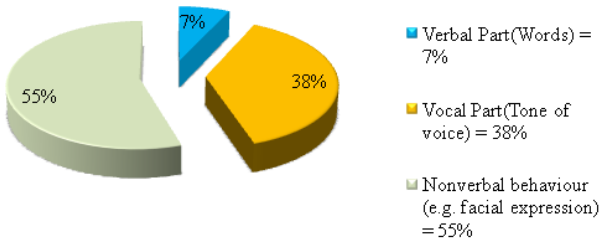


Fig. 1: 7%-38%-55% rule by A. Mehrabian (1968)

action units (AUs), each of which is related to the facial muscle movements. 44 AUs can give up-to 7000 different combinations, with wide variations due to age, size and ethnicity. M. Pantic *et al.* (2000) gave detail survey on facial expression recognition in their paper. Most of the research works on facial expression recognition (FER) are grounded on still images. The psychological experiments

by J.N. Bassili (1979) have proposed that facial expressions are more precisely recognized from video than single static image. I. Kotsia *et al.* (2007) applied facial wire frame model and a Support Vector Machine (SVM) for classification. Y. Zhang *et al.* (2005) proposed IR (Infra Red) illuminated camera for facial feature detection, tracking and recognized the facial expressions using Dynamic Bayesian networks (DBNs). Y.L. Tian *et al.* (2001) proposed multi state face component model of AUs and neural network for classification. M. Yeasin *et al.* (2007) created discrete hidden Markov models (DHMMs) to recognize the facial expressions. K. Anderson *et al.* (2006) used the multichannel gradient model (MCGM) to determine facial optical flow in videos. The motion signatures achieved are then classified using Support Vector Machines. I. Cohen *et al.* (2003) employed Naive–Bayes classifiers and hidden Markov models (HMMs) together to recognize human facial expressions from video sequences. M. Pantic *et al.* (2006) applied face-profile-contour tracking and rule-based reasoning to recognize 20 AUs taking place alone or in a combination in nearly left-profile-view face image sequences and they achieved 84.9% accuracy rate. T. Ahonen *et al.* (2006) proposed a new facial representation strategy for still images based on Local Binary Pattern (LBP). The basic idea for developing the LBP operator was that two-dimensional surface textures can be identified by two complementary measures: 2D local spatial patterns and the gray scale difference. The original LBP operator by T. Ojala *et al.* (1996), labels for

a(1)	a(2)	a(3)
a(8)	e	a(4)
a(7)	a(6)	a(5)

Fig. 2: Local 3x3 pixels Image region

the image pixels by comparing the 3 x 3 neighborhood of each pixel with the center pixel value and transforming the result as a binary number.

$$LBP = \sum_{i=1}^{P} 2^{i-1} f(a(i) - e) \qquad (1)$$

$$f(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases} \qquad (2)$$

Where (e) denotes the gray value of the center pixel, a(i) is the gray value of its neighbors, P stands for the number of neighbors, see Fig. 2. Fig. 3 shows an example of

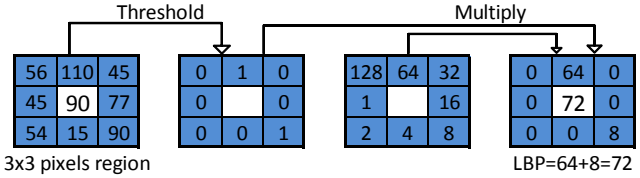


Fig. 3: Example of obtaining LBP from a 3x3 local region.

obtaining an LBP from a given 3×3 pattern. The histogram of these patterns for a local block of an image represents a local feature for the block. The histograms for all blocks can be concatenated to represent the feature vector for the image. G. Zhao *et al.* (2007) applied facial dynamic texture data in conjunction with Local Binary Pattern on the Three Orthogonal Planes (LBP-TOP) and Volume Local Binary patterns (VLBP) to combine motion and appearance. In her earlier work (G. Zhao *et al.*, 2004), she tested with the two-dimensional (2-D) discrete cosine transform (DCT) over the entire face image but got less accuracy on the facial expression recognition. The FACS approaches involve more complexity in facial feature detection and extraction procedures while the appearance-based approaches using local features such as LBP are less complex but still need to be improved to get higher recognition rates. Hence, this paper proposes an alternative local feature extraction technique that would be simple and more effective for facial expression recognition.

The rest of the paper is organized to explain the proposed methodology in section 2, results and analysis in section 3, and conclusion in section 4.

2. Proposed Methodology

2.1 Local Minima (LM)

The proposed method computes the local feature for a pixel from the gray scale value of its neighboring pixels. From a 3x3 local pattern, shown in Fig. 2, the center pixel of the pattern is surrounded by 8 neighboring pixels in 8 possible directions. The directions are denoted by 0°, 45°, 90°, 135°, 180°, 225°, 270° and 315°. The direction of the neighboring pixel with the minimum, the second minimum and so on for the gray scale values can be considered as the local feature for the given pixel. To identify the neighboring pixels with the first minima, second one and so on, the gray scale color values of all the eight

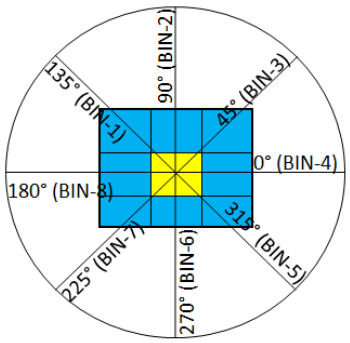


Fig. 4: 8 possible BINS denoted as 0°, 45°, 90°, 135°, 180°, 225°, 270° and 315°

neighboring pixels can be sorted in ascending order. If there are several of the pixels with the same gray color value, the positions of the pixels starting from the northwest one in clockwise direction can be considered to break the ties. The direction would represent the changing direction of the gray scale color values at the particular center pixel. Thus, eight possible bins are needed to build the histogram on the numbers of pixels in a block for the possible 8 directions as shown in Fig. 4**.** The histograms for all blocks for an image can then be concatenated to form the feature vector for the whole image. Notice that the direction is insensitive to light changes since the light changes would change the gray scale color values of all the pixels by nearly same amount but not the direction of the minima for each of the pixels.

2.2 Experimental Setup

The experiments for the facial expression recognition include three distinguished phases. i.e. facial feature extraction, SVM training and facial expression determination. The Extended Cohn-Kanade Dataset (CK+) (P.Lucey *et al.*, 2010) is used for both training and testing images. There are 326 peak facial expressions of 123

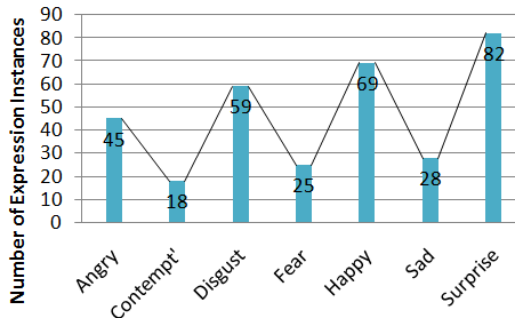


Fig. 5: CK+ Dataset, 7 expressions and numbers of instances of each expression

subjects. Seven emotion categories are in this dataset. They are 'Anger', 'Contempt', 'Disgust', 'Fear', 'Happy', 'Sadness' and 'Surprise'. No subject with the same emotion has been collected more than once. The data distribution of the dataset is shown in Fig. 5**.** This is the most common dataset used in FER (Facial Expression recognition). All the images are posed in this dataset. Facial feature extraction phase is illustrated in Fig. 6, which includes detecting face, masking the face, dividing the cropped face into equal blocks, calculating feature histogram for each block and concatenating all histograms to build feature vector.

Face detection is done using **fdlibmex**, free code is available for Matlab. The library consists of single mex file with a single function that takes an image as input and returns the frontal face. It is then resized to 180x180 resolutions and masked using a round shape, outside which all the pixels are removed from the consideration. In experiment, the 180x180-size face is equally divided into 9x9=81 blocks of 20x20 resolutions each. Feature is extracted from each block using the proposed method, concatenating histograms of all the blocks into a unique feature vector. Therefore, the length of the feature vector is 8x9x9=648. In the training phase, LIBSVM, by C.C. Chang *et al.* (2011) is used to train a multiclass Support Vector Machine to classify the facial expression for an image. The 10-fold cross validation was used to evaluate the performance of the classifier when the proposed local feature was used. Each expression instances are divided into equal size 10 folds. Ten rounds of evaluations were conducted. For each round, nine alternative folds (90% for each expression) are used for training and the rest one fold are used for testing. The kernel parameters are set to: **(-s 0 -t 1 -c 100 -g 0.00125 -b 1),** where s=0 for SVM type C-Svc, t=1 for polynomial kernel function, c=100 is the cost of SVM, g=0.00125 is the value of 1/ (length of feature vector), b=1 for probability estimation. The kernel

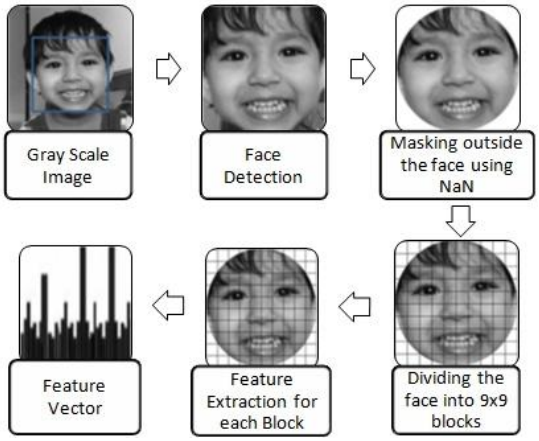


Fig. 6: Facial Feature Extraction

parameters are tuned so that it can produce optimal results during this phase.

3. Experimental Results and Analysis

Table 1 shows the achieved classification accuracy rates when the directions of the 1st to 8th local minima are individually considered as the local feature. The 2nd minima (LM-2) gives the peak accuracy of 90.1%±3.8. This is the average accuracy of 10-fold cross validation. The highest accuracy achieved by any one fold from the tenfold is 93.9% and the lowest is 87.1%. We also tried the

same experiment with different block sizes as shown in Table 2. However, the block size of 15x15 pixels gives the highest 90.33% accuracy rate but there is a penalty in feature vector length.

Table 1: Classification Accuracy of the 1st to the 8th Local Minima

Local Minima	Classification Accuracy
1st (LM-1)	88.8%
2nd (LM-2)	**90.1%**
3rd (LM-3)	89.8%
4th (LM-4)	88.9%
5th (LM-5)	88.9%
6th (LM-6)	89.2%
7th (LM-7)	89.5%
8th (LM-8)	86.4%

Table 2: Classification Accuracy Vs Block Dimension.

Face Dimension (Pixels)	Number of Blocks	Block Dimension (pixels)	Classification Accuracy (%)	Feature Vector Length
180x180	6x6	30x30	88.28	288
180x180	**9x9**	**20x20**	**90.1**	**648**
180x180	10x10	18x18	88.8	800
180x180	**12x12**	**15x15**	**90.33**	**1152**
180x180	15x15	12x12	88.25	1800
180x180	18x18	10x10	87.63	2592

Table 3: Confusion Matrix for LM-2

LM-2 (Local Second Minima) = 10-fold validation
Feature Extraction time for 326 Image = 96 Seconds
Average Classification Accuracy = 90.1 ± 3.8%
Kernel parameter: = (-s 0 -t 1 -c 100 -g 0.0015 -b 1)
Confusion Matrix:

C : Actual

prediction	Angry	Contempt	Disgust	Fear	Happy	Sad	Surprise
Angry	38	1	2	0	0	4	0
Contempt	2	15	0	0	0	1	0
Disgust	2	0	54	2	1	0	0
Fear	1	1	0	21	1	0	1
Happy	0	0	0	0	69	0	0
Sad	4	1	1	1	0	20	1
Surprise	0	1	0	1	0	0	80

The confusion matrix using proposed feature extraction method of LM-2 is shown in Table 3. The feature extraction takes 96 seconds including 30 seconds for face detection and round masking (Preprocessing) for the 326 images, or 0.29 seconds per image. Table 4 shows comparisons of the individual expression accuracy and the average all 7 class expression accuracy achieved by the proposed method and the other recent methods using shape or combination of shape and texture information.

Table 4: Comparison of individual expression accuracy and the average accuracy (∑ (Accuracy of all 7 expressions/7)) of different methods. [S: shape based method, T: texture based method. S + T: both shape and texture based method. (CLM-Constrained Local Model, AAM-Active Appearance Model, An.= Anger, Co.= Contempt, Di.= Disgust, Fe.=Fear, Ha.=Happy, Sa.=Sad, Su.=Surprise, Avg.=Accuracy of all expressions/7)]

Authors	Method	T/S	An.	Co.	Di.	Fe.	Ha.	Sa.	Su	Avg.
P. Lucey *et al.* (2010)	AAM + SVM	S	35.0	25.0	68.4	21.7	98.4	4.0	100.0	50.3
	AAM + SVM	T	70.0	21.9	94.7	21.7	100.0	60.0	98.7	66.7
	AAM + SVM	T + S	75.0	84.4	94.7	65.2	100.0	68.0	96.0	83.3
S.W. Chew *et al.* (2011)	CLM + SVM	T	70.1	52.4	92.5	72.1	94.2	45.9	93.6	74.4
L.A. Jeni *et al.* (2012)	CLM + SVM (AU0 norm.)	S	73.3	72.2	89.8	68.0	95.7	50.0	94.0	77.6
	CLM + SVM (personal mean shape)	S	77.8	94.4	91.5	80.0	98.6	67.9	97.6	86.8
Proposed Method (LM-2)	**No Registration + SVM**	**T**	**84.4**	**83.3**	**91.5**	**84.0**	**100.0**	**71.4**	**97.6**	**87.0**

It should be noted that the results are not directly comparable due to different experimental setups, version differences of the CK (T. Kanade *et al.*, 2000) dataset with different emotion labels, preprocessing methods, the number of sequences used, and so on, but they still point out the discriminative power of each approach. It is clearly

mentioned by L.A. Jeni *et al.* (2012) that aligned faces can give an extra 5-10% increase in the facial expression recognition accuracy and leave-one -subject-out validation can increase the accuracy by 1-2% , (M.S. Bartlett *et al.*, 2003) and incorporation of adaboost algorithm can also

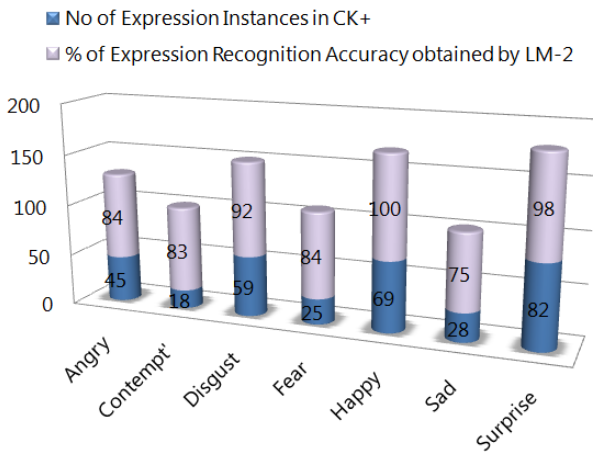


Fig. 7: Number of Instances Vs percentage of individual expression recognition Accuracy.

increase the accuracy by 1-2% on CK+ dataset. So overall an extra 7-12% accuracy can be obtained using proper alignment, increasing training data size and adding boosting algorithm along with the classifier.

A facial expression can be spontaneous or caused externally. In general, cases boundaries for spontaneous expressions are tough to determine. The dataset has only one peak expression for a particular subject. Some subjects do not contain all seven expressions. Training with multiple instances of the same expression and subject can increase accuracy. Fig. 7 clearly shows that in CK+ dataset number of 'Sad', 'Contempt' or 'Fear' instances are less in compare with the other expressions. Increasing these instances can increase accuracy like others.

4. Conclusion

A novel technique for facial feature extraction is proposed for facial expression recognition. It extracts from a gray scale image the direction of the neighboring pixel with local minima on the gray scale color value among those of the eight neighboring pixels. Eight possible Minima neighboring pixels can be considered as a local feature for a given pixel; however, the direction of the second minima yields the highest facial expression recognition rate in the experiment. Further techniques such as AdaBoost or SimpleBoost algorithms can be incorporated with the SVM classifier to increase the accuracy rate substantially.

References

[1] A. Mehrabian. "Communication without words." Psychology Today, 2, 4 (1968), 53-56.

[2] C.C. Chang and C.J. Lin." LIBSVM: a library for support vector machines". ACM Transactions on Intelligent Systems and Technology (2011).

[3] G. Zhao and M. Pietikainen." Dynamic texture recognition using local binary patterns with an application to facial expressions." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)., 29, 6 (2007), 915–928.

[4] G. Zhao and M. Pietikainen. "Facial Expression Recognition Using Constructive Feed forward Neural Networks." IEEE Transactions on Systems, Man, and Cybernetics., 34, 3 (2004), 1588–1595.

[5] I. Cohen, N. Sebe, S. Garg, L. S. Chen and T. S. Huanga. Facial expression recognition from video sequences: temporal and static modelling. Computer Vision and Image Understanding, 91 (2003), 160-187.

[6] I. Kotsia & I. Pitas. Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines. IEEE Transaction on Image Processing, 16, 1 (Jan 2007).

[7] J.N. Bassili. Emotion Recognition: The Role of Facial Movement and the Relative Importance of Upper and Lower Area of the Face. J.Personality and Social Psychology, 37 (1979), 2049-2059.

[8] K. Anderson and Peter W. McOwan. A Real-Time Automated System for the Recognition of Human Facial Expressions. IEEE Transactions on Systems, Man, and Cybernetics, 36, 1 (2006), 96-105.

[9] L. A. Jeni, András Lőrincz, Tamás Nagy, Zsolt Palotai, Judit Sebők, Zoltán Szabó & Dániel Takács. 3D shape estimation in video sequences provides high precision evaluation of facial expressions. Image and Vision Computing, 30, 10 (October 2012), 785-795.

[10] M. Pantic and Ioannis Patras. Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments From Face Profile Image Sequences. IEEE Transactions on Systems, Man, and Cybernetics, 36, 2 (2006), 433-449.

[11] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: the state of the art. IEEE Trans. Pattern Analysis and Machine Intelligence., 22, 12 (2000), 1424–1445.

[12] M. Yeasin, B. Bullot and R. Sharma. Recognition of Facial Expressions and Measurement of Levels of Interest From Video. IEEE Trans. Multimedia, 8, 3 (2006), 500-508.

[13] M.S. Bartlett, G. Littlewort, I. Fasel & R. Movellan. Real Time Face Detection and Facial Expression Recognition: Development and Application to Human Computer Interaction. In Proc. CVPR Workshop Computer Vision and Pattern Recognition for HumanComputer Interaction(2003).

[14] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar & I. Matthews. The Extended Cohn-Kande Dataset (CK+): A complete facial expression dataset for action unit and emotion-specified expression. Paper presented at the Third IEEE Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010) (2010).

[15] S.W.Chew, P.Lucey, S. Lucey, J. Saragih, J.F. Cohn & S. Sridharan. Person-independent facial expression detection using Constrained Local Models. In 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), (2011), 915-920.
[16] T. Ahonen, A. Hadid and M. Pietikainen. Face Description with Local Binary Patterns: Application to Face Recognition. IEEE Trans. Pattern Analysis and Machine Intelligence, 28, 12 (2006), 2037-2041.
[17] T. Kanade, J. F. (2000). Comprehensive database for facial expression analysis. Fourth IEEE International Conference on Automatic Face and Gesture Recognition.
[18] T. Ojala, M. Pietikäinen & T. Mäenpää. Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns. IEEE Trans. Pattern Analysis and Machine Intelligence, 24, 7 (2002), 971-987.
[19] Y. L. Tian, T. Kanade and J. F. Cohn. Recognizing action units for facial expression analysis. IEEE Trans. Pattern Anal. Mach. Intell, 23, 2 (2001), 97-115.
[20] Y. Tong, W. Liao, and Q. Ji. Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships. IEEE Trans. Pattern Anal. Mach. Intell, 29, 10 (2007), 1-17.
[21] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Trans. Pattern Anal. Mach. Intel, 27,5(2005),699-714.
[22] Z. Zeng, M. Pantic, G. Roisman & T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31, 1 (2009), 39–58.

Mohammad Shahidul Islam received his B.Tech. degree in Computer Science and Technology from Indian Institute of Technology-Roorkee (I.I.T-R), Uttar Pradesh, INDIA in 2002, M.Sc. degree in Computer Science from American World University, London Campus, U.K in 2005 and M.Sc. in Mobile Computing and Communication from University of Greenwich, London, U.K in 2008. He is currently pursuing the Ph.D. degree in Computer Science & Information Systems at National Institute of Development Administration (NIDA), Bangkok, THAILAND. His field of research interest includes Image Processing, Pattern Recognition, wireless and mobile communication, Satellite Commutation and Computer Networking.

Surapong Auwatanamongkol received a B.Eng. (Electrical Engineering) from Chulalongkorn University, THAILAND, in 1978 and M.S.(Computer Science) from Georgia Institute of Technology, U.S.A. in 1982 and Ph.D.(Computer Science) from Southern Methodist University, U.S.A. in 1991. Currently, he is an Associate Professor in Computer Science at the School of Applied Statistics, National Institute of Development Administration (NIDA), Thailand. His research interests include Evolutionary Computation, Pattern Recognition, Image processing and Data Mining.

The Simulation of Direct Spread Spectrum System based on Transmitted Reference Signal

Wu Guoqiang[1], Bai Yuguang[1] and Zhao Dongsheng[2,*]

[1] School of Aeronautics and Astronautics, Dalian University of Technology
Dalian, Liaoning 116024, China

[2] School of Naval Architecture Engineering, Dalian University of Technology
Dalian, Liaoning 116024, China

Abstract

Code synchronization is indispensable in the direct spread spectrum system because it can influence the incepting capacity directly. Transmitted reference is proposed in this paper to predigest the code synchronization circuit of the incepting machine in order to reduce the cost of time, energy and money for the development of the code synchronization technology. The software named Systemview is employed to simulate the transmitted reference direct spread spectrum system. The simulation results were presented with the condition of gauss noise and temperature. It demonstrates that the proposed simulation has significant effect and benefit in engineering.

Keywords: *Direct Spread Spectrum System, Systemview, Gauss Noise, Transmitted Reference*

1. Introduction

Spread spectrum communication is an important embranchment in communication fields. It represents one of the developmental direction of channel communication system [1,2]. The Spread spectrum technology has many advantages, e.g. strong anti-jamming ability, good quality of keeping secret and convenient multitudinous address communication. Therefore the Spread spectrum technology cannot only possess important status in martial communication, but also infiltrates into the civilian domain of personal communication and computer communication. [3,4].

Recently, the Spread spectrum technology has become one of the most potential communication technologies [5]. The direct spread spectrum system is widely used at present, of which the best advantage can include anti-jamming ability, secret keeping, multitudinous address communication and compose net and etc [6]. Usually, when we analyze the capability of the spread spectrum system, it was assumed that the synchronization between transmitter and receiver is good. Actually, we must use very accurate oscillator and code synchronization circuit in order to assure the PN code between the transmitter and receiver for accurate synchronization. These processes are accompanied by great cost and complex degree of technology. Though we can use the frequency with skyscraping stabilization degree, it is still impossible to eliminate all instability factors due to the Doppler shift and multipath fading which can bring significantly blight to the synchronization of system [7]. Even for the fixed position of receiving and transmitting system, the change of transmission channel also brings on the change of code phase and carrier wave frequency. Therefore accurate synchronization is quite difficult in actual system. The synchronization of spread spectrum sequence can be separated to two phases: capture and trace. The rough synchronization can be achieved in capture phase. It confines the spread spectrum sequence phase to be different from the receiver and transmitter in a code patch or little scope of a code patch. How to achieve the celerity capture of the spread spectrum sequence reliably is a key problem that influence the capability of system significantly [8,9]. The methods of synchronization capture include glide correlation, synchronization head, leap frequency synchronization, matching filter synchronization and etc.

Systemview is designed by ELANIX Company of USA, it is a full scale and dynamic system analysis software which can be used specially to simulate and design for engineering and science system [10]. Systemview provides an advanced system analysis engine. The analytical objects are very comprehensive, including: simulation of digital signals dispose, filter, control system; the design and analysis of communication system; various currency mathematics model simulation and validation, and etc. In this paper, Systemview simulation software is employed to construct the direct sequence spread spectrum system with a transmitted reference signal.

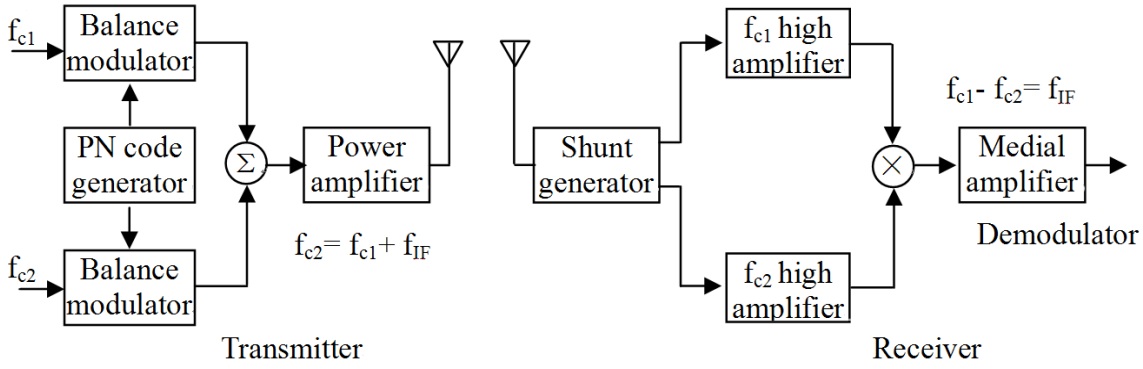


Fig. 1 Theoretical roadmap of the direct sequence spread spectrum

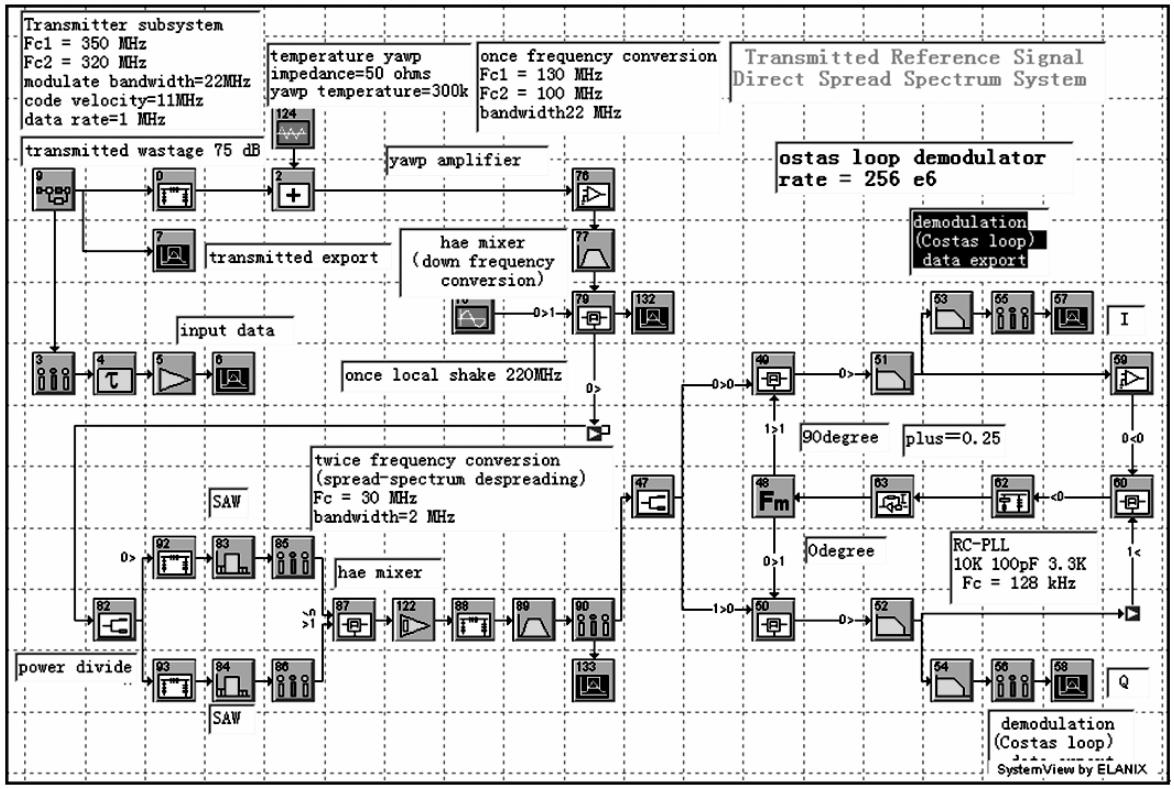


Fig. 2 The simulation model of direct spread spectrum system based on transmitted reference signal

2. Simulation Design of Transmitted Reference Signal of the Direct sequence Spread Spectrum System

The method of transmitted reference signal can be used to identify synchronization capture and trace. The receiver of transmitted reference signal cannot use the code generator and any other local reference oscillator. The direct sequence spread spectrum code reference signal is produced by transmitter, and transmitted with practical information signal at the same time.

As shown in Figure 1, fc1 and fc2 use the same spread spectrum code to modulate. Here fc1 is used to transmit information, whereas fc2 does not take information. After mixing two signals at the end of receiver, the signals can express intermediate frequency without spread spectrum. The working course of a transmitted reference receiver is the same as some other receivers that use local reference signals. The difference between transmitted reference and

local reference is that the local spread spectrum codes are produced by transmitter and demodulated by receiver. As shown in Figures 2 and 3, the GOLD code is used as spread spectrum code. The departure of two carrier wave frequency is equal to the first middle amplificatory frequency. The correlation intermediate frequency signals are produced by mixing. The cost, weight and size must be limited in a receiver. The method of transmitted reference has obvious superiority, because it do not need spread spectrum code sequence generator, the circuit of code capture, the circuit of code synchronization, the circuit of code trace, and any other circuits which are correlated with code. The transmitted reference receiver has a thick skin due to the influence of Doppler shift, and can be compatibly used in the objects which have fast movement speed. However, the method of transmitted reference still has some disadvantages: its anti-jamming capability decreases; the reference signals of a receiver are produced by transmission, thus the yawp which can degrade system performance can be drew into. In this paper, the simulation system engrosses the bandwidth 22MHz. When the date rate is 1MHz, the theoretical plus is 13.4 dB. The reference signals are transmitted by channel, so the yawp is drew into. In current disturbance, the system execute wastage of reference receiver is not twice larger than the direct sequence spread spectrum system with local reference signal. Commonly, the wastage is 1-2 dB, the result of this simulation system is 3 dB. If we can choose appropriate code counts, we can get higher spread spectrum plus. When comparing this plus with profit produced by predigestion system, this wastage is inessential.

The worst condition is the occurrence of intermediate frequency interference which is the difference of two frequencies falls into the frequency band. When two interferential signals exceed half disturb tolerance of the system, the system will be blocked and cannot work. If the disturbance is eliminated, the system can resume to work and do not need any other direct sequence spread spectrum system synchronism to set up the process again. In order to handle the artificial disturbance, the best way is to protect intermediate frequency and make the intermediate frequency being changed in a certain range. When the system suffer from the artificial disturbance, the routine analysis of correlative spread spectrum system capability cannot be applicable for the system.

Therefore, when we design analogous system, we must adopt the high value of the intermediate frequency as quickly as possible. The intermediate frequencies cannot be less than the bandwidth of spread spectrum signals. At the same time, it is proper to use heterodyne correlator, and it can make the disturbance signals to act ahead of correlator, and it is unable to divulge to back circuit through correlator. The frequency answer characteristic of a SAW filter in the transmitter and receiver usually take the finite impact answer filter (FIR) to simulate. In order to improve operation speed of computer simulation, the pigtail counts cannot be selected accurately, so the parameter of simulative filter is not as good as the characteristic of SAW made in practical engineering.

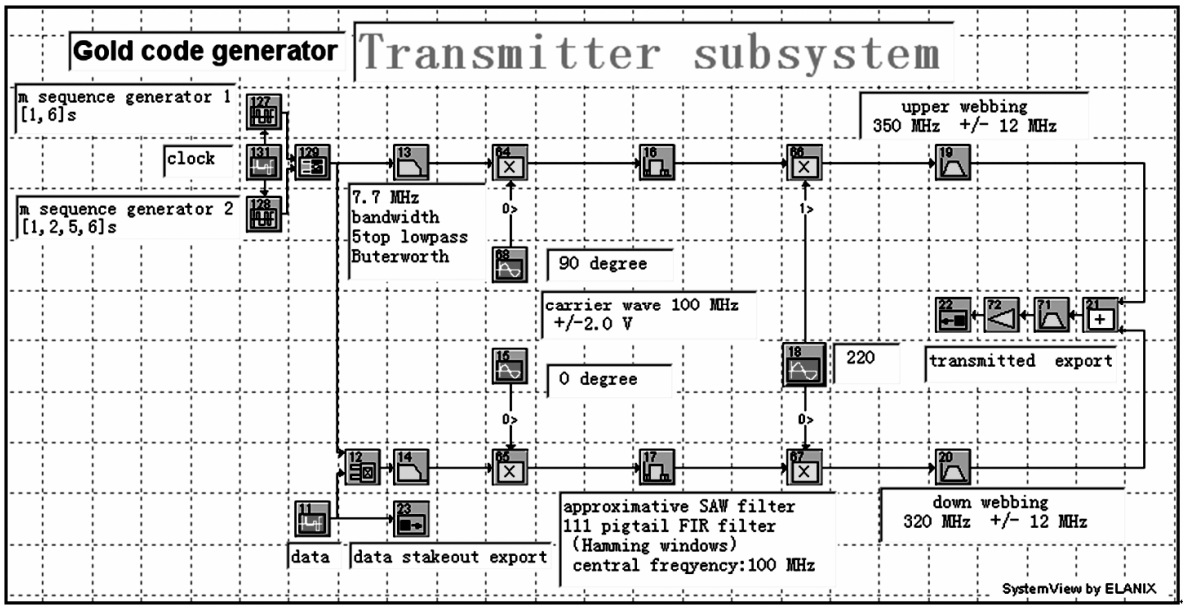


Fig. 3 Transmitter subsystem

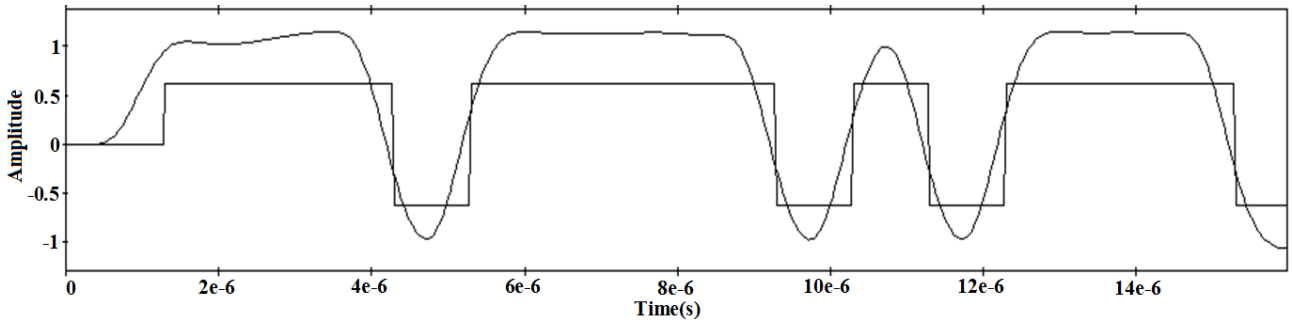


Fig. 4 The export signals add with original signals after Coatas loop demodulation

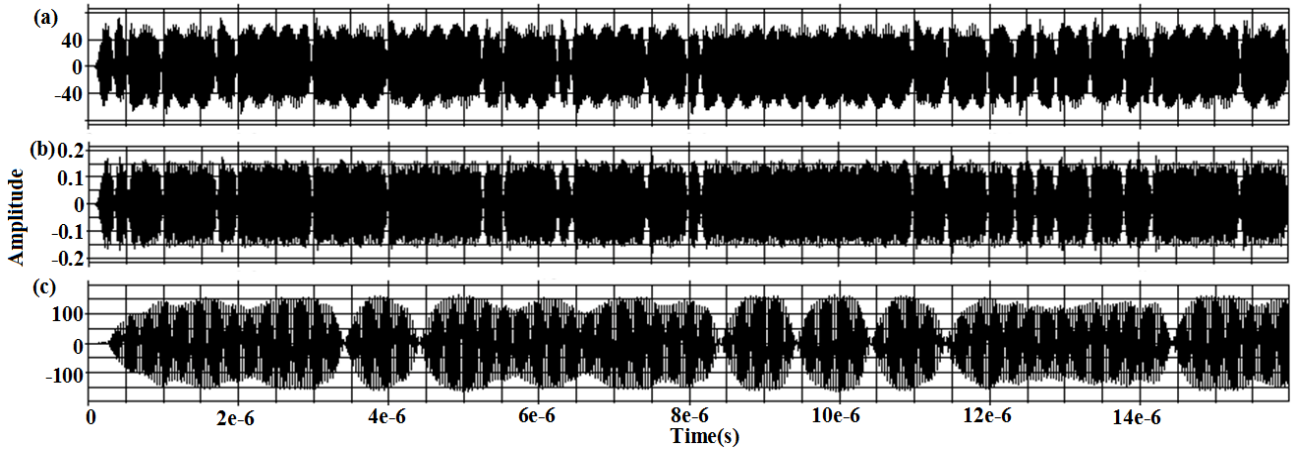


Fig. 5 The export wave: (a) of transmitter; (b) after once frequency conversion; and (c) after twice frequency conversion. Note that (a), (b) and (c) have the same horizontal axis.

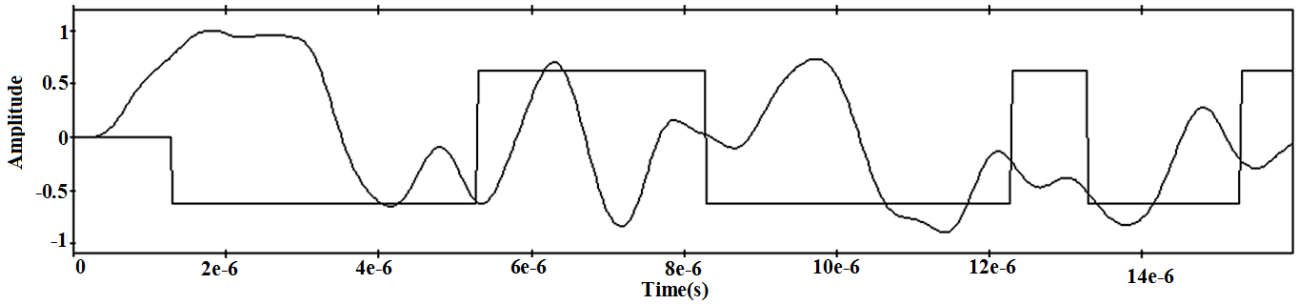


Fig. 6 The export signals add with original signals after Coatas loop demodulation

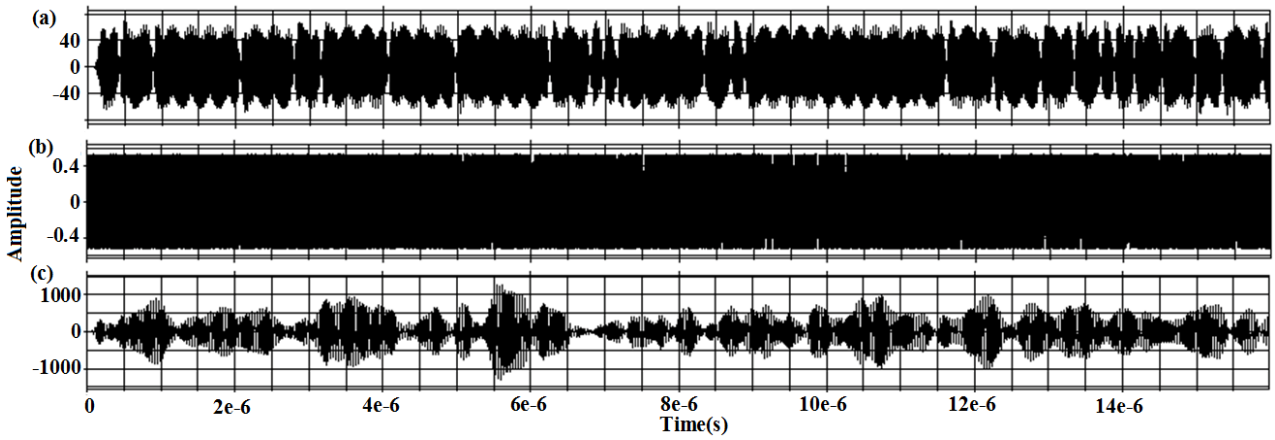


Fig. 7 The export wave: (a) of transmitter; (b) after once frequency conversion; and (c) after twice frequency conversion. Note that (a), (b) and (c) have the same horizontal axis.

3. Result and discussion

3.1 Simulation Result of Transmitted Reference Direct Sequence Spread Spectrum System at Temperature Yawp

From the Figures 4 and 5, it can be found that the input wave matches preferably with the output wave. Though the usage rate of available frequency is immolated under the transmission, the structure of transmitter can be predigested greatly. And the precision require of fake code generator decreases greatly too. By using high speed short list code, the higher spread sequence plus can be gained.

3.2 Simulation Result of Transmitted Reference Direct Sequence Spread Spectrum System at Gauss Yawp

As shown in Figures 6 and 7, the simulation results demonstrate that the transmitted reference direct sequence spread spectrum system can work normally under temperature yawp, but the output signals are already anamorphic under gauss yawp.

4. Conclusions

In this paper, Systemview simulation software is employed to simulate the direct sequence spread Spectrum system based on transmitted reference, and the simulation results of the system are presented under temperature and gauss yawp. The simulation results demonstrate: the direct sequence spread Spectrum system based on transmitted reference can work normally under temperature yawp, though the usage rate of practicable frequency is immolated through transmission; the structure of receiver is predigested; the precision of bogus code generator is decreased; and the higher spread Spectrum plus can be achieved by using high speed short sequence codes. The direct sequence spread Spectrum system based on transmitted reference cannot work normally under gauss yawp, so it is inadvisable to be used under gauss yawp.

Acknowledgments

This work is supported by the National High Technology Research and Development Program of China (No. 2012AA120601), National Natural Science Foundation of China (No. 11202044, No. 11072044) and the Fundamental Research Funds for the Central Universities.

References

[1] Q. Wang, L.J. Liu, X. T. Zhang, J. L. Liu, and Y. Z. Zhang, "Design and implementation of a FPGA based low complexity underwater acoustic direct sequence spread spectrum communication system", Chinese High Technology Letters, Vol. 19, No. 10, 2009, pp. 1006-1013.

[2] R. Gharsallah, and R. Bouallegue, "Comparison between MC-CDMA and CDMAOFDM/OQAM systems in presence of MIMO channel", International Journal of Computer Science Issues, Vol. 9, No. 4, 2012, pp. 103-109.

[3] K. Jayanthi, and P. Dananjayan, "Improving antijamming characteristics of spread spectrum communication systems", International Journal of

Autonomous and Adaptive Communications Systems, Vol. 5, No. 1, 2012, pp. 77-87.
[4] N. Larbi, F. Debbat, and S. A. Boudghen, "MC-CDMA Scheme in WiFi Environment", International Journal of Computer Science Issues, Vol. 9, No. 1, 2012, pp. 243-247.
[5] R. Skaug, "spread spectrum radio systems-technological implementations and the compatibility issue", Institution of Electronic and Radio Engineers, No. 68, 1986, pp. 171-179.
[6] Q. Lin, and L. L. Guo, "BER performance study of non-equal probability UWB system based on parallel combinatory spread spectrum", Systems Engineering and Electronics, Vol. 33, No. 3, 2011, pp. 659-664.
[7] S. Saleemb, and I. Qamar, "On comparison of DFT-based and DCT-based channel estimation for OFDM system", International Journal of Computer Science Issues, Vol. 8, No. 3, 2011, pp. 353-358.
[8] L. C. Wung, S. L. Su, and C. F. Jhun, "Novel signaling and detection schemes for ultra-wideband transmitted reference systems", Telecommunication Systems, Vol. 2011, 2011, pp. 1-12.
[9] M. Farhang, and J. Salehi, "Optimum receiver design for transmitted reference signaling", IEEE Transactions on Communications, Vol. 58, No.5, 2010, pp. 1589-1598.
[10] X. H. Xu, Z. X. Zhang, and L. Q. Yin, "The teaching application of systemview in error controlling", Advanced Materials Research, Vol. 542-543, 2012, pp. 1413-1417.
[11] C. W. Chow, and L. Xu, "Mitigation of signal distortions using reference signal distribution with colorless remote antenna units for radio-over-fiber applications", Journal of Lightwave Technology, Vol. 27, No. 21, 2009, pp. 4773-4780.
[12] W. Y. Luo, L. Jin, Y. S. Li, "A subcarrier-reference scheme for multiuser MISO-OFDMA systems with low probability of interception", IEICE Transactions on Communications, Vol. E94-B, No. 10, 2011, pp. 2872-2876.

First Author: Guoqiang Wu is a Lecturer in Dalian University of Technology, P.R.China. He received a Ph.D. degree from Astronautics School in Harbin Institute of Technology (HIT), Harbin, P.R.China, also received his BE and ME degrees in Solid Mechanics in HIT. His research interests are in area of multidisciplinary modeling and simulation, micro-satellite communication, and channel decoding.

Second Author: Yuguang Bai is a Lecturer in Dalian University of Technology, P.R.China. He received a Ph.D. degree in 2011 from Faculty of Vehicle Engineering and Mechanics in Dalian University of Technology (DUT), Dalian, P.R.China, also received his BE degrees in 2004 from Faculty of Vehicle Engineering and Mechanics in DUT. His research interests are in area of multi-physics modeling and simulation, high performance computing, and computational fluid dynamics.

Third Author: DongSheng-Zhao is an assistant professor in the School of Naval Architecture Engineering in Dalian University of Technology, Dalian, Liaoning, P.R. China. He received his Ph.D in the Department of Welding Science and Engineering from Harbin Institute of Technology, Harbin, Heilongjiang, P.R. China in 2009. His current research interests are in welding residual stress.

Using Chinese Natural Language Interfaces for Navigation in Mobile GIS

Jiangfan Feng[1], Nan Xu[2]

[1] College of Computer Science and Technology, Chongqing University of Post and Telecommunications, Chongqing 400065, China

[2] College of Computer Science and Technology, Chongqing University of Post and Telecommunications, Chongqing 400065, China

Abstract

The combination of "voice technology" and "Mobile GIS", has greatly improved the intelligent degree of mobile GIS. Recently, significant attention in the field of scientific research is turning to how to realize quick conversion between natural language and GIS commands. However, the current study mostly concentrates on rules of conversion between natural language sentences and GIS commands from the angle of artificial summary. From the perspective of intellectualization, this paper does a study of natural language understanding method of mobile voice GIS, which makes a conversion between natural language and GIS commands based on machine learning methods, reorganizes existing knowledge structure to acquire new knowledge for the purpose of identifying unrecognized sentence patterns. Our experimental results prove that the preliminary conversion mechanism between natural language and GIS commands could be formed on the basis of self-learning which based upon the current research results by utilizing BP algorithm based on artificial neural network.

Keywords: *Mobile GIS, Natural language understanding, Machine learning, Algorithm*

1. Introduction

With the development of Internet and the mobile communication technology, the mobile geographic information system, with mobile Internet as center, has become one of the most popular research directions of GIS applications. The intelligent degree of GIS applications can be improved by combining voice technology and mobile GIS [1]. For the reason that natural languages are complicated, colloquial, and diversified, there are still some obstacles in the process of mobile terminals handle the natural languages. Therefore, how to eliminate obstacles in natural language understanding and convert the natural languages to GIS commands smoothly has become the most popular research in the field of mobile GIS.

At present, there are several research methods as follows: (1) Map the natural languages to E-R models. (2) Map the natural languages to SQL sentences. (3) Map the natural languages to functions. Although these methods could realize the conversion natural languages to GIS commands, there is not an integrated mechanism of natural language understanding. It still in artificial summary stages with a certain limitation.

In the view of artificial intelligence, this paper utilizes the BP algorithm based upon artificial neural network, and reorganizes knowledge structures based on the current research results, to set up an integrated mechanism of conversion natural languages to GIS commands.

2. Research Status

2.1 The Conversion Technology of Natural language and GIS command

All the time, the natural language used in mobile GIS, reduce the complexity of operation, is the focus in the field of GIS target. Combination of speech technology and mobile GIS applications for mobile GIS, it provides users a good human-computer interaction and intelligent user-experience.

In recent years, the research of natural language understanding has made a series of achievements: The GIS query language based on Chinese is put forward after an explanation of the characteristics and merits of the Chinese language. (Zhou Yankun, Li Manchun, 2001) [2]. On the condition of restricted applied range, it proposed a thinking method that uses the keywords of natural language to structure a model library to mapping the query sentences (Zhang Lianpeng, 2002) [3]. And the basic spatial relation query form and spatial semantic in

query sentences in the light of the basic spatial relation i.e. measure relations, direction relations, and topology relations among geographical features (Ma Lin-bing, 2002) [4]. A computational model of ER-model-based restrictive-Chinese query language of relational database is put forward which simulates the language process mechanism of human (Yang Dong-qing et al, 2001) [5]. The semantic information extraction of the query conditions are studied based on the principle of information extraction. She proposed an intermediate language way of semantic query trees, designed a set of conversion algorithm to achieve the conversion of GIS query language in Chinese into SQL statements (Xu Ai-ping, 2007) [6]. A paper proposed an analytic method of GIS command based on restricted natural language, and they classification marked the restricted natural language, and designed corresponding GIS command functions to realize the geographic information service intelligently (Long Yi et al. 2009) [7].

Although it has made progress in the field of natural language understanding at present, but there still has the following questions:

(1) The current research mainly concentrates on conversion restricted natural languages to SQL sentences or E-R models, instead of establishment a complete natural languages understanding mechanism of application-oriented mobile GIS. The research achievements are still in design stage of simple prototype systems.

(2) Due to the fact that natural languages are complicated and artificial summary is insufficient, there are still some obstacles in conversion natural languages to GIS commands.

Based on machine learning methods, this paper utilize BP algorithm to self-learning current research results and to reorganize existing sentence structures, for the purpose of identifying unrecognized sentence patterns, then converting these sentences patterns to GIS commands.

2.2 Machine Learning in Natural Language Understanding

Machine learning is the core of artificial intelligence, whose main application areas contain expert system, pattern recognition, intelligent robots, automated reasoning, computer vision, and natural language understanding etc.

In the field of natural language understanding, the main source of Machine Learning is primitive linguistic data. We should preprocess the linguistic data through words segmentation and text markup firstly. Then machine learns the existing text samples, does some training and analysis, and applies to the unrecognized texts, for the purpose of learning the meaning of unrecognized texts and solving practical problems.

The machine learning methods applied in natural language understanding could mainly be divided in three categories:

(1) The Symbolic machine learning, such as the learning of decision trees etc.

(2) The Statistical machine learning, such as artificial neural network, Bayesian Learning, Genetic Algorithm, support vector machine etc.

(3) The learning based on the cases.

These methods could be used in the different aspects of natural language understanding, such as part-of-speech tagging, clause identification, speech recognition, word sense disambiguation, vocabulary acquisition, and grammar inference etc [8]. Among them, there is a very good performance in the artificial neural network method applied to study of the grammar inference. Consequently, combined the method of artificial neural network, this paper converts the natural languages to GIS commands.

In a word, Machine Learning is an important way and the key method of mobile GIS intelligence. The study of Machine Learning methods in the field of natural language understanding will promote the further development of natural language understanding and mobile GIS intelligence.

3. Overall Design of System

The most direct application of voice in mobile GIS is navigation in LBS. Hence this paper designs a system model combine the voice with the navigation in mobile GIS. As the figure:

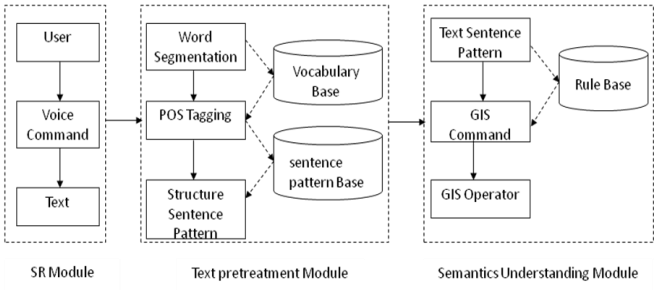


Fig.1 overall structure chart of system model

The system consists of three parts: the voice recognition module, the text preprocessing module and the semantic understanding module. Among them, the semantic understanding module is the core part of the system. Therefore transformation rules of natural languages to GIS commands is the research focus of this paper. The voice recognition module can identify user's voice commands by the voice recognition interface of an intelligent terminal and convert these commands to text languages. The text preprocessing module divides text sentences into small ones on the basis of the established lexicon, and constructs the input-pattern rules. The specific patterns include command operation, spatial query and spatial analysis. By this sentence pattern rules, we analyze segmented text languages so as to get the sentence patterns and do a standardizing disposal to the text languages. The semantic understanding module converses the treated sentences into corresponding GIS commands by the conversion rules to execute GIS operations, and achieve navigation for users.

4. Algorithm Steps

The artificial neural network could combine with a given learning sample to adjust the interconnection weights among neurons by faster speed and higher accuracy. And this method could make the system achieve stable state to content learning requirements. Moreover this algorithm also has some certain ability of self-study, promotion and summary. Combined with the BP algorithm of artificial neural network, this paper learns and trains the current syntactic structures, and builds an artificial neural network model of conversion natural languages to GIS commands.

The concrete steps as followed:

(1) To classified mark the vocabularies. As the table shows:

Table 1: Vocabulary classification and Numbers show

Example	label	number
Zoom out/in	Z	111
move	M	112
Open/save	O	120
gather	C	130
Look up/list	F	140
"From…to…"	SR、ST	151、152
"CQUPT"	I	210
"supermarket"	ES	220
"in/be apart from"	SV	310
"in the east/west"	SD	320
"adjacency"	SA	330

(2) To form digital sentence patterns, our experiment replaces the vocabulary in the table included in the natural language query sentence patterns into numbers and. For example, the sentence pattern of "from Chongqing University of Posts And Telecommunications to Jiefangbei" is "from... to ...". And according to the above-mentioned table, we mark the vocabularies and get the digital patterns as "151/210/152/210".

(3) To extract the keywords in the sentence patterns which can distinguish the sentence pattern, and make the keywords be the neurons of a three-layer BP neural network model. For instance, the key words distinguished the sentence pattern of "from Chongqing University of Posts And Telecommunications to Jiefangbei" are "from" and "to". Therefore, the neuron of this model is "151/152". By the same token, the other sentence patterns also can extract the neuron to set up the BP neural network model. Combine with these neurons, the input layer of the neural network model is formed.

(4)To mark the GIS commands. Such as "Draw the path from... to... " marks as 100, so do the other GIS commands. Combine with these marked commands, the output layer of the neural network model is formed. There is almost a one-to-one correlation between the input layer and the output layer.

(5) To Learn and train the input layer and the output layer by function $y = f(\sum_{i=1}^{m} w_i x_i)$, the weight coefficient of the network model is determined by function $\min \sum_{i=1}^{n} (d_{pi} - o_{pi})^2$.

(6) When the BP neural network model is stable, the conversion model of natural languages to GIS commands is formed preliminarily.

5. Experiment and analysis

(1) The experimental environment

The experiments of this paper is using client/server mode, the development environment as follows:

Server-side: MyEclipse8.5+Tomcat6.0

Client-side: Eclipse

(2) The experimental results

Based on learning and training the artificial neural network model set up by natural language sentence patterns, and then input the test sample data as:

$\begin{bmatrix} 140 & 342 & 0 & 0 \\ 140 & 341 & 0 & 0 \\ 140 & 330 & 0 & 0 \end{bmatrix}$, then get the output result: $\begin{bmatrix} 190 & 0 & 0 & 0 \\ 190 & 0 & 0 & 0 \\ 160 & 0 & 0 & 0 \end{bmatrix}$. To restore the data of the result, it will get three sentence patterns as: 140+342+E1+AU+E2, 140+E1+CJ+E2+341, 140+E1+340+E2, the corresponding GIS commands are: Judgment E1, E2 whether intersected and Judgment E1, E2 whether adjacent.

Therefore, the experimental results show that the method of this paper could achieve the purpose of identifying unrecognized sentence patterns by constructing artificial neural network model and study existing knowledge structures, and set up a conversion model of natural languages to GIS commands.

6 Conclusions

This paper utilizes the BP algorithm of artificial neural network to do research on natural language understanding in mobile GIS, and has got the artificial neural network model of conversion natural language to GIS commands preliminarily. But due to the limitation of sample data, the conversion accuracy of the network model needs to be improved. At the same time, the BP algorithm has the following disadvantages: (1) The weight value got from the BP algorithm is the local optimal value, not the global optimal value. (2) Model training continues for a long time. (3) It needs time and energy to determine the appropriate model, algorithm and parameter settings. Therefore, our next step work can utilize GA algorithm to optimize BP algorithm for an auxiliary, and improve the learning and training rate of network model.

Acknowledgments

The Project is supported by the National Nature Science Foundation of China (No.41101432), the Natural Science Foundation Project of Chongqing CSTC(No.2010BB2416)

References

[1] Zhang Xue-ying, and Lv Guo-nian, "Natural-language Spatial Relations and Their Applications in GIS", GEO-INFORMATION SCIENCE, Vol. 9, No. 6, 2007, pp. 77-81.
[2] Zhou Yan-kun, and Li Man-chun, "The Research of GIS Query Language Based on Chinese", BULLETIN OF SCIENCE AND TECHNOLOGY, Vol. 17, No. 1, 2001, pp. 35-40.
[3] Zhang Lian-peng, Liu Guo-lin, and JIANG Tao, "The Application of the Limited Natural Language Query on GIS", JOURNAL OF INSTITUTE OF SURVEYING AND MAPPING, Vol. 19, No.4, 2002, pp. 284-289.
[4] Ma Lin-bing, Gong Jian-ya. "Research on Spatial Database Query Oriented Natural Language", COMPUTER ENGINEERING AND APPLICATIONS, Vol. 39, No. 22, 2003, pp. 16-19.
[5] Cui Zong-jun, Tang Shi-wei, and Yang Dong-qing, "Studies on Er-model-based Restrictive-Chinese Query Language of Database", JOURNAL OF CHINESE INFORMATION PROCESSING, Vol. 15, No. 4, 2001, pp. 7-13.
[6] Xu Ai-ping, Xiong Hao, and Huang Yuan, "Semantic Information Extraction of the Query Conditions in GIS Chinese Query Sentences", COMPUTER ENGINEERING AND SCIENCE, Vol.29, No.8, 2007, pp. 99-101, 126.
[7] Ming Xiao-na, Long Yi, and Qian Chen-yang, "The Analytic Method of GIS Command Based on Constrained Natural Language", JOURNAL OF GEO-INFORMATION SCIENCE, Vol. 11, No. 2, 2009, pp. 183-188.
[8] Xu Rui, Wang Huilin, "Machine Learning Techniques Applied to Natural Language Processing", NEW TECHNOLOGY OF LIBRARY AND INFORMATION SERVICE,10.3969/j.issn.1003-3513.2008.z1.007.

Jiang-Fan Feng He received his B.S. degree from Southwest Agricultural University, and his Ph.D. degree from Nanjing Normal University, in 2002 and 2007. He works as associate professor of Chongqing University of Posts and Telecommunications. His main research area include spatial information integration and multimedia geographical information system.

Nan Xu She received the B.S in information management from ShiJiaZhuang Tiedao Univ. in 2009. She is currently working towards her M.S. in the Chongqing University of Posts and Telecommunications. Her current research interests include mobile GIS and multimedia geographical information system.

GIS Based Construction Land Layout in Ecological Area

Xiaolei Wu[1], Yinghong Wang [2] and Weixing Mao[3]

[1] Jiangsu key laboratory of Resource and Environmental Information Engineering , School of Environment Science and Spatial Informatics, China University of Mining and Technology
Xuzhou, Jiangsu, 221008, China
Henan Mechanical and Electrical Engineering College,
Xinxiang, Henan, 453002, China

[2] Jiangsu key laboratory of Resource and Environmental Information Engineering, School of Environment Science and Spatial Informatics, China University of Mining and Technology
Xuzhou, Jiangsu, 221008, China

[3] Environment and Planning College，Henan University
Kaifeng, Henan, 475002, China

Abstract

This paper explores the rational construction land layout in Qi river ecological area of Hebi city so as to provide references for the construction land layout in other similar domestic area. We take the geographical information system software (ARCGIS) and statistical software (SPSS) as technical support. We adopt qualitative analysis with quantitative calculation, data analysis with graphical analysis as research methods. The paper evaluates the suitability and the ecological sensitivity of the construction land in Qi River ecological area, then overlays the evaluation results by ARCGIS software, finally gets the layout of construction land in the ecological area. Conclusions are that: Rational layout of construction land in ecological relates area not only to the suitability evaluation but also to the sensitivity evaluation; there are strong correlations among ecological sensitivity and water system, land cover type, elevation, special value.

Keywords: *Ecological Area, Construction Land, Rational Layout, ARCGIS.*

1. Introduction

Healthy and stable ecological environment is the premise of survival and development of human society, and the layout of the construction land is an important regional economic foundation. Hebi is a both resource and tourist city. At present, the city is in the accelerated development period of industrialization, urbanization and modernization. Due to coal mining, environmental, pollution of Hebi spread from the point to the whole. The resource destruction is becoming more and more serious, which is a serious threat to the sustainable development of social economy. Therefore, the study on reasonable layout of construction land in Hebi ecological area has a strong practical significance to the rational utilization of ecological natural resources, human resources and tourism resources.

2. Research Methodologies

Taking the sustainable development and recycling economy theory as the guide, construction land of ecological area is evaluated under the suitability assessment and the sensitivity assessment. Then we overlay the two evaluation results by ARCGIS software to obtain the construction land layout of ecological area [1, 2].

3. Data sources

The research data is cited in Hebi official data, including the statistical yearbook of Hebi from 2000 to 2010, agricultural economics report of villages and towns of Hebi from 2000 to 2010, statistics of natural disaster of Hebi, land use maps of Hebi in 2009(1:1million), topographic maps of Hebi(1:5 million), traffic map of Hebi in 2011. We use the analysis, experiment and investigation method to obtain the data which is unable to access directly.

4. Empirical analyses

4.1 Construction land analysis of ecological area

Qi River ecological area is established in 2007, approved by Hebi People's Government, with total area of 37 square kilometers and population about 16000 people; Qi River flows from northwest to southeast, about 18 kilometers length, and the water quality is better. According to the land use map of ecological area of Qi River, we get the current land use situation: mountain and forest land is large, and the forest coverage is high; the town and rural residential land are relatively dispersive, local population density is less; there are some land use types, for example, traffic land, industrial and mining land, tourism land.

Ecological problems are summed up in the following aspects: the infrastructure was weak, and the existing road grade is in poor quality; construction land was dispersive and low using efficiency; municipal infrastructure corridor across the ecological area and the area was divided into pieces, with a high degree of landscape fragmentation; unauthorized reclamation and construction in the ecological region did big harm to the wetland resource of Qi River.

4.2 Suitability evaluation of construction land layout in the ecological area

According to the condition and features of Qi River ecological zone and some recognized indicators of construction land suitability evaluation. We built the construction land suitability evaluation system from 5 aspects (containing 15 impact factors): the engineering geological conditions, terrain conditions, geographical conditions, natural disasters and fundamental condition [3, 4, 5, 6]. Analytic hierarchy process was used to determine the index weight of construction land suitability in Qi River ecological area.

After using this method to calculate the weight of each influencing factor, we use the expert scoring method to amend the weight.

Table 1: Index weight of influence factor of suitable assessment system

Influence factor	***Weight***	***Specific factor***	***Weight***	***Index Weight***
Terrain conditions	0.150	Slope	0.200	0.030
		Aspect	0.150	0.023
		Elevation	0.350	0.053
		Topography types	0.300	0.045
Engineering geological conditions	0.100	Components of the earth's surface	0.300	0.030
		Bearing capacity of foundation (t/m^2)	0.250	0.025
		Depth of Groundwater	0.350	0.035
		Water and soil erosion	0.100	0.010
Geographical conditions	0.450	Hebi city circle radiation zone(Km)	0.200	0.090
		Radiation of ecological area and nearby towns (m)	0.450	0.203
		Traffic location(m)	0.350	0.158
Natural disaster	0.050	Flood buffer and water ecological isolation zone	0.400	0.020
		Geological stability	0.600	0.030
Fundamental condition	0.250	Land utilization	0.700	0.175
		Communicati-on, electric and water conditions	0.300	0.075

Using ARCTOOLBOX Union and Buffer functions for data spatial analysis in ARCMAP, we get the map of impact factors.

According to the respective weight of each topography factor in the graph based on slope, aspect, elevation and topography types, we overlay four maps to obtain influence graph of topography factors.

Construction land suitability assessment graph is obtained by the ARCGIS software according to the weight of each factor corresponding to the overall goal in the influence graph of topography factors, flood buffer and water ecological isolation area factor, traffic location

factor, land utilization factor, ecological area and nearby towns' radiation factor. By the construction land suitability assessment graph of Qi River ecological area, the appropriate and suitable area for construction was obtained. Number 1 to 5 in figure 1 represents suitable degree of ecological area from low to high.

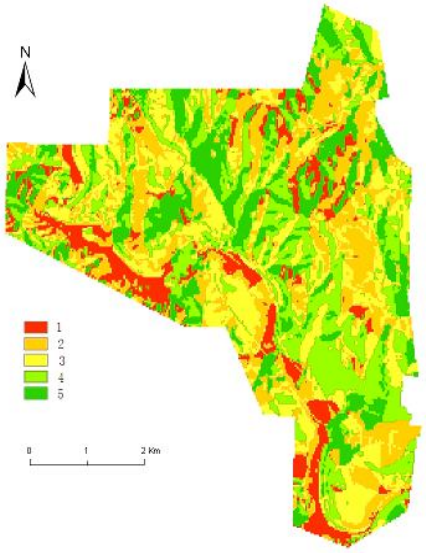


Fig.1 Suitability evaluation of construction land graph

4.3 Ecological sensitivity assessment of construction land

On the analysis of current situation of Qi River ecological area in Hebi, We built the construction land ecological sensitivity assessment system from5 aspects: elevation, slope, water, land and vegetation, special value factor [7, 8, 9, 10, 11].

The Delphi method is used to determine the sensitivity value of the internal components of single factor, and the assessment standards is divided into five levels,respectively"1, 2, 3, 4, 5". "1" represents non-sensitive index, its index is corresponding to regional non-sensitive area; "2" represents low sensitivity, its index is corresponding to the area for the low sensitive area; "3" represents middle sensitive, its index is corresponding to the area of the middle sensitive area; "4" represents sensitive, its index is corresponding to the area of sensitive area; "5" represents high sensitive, its index is corresponding to the region of high sensitive area. The higher ecological sensitivity of the area there is, the fewer suits for the construction land layout, and vice versa.

Table 2：Specific assignment of ecological sensitivity evaluation standard

Factor	*Classification*	*Index*
Elevation	90—120	1
	120—150	2
	150—180	3
	180—210	4
	210—495	5
Slope	0° —5°	1
	5° —10°	2
	10° —15°	3
	15° —20°	4
	＞20°	5
Water	200 meters outside of the river buffer	1
	150—200 meters outside of the river buffer	2
	100—150 meters outside of the river buffer	3
	50—100meters outside of the river buffer	4
	50 meters of water around	5
Land and vegetation	Current construction land	1
	Farmland	2
	Woodland	3
	Natural scenery protection areas, rivers and wetlands	5
Special value	Tai chi natural scenic area	5
	Qi River coast and core wetland area	5
	Jinshan Temple Scenic Area, Luo Guanzhong Literature Research Institute	5

In the ecological sensitivity assessment of Qi River in Hebi, the analytic hierarchy process method is used again to determine the weight of influence factor. After calculating the each weight of influence factor, the expert scoring method is used to amend the weight, finally we get the index weight of impact factor of ecological sensitivity assessment system. (as the table below).

Table3：Index weight of impact factor of ecological sensitivity evaluation system

Factor	*Water*	*Elevation*	*Slope*	*Land cover types*	*Special value*
Weight	0.3	0.20	0.10	0.2	0.2

Again using ARCTOOLBOX Union and Buffer functions for data spatial analysis in ARCMAP, the thematic map of impact factors was obtained.

According to the respective weight of each topography factor in the factor graph based on water factor, special value factor, elevation factor, slope factor, land cover types factor and topography types, four maps were superimposed to obtain ecological sensitivity factors influence graph. Number 1 to 5 in figure 2 represents sensitive degree of ecological area from low to high.

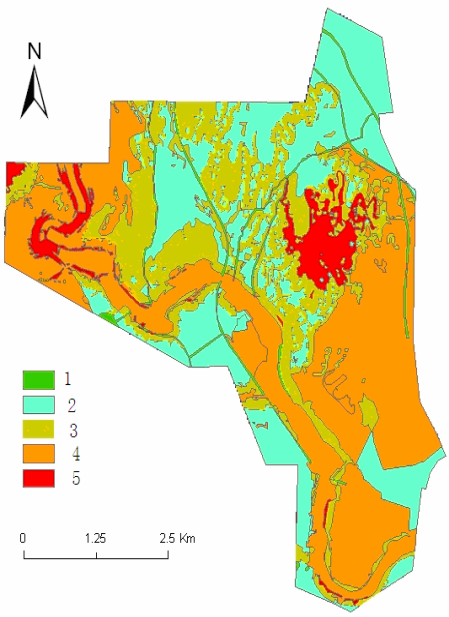


Fig.2 Ecological sensitivity evaluation graph

4.4 Comprehensive assessment of the construction land in ecological area

The suitability assessment and ecological sensitivity assessment graph is superimposed by ARCGIS software, and then comprehensive assessment of construction land in the ecological area is obtained. We get the comprehensive map of ecological sensitivity and suitability assessment by the weights of suitability and ecological sensitivity assessments are accounted for 0.3 and 0.7 (by the expert consulting results). Number 1 to 5 in figure 3 represents suitable degree of ecological area - from low to high.

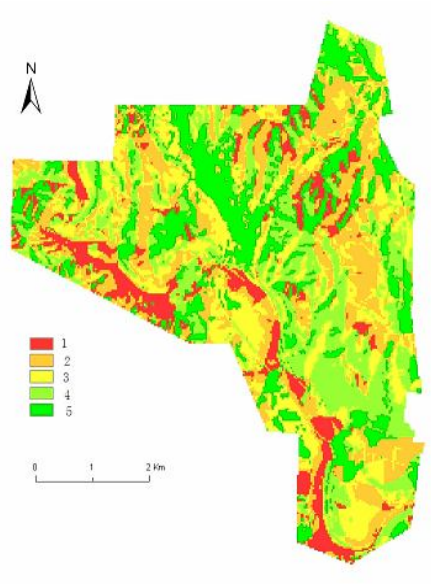


Fig. 3. The comprehensive map of ecological sensitivity and suitability assessment of construction land

Number 1 to 5 in figure 3 represents suitable degree of ecological area from low to high. Number 4 and 5 represents very appropriate and suitable area, mainly for the existing urban construction land, including the original rural residential, industrial and mining land distribution. These areas can withstand a certain degree of human interference, but are easy suffer from soil erosion and other natural disasters, with slower ecological restoration. Unsuitable areas for the construction land are the areas with fragile ecological environment and which are vulnerable to human disturbance. If the land is used improper, it can result ecosystem instability. The unsuitable areas for construction are mainly about 50 meters vertically away from the river, and some areas which are being ecological restoration construction now. The rest areas are middle unsuitable for construction, mainly for woodland distribution areas which are have large slope.

5. Suggestion of the reasonable layout of Qi River construction land in ecological area

5.1 Suggestion for urban construction land

The original urban construction land in ecological area is very suitable, so we can expand the scale and improve the urban population density and land utilization to make eco-town township form a centralized sheet pattern area based on the original building scale.

5.2 Suggestion for rural settlements

We should merge the existing rural settlements, move them to the four suitable settlements gradually, and reduce the current scale and improve the land use intensity. We should respect ecological resources and pay equal attention to the protection and development in the village construction, so as to build the rural ecological community, which is of green environmental protection with leisure and tourism.

5.3 Suggestion for the independent industrial and mining land

We can continue to arrange the independent industrial and mining land in situ site, and take control of the area scale. For those industrial and mining enterprises, which did serious damage to the ecological environment, should be closed, then take measures to reconstruct the ecological restoration immediately and accelerate the construction of water conservation forest.

5.4 Suggestion for traffic land

Considering the results of the construction land suitability assessment, skeleton of road system in ecological area planning should be cooperated with the urban space development. The improper original road construction should be adjusted. The road in the ecological area should avoid crossing the sensitive area; secondary road should set traffic road for walk, bicycles and other traffic tools.

5.5 Suggestion for waters and water conservancy facilities

The area along Qi River is not suitable for construction land, so it should be defined strictly protected areas. We should establish banning digging, mining, lumbering, grazing, reclamation areas and phosphorus prohibition area. The contaminative and poisonous industries should be banned in the area .But some constructions which are mainly for the purpose of water resources protection and wetland development should be arranged along the Qi River ecological protection buffer area, such as wetland science base, scientific research and observation point.

5.6 Suggestion for scenic spots and special use land

Unsuitable areas should be prohibited the construction activity. We should carry out the strict measures of protecting biological species, ecological environment and natural landscape to ensure the whole regional ecological safety. The land in core area is limited to scientific research and observation using. The land out of the core area is proper to develop the tourism and health with the main content of the accommodation, medical and other industries to improve the economic value of land.

6. Conclusion

The layout and assessment of construction land in ecological area is a comprehensive, highly relevant research topic. The following conclusions have been obtained through this study.

Study on the reasonable layout of construction land in ecological area should relate not only to the ecological sensitivity but also to the ecological suitability assessment of construction land. So we can characterize the suitable layout of the construction land fully and arrange the construction land reasonably.

Ecological sensitivity is great influenced by water system, land cover types, elevation and special value. Study on the ecological sensitivity assessment has shown that the weight of water system, land cover type, elevation and special value are higher. Therefore, their influences on ecological sensitivity are greater. We should strengthen the development and protection on these factors reasonably in order to promote the sustainable development of Qi River ecological area.

Acknowledgements

This work was supported by Henan government decision-making research funded projects NO.A238 (2010). The author wishes to express her most sincere appreciation to Prof. Changyou Chen, who read the manuscript carefully and gave valuable advice. Tremendous thanks are owned to Dr. Xi Wang for helping her with the data analysis. The author is also indebted to Hebi Bureau of Land and Resources for offering some data.

References

[1] L. Zhang and Y.G.Zong "Ecological suitability assessment of urban construction land use based on GIS——the case study of Liancheng county of Fujian province", Journal of Shandong Normal University (Natural Sciences), Vol.9, No.23, 2008, pp. 95-98.

[2] J.F.Zhou and G.M.Zeng, "The ecological suitability evaluation on urban expansion land based on uncertainties", Acta Ecologica Sinica, Vol.2, No.2, 2007, pp. 774-781.

[3] C.G.Wang, Y.G. Zong, "GIS-based ecological suitability evaluation for town development used-land in Dalian city". Journal of Zhejiang Normal University (Natural Sciences), Vol.30, No.1, 2007, pp. 109-114.

[4] Y.F.Chen and P.F.Du, "Evaluation on ecological applicability of land construction in Nanning city based on GIS",Journal of Tsinghua University (Science and Technology), Vol.46, No.6, 2006, pp. 80l-804.

[5] A.González and A.Gilmer, "Applying geographic information systems to support strategic environmental assessment: Opportunities and limitations in the context of Irish land-use plans", Environmental Impact Assessment Review, Vol.31, No.3, 2011 pp.368-381.

[6] A. Raizada and B.L. Dhyani, "Assessment of a multi-objective decision support system generated land use plan on forest fodder dependency in a Himalayan watershed", Environmental Modeling & Software, Vol.23, No. 9, 2008,pp. 1171-1181.

[7] J.Liu and J.Ye "Environmental Impact Assessment of Land Use Planning in Wuhan City Based on Ecological Suitability Analysis", Procedia Environmental Sciences, Vol.2, No.1, 2010, pp.185-191.

[8] M.Barral and M. Oscar, "Land-use planning based on ecosystem service assessment: A case study in the Southeast Pampas of Argentina, Agriculture", Ecosystems & Environment, Vol.154, No.7, 2012, pp.34-43.

[9] I.Santé-Riveira, R.C.Maseda and D.M.Barrós, "GIS-based planning support system for rural land-use allocation", Computers and Electronics in Agriculture, Vol. 63, No.2, 2008, pp.257-273.

[10] H.Nuissl and D.Haase, "Environmental impact assessment of urban land use transitions—A context-sensitive approach", Land Use Policy, Vol.26, No.2, 2009, pp.414-424.

[11] K.R. Manjula1, S. Jyothi2 and S. Anand Kumar Varma, "Digitizing the Forest Resource Map Using ArcGIS ", International Journal of Computer Science Issues, Vol. 7, No. 6, 2010, pp.300-305.

Xiaolei Wu attended China University of Mining and Technology in 2011. Currently, she is a Ph.D. student at School of Environment Science and Spatial Informatics and Jiangsu key laboratory of Resource and Environmental Information Engineering. Her major is land management. Her research interests focus on land reclamation, land use and planning.

Yinghong Wang is a professor in China University of Mining and Technology. His current research is land use and planning.

Weixing Mao received MS degree in Henan University in 2011.

A Comparative Study on Contamination Deposited Characteristics of ±800kV DC Line Insulators

Fangcheng Lv[1], Chunxu Qin[1], Yunpeng Liu[1], Wenyi Guo[2], Ruihai Li[2]

[1] Hebei Provincial Key Laboratory of Power Transmission Equipment Security Defense, North China Electric Power University
Baoding, 071003, China

[2] Electric Power Research Institute, China Southern Power Grid Corporation Limited
Guangzhou, 510000, China

Abstract

This report describes the natural contamination test results of ±800kV line insulators at high altitudes. Natural exposure tests with -816kV DC voltage energization were carried out at National Engineering Laboratory for UHV Engineering Technology (Kunming).The results showed that the strain type and V-type insulator strings show a lighter pollution and the insulators with a better aerodynamic shed shape show a better contamination deposited characteristic. Also the ratios of contamination on top surface to that on bottom surface in terms of the equivalent salt deposit density (ESDD) and the nonsoluble deposit density (NSDD) were influenced by the insulator shed shapes and the string types. The ESDD and the NSDD along the insulator strings presented decreasing trend from the earth side to the line side.

Keywords: *±800kV, Comparison, ESDD, NSDD.*

1. Introduction

Due to the economy development, energy demand and resource distribution in China, the ultra-high-voltage direct current (UHVDC) transmission lines have been built since 2006. The ±800kV Yunnan-Guangdong Line built by the China Southern Power Grid Corporation (CSG), has been in bipolar operation since June 2010. Several UHVDC transmission lines will be constructed in China during the Twelfth Five-Year Plan Period (2011-2015). Most of these lines will cross over high altitude regions, for example, the partial Nuozhadu-Guangdong ±800kV UHVDC transmission line will be situated in areas over 2000 meters above sea level[1-4].

A series of studies about the natural contamination deposited characteristics of the HVDC insulators have been studied in the USA, Japan, Sweden, France, and China since 1980s [5-13]. Several investigations have indicated that more contamination accumulated on DC insulators is more than on AC insulators. The amount of contamination collected is not uniformly distributed along a string. Insulators near the line and earth sides of a string accumulate more contaminant than those of middle part when DC voltage is applied. However, some measured the highest (ESDD) values measured in some researches occurred at the earth sides while the highest ESDD values measured in others occurred at the line sides. The investigations have showed that Estimations of contamination severity for inland areas should be made using data obtained from insulators with exposure periods of at least one year. And some investigations about the contamination deposited characteristics of ± 500kV transmission line have been reported in China [14-16].

However, in these researches, the range of the tested voltage levels was from 20kV to ±600kV, such as ±250kV, ±400kV and ±500kV, the voltage level is much lower than ± 800kV. The tested sites usually located in low altitudes, and also the insulators of ± 800kV are different from before. To ensure the safe and reliable operation of UHVDC transmission lines, the insulators with mechanical failing load of no less than 300kN are proposed to satisfy the electrical requirements [17-21]. However, there is no study on the DC natural contamination deposited characteristics of all the kinds of insulators with the minimum mechanical failing load of 300kN.

So, in order to provide important reference to the outdoor insulation design of the UHVDC transmission line at high altitudes, in this paper, the natural contamination deposited characteristics of all the kinds of ±800kV full scale insulators under -816kV energized for one year exposure are investigated and compared in high altitude areas.

2. Test Site and Specimen Insulators

2.1 Test Site

Exposure test was conducted at the National Engineering Laboratory for UHV Engineering Technology (Kunming), which is located in the Songming Town, Kunming City,

Yunnan Province of China, where at an altitude of 2100 meters. The insulators were energized with -816kV DC voltages, the highest operating voltage of the Yunnan-Guangdong ±800 kV DC transmission project in China. The power supply for the test came from the ±1200kV/500mA bipolar DC voltage generator with ripple factor less than 3%, in the National Engineering Laboratory.

2.2 Specimen Insulators

The specimen insulators consisted of nine insulator strings, which include seven shed shapes (type A~G) insulators as shown in Fig. 1 and Fig. 2. And the more detail of the insulators are listed in Table1.

Fig. 1 The tested insulators

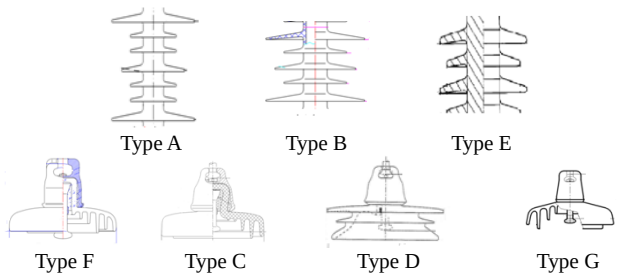


Fig. 2 Shed shapes of specimen insulators. The type A, B, E is long rod insulators. The type C, F, G is DC fog disc insulators. The type D is 3 outer-rib type insulators.

3. Exposure Test Periods and Contamination Test

3.1 Test Period

Located at an area of 2,100 meters above sea level on the Yungui Plateau with low latitude and high elevation, Somgming has one of the mildest climates in China, characterized by short, cool and dry winters with mild days and crisp nights, and long, warm and humid summers, but much cooler than the lowlands. The period from May to October is the rainy season and the rest of the year is dry. The rainfall in the rainy season is about 85% of annual precipitation. The exposure test period extended for one year from April 2011 to April 2012. So there were twice contamination measurement tests for the insulators in this study. The first test was in October 2011 before the dry season, and the second test was in the beginning of April 2012 before the rainy season.

3.2 Contamination Test

The test of contamination was referred to DL/T374-2010, GB/T16434-1996, Q/GDW152-2006 and other papers [22-27]. Test unit was defined as follows: one piece of insulator disc as one test unit for the type C, D, F and G insulators, and one group of sheds as one test unit for the type A, B and E insulators. In this study, the contamination on the top surface and the bottom surface of the insulator disc was tested respectively for the type C E F and G insulators, and the contamination on the top, the middle and the bottom surface of the insulator disc was tested separately for the type D insulator. There was not separately test on the top and bottom surfaces of the sheds for the type A and B insulators, however, a supplementary test was carried out for getting the contamination difference on the top and bottom surfaces of the sheds of the type A and B insulators. In order to get the contamination distribution along the insulator, three test units including one near the line side, the other in the middle and another near the earth side, were selected in every tested insulator string.

Table 1.Dimensions of insulators

Insulator String No.	*Materials*	*Shape*	*Shed Diameter (mm)*	*String Length (mm)*	*String Type*	*Insulator Type*	*MFL (kN)*
#1	SIR	A	218/166/121	8160	I	FXBZ-800/400	400
#2	SIR	A	218/166/121	12000	V	FXBZ-800/400	400
#3	SIR	B	218/178/138	8160	I	FXBWZ-800/400	400

#4	SIR	B	218/178/138	12000	V	FXBWZ-800/400	400
#5	porcelain	C	400	58×205	I	XZP-400	400
#6	porcelain	D	400	53×240	I	CA-779EX	550
#7	porcelain	E	265/235	8×1790	I	LG115/21+20/1790	400
#8	glass	F	390	52×240	I	LXZY3-550	550
#9	glass	G	380	47×195	strain	FC300P/195DC	300

Note:*MFL stands for mechanical failing load MFL

Table 2. Ratios of contamination on bottom surface to that on top surface in terms of the ESDD and the NSDD and the ratio of the NSDD to the ESDD after one year exposure period.

No.	*Shed shape*	*String type*	R_{ESDD}		R_{NSDD}		*NSDD/ESDD*	
			1^{st}	2^{nd}	1^{st}	2^{nd}	1^{st}	2^{nd}
#1	A	I	/	2.56:1	/	1.57:1	5.1:1	5.16:1
#2	A	V	/	1.27:1	/	1.1:1	5.15:1	5.96:1
#3	B	I	/	2.0:1	/	1.41:1	5.1:1	4.0:1
#4	B	V	/	0.92:1	/	1.0:1	6.6:1	5.71:1
#5	C	I	9.7:1	3.8:1	32:1	8.0:1	3.4:1	6.5:1
#6	D	I	11:1	2.7:1	37:1	1.8:1	3.2:1	4.17:1
#7	E	I	2.5:1	1.0:1	5.3:1	1.2:1	0.7:1	3.66:1
#8	F	I	16:1	5.8: 1	19:1	3.8:1	3.2:1	3.77:1
#9	G	strain	7.5:1	3.3:1	22:1	2.6:1	3.1:1	4.69:1

4. Results and Analysis

The typical contamination of the tested insulator strings is shown in Fig. 3. The measurement results of the ESDD and NSDD of the measured insulator strings exposed for one year under an energized condition are shown in Figures 4 and 5.From Figure 4, the values of the ESDD are more than 0.04 mg/cm2 except the strain insulator after one year exposure. The distribution of the contamination is uneven along the insulator from the Figure 3and Figure 5.

4.1 SIR Composite Insulators

Figure 4 shows the twice tests results of the ESDD and the NSDD of the composite insulators. From this figure the influence of the shed shape and the insulator string type on the contamination deposited is summarized as follows:
(1)After one year exposure, the ESDD of the I-type insulator strings (#1and #3) is about 25% more than that of the V-type (#2 and #4) ones. The NSDD of the #1 I-type insulator string is only a little more than that of the #2 V-type one, and the NSDD of the #3 I-type insulator string is less than that of the #4 V-type one. As for the ESDD and the NSDD of the first test, the ESDD and the NSDD of the I- type insulator strings are more than that of the V-type. It can be concluded that the I-type composite insulator shows a heavier contamination than the V-type one in this study.
(2)The 2nd test ESDD of the type A sheds insulator strings(#1and #2) is a little more than that of the type B sheds strings(#3and #4). And the ESDD growth of the #2 insulator string is more than that of the #4 one. The NSDD of the type A sheds insulator strings is obvious more than that of type B. And the NSDD growth of the type A sheds insulator strings is more than that of type B. The sheds spacing of the type A insulator is 26 mm, while that of type B insulator is 35 mm and the type B insulator shows a better aerodynamic type configuration. With a better aerodynamic shed, the type B insulator collected less contamination than the type A. It can be seen that the contamination deposited of the type B insulator is less than that of the type A insulator in the exposure time.

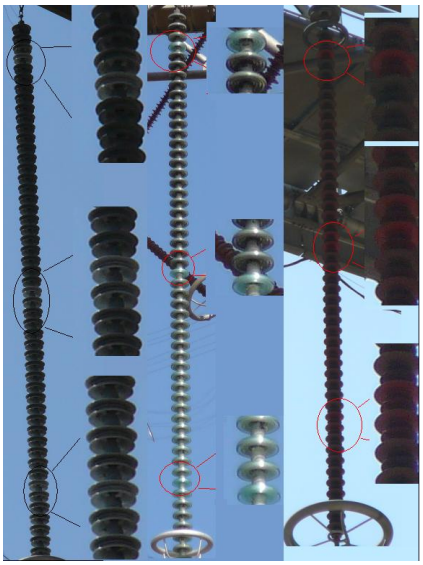


Fig. 3 The typical contamination of the test insulator strings. The left one is the porcelain insulator. The middle is the glass insulator. The right is SIR composite insulator.

4.2 Glass and Porcelain Insulators

Figure 4 also shows the twice tests results of the ESDD and the NSDD of the porcelain and glass insulators. From the figure the influence of shed shape and the insulator string type on the contamination deposited is summarized as follows:

(1) As for the glass DC fog disc insulator strings, the ESDD of the strain (#9) insulator string is about 1/3 that of the I-type #8 one. The NSDD almost has the same regulation as the ESDD. It can be concluded that the I-type insulator strings shows much heavier contamination than the strain-type ones.

(2) The #5 #6 #7 porcelain insulator strings with different shed shapes show the influence of the disc shed shape on the contamination deposited of the insulator strings. Among the three insulator strings, the ESDD of DC fog disc insulator string (#5) is the heaviest one, and the ESDD of the 3 outer-rib type insulator string (#6) is a little less than that of the #5 one. And the ESDD of the long rod insulator string (#7) is about 80% of that of the #5 one. The NSDD of the #5 string is the most serious among the three ones, and the NSDD of the #6 is about 60% of that of the #5. And the NSDD of the #7 string is the least one among the three, about 55% less than that of the #5. As for the comparison of the ESDD and the NSDD among the three insulator strings, the long rod insulator (type E) shows less contamination than the others (type C and D), and the 3 outer-rib type(type D) insulator shows less contamination than the DC fog disc insulator(type C) due to the aerodynamic type configuration. But the leakage distance of the long rod insulator string is the shortest among the three strings.

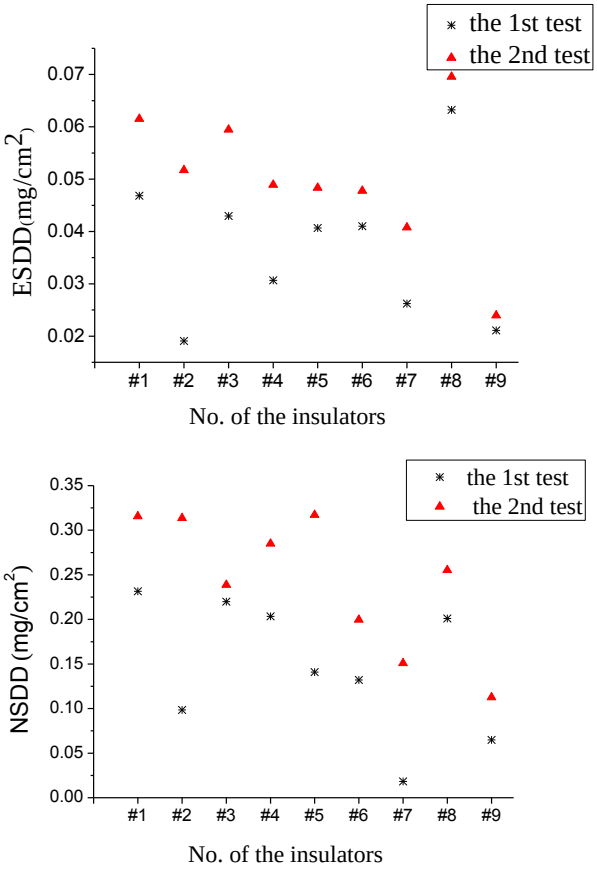


Fig.4 The values of ESDD and the NSDD of the twice tests results

4.3 Ratios of Contamination on Bottom Surface to That on Top Surface and Ratio of NSDD to ESDD

Table II shows the ratios of contamination on top surface to that on bottom surface in terms of the ESDD and the NSDD of all the insulators in the exposure period.

From the Table II, the ratios of the ESDD range from 0.92:1 to 5.8:1, and the ratios of the NSDD range from 1.0:1 to 8.0:1. The data are dispersive. The ratios of the RESDD and the RNSDD of the long rod insulator strings (#2 #4 and #7) are very close to each other after one year exposure. These ratios of the DC fog disc insulator strings (#5 #8 and #9) are quite close to each other, and that of the 3 outer-rib type insulator string (#6) is somewhere in between the two shape type insulators .

As for the comparison of these ratios between I-type glass DC fog disc insulator string and the strain type one, the former is greater than the latter. And these ratios of the I-type composite insulators (#1and #3) are greater than those values of the V-type insulators (#2 and #4).

The RESDD of the 1st test at rain season is much greater than that of the 2nd test at the dry season. And the RNSDD is similar to the RESDD. This is caused by the rain washing. The contamination on top surface of the insulator shed was washed more easily in the rain season. In the dry season, the rainfall has little effect on the contamination.

From the above, the shed shape of the insulator and the insulator string type are the influencing factors of the ratios of the ESDD and the NSDD on the bottom surface to that on the top surface. The results show that: these two ratios of insulators with the aerodynamic shed shape (Type A, B, D and E) are lesser than that of the DC fog disc insulators, and these ratios of the I-type insulator strings is greater than those of the V-type and strain type ones.

The ratios of the NSDD to the ESDD of all the insulators of the 2nd test are listed in the Table II, the maximum value is 6.5:1, while the minimum value is 3.66:1, and the average value is about 4.95:1. From the ratios, there is no evidence difference among the various type tested insulators. The data also show that the shed shape and the string type have no clear effect on the ratio of NSDD to ESDD. The data of the 1st test are not the same as that of the 2nd test, but there is no obvious difference between the two tests. It can be seen from these data that the ratio of the NSDD to the ESDD has nothing to do with the insulator shed and string type.

4.4 The ESDD and the NSDD along the Insulator Strings

The results of the ESDD and the NSDD along the insulator strings are shown in Figures 3 and 5. The distribution of the ESDD and the NSDD along the insulator

strings is uneven shown in Figures 3 and 5.It was observed that the earth side of insulator became darker and the color of surface contaminant on the earth side of insulator was different from that on the middle and line side ones from the Figure 3. The distribution of the ESDD and the NSDD along the insulator strings presented decreasing trend from the earth side to the line side. It can be seen from the Figure 5 that for all the tested insulators, the ESDD of the insulator sheds near the earth side were the most serious; the ESDD values of most strings were around 0.08 mg/cm2. The ESDD values in the middle were much smaller than those of the earth side, only around 0.04 mg/cm2. And the data near the line side were a little less than those in the middle. The NSDD of the entire insulator strings also appeared the same phenomenon. This phenomenon of the distribution of the ESDD and the NSDD is different from the former reports, which at lower voltage levels. The distribution of contamination along the insulator string is influenced by kinds of factors such as electric filed, the mass of dust particle, the diameter of dust particle and et al. In DC electric filed, the dust particles can be charged by multiple ways and move in a definite direction by the electrostatic force. And the wild force and gravity also can influence the dust particles movement. However, how the factors effect on the distribution of the ESDD and the NSDD still needs more studies.

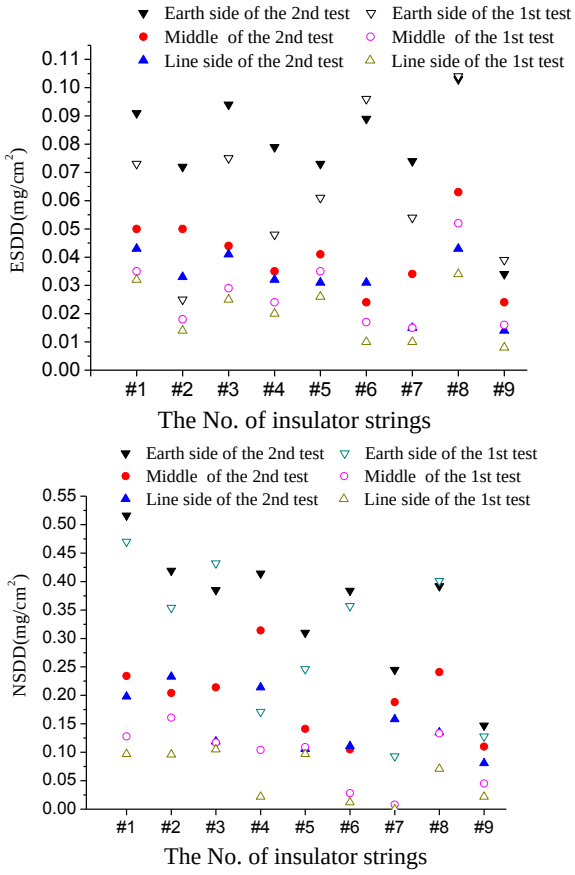


Fig. 5.The distribution of the ESDD and the NSDD along the insulator stings

4.5 Results of Chemical Analyze of Contaminants

According to the results of quantitative analysis of contaminants as shown in Fig. 6, the Ca^{2+} occupies the maximum component of the cations, and the SO_4^{2-} is the maximum one of the anions. It is recognized that calcium sulfate corresponds to $CaSO_4$ (Bassanite). So about 75% of soluble contaminants was $CaSO_4$, and also about 17% of soluble contaminants was sodium salt (NaCl). The results influenced the insulator withstand voltages according to [28].

The major compositions of insoluble contaminant were silica oxide (SiO_2), alumina (Al_2O_3) and ferrite (Fe_2O_3)

which were contained within the soil, and calcium oxide (CaO), which came from limestone near the test site. And else (mainly carbon) accounting of 31.44% of the total weight, which is the maximum component of the nonsoluble contaminants. It is found that the content of carbon increase slightly with dc voltage energization in [3], however, in this study the content of carbon increase greatly. This phenomenon may be noticed in the ±800kV projects.

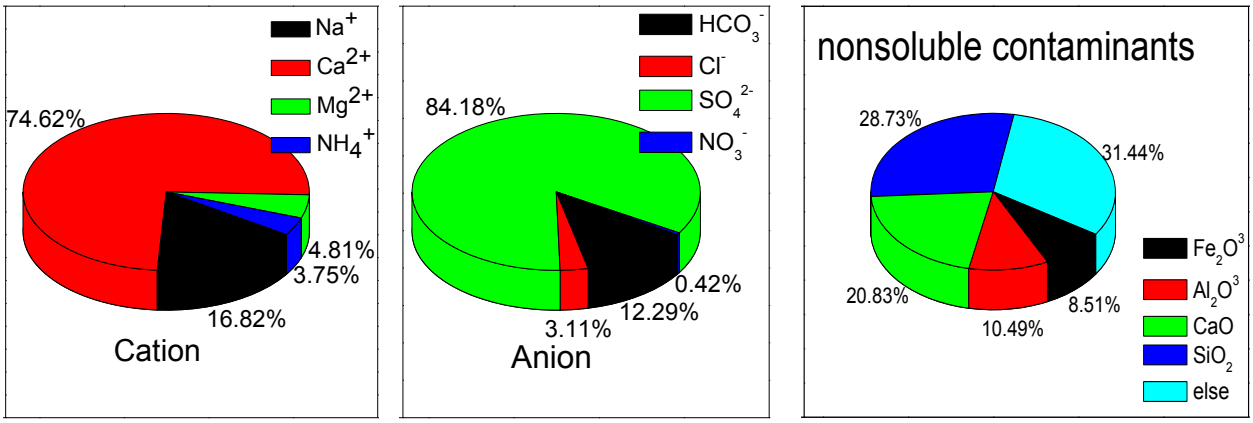


Fig. 6 Quantitative analysis results of contaminants

5. Conclusions

The ESDD and the NSDD of the tested insulator strings were measured after one year exposure under an -816kV energized condition, and the results of the ESDD and the NSDD have been analyzed in this study, the following conclusions have been reached:

1 The insulator string type influence on the contamination deposited of the insulators. The V-type insulator string has lighter contaminants than that of I-type ones, and the strain type insulator string has much lighter contaminants than that of the I-type ones.

2 The insulator shed shape was an important factor that influenced the contaminants of the insulator strings. The shed shapes with a better aerodynamic type configuration show a better contamination deposited characteristic.

3 The ratios of contamination on top surface to that on bottom surface in terms of the ESDD and the NSDD were influenced by the insulator shed shape and the string type. It showed that the insulator with an aerodynamic type configuration has a uniform contamination on both top surface and bottom surface, and the V-type insulator string also has a uniform contamination. And the shed shape and the string type have no clear effect on the ratio of NSDD to ESDD.

4 The distribution of the ESDD and the NSDD along the insulator strings presented decreasing trend from the earth side to the line side. This phenomenon was different from the former low voltage results, and more studies need making for the reason.

5 $CaSO_4$ was the dominant component of soluble contaminants by the chemical analysis. The content of carbon increased greatly than the former reports.

Acknowledgments

This work was supported in part by the National Natural Science Fund of China (51077054) and scientific and technological projects of China Southern Power Grid (K201016).

References

[1] Yin-biao SHU, "Development and execution of UHV power transmission in China," Electric Power, Vol.38, No.11, 2005, pp.1-8.
[2] Chun SHANG, "Development of ultra-high voltage transmission technology in China southern power grid ", High Voltage Engineering, Vol.32, No.1, 2006, pp.35-37.
[3] Qing-yun YUAN, "Present state and application prospect of ultra HVDC", Power System Technology, Vol.7, No.14, 2005, pp.1-3.
[4] Hao ZHOU, Yu-hong YU, "Discussion on several important problems of developing UHV AC transmission in China", Power System Technology, Vol.29, No.12, 2005, pp.1-9.
[5] JW. Lampe, T. Höglund, C. Nellis, P. Renner and R. Stearns, "Long term tests of HVD insulators under natural pollution conditions at the Big Eddy Test Center", IEEE Trans. Power Delivery, Vol.4, No.1, 1989, pp.248-259.
[6] C. Y. Tang and X.D. Liang, "A brief introduction to service performance and natural contamination test on abroad DC polymeric insulators", Power System Technology, Vol.23, No.9, 1999, pp.50-53.

[7] Kazuhiko Takasu, Takatushi Shindo and Noburo Arai, "Natural contamination test of insulators with DC voltage energization at inland areas", IEEE Trans. Power Delivery, Vol.3, No.4, 1988, pp.1847-1853.
[8] T. C. Cheng, C. T. Wu, J. N. Rippey and F. M. Zedan, "POLLUTION PERFORMANCE OF DC INSULATORS UNDER OPERATING CONDITIONS", IEEE Transactions on Electrical Insulation, Vol. EI-16, No. 3, 1981, pp.154-164.
[9] T. Kawamura, K. Nagai, T. Seta, K. Naito, "DC pollution performance of insulators", CIGRE, Paris, France, Report 33-10, 1984.
[10]Z.Y. Su, "Survey of Insulators Gezhouba-Nanqiao HVDC Transmission Line and Nanqiao HVDC Converter Station in East China with Regard to Natural Pollution", Power System Technology, Vol.17, No. 8, 1993, pp.9-15.
[11]Z.Y. Su and Y. S. Liu, "Comparison of natural contaminants accumulated on surfaces of suspension and post insulator with DC and AC stress in northern China's inland areas", Power System Technology, Vol.28, No.10, 2004, pp.13-17.
[12]Z.Y. Li, X.D. Liang, B. Wang and Yuan-xiang Zhou, "Natural pollution deposit test of polymeric insulators operated under DC voltage", Power System Technology, Vol.37, No.14, 2007, pp.10-14.
[13]G.L. Wang, R.H. Li, G.Q.Lu and Xi-dong Liang, "Design and Function of Long Term Live Examination Field for UHVDC Equipments", SOUTHERN POWER SYSTEM TECHNOLOGY, Vol.3, No.6, 2009, pp.22-26.
[14]H.F. Gao, L.M. Fan, Q.F. Li and et al, "Comparative analysis on pollution deposited performances of insulators on the ±500 kV Gao-Zhao DC transmission line", High Voltage Engineering, Vol.36, No.3, 2010, pp.672-677.
[15]W. Cai, Y. Xiao, Guang ya Wu, "Elementary Analysis o f Pollution Rules of Insulators on ±500 kV DC Transmission lines", High Voltage Engineering, Vol.29, No.6, 2003, pp.4-4, 40.
[16]Guang ya Wu, X.S. Guo, R. Zhang, "Pollution External Insulation Design and Arrangement of UHVDC Transmission Line", High Voltage Engineering, Vol.34, No.5, 2008, pp.862-866.
[17]Guang ya Wu, B. Luo, H. Wang, "Reliability Analysis of ±800 kV UHVDC Insulators", High Voltage Engineering, Vol.34, No.9, 2008, pp.1082-1086.
[18]Guang ya Wu, W.CAI, Y.L. LU, Y. Xiao, Q. J. Zhao and C.L. Xue, "Selection of numbers of insulator of pollution insulator strings for DC transmission line", High Voltage Engineering, Vol.27, No.6,2001, pp. 51-53.
[19]GUAN Zhi cheng, ZHANG Fu-zeng, WANG Xin, et al, "Consideration of external insulation design and insulator selection of UHVDC transmission lines", High Voltage Engineering, Vol.32, No.12, 2006, pp.120-124.
[20]Guang ya WU and Rui ZHANG, "Reliability analysis of insulators technology and economy for UHV transmission and distribution equipment", Electrical Engineering, No.6, 2006, pp.54-58.
[21]Guang ya Wu, "Domestic Development Situation of Insulator and Problems Needing Consideration", Electric Power Technology, Vol.19, No.3, 2010, pp.1-4.
[22]DL/T374-2010, "Drawing method of pollution distribution map for electric power system", 2010.
[23]GB/T16434-1996, "Environmental pollution classification and external insulation selection for high voltage transmission line, power plant and substation", 1996.
[24]Q/GDW152-2006, "Construction technology guide for tension stringing of 1000V overhead transmission line", 2006.
[25]J.Q. Wang, Q.H. Shen, K. Liu and X. Cheng, "Test and Research for the Natural Contamination on 500kV Transmission Lines", China Power, Vol.18, No.6, 1994, pp.46-50.
[26]T. Wang, Q.H. Ou, H. Jiang, J. Li and S. Yao, "TEST AND RESEARCH ON CLEANING OF HIGH VOLTAGE TRANSMISSION LINE BASED ON SALT DENSITY", Power System Technology, Vol.28, No.4, 2004, pp.22-26.
[27]Yanming CAO and Weimin MA, "Pollution Measurement Technology of Insulators for UHVDC", High Voltage Engineering, Vol.33, No.1, 2007, pp.22-25.
[28]X. Lin, Z. Chen, X. Liu and et al, "Natural Insulator Contamination Test Results on Various Shed Shapes in Heavy Industrial Contamination Areas", IEEE Transactions on Electrical Insulation, Vol.27,No.3,1992, pp.593-599.

Fangcheng Lv was born in P. R. China in 1963. He received the M. Sc. degree in 1989 and Ph. D. degree in 1999, both from North China Electric Power University (NCEPU), Baoding City, Hebei Province, P. R. China. Since 1987, he has been with the NCEPU as a teacher. Since 2003, he is engaged with teaching and research as a Professor and now he is the Director of the Electrical Engineering Department of NCEPU. His research interests include the study of monitoring and diagnosis of insulation.

Towards an Intelligent Project Based Organization Business Model

ALAMI MARROUNI Oussama[1], BOUKSOUR Othmane[2] ,BEIDOURI Zitouni[3]

[1] *Department of Mechanical & Industrial Engineering Production Engineering (LMPGI), University Hassan II Ain Chock, School of Technology, Km 7 Route El Jadida, Casablanca, 20100, Morocco*

[2] *Department of Mechanical & Industrial Engineering Production Engineering (LMPGI), University Hassan II Ain Chock, School of Technology, Km 7 Route El Jadida, Casablanca, 20100, Morocco*

[3] *Department of Mechanical & Industrial Engineering Production Engineering (LMPGI), University Hassan II Ain Chock, School of Technology, Km 7 Route El Jadida, Casablanca, 20100, Morocco*

Abstract

Global economy is undergoing a recession phase that had made competition tougher and imposed new business framework. Businesses have to shift from the classical management approaches to an Intelligent Project Based Organization Model (IPBOM) that provides flexibility and agility. IPBOM is intended to reinforce the proven advantages of Project Based Organization (PBO) by the use of suitable Enterprise Intelligence (EI) Systems. The goal of this paper is to propose an IPBOM that combines benefits of PBO and EI and helps overcoming their pitfalls.

Keywords: *Intelligent Project Based Organization (IPBO), Enterprise Intelligence (EI), Project Based Organization (PBO), Project Management (PM)*

1. Introduction

Fundamental changes in global and regional economies are driving the need for high quality information and knowledge [1, 2], on one hand, and flexible organizational structures on the other hand [3].Businesses have to prepare their structures and their management styles to sustain differentiating competitive advantages. They are urged to readapt the way they think, make decisions and operate to the new economic framework characterized by a shift to Intelligent Project Based Organization Model (IPBOM). This model is intended to reinforce the proven advantages of PBO [4, 5] by the use of suitable (EI) Systems [6]. PBO and EI had been extensively but separately discussed in literature and many advantages had been used to support the idea that both approaches are necessary [6, 7, 8]. However we still need to analyze in depth the conjunction of the two aspects and to propose a model that combines their benefits, overcomes their pitfalls and helps executives assimilate the relevance of IPBOM in nowadays turbulent environment.

2. Enterprise Intelligence (EI) Systems

2.1Concept and definition

EI is the ability to transform and valuate business information with regard to its currency and relevance [6]. It is a broad category of systems, applications and technologies for gathering, providing access to, and analyzing data for the purpose of increasing the organization intelligence and therefore helping enterprise users make better business decisions [6]. The term EI was chosen instead of the "strategic intelligence" used by] Liebowitz6] because the latter seems to be limited to strategic issues while the first covers tactical ,operational .and strategic intelligence

In EI system CI, BI and KM are utilized in conjunction to feed organization intellige ,with each otherncе.

2.2 Business Intelligence (BI)

2.2.1Concept and definition

Brought up by Gartner group since 1996 [9], BI is defined as a set of tools and processes that gather internal data from several sources, organize them, process, store and present them to end users in order to improve the decision making in the organization and generate value through information and knowledge [6].

2.2.2 BI System

A BI system is generally composed of three major phases (Figure 1)

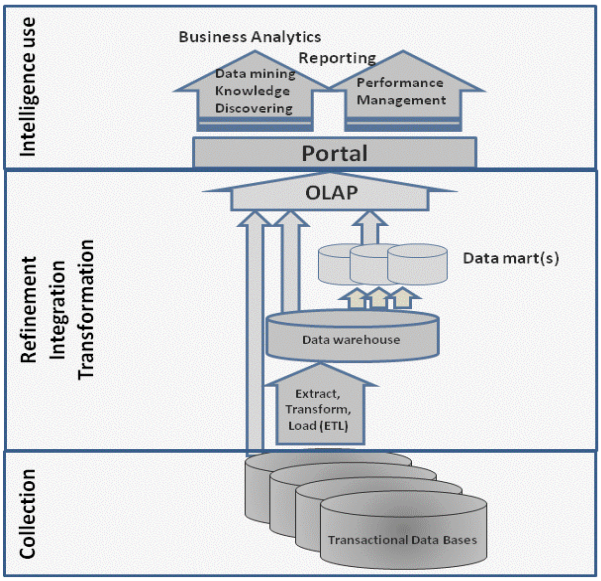


Fig. 1: BI system (source: Adapted by authors from [10, 11, 6]

Collection phase
Data is collected from different data sources (operational databases, historical data, or information from the already existing data warehouse environment) within organization.

Refinement, integration and transformation
ETL: The extraction transformation and load process of required data from specific data sources in the organization [10].
The data warehouse :(DW) Subject oriented and integrated, DW supports the physical propagation of data by handling the numerous enterprise records for integration, cleansing, aggregation and query tasks.
Data mart: A data mart as described by [11] is a collection of subject areas organized for decision support based on the needs of a given department.

Intelligence phase
In this phase intelligence user can use the basic retrieval level trough reporting or go further to the next and very high value added levels such as advanced analytics or corporate performance management [10].

2.2.3 BI and data Architecture
Business intelligence goal of transforming internal data into actionable information is only achievable if it is built on data of a guaranteed quality, which is relevant to the business. In order to enable this, a successful data architecture framework is vital.

Data architecture covers the provision of a structured framework for an organization's data, enabling that organization to develop and evolve its systems and processes in order to support its current business activity and, most importantly, allowing it to change in order to achieve its strategic goals in a cost-effective manner [12].

2.2.4 Information quality improvement
The following information gap can easily be fulfilled thanks to a BI system [13]:

- Data required for analysis is located in different sources that are hard to integrate.
- Data sources are inconsistent.
- Management gets extensive reports that are rarely used or inappropriate.
- Data within operational databases is not properly arranged to support management's decision.

2.3 Competitive Intelligence
2.3.1 Concept and definition
Competitive Intelligence consists of the analysis of information gathered from the market place, in contrast with BI dealing internal data [6,14] and the generation of recommendations for decision makers, done in an ethical and legal manner. It is involved with the development of a systematic program for capturing, analyzing, and managing external information and knowledge to improve organizational decision-making capabilities [6, 15, 16].

2.3.2 Benefits of CI
To achieve CI goal, organizations need to create a competitiveness corporate culture, allowing for the exchanging of knowledge and ideas among individuals and departments [17].
CI serves the following primary purposes [15]:
•Market, industry, political, customer, supplier, and technological profiling, benchmarking, and assessment;
•Early warning of opportunities and threats;
•Support for strategic planning and implementation; and
•Support of strategic decision making.

2.3.3 CI system
An adapted model of CI 4C-cycle proposed by A. Weiss [17,18] is presented in Figure 2

2.3.4 CI analysis
Analysis is essential and should explore the entire external environment, including the general and task environment, and not be limited to competitors only. A complete analysis will assist in shaping the appropriate strategy for the organization by detecting trends that should be monitored and assessed [19,20].

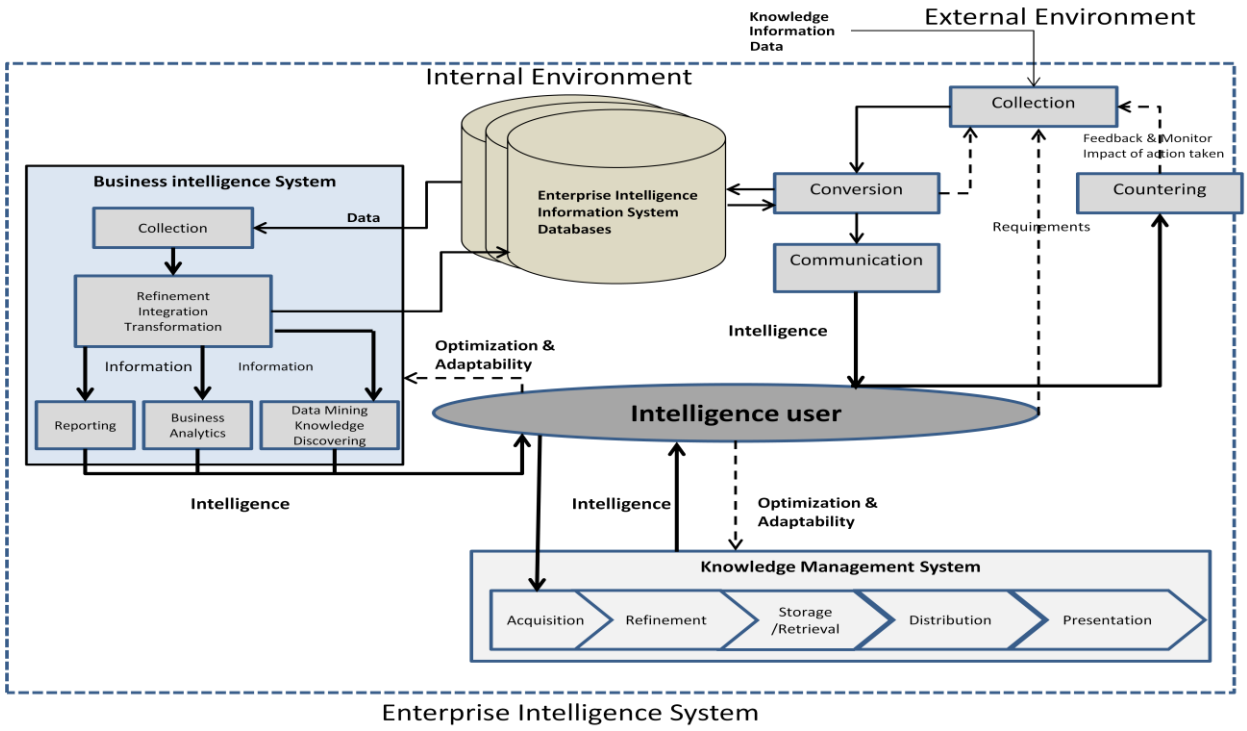


Fig. 2: EI system (source: adapted by authors from [10, 11, 6, 17, 22]

2.4 Knowledge management

2.4.1Concept and definition

Knowledge is present in ideas, judgments, talents, root causes, relationships, perspectives and concepts. Knowledge is stored in the individual brain or encoded in organizational processes, documents, products, services, facilities and systems. Knowledge is action, focused innovation, pooled expertise, special relationships and alliances. Knowledge is value-added behavior and activities [21].

KM can be defined as the collection of processes that govern the creation, dissemination, and utilization of knowledge.

2.4.2 Knowledge transformation

Knowledge encompasses both tacit and explicit knowledge. It is not static; instead, it changes and evolves] during the life of an organization22].

2.4.3 KM system

KM process model that could be used for knowledge capture, creation, and distribution and sharing is shown in Figure 2.

2.5 Enterprise Intelligence (EI)

2.5.1 EI system

As conjunction of the BI, CI and KM systems, EI system can be approached trough the following scheme (see Figure 2)

This architecture relies on the crucial role of the intelligence user that should:

- Identify the strategic, tactical and operational need of information
- Define the quality and the scope of required information
- Be able to utilize the gathered information in its various forms
- Adapt and optimize BI,CI and KM systems according to, but not limited :
  - ✓ The new external environmental factors: new market trends, new competitor's new regulations etc.
  - ✓ Information technology Immersion degree: to what extent is the organization familiar with advances in information technology
  - ✓ The strategic orientations

2.5.2EI success criteria

To successfully handle EI initiatives three relevant issues must be considered:

- The central role of intelligence-user in the whole Intelligence cycle and, in particular, in direction and use phases;
- The importance of suitable EI processes that consider specificities of the company
- The effective and efficient EI infrastructure that is necessary to achieve expected EI results

3. Project Based Organizations (PBOs)

3.1 Definitions

Project

A project is a temporary organization to which resources are assigned to undertake a unique, novel and transient endeavor that involves managing the inherent uncertainty and need for integration in order to deliver beneficial objectives of change [4].

Project Based Organization (PBO)

PBOs are organizations in which the majority of products are made against bespoke design for customers. These types of organizations may be stand-alone, making products for external customers, or subsidiaries of larger firms, producing for internal or external customers. They may also be consortiums of organizations that collaborate in order to serve third parties [5] Project-based companies are often involved in several projects simultaneously.

Definitions of PBO vary, but a key point is that PBOs possess all internal and external resources, as well as individual functions such as development, production established organizations are structured to and ,and sales execute business as individual projects .

The structure of PBOs has come to be applied to a range of industries, especially construction, IT, communications, automobiles, the media and consulting and professional services [4].

PBO can refer either to the entire company or to a department within a company.

In a PBO structure, a company's departments and personnel are organized around each particular project. For example, many PBOs have project managers that run teams of employees. These employees are often from different departments and have different job titles, but all are needed to get the project done. Typically, there are many teams operating at once, but they have no need to interact with each other because each team is focused on completing its project [5].

Project Management

In competitive environment businesses in general find themselves in search of disciplined approach to gaining market share or even surviving. Project management (PM) as management discipline involving, planning, organizing, and managing the resources needed to bring about a successful conclusion is the ultimate solution. For non PBOs, this approach simply helps getting better organized [8]. But there are other specific reasons to use PM [23]:

- It establishes a single point of contact and accountability for the overall success of the project.
- It focuses on meeting customer needs and expectations.
- It improves performance in time, cost, and technical areas
- It obtains consistent results through the definition and application of a process across the business unit. It focuses on managing project scope and controlling change.
- It helps avoiding disasters by managing risk.
- It strengthens project teams and improves morale.

3.2 Benefits of PBO model

In PM literature many benefits of PBO are pointed out:

- In the current knowledge-based economy, PBO model allows management to integrate advanced knowledge from multiple viewpoints dispersed within and outside the organization[3].
- PBOs are flexible and autonomous enough to be optimal and to generate business models for new products and services [4].
- PBO is suitable organization structure for large companies to implement the most important themes, such as projects to enhance management efficiency or develop new products[4].
- Research into inter firm alliances has emphasized the importance of PBO in collaborative ventures [3].
- PBO represents an important complement to formal organizational structures, dedicated .g.e departments[3].
- PBOs are very beneficial for inward and outward knowledge transfer [3].
- Projects most likely contribute to a firm’s proficiency in conducting the critical tasks throughout] transfer process the knowledge3].

3.3 Knowledge challenges for PBOs

The discretion required to manage every project with regard to its own objective and constrains can be a serious obstacle facing knowledge sharing effort between interrelated projects. The PBO tend to suffer from weaknesses in company-wide development and learning, and difficulties in linking projects to firm-level business processes. Furthermore, projects typically comprise a mix of individuals with highly specialized competences, belonging to functionally differentiated worldviews

making it difficult to establish shared understandings and common knowledge base [5].
Also relevant pieces of knowledge are distributed into a multitude of local settings and a great amount of knowledge resides in individual members [5]. Finally, the time and money limits of individual projects may cause problems when it comes to knowledge sharing [5].

3.4 Limitations of PBO model

Despite the fact PBOs operate trough projects, not all projects have a well-defined objectives and sufficient resources to carry out all the required tasks [5].
PBOs tend to be not only strongly decentralized, but also quite loosely coupled, and the division between functional and PBOs is not at all clear-cut. The functional organizations appear to be growing more project-based and the PBOs growing more routinized [5].

3.4 Success criteria for PBOs

PBO model can be beneficial and bring all the benefits cited above in terms of flexibility and knowledge transfer if projects are consistently managed following effective PM processes. Such processes are meant, in addition to increase projects success probability by implementing the best practices, to preserve and capitalize project knowledge to avoid wasting time in reinventing the wheel and loosing the main advantage brought by PBO model with regard to knowledge transfer.
Meanwhile PM adds value-but only when applied in the proper dosage [7].It is much wiser to apply a little, measure the success, then build up where needed into more sophisticated approaches than to drown your best and brightest in paperwork. Adopting PM processes should take into consideration the organization ability and the relevance of gradual organizational. That is the goal of maturity models in general and those dealing with project governance in particular.

4. IPBO Business Model

4.1 Model description

PBOs use projects to execute their strategies therefore they have to:

- Identify and undertake adequate projects portfolio with regards to its strategy or in other words strategic management -doing the right things
- Succeed these projects trough effective project management processes or doing things right-performance management

EI system will impact both strategy and performance as :shows 3the Figure

- portfolio identification and selection trough CI tools
- Performance Management troughBI tools
- PM and project knowledge capitalization through KM tools
- Broken lines highlight the complementary role of adaptive and restructuring feedbacks to ensure the required agility.

4.2 Business value of IPBO for PBOs

4.2 .1 Managing project information

PM is an information-intensive activity, and information or knowledge generated during the project is either archived or by default often destroyed, or at best it becomes difficult to retrieve tacit information locked away in a silo [21].

Two relevant types of information are valued thanks to EI system [24]:

- External information to project but internal to organization.
- External information to the project, referred to as environmental factors

Information management during the project life cycle is extremely important. But information is not knowledge unless that information is organized and processed in a meaningful way.

4.2 .2 Managing project knowledge

While technology facilitates generating and organizing information, nowadays information overload can also impede efficiency and affect productivity if not properly managed trough an adequate EI system [8].To be successful, the manager will be required to “deploy” the knowledge resource (knowledge worker) where that worker’s specialized knowledge can make the greatest contribution.
In PBOs, the task of managing knowledge is even harder the because of typically discrete nature of projects and their lack of continuity. One of the main benefits of EI to PM is the project knowledge gradual master it allows.
Indeed without EI contributions it will be difficult to take advantage of tremendous amounts of data resulting from project activities.

4.2 .3 Benefits of EI for Project Managers

EI provides Project Managers with decision support in three dimensions:
•Structural dimension: enhance its ability to define relationships of organizations & people (stakeholders) so decisions can be implemented

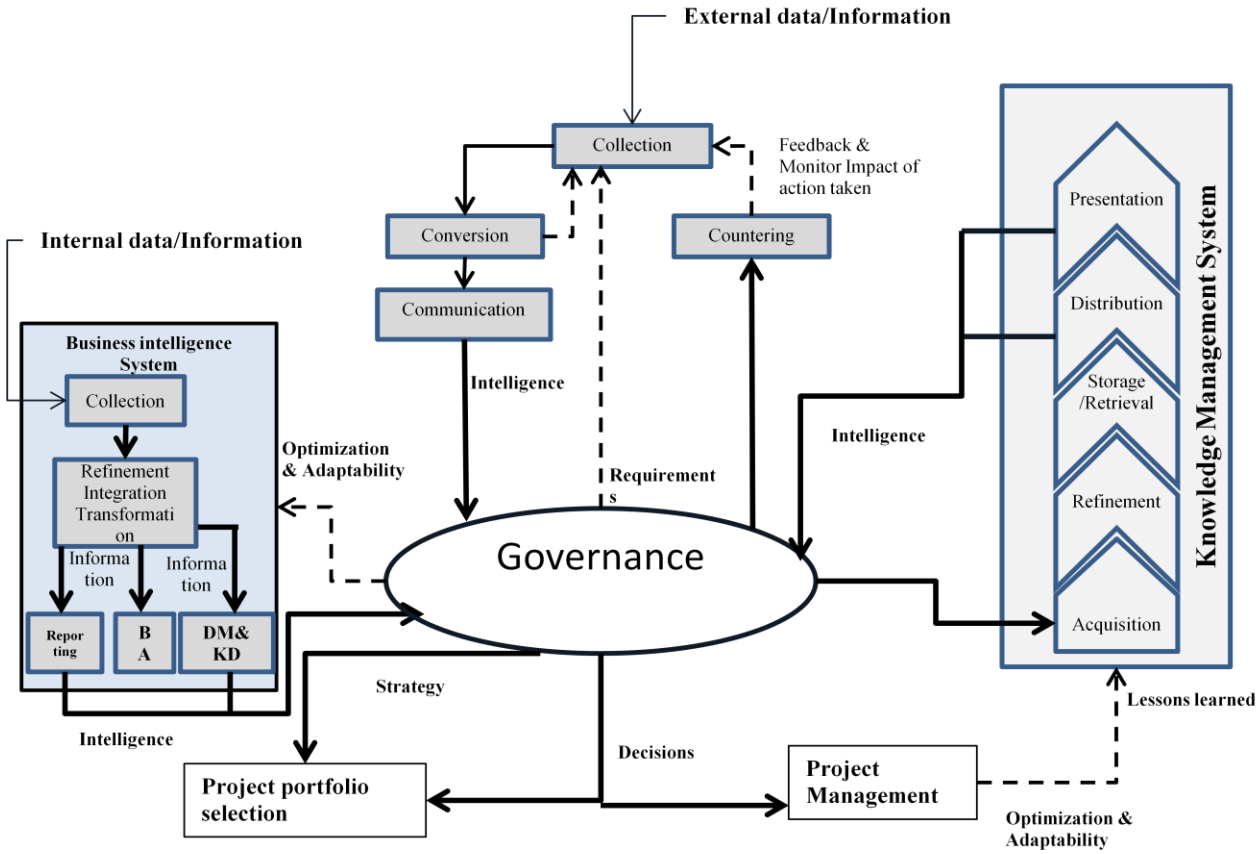


Fig. 3: Intelligent Project Based Organization Model (IPBO) (source: Authors)

•Operational dimension: develop its competency to respond to disruptive events

•Strategic dimension: provides analytic approach to deal with and to respond to trends and significant changes

4.2 .4 Benefits for Stakeholders

Stakeholders' management is a key factor to succeed a project and good project governance should consider the importance of their adhesion and support. In this perspective EI is proven to [24] :

- Keep sponsors motivated and thus more support
- Help satisfying customers thus bring more business
- Motivate and valuate employees' contribution so it ensures higher morale and productivity for team and employees
- Keep suppliers aware of the real need of the company so they are able to propose adequate services with lower prices

5. Conclusion

Businesses need more than ever business models that provide them with agility and flexibility required for survival and growth in such tougher economic framework. This paper confirms that a combination of Project Based Organization Model and Enterprise Intelligence System will help companies sustain differentiating competitive advantages by a strategic use of projects and project management through the transformation of valuable internal and external data into actionable information and a methodologically capitalized knowledge. This Paper also proposes a practical and useful Intelligent Project Based Organization Model (IPBOM) that can reinforces benefits of PBOs and EI systems and help overcoming their Pitfalls.

References

[1] Jones, R.B., It's Eleven O'Clock... Do you know what your Competition is doing?. Adhesives & Sealants Industry. Issue 4. pg 14 – 19,2009

[2] Ales popovic & jurij jaklic, Benefits of business intelligence system implementation: an empirical analysis of the impact

of business intelligence system maturity On information quality; European, Mediterranean & Middle Eastern Conference on Information Systems 2010

[3] Lichtenthaler, Outward knowledge transfer: the impact of project-based organization on performance, Industrial and Corporate Change, ISSN : 0960-6491,Volume : 19,Numéro :6,2010

[4] Kodama Mitsuru , Project Based Organization In The Knowledge Based Society, 284pp, imperial college press, Jun 2007

[5] Koskinen Kaj U. & Pihlanto Pekka, Knowledge Management in Project-Based Companies An Organic Perspective, Palgrave and Macmillan 2008

[6] Liebowitz, J, Strategic Intelligence: Business Intelligence, Competitive Intelligence, and Knowledge Management. Boca Raton, FL.: Auerbach Publications. Taylor & Francis Group2006.

[7] Deborah S. Kezsbom, Katherine A. Edward,The new dynamic project management W inning Through the Competitive Advantage, Wiley, 30 janv. 2001 - 574 pages, JOHN WlLEY & SONS, INC

[8] Hawamdeh Suliman, T. Kanti Srikantaiah, and Michael E. D. Koenig, Convergence of Project Management and Knowledge Management , THE SCARECROW PRESS, INC.Lanham ,Toronto ,Plymouth, UK 2010 (2)Curtis R. Cook , Just Enough Project Management: The Indispensable Four-Step Process for Managing Any Project Better, Faster, Cheaper , McGraw-Hill ,2005

[9] Pirttimäki, V., Lönnqvist, A. and Karjaluoto, A. Measurement of business intelligence in a Finnish telecommunications company. Journal of Knowledge Management, 4(1), 83–90, 2006

[10] Carlo Vercellis, Business Intelligence : Data Mining and Optimization for Decision Making John Wiley & Sons 2009

[11] Inmon, W. H., Building the data warehouse (3ed.). New York, NY: John Wiley & Sons. 2002.

[12] Clive Ellen &John Jordan, Business need, data and business intelligence, Journal of Digital Asset Management, (2009) 5, 10 – 20. doi: 10.1057/dam.2008.53, 2008

[13] Curtis, B., Hefley, W.E., Miller, S.A.: The People Capability Maturity Model – Guidelines for Improving the Workforce, 2 ed., SEI Series in Software Engineering, Boston, MA: Addison-Wesley .2010

[14] Britt, P., the New Competitive Intelligence: Raising the Confidence Quotient. KM World. pg 10, 11, 24,2006

[15] Calof, J.L & Wright, S, Competitive Intelligence: A Practitioner, Academic and Inter-disciplinary Perspective. European Journal of Marketing. Vol. 42. Issue 7/8. pg 717 – 730,2008

[16] Viviers, W. and Saayman, A. and Muller, M.. Enhancing a Competitive Intelligence Culture in South Africa, International Journal of Social Economics. Vol. 32. Issue 7. pg 576, 2005

[17] Weiss, Arthur, A brief guide to competitive intelligence , Business information Review , (ISSN 0266-3821), vol 19, 2002

[18] Becker, J., Niehaves, B., Pöppelbuß, J. and Simons, A. Maturity Models in IS Research. In Proceedings of the European Conference on Information Systems (ECIS), Pretoria. 2010.

[19] Buhler, P.M.,Managing in the New Millennium. Super Vision. Vol. 64. Issue 8. pg 20, 2003

[20] Murphy, C., Competitive intelligence: gathering, analyzing, and putting it to work, Aldershot, Hants, England.: Gower, 2005

[21] Pfeffer, J. & Sutton, R, IThe Knowing-Doing Gap: How Smart Companies Turn Knowledge into Action, 2000.

[22] Jatinder N. D. Gupta,Sushil K. Sharma,Jeffrey Hsu, An Overview of Knowledge Management; Fundamental Concepts and Theories in Knowledge Management ,Knowledge Management: Concepts, Methodologies, Tools, and Applications , Information Science reference; Hershey New York ,IGI Global 2008

[23] Curtis R. Cook , Just Enough Project Management: The Indispensable Four-Step Process for Managing Any Project Better, Faster, Cheaper , McGraw-Hill ,2005

[24] Pells, David L., "Aftershocks: How Significant Global Events Can Affect the Project Management Profession;" Proceedings of the PMI South Africa International Project Management Conference; Gauteng, Johannesburg, South Africa; November 1999. Republished in 2008 at http://www.pmforum.org/library/second-edition/2008/PDFs/Pells-10-08.pdf

Neighborhood covering rough set model of fuzzy decision system

Bai-ting Zhao[1], Xiao-fen Jia[2]

[1] **Electrical and Information Engineering College, Anhui University of Science and Technology, Huainan, China, 232001**

[2] **Electrical and Information Engineering College, Anhui University of Science and Technology, Huainan, China, 232001**

Abstract

The neighborhood covering rough set model is established for the fuzziness of decision system. Firstly, the knowledge representation of fuzzy decision system is analyzed. Then fuzzy neighborhood relation is proposed to measure the fuzziness of the decision system. Finally prove that the classical indiscernibility relations and classical neighborhood relationship are special case of fuzzy neighborhood relations. Fuzzy neighborhood covering rough set model solved the problem of processing fuzzy attribute, and deal with the hybrid decision system which including both of fuzzy attributes and numerical attributes. The proved model expanded the applications of rough set.

Keywords: Fuzzy, Decision System, Neighborhood, Rough Set.

1. Introduction

The real world is diverse, complex and changeing, and the expressions of people to information are often imprecise, uncertain and fuzzy. The fuzziness is the basic characteristic of information uncertainty. The representation and processing of the fuzzy knowledge in fuzzy hybrid decision system has become one of the most important key issues in the research of artificial intelligence. Pawlak rough sets are based on the equivalence relations, to research rough approximation problems of distinct sets. The classical rough sets cannot deal with the fuzzy decision system which contains fuzzy attributes.

Fuzzy rough sets and rough fuzzy sets are important promotion of Pawlak rough sets model. Both rough set and fuzzy set can be used to deal with uncertainty and imprecision problems, therefore the organic combination of them is a good tool to process fuzzy information system. Dubois proposed a fuzzy rough sets model based on the promotion of rough fuzzy model to fuzzy approximation space. Mi improved axiomatic definition of the ambiguity, Feng proposed a new definition of the ambiguity. The different equivalence relations may have the same hierarchical structure are researched by Zhang ling. Zhang qinghua proposed information entropy sequence based on the general fuzzy relations, and the relations between fuzzy similarity relation, fuzzy equivalence relation, the hierarchical structure and the entropy sequence of the hierarchical structure.

2. KNOWLEDGE REPRESENTATION OF FUZZY DECISION SYSTEM

Knowledge is represented by knowledge system, the basic ingredient is the set of the studied object. The decision system can represent knowledge and describe the object by the attributes and attribute values.

Definition 1. Definite an knowledge system as $KRS = (U, A, V, f)$. $U = \{X_1, X_2, \cdots, X_n\}$ is a nonempty finite set, which be called the universe. $A = \{a_1, a_2, \cdots, a_N\}$ is a nonempty finite set of attributes. $V = \bigcup_{a \in A} V_a$, where V_a is a domain value of the attribute a, and $f : U \times A \to V$ is called the information function such that $f(x_i, a) \in V_a$ for each $a \in A, x_i \in U$.

Definition 2. Definite an decision system as $DT = (U, C \bigcup D, V, f)$. If $A = C \bigcup D$ and $C \bigcap D = \varnothing$, C is a finite set of condition attributes and D is a finite set of decision attributes, the information system is called decision system.

Definition 3. Definite fuzzy decision system as $FDS = (U, A, V, f, D)$. $U = \{x_1, x_2, \cdots, x_n\}$ is the set of objects, $A = \{a_1, a_2, \cdots, a_m\}$ is the set of attributes, $f = \{f_l : U \to V(l \le m)\}$ is the mapping of objects set to attributes set, and $D = \{D^{\sim}_{\ j} : U \to [0,1](j \le r)\}$ is fuzzy decision set.

The the fuzzy decision system is the fuzzy knowledge representation of information systems, object attributes and attribute values in the system can be a good description of the object.

3. NEIGHBORHOOD ROUGH SET MODEL OF DECISION SYSTEM

Lin proposed neighborhood model in 1988. Hu build neighborhood rough set model and proposed hybrid reduce algorithm.

Definition 4. Let U is a nonempty finite set, if $\exists\Delta$ which is a real function corresponding to $\forall x \in U$ and satisfy the follows:

(1) $\Delta(x_i, x_j) \geq 0$,and $\Delta(x_i, x_j) = 0$ only when $x_i = x_j$;

(2) $\Delta(x_i, x_j) = \Delta(x_j, x_i)$;

(3) $\Delta(x_i, x_k) \leq \Delta(x_i, x_j) + \Delta(x_j, x_k)$.

$\triangle$is distance function of U, and $<U, \Delta>$ is distance space which also called metric space.

In Euclidean space of dimension N, to random two points of $x_i = (x_{1i}, x_{2i}, \cdots, x_{Ni})$ and $x_j = (x_{1j}, x_{2j}, \cdots, x_{Nj})$, the distance of them is:

$$\Delta(x_i, x_j) = \left(\sum_{i=1}^{N} (x_{li} - x_{lj})^2 \right)^{1/2}$$

Definition 5. Let $<U, \Delta>$ is a nonempty metric space, $x \in U$, $\delta \geqslant 0$, the points set $\delta(x) = \{y \mid \Delta(x, y) \leqslant \delta, y \in U\}$ is a closed sphere and is called δ neighborhood of x.

When the attributes both include numerical and characteristics values, let $B_1 \subseteq A$ is numerical attributes and $B_2 \subseteq A$ is characteristics attributes, then the neighborhood of sample x is:

(1) $\delta_{B_1}(x) = \{x_i \mid \Delta_{B_1}(x, x_i) \leq \delta, x_i \in U\}$

(2) $\delta_{B_2}(x) = \{x_i \mid \Delta_{B_2}(x, x_i) = 0, x_i \in U\}$

(3) $\delta_{B_1 \cup B_2}(x) = \{x_i \mid \Delta_{B_1}(x, x_i) \leq \delta \wedge \Delta_{B_2}(x, x_i) = 0, x_i \in U\}$

Let $NAS = (U, N)$ is a neighborhood approximate space and $X \subseteq U$, the lower and upper approximations of X in $NAS = (U, N)$ can be defined as follows,

$$\begin{cases} \underline{N}X = \{x_i \mid \delta(x_i) \subseteq X, x_i \in U\} \\ \overline{N}X = \{x_i \mid \delta(x_i) \bigcap X \neq \varnothing, x_i \in U\} \end{cases}$$

4. Neighborhood covering rough set model (NCRSM) for fuzzy decision system

Let U is a nonempty finite set, $C = \{C_1, C_2, \cdots, C_k\}$ is fuzzy covering of U , $\mathbf{I}$ is arbitrary contain, $D^{\mathbf{I}}$ is fuzzy inclusion degree of $\mathbf{I}$. $A=(\left(F(\mathrm{U}), C, D^{\mathbf{I}}\right)$ is fuzzy contains approximation space of $\mathbf{I}$. To $\forall X \in F(U)$ and $0 \leq l < u \leq 1$, the lower approximation $\underline{A}_u(X)$ and upper approximation $\overline{A}_l(X)$ of X about A can be defined as follows:

$$\underline{A}_u(X) = \bigcup \left\{ C_i \in C \mid D^{\mathbf{I}}(C_i, X) \geq u \right\}$$

$$\overline{A}_l(X) = \bigcup \left\{ C_i \in C \mid D^{\mathbf{I}}(C_i, X) > l \right\}$$

Definition 6. When the attributes both include clear and fuzzy values, numerical and characteristics values, let $B_1 \subseteq A$ is numerical attributes and $B_2 \subseteq A$ is characteristics attributes, let $B_3 \subseteq A$ is fuzzy attributes, then the neighborhood of sample x is:

1) $\delta_{B_1}(x) = \{x_i \mid \Delta_{B_1}(x, x_i) \leq \delta, x_i \in U\}$

2) $\delta_{B_2}(x) = \{x_i \mid \Delta_{B_2}(x, x_i) = 0, x_i \in U\}$

3) $\delta_{B_3}(x) = \{x_i \mid \Delta_{B_3}(f_l(x), f_l(x_i)) = 0, x_i \in U\}$

4) $\delta_{B_1 \cup B_2 \cup B_3}(x) = \{x_i \mid \Delta_{B_1}(x, x_i) \leq \delta \wedge \Delta_{B_2}(x, x_i) = 0 \wedge \Delta_{B_3}(f_l(x), f_l(x_i)) = 0, x_i \in U\}$

Further we can get the properties as follows:

1) $\delta(x_i) \neq \varnothing$,for $x_i \in \delta(x_i)$

2) $\bigcup_{i=1}^{n}\delta(x_i)=U$

Neighborhood Particle tribe $\{\delta(x_i)\mid i=1,2,\cdots,n\}$ can constitute a covering of U. However, due to $x_i \neq x_j$ can not ensure $x_j \notin \delta(x_i)$, so $\{\delta(x_i)\mid i=1,2,\cdots,n\}$ generally does not constitute a partition of U.

The neighborhood equivalence relation extended to fuzzy and clear attributes, characteristics and numeric attributes coexist fuzzy decision system, we can get fuzzy neighborhood relations:

$$\begin{aligned}R(X)=\{(x,y)\in U^2:\forall a\in X\\ \bigcap a(x)\neq ?\bigcap f_l(x)=f_l(y),\\ a(x)\in\delta(y,a)\bigcup a(y)\in\delta(x,a)\}\end{aligned}$$

From the basic nature of the neighborhood, the neighborhood relations satisfy reflexivity, symmetry, transitivity, so fuzzy neighborhood relations can be further simplified as:

$$\begin{aligned}R(X)=\{(x,y)\in U^2:\forall a\in X\\ \bigcap a(x)\neq ?\bigcap f_l(x)=f_l(y),\\ a(x)\in\delta(y,a)\}\end{aligned}$$

The simplified fuzzy neighborhood relations meet reflexivity but not necessarily to meet symmetry and transitivity. The relaxation of the requirements to symmetry and transitivity can avoids overly conservative, and broaden the range of applications.

Theorem 1: The classic neighborhood relation is a special case of fuzzy neighborhood relation.

When the decision system is completely clear, fuzzy neighborhood relation degenerate to the classic neighborhood relation as follows:

$$\begin{aligned}R(X)=\{(x,y)\in U^2:\forall a\in X\\ \bigcap a(x)\neq ?,a(x)\in\delta(y,a)\\ \bigcup a(x)=*\bigcup a(y)=*\}\end{aligned}$$

Theorem 2: The classic Pawlak indiscernibility relation is a special case of fuzzy neighborhood relation.

Let $FDS=<U,A,D>$ is FHDS, U is divided into N equivalence class by D as follows: $X_1,X_2,\cdots,X_N$, $B\subseteq A$ generate the neighborhood relation of U, so the upper approximation and lower approximation of decision D to B is :

$$\begin{cases}\underline{\underline{N_B}}D=\{\underline{\underline{N_B}}X_1,\underline{\underline{N_B}}X_2,\cdots,\underline{\underline{N_B}}X_N\}\\ N_BD=\{N_BX_1,N_BX_2,\cdots,N_BX_N\}\end{cases}$$

5. Simulation

The "Wpbc" and "Segmentation" of UCI machine learning database are used to verify the reasonable and validation of NCRSM model. Reduction and forecasting simulation compared with the Dubois fuzzy rough set model and fuzzy Radzikowska rough set model, and results as follows.

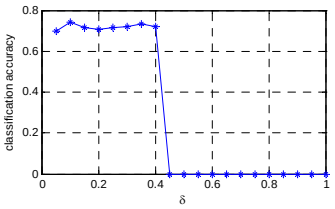
The data sample is divided into a training set and a test set, and then attributes are reduced by the algorithm based on attributes significant.

The SVM classifier is used as the evaluation function, in which used spline function. Raw data and reduction data are used to train the SVM separately, the prediction accuracy is used to evaluate the quality of the reduction. Shown form the results the fuzz neighborhood rough set model is better than Dubois and Radzikowska model.

Table 1 Prediction accuracy

Data	*Raw*	*Dubois*	*Radzikowska*	*NCRSM*
Wpbc	84.6%	$81.37\pm 6.02\%$	$78.26\pm 3.89\%$	$80.24\pm 3.93\%$
Segmentation	96.37%	$94.26\pm 5.74\%$	$95.41\pm 4.59\%$	$94.74\pm 3.63\%$

In order to study the affection of the selection and the size of neighborhood operator to classification accuracy and attributes reduction number, the "Image" are used in reduction and prediction simulation. The results are shown in Fig.1.



(a)

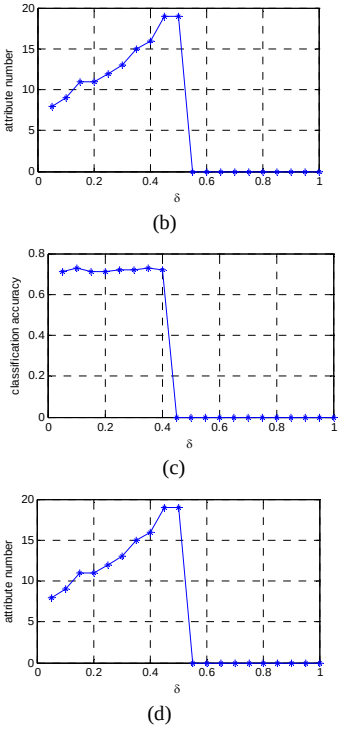


Fig.1 Variation of accuracies and attribute numbers in reduction with δ of image

Figure.1(a) shows the classification accuracy changes with the different sizes δ which uses the 2-norm distance function. Figure.1(b) shows the reduction number of attributes changes with the different sizes δ which uses the 2-norm distance function. Figure.1(c) shows the classification accuracy changes with the different sizes δ which uses the infinite-norm distance function. Figure.1(d) shows the reduction number of attributes changes with the different sizes δ which uses the infinite-norm distance function.

Shown from Figure1.(a) and (c), when the value of δ is small, the particle size of classification is small. So fewer features can distinguish the decision attributes. When the value of δ is larger, the required characteristics to distinguish the decision attributes is also more. When the δ exceeds a certain degree, any characteristics will not distinguish any samples.

From Figure.1(b) and (c) we can see, the attributes number of reduction will increase with the increase of δ, but not monotonic changes. When the neighborhood operator exceeds a certain value, any feature is not sufficient to distinguish any samples and cannot obtain any reduction attributes. On the other hand, the increase of the attributes number of reduction may not improve classification performance. When the neighborhood operator is smaller, the increase of attributes can improve the accuracy of classification. Nevertheless, if attributes continue to increase, the accuracy of classification does not continue to be improved and even be declined slightly. Therefore the reasonable choice of neighborhood operator can obtain better classification effect.

Comparing Figure 1(a) (c) with Figure 1 (b) (d) we can see, using different distance functions has less affect to classification accuracy. The main factors affecting the classification accuracy is the size of the neighborhood. The value of δ depends on the specific classification problems. The general value is: [0.1 0.2].

6. Conclusion

Pawlak rough sets are based on the strict equivalence relation, which can only deal with characteristic attributes. But in practice, many information systems are hybrid systems which including characteristic attributes, numerical attributes and fuzzy attributes. In this paper a fuzzy neighborhood rough sets model is established for fuzzy hybrid decision system. The model uses the fuzzy neighborhood relation to measure the indiscernibility relation of the classical rough sets, and uses granulation to approximate the universe space, which can deal with fuzzy attributes directly. The neighborhood relation is a special case of the fuzzy neighborhood relation, and the fuzzy neighborhood relation can deal with the hybrid decision system include both numerical and fuzzy attributes.

The UCI machine learning databases are used in simulation experiments, and the simulation results proved the validity of the fuzzy neighborhood rough sets model which better than Dubois fuzzy rough sets model and Radzikowska rough sets model. Finally, we analyze the influence of the selection and size of neighborhood operator to the classification accuracy and the attributes number in reduction of fuzzy neighborhood rough set model. Fuzzy neighborhood rough set model is the promotion of Pawlak rough sets which provides effective solutions to the classification of hybrid decision system in the practical application.

Acknowledgement

The research is supported by the Youth Foundation of An hui University of science & technology of China under Gr ant No.12257, No.2012QNZ06, the Doctor Foundation of Anhui University of science & technology of China under Grant No.11223, and the Guidance Science and Technolog y Plan Projects of Huainan under Grant No.2011B31

References

[1] WANG Guo-Yin, ZHANG Qing-Hua, MA Xi-Ao, YANG Qing-Shan. "Granular Computing Models for Knowledge Uncertainty", Journal of Software, Vol.22, No.4, 2011, pp. 676-694

[2] Dubois D, Prade H. "Rough fuzzy sets and fuzzy rough sets", Journal of General Systems, Vol.17, No.3,4, 1990, pp.191-208

[3] Dubois D, Prade H. "Putting rough sets and fuzzy sets together", In: Slowinski R, ed. Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory. Boston: Kluwer Academic Publishers, 1992, 203-222.

[4] HU Jun, WANG Guo-Yin, ZHANG Qing-Hua. "Covering Based Generalized Rough Fuzzy Set Model", Journal of Software, Vol.21, No.5, 2010, pp. 968-977

[5] LUO Gong-zhi, YANG Xi-bei, YANG Xiao-jiang. "Limited dominance-based rough fuzzy set and knowledge reductions", Systems Engineering and Electronics, Vol.32, No.8, 2010, pp.1657-1661

[6] Huang bing, Wei da-kuan. "Distance-based rough set model in intuitionistic fuzzy information systems and its application", Systems engineering theory & practice, No.7, 2011, pp. 1356-1362

[7] Li Bai, Min Liu. "Fuzzy sets and similarity relations for semantic web service matching", Computers & Mathematics with Applications, Vol.61, No.8, 2011, pp. 2281-2286

[8] Degang Chen, Yongping Yang, Hui Wang. "Granular computing based on fuzzy similarity relations", Soft Computing, Vol.15, No.6, 2011, pp.1161-1172

[9] Hu Q. H, Liu J. F, Yu D. R. "Mixed Feature Selection Based on Granulation and Approximation", Knowledge Based systems, Vol.21, No.4,2008, pp.294~304

[10] Tripathy, B.K. Panda, G.K. " Approximate equalities on rough intuitionistic fuzzy sets and an analysis of approximate equalities ", International Journal of Computer Science Issues, Vol.9, No.2 2-3, 2012, pp.371-380.

[11] Li, Ju. Liu, Xiao-Ping. Du, Hui. " Research of image recognition based on rough set ", International Journal of Digital Content Technology and its Applications, Vol.6, No.9, 2012, pp.141-146,

Dr. Zhao received the Master. degree in control theory and control engineering from the Qingdao University of Science & Technology, in 2005. He received the Ph.D. degree in control science and engineering, from the Harbin Institute of Technology. Currently, he is a lectorate at Anhui University of Science & Technology, Electrical and Information Engineering College. His research interests include intelligent control and Rough sets.

Mrs. Jia received the Master. degree in control science and engineering, from the Harbin Institute of Technology. Currently, she is a lectorate at Anhui University of Science & Technology, Electrical and Information Engineering College. Her research interests include Rough sets and Image processing.

An Autonomic Intrusion Detection Model with Multi-Attribute Auction Mechanism

Qingtao Wu, Xulong Zhang, Ruijuan Zheng and Mingchuan Zhang

Electronic & Information Engineering College, Henan University of Science and Technology
Luoyang, Henan Province 471023, China

Abstract

We present an innovative intrusion detection model based on autonomic computing to extend the passive detection mechanism in a traditional intrusion detection system (IDS). Centered on an autonomic manager, this model introduces a multi-attribute auction mechanism in the agent coordination layer to perceive environmental changes, manage and allocate resources, and achieve an active response to intrusions or attacks. Experimental results show that the model can improve the adaptability and detection accuracy of the IDS effectively, through its rational parameter configuration capability.

Keywords: *Intrusion detection, Autonomic computing, Auction mechanism, Agent coordination.*

1. Introduction

Intrusion detection is a widely used and important network security technology that can improve safety greatly and reduce security threats to a system by creating a dynamic safety cycle. With the development of large-scale networks and the establishment of complex requirements involving network intrusion, there are now many demands on intrusion detection technology. Existing intrusion detection systems (IDSs) can offer only passive detection mechanism, wherein, only when an intrusion or attack has occurred can the IDS respond. An intrusion or attack may therefore still cause local or widespread compromises to system safety. In essence, IDS is a post-mortem mechanism that can identify an event only after it has already occurred. It can report the event, but has no adaptive ability. Artificial intelligence, mobile agents, data fusion, information correlation [1–3], and other technologies and methods have been introduced by researchers into continuous intrusion detection, aiming to identify an attack in a timely and effective manner.

Autonomic computing can overcome the heterogeneity and complexity of computing system, has been regarded as a novel and effective approach to implementing autonomous systems to address system security issues. The “autonomic” is inspired by the autonomic nervous system of the human body, which can manage several key functions via involuntary control. Autonomic system is the adjustment of the software and hardware resources of a system to manage its operation, driven by changes in internal and external demands. It has four main characteristics, namely self-configuration, self-healing, self-optimization, and self-protection. The core of an autonomic system enables the computer system to realize high reliability, availability, and service performance.

However, studies of security technology based on autonomic computing have focused only on safety technology. In this case, action is delayed until after the system is attacked, with system safety being compromised and the intrusion not being detected in time. For the purposes of system safety, an autonomous system combined with intrusion detection technology that enables dynamic adaptation to environmental changes, thereby achieving a timely detection of intrusion, should be investigated. The present study combines intrusion detection with autonomic computing to improve the poor adaptive capability in the passive detection mechanisms of traditional intrusion detection technology. To achieve this, we propose an autonomic characteristic intrusion detection model (ACIDM) with auction mechanism.

2. Autonomic intrusion detection model

The proposed autonomic intrusion detection model is shown in Fig. 1. In this model, the managed resource (MR) [4] covers all types of physical and virtual resources, such as databases, servers, routers, application modules, Web servers, virtual machines, host logs, network packets, and firewall alarm messages. These resources must be manageable, observable, and adjustable. The state of the resources refers to all data (events) reflecting the existing resource state, including log and real-time events, such as the operative and performance status (throughput and availability of resources) of the resources, and anomalous events. The MR is uniformly distributed and managed by an agent coordination layer.

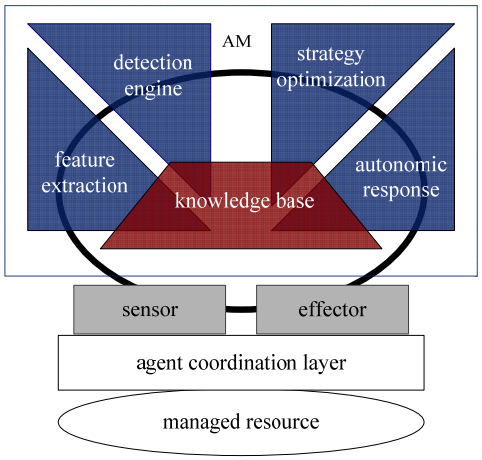


Fig. 1 Autonomic intrusion detection model

The autonomic manager (AM) in Fig. 1 consists of feature extraction, detection engine, strategy optimization, autonomic response, and knowledge base. The line connecting these four parts indicates the sharing information and message. Strategy optimization requires the detection engine to collect additional information before operating. Coordination among these four elements can be implemented via asynchronous communication.

2.1 Agent Coordination Layer

The agent coordination layer uses different intelligent agents for the various MRs to provide data support for the AM. These agents are entities that can operate independently. Agents capture MR information and remove redundancy by preprocessing before its final submission to the AM. Another major function of the agent coordination layer is to receive feedback regarding AM information and adjust the system environment autonomously to adapt to changes. In addition, the agent coordination layer realizes the dynamic configuration of resources, the synthesis of services, and the calibration of system parameters. For example, when the system detects intrusion, the agent controlling the firewall will update the blocking strategy based on the intrusion alarm and will control information from the AM to block subsequent attacks over a certain time according to IP address, interface, and other information. This process can be described as IDS with dynamic self-adaption.

The agents in the agent coordination layer work synergistically to form a multi-agents system. An auction mechanism is introduced by the multiagent system to resolve task allocation, resource configuration, and system performance optimization [5–6]. A variety of auction methods serve the different environments. The multiattribute auction method defined below was used in the present paper.

Definition 1. Multiattribute auction model
In this model, $M = \langle A, B, S, V, C, \mathrm{Re}s \rangle$, where A refers to attribute space and $A = A_1 \times \cdots \times A_m$. Every auctioned event includes the m attributes (i.e., $a_1, \cdots, a_m$). The value range is $A_1, \cdots, A_m$, and we specify $a = (a_1, \cdots, a_m)$ as an attribute vector of the event, so $a \in A$.

B refers to a unique buyer at auction who needs to purchase an event. S refers to a seller set, which includes n buyers, i.e., $S = \{s_1, \cdots, s_n\}$. The buyers can provide events with different attributes.

$V: A \to R$ refers to the attribute assessment function of buyer B (R is the set of real numbers). The assessment value that buyer B made for an event with an attribute of a is $V(a) \in R$.

In this model, $C = \{C_1, \cdots, C_n\}$, where C_i refers to the cost function for seller i. The amount that seller i receives for an article with an attribute of a is $C_i(a) \in R$.

$\mathrm{Re}s$ refers to a transaction program. $\mathrm{Re}s = (P, a)$, where the knockdown price is $P \in R$ and the transaction attribute vector is $a \in A$. At this time, the benefit of the buyer B is $U = V(a) - P$. The benefit of the seller S_i is $U_i = P - C_i(a)$.

The process of the auction is divided into four steps:

(1) The buyer publishes an evaluation function $V'(a)$ (V' may differ from V).

(2) Each seller i makes a sealed bid $B_i \geq 0$.

(3) The transaction seller is confirmed. First, the buyer decides on an optional seller set $W = \{\omega \mid (\omega \in S) \wedge (B_\omega = \max_{i \in S}(B_i)) \wedge (B_\omega > 0)\}$

where B_ω refers to a bid for ω. If $W = \varnothing$, no transaction seller exists, and the auction ends. If $W \neq \varnothing$, $\omega \in W$ is generated randomly as a transaction seller. We define

$$B^* = \max_{i \in S}{}_2(B_i)$$

where $\max_{i \in S}{}_2(B_i) \overset{def}{=\!=} \min_{i \in S}(\max_{j \in S - \{i\}}(B_j))$. Here, $\max_2$ is the maximal value of the residual element after removing the maximal element (i.e., $\max_2(1,2,3) = 2$, $\max_2(1,2,3,3) = 3$). The quantity B^* is then the highest bid of other sellers, except for the seller ω.

(4) The transaction process is proposed by the transaction seller (P_t, a_t). The legal proposal should satisfy $V^{'}(a_t) - P_t = B^*$, based upon which the transaction seller reaches a deal with the buyer. The auction then ends.

2.2 Sensor and Effector

The hardware and software of the distributed system may be sourced from different service suppliers. A standard interface is therefore needed to prevent the heterogeneity with respect to the resources. The fundamental method for resolving this issue is to establish a sensor and an effector through standardization and semantic technology. Based on the theory related to the sensor, a formalized definition can be obtained, as follows.

Definition 2. Sensor
Let $T = \{t_1, ..., t_n\}$ be a group of feature sets that can reflect the existing MR state and let $V = \{v_1, ..., v_m\}$ be a group of event sets that reflects an MR state change. Let $O = (C, R)$ be the domain ontology, where C is the domain concept set and R is related to C. Let $\xi = \{get, report\}$ be a group of operation sets. *Sensor* can then be defined as a 4-tuple: $Sensor = (T, V, O, \xi)$, where $\forall t_i, v_j (1 \le i \le n, 1 \le j \le m)$, $t_i \in C, v_j \in C$.

The sensor supports automatic interpretation and reasoning and realizes self-awareness. The MR feature and event sets comply with the specific domain ontology expressed in an ontology language with clear semantics. The operation *get* was used to capture MR state features, with $get(t_i)$ indicating that the AM obtains the characteristic t_i from MR via the sensor. The operation *report* was used for reporting MR state changes, with $report(v_j)$ indicating that the MR reports v_j to the AM.

Definition 3. Effector
Let $A = \{a_1, ..., a_n\}$ be a group of executable action sets that can operate the MR state and let $Q = \{q_1, ..., q_m\}$ be a group of action sets released by the AM for MR application. Let $O = (C, R)$ be the domain ontology, where C is the domain concept set and R is related to C. In this study, $\psi = \{set, request\}$ is a group of operation sets. *Effector* can then be defined as a 4-tuple: $Effector = (A, Q, O, \psi)$, where $\forall a_i, q_j (1 \le i \le n, 1 \le j \le m)$, $a_i \in C, q_j \in C$.

In Definition 3, the operation *set* was used for execution action, with $set(a_i)$ indicating that the AM executes action a_i through the effector. The operation *request* triggers the MR to send a request (e.g., for help or consultation) to the AM, with $request(q_j)$ indicating that the MR executes the request action q_j to the AM.

2.3 Data Normalization

The collected data should be preprocessed to resolve heterogeneity. Normalization theory [7] is adopted to unify the type and format of the data. In an IDS, the Euclidean distance between characteristic vectors must be calculated. This distance should normalize the process, because leading one numeric data item to affect another is easy for the sake of the difference in value ranges. The steps of the processing method are as follows. First, the mean and standard deviation for training each characteristic attribute of sample are calculated as follows.

$$mean[j] = \frac{1}{n}\sum_{i=1}^{n} ins\tan ce_i[j] \quad (1)$$

$$s\tan dard[j] = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(ins\tan ce_i[j] - mean[j])^2} \quad (2)$$

where $ins\tan ce_i[j]$ is attribute j in the training sample i, and n is the number of samples. This sample from the training collection is transferred as follows.

$$newins\tan ce[j] = \frac{ins\tan ce[j] - mean[j]}{s\tan dard[j]} \quad (3)$$

Formula (3) can be used to transfer the value of the attribute to multiple standard deviations. Considering that this value deviates from the mean, it can map the attribute value of a sample from its value space to a standard value space.

2.4 The AM

Knowledge base: "Knowledge" refers mainly to state determination (KD), strategy knowledge (KP), problem solving knowledge (KS), and detection rules (KR). That is, the knowledge base is K=KD+KP+KS+KR. In this equation, KD is used mainly to monitor the state parameters of the managed resources and the internal and external environments. KP mainly includes the strategy defined by the IT manager and the strategy obtained through machine learning (i.e., mapping from state to action). KS mainly includes rules, configurations, and optimizations, and how to solve problems when the running state of the system deviates from expectations. KR mainly includes the characteristic base derived from the misuse of intrusion detection, and the behavioral model base derived from abnormal intrusion detection. Other subsystems run with the support of the knowledge base.

Feature extraction: The sensor obtains the data captured by the agent coordination layer. Expansion matrix theory is used to extract the intrusion characteristics [8] through analysis, relationships, and data integration. This method

establishes an integer programming model selected by its optimal characteristic subset through the creation of an expansion matrix of intrusion and normal subsets. In addition, this method can generate an optimal rule for detecting a specific type of attack using a simple genetic algorithm.

Detection engine: The detection engine is a functional component, performing detection for the AM. It can identify the intrusion intention using mixed detection technology.

Strategy optimization: Strategy optimization is realized by adopting machine learning, intelligent planning, and other related technologies that can adapt to environmental change.

Autonomic response: Autonomic response completes the response to intrusion according to the strategy knowledge in the database.

3. Simulation Experiments and Performance Analysis

3.1 Experimental Data Set and Design

We adopted the KDD Cup 1999 data set [7], which has been approved and adopted widely in the intrusion detection research field as a benchmark for detection, to validate the experiment. This data set includes approximately 4,900,000 data records. The records were extracted from original network data obtained by a simulated attack on a military network environment. The data are based on a set of 41 characteristic vectors describing statistical information about network connections that include five kinds of data. Among these data types are four kinds of attack data (namely Dos, Probe, R2L, and U2R, with 24 kinds of attachment types in total) and one type of normal data. The 41 characteristics of this data set are mainly categorized into two data types: numerals and nouns. The numeric data are processed first. The noun attributes in the data set, including protocol and service types, are processed using data normalization based on the occurrence frequency of each value in the value range. Therefore, the value of an attribute ranges from 0 to 1. In the current experiment, 10% of the selected data set was used as experimental data.

The following two experiments were designed to investigate the feasibility and effectiveness of the proposed model:

Experiment 1: A comparison of the performances of AM detection engines with respect to detection accuracy, using mixed and misuse detection technologies.

Experiment 2: A comparison of the performances of ACIDM and two intrusion detection models with respect to detection accuracy and time. The two other models were an artificial neural network (ANN) and a support vector machine (SVM).

3.2 Experimental Results and Analysis

First, the same data set was adopted for the two experiments. The comparison of AM detection engine accuracies using mixed and misuse detection technologies is shown in Table 1.

Table 1: Comparison of misuse and mixed detection technology performances

ttack method	Detection rate (%)	
	Misuse detection	Mixed detection
ormal	93.26	98.35
Dos	83.47	98.64
U2R	76.68	95.28
R2L	74.57	94.75
Probe	86.69	98.43

Table 1 shows that the detection performance using a mixed detection technology is significantly better than that using misuse detection technology.

The detection accuracy and time performance of the ACIDM were compared with those for the ANN and SVM intrusion detection models. The experimental results are shown in Table 2 and Fig. 2.

Table 2: Detection accuracy of ACIDM, ANN, and SVM

Attack method	Detection rate (%)		
	ANN	SVM	ACIDM
Normal	82.21	93.26	98.35
Dos	67.35	83.47	98.64
U2R	64.28	76.68	95.28
R2L	69.57	74.57	94.75
Probe	76.24	86.69	98.43

The ACIDM performed substantially better with respect to detection accuracy than did the ANN and SVM models, as

shown in Table 2. The performance comparison for detection time values is shown in Fig. 2.

Table 2 shows that the ACIDM performs better than the ANN and SVM models in terms of detection accuracy. However, the ACIDM fared poorly with respect to detection time because it adopts a mixed intrusion detection technology in the detection process and adds autonomic-response and response-strategy optimization to improve the self-adaptability of the system, thereby extending the detection time. The detection time for the SVM model was the shortest.

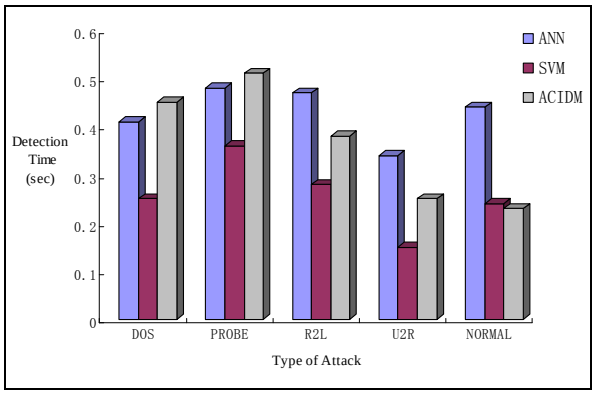


Fig. 2 Detection times for ACIDM, ANN, and SVM

The following conclusions can be derived from these experimental results:

(1) The detection engine should adopt a mixed detection technology. Its detection accuracy was considerably better than that for the misuse detection technology.

(2) The ACIDM performs better than ANN and SVM in terms of detection accuracy. However, its detection time is greater.

4. Conclusions

Focusing on resolving the drawbacks of the passive detection mechanisms in traditional IDSs, an autonomic intrusion detection model with auction mechanism was proposed in this paper. The model is centered on the AM and integrates a multiattribute auction mechanism, which can perceive changes in the system environment, into the agent coordination layer. This can perceive changes in the system environment, and adapt the configuration management accordingly. The experimental results show that the model can enhance the self-adaptive performance of a system and obtain high detection accuracy with appropriate settings. Although the ACIDM demonstrated a high detection performance, the detection time was relatively long, which should be the focus in further research.

Acknowledgments

The authors thank the anonymous reviewers for their valuable comments and suggestions. This work is sponsored partially by the National Natural Science Foundation of China (No. 61003035) and the Plan for Scientific Innovation Talent of Henan Province (No. 124100510006)

References

[1] Y.Y. Zhang, W. Nurbol, J.Q. Cheng-ming, and L. Hu. “Status of Intrusion Tolerance”, Journal of Jilin University (Information Science Edition), Vol. 27, No. 4, 2009, pp. 389–394.

[2] Shakhatreh. Ala' Yaseen Ibrahim, and Bakar. Kamalrulnizam Abu, “A Review of clustering techniques based on machine learning approach in intrusion detection systems”, International Journal of Computer Science Issues, Vol. 8, No. 2, 2011, pp. 373-381.

[3] Qingtao Wu, Ruijuan Zheng, Guanfeng Li, Juwei Zhang. “Intrusion Intention Identification Methods Based on Dynamic Bayesian Networks”, Procedia Engineering, Vol.15, 2011, pp.3433-3438.

[4] LI Bing-yang, WANG Hui-qiang, FENG Guang-sheng, “Model construction and quantitative analysis of autonomic intrusion tolerance system”, Application Research of Computers. Vol. 26, No.5, 2009, pp. 1883-1887.

[5] D. H. Shih, D. C. Yen, C. H. Cheng and M. H. Shih. “A secure multi-item e-Auction mechanism with bid privacy”, Computers & Security, Vol.30, No.4, 2011, pp.273-287.

[6] Noman Mohammed, Hadi Otrok, Lingyu Wang, Mourad Debbabi, Prabir Bhattacharya. “Mechanism Design-Based Secure Leader Election Model for Intrusion Detection in MANET”, IEEE Transactions on Dependable and Secure Computing, Vol. 8, No. 1, 2011, pp. 89-103.

[7] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. “A Detailed Analysis of the KDD CUP 99 Data Set”, Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009), 2009, pp.1-7

Qingtao Wu got his PhD degree in computer science from East China University of Science and Technology, in 2006, on network and information security. He’s Associate Professor in Computer Science at Electronic & Information Engineering College of Henan University of Science and Technology, China. He’s currently managing and leading 2 projects supported by the National Natural Science Foundation of China to address the autonomic mechanism for the retainment and enhancement of system security. His main research interests include computer system security, intelligent information processing, etc.

Xulong Zhang received his Bachelor's degree in Computer Science and Technology in 2011. He is currently a Master Degree

Candidate directed by Dr. Qingtao Wu in Computer Science at Electronic & Information Engineering College of Henan University of Science and Technology, China. His research is focused on network security.

Ruijuan Zheng got her PhD degree in computer science from Harbin Engineering University, in 2008, on autonomic system security. She's Associate Professor in Computer Science at Electronic & Information Engineering College of Henan University of Science and Technology, China. Her main research interests include computer system security, network security, etc.

Mingchuan Zhang got his master degree in computer science from Harbin Engineering University, in 2005, on intelligent information processing. He's lecturer in Computer Science at Electronic & Information Engineering College of Henan University of Science and Technology, China. Her main research interests include computer system security, intelligent information processing, etc.

Explicit travelling wave solutions in a magneto-electro-elastic circular rod

Xinmou Ma[1], Yutian Pan[1] and Liezhen Chang[2]

[1] College of Mechatronic Engineering, North University of China, Taiyuan, 030051, P.R. China

[2] College of Science, North University of China , Taiyuan, 030051, P.R. China

Abstract

The abstract A modified (G'/G)-expansion method is proposed for constructing exact travelling wave solutions of nonlinear wave equations, and this method finds travelling wave solutions in a straightforward manner and in a neat and helpful form than (G'/G)-expansion method. The abundant exact travelling wave solutions of nonlinear longitudinal wave equation(NLWE) with dispersion caused by the transverse Poisson's effect in a long magneto-electro-elastic(MEE) circular rod are successfully obtained by the modified (G'/G)-expansion method. The relation between solitary wave velocity with wave number are derived strictly. Numerical examples are further presented for the wave in a rod made of five different materials. The obtained results show that the solitary wave not only exists in such rods but also shows different features in different materials, which could have potential applications in non-destructive evaluation of structures made of the advanced MEE material.

Keywords: *magneto-electro-elastic(MEE), modified (G'/G)-expansion method, Riccati like equation, travelling wave solution, exact solution.*

1. Introduction

The In the last two decades, nonlinear elastic effects on solitary waves have received considerable attention in solid mechanics[1-6]. With increasing usage of magneto-electro-elastic (MEE) structures in various engineering fields (such as sensors, actuators, etc), wave propagation in MEE media has also attracted many researchers[7-10]. Very recently, Xue et. al.[11] had derived the longitudinal wave equation with dispersion caused by the transverse Poisson's effect in a MEE circular rod, and the solitary waves had been successfully derived by Jacobi elliptic function method, the NLWE reads:

$$\frac{\partial^2 u}{\partial t^2} - c_0^2 \frac{\partial^2 u}{\partial z^2} = \frac{\partial^2}{\partial z^2}\left(\frac{c_0^2}{2} u^2 + N \frac{\partial^2 u}{\partial t^2} \right), \qquad (1)$$

where c_0 is the linear longitudinal wave velocity for a MEE circular rod and N is the dispersion parameter, both depending on the material properties as well as the geometry of the rod.

Here, assume the infinite homogeneous MEE circular rod is mad of composite $BaTiO_3$-$CoFe_2O_4$ with different volume fractions (v_f) of $BaTiO_3$, The rod has a radius R =0.05m. The material properties of the composite are estimated using the simple rule of mixture according to the volume fraction[11]. Denoting for the composite the volume fraction of $BaTiO_3$ as v_f, and that of $CoFe_2O_4$ as $1-v_f$, we then have $M_C = M_E v_f + M_M(1-v_f)$,where M represents an arbitrary material constant, and the subscripts C, E, and M indicate the composite, piezoelectric phase and piezomagnetic phase, respectively. In the following, we consider three different cases of material combinations, by taking the volume fraction of $BaTiO_3$ as 0% (PM), 50% (MEE) and 100% (PE), respectively. Obviously, when v_f =0, the composite is piezomagnetic (PM), whilst v_f =100% corresponds to a piezoelectric (PE) material[14]. Another two purely elastic materials are also considered. One is the transversely isotropic elastic material (TI) taking from 50% (MEE) only the elastic coefficients. The other one is the effective elastic isotropy (EI) obtained from the TI by making it isotropic (i.e., letting $c_{11} = c_{33}$ and $c_{12} = c_{13}$). Xue et al.[11] have calculated the linear wave velocity c_0, dispersion parameter N, as listed in table 1.

On the other hand, searching for explicit solutions of nonlinear wave equations by using various methods has being a main goal for many authors, and many powerful methods to construct explicit solutions of nonlinear wave equations have been established and developed, such as the tanh-function expansion method, the extended tanh-function method, the F-expansion method, the sub-ODE method, the Jacobi elliptic function expansion method, the homogeneous balance method, the Exp-function method,

the (G'/G)-expansion method, the sine-cosine method [12,15-24], and so on. The above methods derived many exact solutions from most nonlinear wave equations.

Table 1: Linear wave velocity and dispersion parameter for different material

v_f	$c_0(\mathrm{ms}^{-1})$	$N(\times 10^{-4}\mathrm{m}^2)$
0%(PM)	5.2131	1.7350
50%(MEE)	5.1446	1.4890
100%(PE)	5.0498	1.0560
TI	4.8003	1.5700
EI	4.8398	1.6200

Based on the main idea of the (G'/G)-expansion method and the extended tanh-function method, we introduced a new method called modified (G'/G)-expansion method. This new method contains more parameters than the (G'/G)-expansion method, and all the solutions obtained by the (G'/G)-expansion method can be obtained by the modified (G'/G)-expansion method. Moreover, the modified (G'/G)-expansion method can obtain some new exact solutions in a neat and helpful form, and some of them can not be obtained by (G'/G)-expansion method.

So far, however, there has been no report on exact travelling wave solution of nonlinear wave equations of a MEE circular rod, which motivates this study, we will apply the modified (G'/G)-expansion method to construct the exact travelling wave solutions of nonlinear wave equations of a MEE circular rod. Therefore, this paper is organized as follows, in section 2, we describe the basic idea of the modified (G'/G)-expansion method. In Section 3, we apply the modified (G'/G)-expansion method to solve nonlinear wave equations of a MEE circular rod for exact traveling wave solution. Numerical examples are given in section 4 and conclusions are drawn in section 5.

2. Basic idea of the modified (G'/G)-expansion method

In this section, according to Wang's work [22], we describe basic idea of the (G'/G)-expansion method for finding travelling wave solutions of nonlinear partial differential equations. Suppose that a nonlinear equation, say in two independent variables x and t, is given by

$$F(u,u_t,u_x,u_{tt},u_{xt},u_{xx},\cdots)=0, \quad (2)$$

where $u=u(x,t)$ is an unknown function, F is a polynomial in $u=u(x,t)$ and its various partial derivatives, in which the highest order derivatives and nonlinear terms are involved. In the following steps, we give the main steps of the modified (G'/G)-expansion method.

Step 1. Use the travelling wave transformation:

$$u=u(x,t)=u(\xi), \quad \xi=k(x-Vt+\xi_0), \quad (3)$$

where k and V is a constant to be determined later, ξ_0 is an arbitrary constant. The travelling wave variable (3) permits us to reduce (2) to an ODE for $u=u(\xi)$

$$F(u,-kVu',ku',k^2V^2u'',-k^2Vu'',k^2u'',\cdots)=0. \quad (4)$$

Step 2. Suppose that the solution of ODE (4) can be expressed by a polynomial in $(\frac{G'}{G}+\frac{\lambda}{2})$ as follows:

$$u(\xi)=\sum_{i=-m}^{m}\alpha_i(\frac{G'}{G}+\frac{\lambda}{2})^i, \quad (5)$$

where $|\alpha_{-m}|+|\alpha_m|\neq 0$, and $G=G(\xi)$ satisfies the second order LODE in the form

$$G''+\lambda G'+\mu G=0, \quad (6)$$

where prime denotes derivative with respect to ξ, $\alpha_i(i=\pm1,\pm2,\cdots,\pm m)$, λ and μ are constants to be determined later. The positive integer m can be determined by considering the homogeneous balance between the highest order derivatives and nonlinear terms appearing in ODE (4).

From the second order LODE (6), after some manipulation we find that

$$(\frac{G'}{G}+\frac{\lambda}{2})'=h-(\frac{G'}{G}+\frac{\lambda}{2})^2, \quad (7)$$

where $h=(\lambda^2-4\mu)/4$, and the h is determined by λ and μ.

So, $(\frac{G'}{G}+\frac{\lambda}{2})$ now satisfies the Riccati like equation (7). It is found that the Riccati like equation (7) admits several types of solutions[16]

$$\frac{G'}{G}+\frac{\lambda}{2}=\begin{cases}\sqrt{h}\tanh(\sqrt{h}\xi), & h>0;\\ \sqrt{h}\coth(\sqrt{h}\xi), & h>0;\\ 1/\xi, & h=0;\\ -\sqrt{-h}\tan(\sqrt{-h}\xi), & h<0;\\ \sqrt{-h}\cot(\sqrt{-h}\xi), & h<0.\end{cases} \quad (8)$$

Step 3. By substituting (5) into (4) and using first order ODE (7), collecting all terms with the same order of

$(\frac{G'}{G}+\frac{\lambda}{2})$ together, the left-hand side of Eq. (4) is converted into another polynomial in $(\frac{G'}{G}+\frac{\lambda}{2})$. Equating each coefficient of this polynomial to zero, yields a set of algebraic equations for $\alpha_i (i=\pm1,\pm2,\cdots,\pm m)$, V, λ and μ.

Step 4. Assume that the constants $\alpha_i (i=\pm1,\pm2,\cdots,\pm m)$, V, λ and μ can be obtained by solving the algebraic equations in Step 3. And the general solutions of the Riccati like equation (7) has been well known for us, as (8). And then substituting $\alpha_i (i=\pm1,\pm2,\cdots,\pm m)$, V and the general solutions (8) into (5) we have more travelling wave solutions of (2).

3. Exact travelling wave solution of NLWE in a MEE circular rod

To construct exact travelling wave solution of the nonlinear longitudinal wave equation in a magneto-electro-elastic circular rod by the modified (G'/G)-expansion method. By using the transformation

$$u=u(z,t)=u(\xi),\quad \xi=k(z-Vt+\xi_0), \tag{9}$$

where k and V are the wave number and wave velocity, respectively, ξ_0 is an arbitrary real constant. Then Eq. (1) can be converted into an ordinary differential equation (ODE) for $u(\xi)$, we have

$$k^2u''''+\frac{c_0^2-V^2}{NV^2}u''+\frac{c_0^2}{2NV^2}(u^2)''=0, \tag{10}$$

where prime denotes derivative with respect to ξ.

Integrating Eq.(10) twice with respect to ξ, and letting the integral constants be zero, we then have

$$k^2u''+\frac{c_0^2-V^2}{NV^2}u+\frac{c_0^2}{2NV^2}u^2=0. \tag{11}$$

Balancing u'' with u^2 in Eq. (11) gives $m=2$. This means that we can write (5) as

$$u(\xi)=\alpha_2(\frac{G'}{G}+\frac{\lambda}{2})^2+\alpha_1(\frac{G'}{G}+\frac{\lambda}{2})+\alpha_0+\alpha_{-1}(\frac{G'}{G}+\frac{\lambda}{2})^{-1}+\alpha_{-2}(\frac{G'}{G}+\frac{\lambda}{2})^{-2}, \tag{12}$$

where $|\alpha_2|+|\alpha_{-2}|\neq0$.

Substituting (12) into (11), collecting the coefficients of $(\frac{G'}{G}+\frac{\lambda}{2})^i (i=\pm1,\pm2,\cdots,\pm4)$, and solving the resulting system with the aid of MATHEMATICA, we have the following sets of solutions:

The 1*th* solutions set:

$$\alpha_0=\frac{12hk^2N}{1-4hk^2N},\alpha_2=\frac{-12k^2N}{1-4hk^2N},$$
$$V^2=\frac{c_0^2}{1-4hk^2N},\alpha_{-2}=\alpha_{-1}=\alpha_1=0; \tag{13}$$

The 2 *th* solutions set:

$$\alpha_0=\frac{4hk^2N}{1+4hk^2N},\alpha_2=\frac{-12k^2N}{1+4hk^2N},$$
$$V^2=\frac{c_0^2}{1+4hk^2N},\alpha_{-2}=\alpha_{-1}=\alpha_1=0; \tag{14}$$

The 3*th* solutions set:

$$\alpha_0=\frac{12hk^2N}{1-4hk^2N},\alpha_{-2}=\frac{-12h^2k^2N}{1-4hk^2N},$$
$$V^2=\frac{c_0^2}{1-4hk^2N},\alpha_2=\alpha_{-1}=\alpha_1=0; \tag{15}$$

The 4 *th* solutions set:

$$\alpha_0=\frac{4hk^2N}{1+4hk^2N},\alpha_{-2}=\frac{-12h^2k^2N}{1+4hk^2N},$$
$$V^2=\frac{c_0^2}{1+4hk^2N},\alpha_2=\alpha_{-1}=\alpha_1=0; \tag{16}$$

The 5 *th* solutions set:

$$\alpha_0=\frac{24hk^2N}{1-16hk^2N},\alpha_{-2}=\frac{-12h^2k^2N}{1-16hk^2N},$$
$$\alpha_2=\frac{-12k^2N}{1-16hk^2N},V^2=\frac{c_0^2}{1-16hk^2N},$$
$$\alpha_{-1}=\alpha_1=0; \tag{17}$$

The 6 *th* solutions set:

$$\alpha_0=\frac{-8hk^2N}{1+16hk^2N},\alpha_{-2}=\frac{-12h^2k^2N}{1+16hk^2N},$$
$$\alpha_2=\frac{-12k^2N}{1+16hk^2N},V^2=\frac{c_0^2}{1+16hk^2N},$$
$$\alpha_{-1}=\alpha_1=0; \tag{18}$$

where $h=(\lambda^2-4\mu)/4\neq0$, λ and μ are arbitrary real constants.

Substituting (13)-(18) into (12) and recall the general solutions (8), we have the solutions of (11) as follows:

When $\lambda^2-4\mu>0$, we have the hyperbolic function travelling wave solutions

$$u_1(\xi)=\frac{12hk^2N}{1-4hk^2N}\operatorname{sech}^2(\sqrt{h}\xi),$$
$$\xi=k\left[z\pm\left(\frac{c_0^2}{1-4hk^2N}\right)^{1/2}t+\xi_0\right]; \quad (19)$$

$$u_2(\xi)=\frac{4hk^2N}{1+4hk^2N}[1-3\tanh^2(\sqrt{h}\xi)],$$
$$\xi=k\left[z\pm\left(\frac{c_0^2}{1+4hk^2N}\right)^{1/2}t+\xi_0\right]; \quad (20)$$

$$u_3(\xi)=\frac{12hk^2N}{-1+4hk^2N}\operatorname{csch}^2(\sqrt{h}\xi),$$
$$\xi=k\left[z\pm\left(\frac{c_0^2}{1-4hk^2N}\right)^{1/2}t+\xi_0\right]; \quad (21)$$

$$u_4(\xi)=\frac{4hk^2N}{1+4hk^2N}[1-3\coth^2(\sqrt{h}\xi)],$$
$$\xi=k\left[z\pm\left(\frac{c_0^2}{1+4hk^2N}\right)^{1/2}t+\xi_0\right]; \quad (22)$$

$$u_5(\xi)=\frac{48hk^2N}{-1+16hk^2N}\operatorname{csch}^2(2\sqrt{h}\xi),$$
$$\xi=k\left[z\pm\left(\frac{c_0^2}{1-16hk^2N}\right)^{1/2}t+\xi_0\right]; \quad (23)$$

$$u_6(\xi)=-\{8hk^2N+12k^2hN[\tanh^2(\sqrt{h}\xi)+\coth^2(\sqrt{h}\xi)]\}(1+16hk^2N)^{-1},$$
$$\xi=k\left[z\pm\left(\frac{c_0^2}{1+16hk^2N}\right)^{1/2}t+\xi_0\right]; \quad (24)$$

and when $\lambda^2-4\mu<0$, then we have the trigonometric solutions

$$u_7(\xi)=\frac{12hk^2N}{1-4hk^2N}\sec^2(\sqrt{-h}\xi),$$
$$\xi=k\left[z\pm\left(\frac{c_0^2}{1-4hk^2N}\right)^{1/2}t+\xi_0\right]; \quad (25)$$

$$u_8(\xi)=\frac{4hk^2N}{1+4hk^2N}[1+3\tan^2(\sqrt{-h}\xi)],$$
$$\xi=k\left[z\pm\left(\frac{c_0^2}{1+4hk^2N}\right)^{1/2}t+\xi_0\right]; \quad (26)$$

$$u_9(\xi)=\frac{12hk^2N}{1-4hk^2N}\csc^2(\sqrt{-h}\xi),$$
$$\xi=k\left[z\pm\left(\frac{c_0^2}{1-4hk^2N}\right)^{1/2}t+\xi_0\right]; \quad (27)$$

$$u_{10}(\xi)=\frac{4hk^2N}{1+4hk^2N}[1+3\cot^2(\sqrt{-h}\xi)],$$
$$\xi=k\left[z\pm\left(\frac{c_0^2}{1+4hk^2N}\right)^{1/2}t+\xi_0\right]; \quad (28)$$

$$u_{11}(\xi)=\frac{48hk^2N}{1-16hk^2N}\csc^2(2\sqrt{-h}\xi),$$
$$\xi=k\left[z\pm\left(\frac{c_0^2}{1-16hk^2N}\right)^{1/2}t+\xi_0\right]; \quad (29)$$

$$u_{12}(\xi)=\{-8hk^2N+12k^2hN[\tan^2(\sqrt{-h}\xi)+\cot^2(\sqrt{-h}\xi)]\}(1+16hk^2N)^{-1},$$
$$\xi=k\left[z\pm\left(\frac{c_0^2}{1+16hk^2N}\right)^{1/2}t+\xi_0\right]. \quad (30)$$

where $h=(\lambda^2-4\mu)/4$, λ, μ and ξ_0 are arbitrary real constants, wave number k is a arbitrary positive real constant.

The obtained solutions existence must meet the following condition, Eqs. (19), (21), (25) and (27) must satisfied $hk^2N<0.25$, Eqs. (20), (22), (26) and (28) must satisfied $hk^2N>-0.25$, Eqs. (23) and (29) must satisfied $hk^2N<0.0625$, Eqs. (24) and (30) must satisfied $hk^2N>-0.0625$.

Here, comparing our obtained results with Xue's exact solitary wave solution in Ref.[11]. Xue et al. only obtained one solitary wave solution in Ref.[11]. We obtained twelve exact traveling wave solutions of Eq. (1). Not only the solitary wave solutions have been given, but also many other period exact travelling wave solutions of the NLWE in a MEE circular rod are successfully obtained by the modified (G'/G)-expansion method in this work. And the obtained solutions Eqs. (19)-(24) are soliton solutions, and Eqs.(25)-(30) are periodic solutions. All the exact travelling wave solutions reported in this paper have been checked with MATHEMATICA.

4. Numerical results and discussion

In section 3, the exact travelling wave solution have been successfully constructed. In the obtained solutions, the parameters c_0 and N are both depending on the material properties as well as the geometry of the rod, c_0

and N have been listed in table 1, wave number k, $h=(\lambda^2-4\mu)/4$ and ξ_0 are free real constants.

For example, the first exact travelling wave solution (19) is a solitary wave solution of Eq.(1). The solitary wave amplitude A and wave velocity V of Eq. (19) can be expressed as:

$$A=\frac{12hk^2N}{1-4hk^2N}, V=c_0(1-4hk^2N)^{-1/2}. \quad (31)$$

Furthermore, from Eq.(31) we can obtain that the maximum wave number $k_{\max}$ must satisfy

$$k_{\max}=(2\sqrt{Nh})^{-1}. \quad (32)$$

Hence, the maximum wave number $k_{\max}$ for five different materials in table 1 sequence are 37.959, 40.975, 48.656, 39.904 and 39.284, respectively. For a solitary wave solution, the wave number k must satisfy $k<k_{\max}$.

If parameters k, h and ξ_0 are given special value. To facilitate our study, we set $\xi_0=0$, h =1 or 2, k =5 or 6. According to the data of table 1 and Eq. (31), we can obtain the solitary wave amplitude and wave velocity, the obtained solitary wave amplitude and wave velocity are list in table 2.

The relations between the solitary wave velocity V and wave number k for the five different materials when h =1 are plotted in Fig.1. If wave number $k>k_{\max}$, the solitary wave velocity V will break.

It is observed that when the wave number k is small, the wave velocity in the coupled class (PM, MEE, and PE) is higher than that in the purely elastic class (EI and TI). However, with increasing wave number k, these five materials form three new classes: PM is the first with the highest velocity; in the middle, we have MEE, EI and TI; and finally PE has the lowest velocity.Eq. (19) is a bell-shaped sech^2 solitary wave solution, and it is a soliton solution.

Solitons are special kinds of solitary wave. The soliton solution is spatially localized solution, hence $u'(\xi)$, $u''(\xi)$ and $u'''(\xi)\to\pm\infty$, $\xi=k(z-Vt)$. Soliton have a remarkable soliton property in that it keeps its identity upon interacting with other soliton. And soliton's graph is a bell-shaped sech^2 soliton solution characterized by infinite wings or infinite tails. Fig.2 shows the solitary wave u in the 50% MEE rod versus the variable time t and z of Eq. (19) with $h=1$ and $k=5$. It is clear that the maximum of u is reached at the center $t=0$ and $z=0$. Obviously, the solitary wave amplitude is symmetrical about the center.

In the same way, Fig.3 shows the solitary wave u in the 50% MEE rod versus the variable time t and z of Eq. (24) with $h=1$ and $k=5$. It is clear that the minimum of u is reached at the center $t=0$ and $z=0$. And it shape like a cone, obviously, the solitary wave amplitude is symmetrical about the center. To the best of our knowledge, the obtained exact solitary solution (24) is a new solution, it have not been reported.

Similarly, we can give the numerical result and figure of other four materials and other exact travelling wave solution of Eq. (1), which are omitted for convenience.

Table 2: The solitary wave amplitude and wave velocity for different k and h

v_f	h	k	0%(PM)	50%(MEE)	100%(PE)	TI	EI
A(m)	1	5	0.05297	0.04535	0.03202	0.04785	0.04940
V(m/s)	1	5	5258.92	5183.33	5076.68	4838.43	4879.49
A(m)	2	5	0.10784	0.09208	0.06473	0.09725	0.10045
V(m/s)	2	5	5305.97	5222.96	5103.99	4877.49	4920.16
A(m)	2	6	0.15779	0.13441	0.09410	0.14207	0.14682
V(m/s)	2	6	5348.44	5258.59	5128.39	4912.65	4956.81
A(m)	1	6	0.07687	0.06573	0.04632	0.06939	0.07166
V(m/s)	1	6	5279.47	5200.66	5088.64	4855.50	4897.26

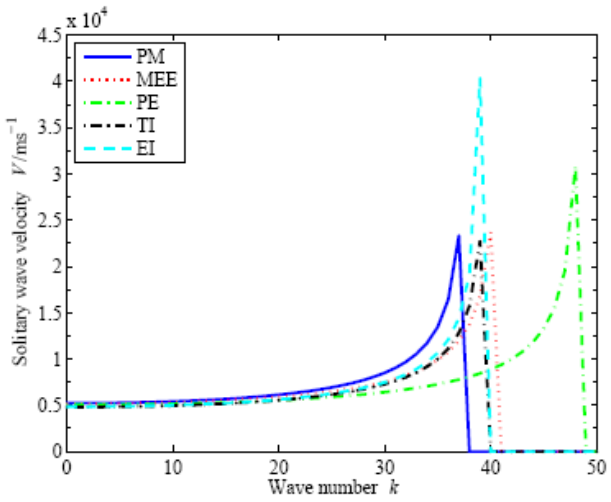


Fig. 1 Wave velocity V governing by (31) versus wave number k in a rod with $h=1$.

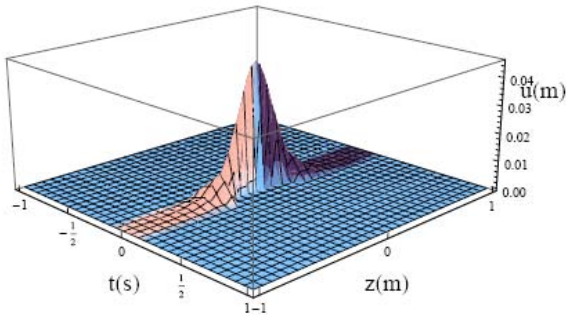


Fig. 2 Soliton versus t and z in a 50% MEE rod of Eq.(19) with $h=1$ and $k=5$.

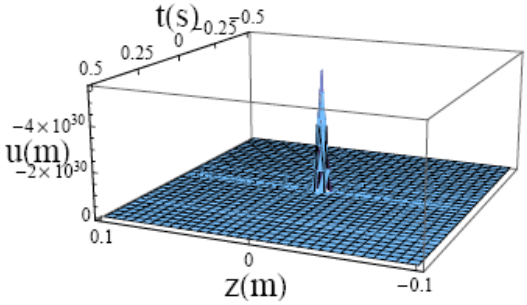


Fig. 3 Solitary waves versus t and z in a 50% MEE rod of Eq.(24) with $h=1$ and $k=5$.

5. Conclusion

In this paper, we present a modified (G'/G)-expansion method based on (G'/G)-expansion method. We have applied the new method to find the exact travelling wave solutions of nonlinear solitary wave equation in a long MEE circular rod. And the obtained exact travelling wave solutions are expressed by the hyperbolic functions, the rational functions and the trigonometric functions. When the parameters are taken as special values, the solitary wave solutions are derived from the hyperbolic functions. The obtained results show the modified (G'/G)-expansion method is direct, concise and effective with the help of MATHEMATICA, and this method can be applied to many other nonlinear partial differential equations in mathematical physics. Some numerical examples are further presented for the wave in a rod made of five different materials: the three-phase fully coupled MEE, coupled piezoelectric PE, coupled piezomagnetic PM, purely elastic but transverse isotropy TI and purely elastic isotropy EI. It is demonstrated that the solitary wave not only exists in such rods but also shows different features in different materials, which could have potential applications in non-destructive evaluation of structures made of the advanced MEE material.

Acknowledgments

This work is financial supported by the Science Foundation of North University of China (2010, 2011).

References

[1] Porubov A V, Velarde M G, "Dispersivedissipative solitons in nonlinear solids", Wave Motion, Vol. 31, No. 1, 2000, pp. 197–207.
[2] Samsonov A M, "Strain Solitons in Solids and How to Construct Them", New York: Chapman and Hall/CRC, 2001
[3] Guo J G, Zhou L J, Zhang S Y, "The geometrical nonlinear waves in finite deformation elastic rods", Appl. Math. Mech., Vol. 26, 2005, pp. 667–674.
[4] Zhang S Y, Liu Z F, "Three kinds of nonlinear dispersive waves in elastic rods with finite deformation", Appl. Math. Mech., Vol. 29, 2008, pp. 909–917.
[5] Liu Z F, Zhang S Y, "Nonlinear waves and periodic solution in finite deformation elastic rod", Acta Mech. Solida Sin., Vol. 9, No. 1, 2006, pp. 1–8.
[6] Christov C I, Marinov T T, Marinova R S, "Identification of solitary wave solutions as an inverse problem: application to shapes with oscillatory tails", Math. Comput. Simul., Vol. 80, No. 1, 2009, pp. 56–65.
[7] Chen J Y, Pan E, Chen H L, "Wave propagation in magneto-electro-elastic multilayered plates", Int. J. Solids Struct., Vol. 44, 2007, pp. 1073–1085.
[8] Chen P, Shen Y, "Propagation of axial shear magneto-electro-elastic waves in piezoelectricpiezomagnetic composites with randomly distributed cylindrical inhomogeneities", Int. J. Solids Struct., Vol. 44, 2007, pp. 1511–1532.
[9] Wu B, Yu J G, He C F, "Wave propagation in non-homogeneous magneto-electro-elastic plates", J. Sound Vib., Vol. 317, No. 1, 2008, pp. 250–264.

[10] Maity Niladri Pratap, Maity Reshmi, "Surface plasmon waves on noble metals at optical wavelengths", Int. J. Comput. Sci. Issues, Vol. 8, No. 3 3-2, 2011, pp. 485–490.
[11] C X Xue, E Pan, S Y Zhang, "Solitary waves in a magneto-electro-elastic circular rod", Smart Mater. Struct., Vol. 20, 2011, pp. 105010-17.
[12] W. Malfliet, "Solitary wave solutions of nonlinear wave equations", Am. J. Phys., Vol. 60, No. 7, 1996, pp. 650–654.
[13] W. Malfliet, "The tanh method: a tool for solving certain classes of nonlinear evolution and wave equation", J. Comput. Appl. Math. Vol. 164, 2004, pp. 529–541.
[14] A.M. Wazwaz, "The tanh method for traveling wave solutions of nonlinear equations", Appl. Math. Comput., Vol. 154, No. 3, 2004, pp. 713–723.
[15] E.J. Parkes, B.R. Duffy, "An automated tanhfunction method for finding solitary wave solutions to nonlinear evolution equations", Comput. Phys. Commun., Vol. 98, 1996, pp. 288–300.
[16] Engui Fan, "Extended tanh-function method and its applications to nonlinear equations", Phys. Lett. A, Vol. 277, 2000, pp. 212–218.
[17] M.L. Wang, X.Z. Li, "Extended F-expansion method and periodic wave solutions for the generalized Zakharov equations", Phys. Lett. A, Vol. 343, 2005, pp. 48–54.
[18] X.Z. Li, M.L. Wang, "A sub-ODE method for finding exact solutions of a generalized KdV-mKdV equation with higher order nonlinear terms", Phys. Lett. A, Vol. 361, 2007, pp. 115–118.
[19] S.K. Liu, Z.T. Fu, S.D. Liu, Q. Zhao, "Jacobi elliptic function expansion method and periodic wave solutions of nonlinear wave equations", Phys. Lett. A, Vol. 289, 2001, pp. 69–74.
[20] M.L.Wang, Y.B. Zhou, Z.B. Li, "Application of a homogeneous balance method to exact solutions of nonlinear evolution equations in mathematical physics", Phys. Lett. A, Vol. 216, 1996, pp. 67–75.
[21] J.H. He, X.H. Wu, "Exp-function method for nonlinear wave equations", Chaos Solitons Fractal, Vol. 30, 2006, pp.700–708.
[22] M.L. Wang, X.Z. Li, J.L. Zhang, "The (G'/G)-expansion method and travelling wave solutions of nonlinear evolution equations in mathematical physics", Phys. Lett. A, Vol. 372, 2008, pp. 417–423.
[23] X.M. Ma, Y.T. Pan, L.Z. Chang. "The Modified (G'/G)-expansion Method and Its Applications to KDV Equation", Int. J. Nonlinear Sci., Vol. 12, No. 4, 2011, pp. 400–405.
[24] A.M.Wazwaz, "A study on nonlinear dispersive partial differential equations of compact and noncompact solutions", Appl. Math. Comput., Vol. 135, 2003, pp. 399–409.

Xinmou Ma Received the B.C. degree in Mechnism design and manufacture from North China Institute of Technology, Taiyuan, China, in 2002. Received the M.S. degree and Ph. D. degree at Gun, automatism weapon and ammunition engineering from North University of China in 2005 and 2012, respectively. Now, I worked in North University of China as a lecture. My reseach interests include nonlinear dynamics, chaos, nonlinear PDE exact solution, approximately analytical solution and numerical solution.

Yutian Pan is a professor in North University of China, My reseach interests weapon design and nonlinear dynamics.

Liezhen Chang Received the B.C. degree in Engineering Mechnics from Taiyuan Heavy Machinery Institute, China, in 2002. Received the M.S. degree in Engineering Mechnics from North University of China in 2007, She is currently working toward the Ph.D. degree at Gun, automatism weapon and ammunition engineering from North University of China. Now, I worked in North University of China as a lecture. My research interests include FEM, nonlinear dynamics and nonlinear PDE exact solution and approximately solution.

Moving Foreground Detection Based On Spatio-temporal Saliency

Yang Xia[1], Ruimin Hu[1], Zhongyuan Wang[1] and Tao Lu[1, 2]

[1] National Multimedia Software Engineering Research Center, Computer School of Wuhan University, Wuhan University, Wuhan, 430072, China

[2] Hubei Province key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan, 430074, China

Abstract

Detection of moving foreground in video is very important for many applications, such as visual surveillance, object-based video coding, etc. When objects move with different speeds and under illumination changes, the robustness of moving object detection methods proposed so far is still not satisfactory. In this paper, we use the semantic information to adjust the pixel-wise learning rate adaptively for more robust detection performance, which are obtained by spatial saliency map based on Gaussian mixture model (GMM) in luma space and temporal saliency map obtained by background subtraction. In addition, we design a two-pass background estimation framework, in which the initial estimation is used for temporal saliency estimation, and the other is to detect foreground and update model parameters. The experimental results show that our method can achieve better moving object extraction performance than the existing background subtraction method based on GMM.

Keywords: *Moving Object Detection, Background Subtraction, Visual Saliency, Gaussian Mixture Model*

1. Introduction

Detection of moving objects in video sequences is essential in many applications, such as visual surveillance [1], object-based video coding [2], etc. The most popular approach for moving object detection is the background subtraction, which maintains an up-to-date model of the background and detects moving objects as those that deviate from the background model. A classical parametric background model with adaptive Gaussian mixture models was introduced by C.Stauffer and W.Grimson [3]. Each pixel is modeled as a mixture of Gaussians (MOG) which can be updated on-line. Object detection is performed by matching luminance and color of every pixel with the most likely background Gaussians distribution. But this kind of parametric approach may fail when the density function is complicated. To solve the problem, Elgammal [4] and Mittal [5] used kernel density estimate to model more complex background. Recently, Zivkovic [6] improved the GMM based approach [3] by adaptively selecting an appropriate number of Gaussian components for each pixel. However, since the update rate of background model parameters are usually slower than the rate of illumination changes, above methods are not robust to illumination changes. At meanwhile, the background models detect foreground pixel only by the temporal information of the pixel, so the background models are inaccurate when the objects move slowly, with only the edges of outstanding objects labeled salient.

Visual saliency is another way to detect objects from video, which can use more spatial information to increase detection robustness. As a representative work of visual saliency, Itti set up computational visual saliency models for still image [7] and video [8] by simulating the pre-attentive selection mechanism. The regions which have high saliency values may be considered as salient objects. However the experiment results [8] show that the method may mistake a lot of background regions for foreground.

It is observed that background subtraction based methods are sensitive to illumination changes, while visual saliency is more robust to illumination changes but mistakes many background areas for foreground in practice. If the background subtraction based methods incorporate some spatial information obtained by visual saliency, the foreground detect performance will be improved. In this paper, a novel object detection algorithm based on spatio-temporal saliency is proposed, which takes advantage of background subtraction and visual saliency. In our approach, spatial saliency based on Gaussian mixture model in luminance space and temporal saliency obtained by background subtraction, are calculated as an auxiliary knowledge to adjust the pixel-wise learning rate of object detection model adaptively, which makes foreground detection more robust when objects move with different speeds and under illumination changes.

The rest of the paper is organized as follows. In Sec. 2 we analyze the drawback of traditional GMM based object detection method. Sec. 3 presents the proposed algorithm, and Sec. 4 shows the experimental results on the CAVIAR's dataset. Finally, Sec. 5 gives conclusions.

2. The Analysis of Traditional Model

Zivkovic [6] used a mixture model consisting of Gaussian distributions to estimate a density distribution from a sequence for a pixel at a location x, which can be denoted by (1)

$$P(I_{t,x}) = \sum_{n=1}^{N} w_{t-1,x,n} * \frac{1}{\sqrt{2\pi\sigma^2_{t-1,x,n}}} \exp\left(-\frac{\left(I_{t,x} - \mu_{t-1,x,n}\right)^2}{2\sigma^2_{t-1,x,n}} \right) \quad (1)$$

where $w_{t-1,x,n}$ is mixture weight for the n th Gaussian model, $\mu_{t-1,x,n}$ and $\sigma^2_{t-1,x,n}$ are the mean and variance of Gaussian model, and N is the number of Gaussian models.

In GMM based object detection method, $w_{t-1,x,n}$, $\mu_{t-1,x,n}$, $\sigma^2_{t-1,x,n}$ must be updated when new data $I_{t,x}$ comes. A learning rate α is used to limit the influence of past data and absorb coming information. When $I_{t,x}$ matches the n th Gaussian model, the $w_{t-1,x,n}$, $\mu_{t-1,x,n}$, $\sigma^2_{t-1,x,n}$ can be update as follow:

$$w_{t,x,n} = w_{t-1,x,n} + \alpha(1 - w_{t-1,x,n}) \quad (2)$$

$$\mu_{t,x,n} = \mu_{t-1,x,n} + (\alpha / w_{t-1,x,n}) * (I_{t,x} - \mu_{t-1,x,n}) \quad (3)$$

$$\sigma^2_{t,x,n} = \sigma^2_{t-1,x,n} + (\alpha / w_{t-1,x,n}) * ((I_{t,x} - \mu_{t-1,x,n})^2 - \sigma^2_{t-1,x,n}) \quad (4)$$

In [6] α is set to 0.001 to guarantee that the objects that move slowly can be correctly detected. But with illumination changes, the adjustment of $w_{t-1,x,n}$, $\mu_{t-1,x,n}$, $\sigma^2_{t-1,x,n}$ cannot keep up with the change of a scene due to the low learning rate, and some parts of background may be considered as foreground. Meanwhile, because the learning rate is low, the misclassification event will last for a long time. On the contrary, if we set α larger, the objects or parts of objects with slow motion speed may be seen as background. Therefore, the performance of object detection based on GMM depends on the selection of learning rate.

3. The Proposed Method

In our opinion, the learning rate of foreground in GMM should be different from that of background. In this paper, we use the semantic information to adjust the pixel-wise learning rate adaptively for more robust detection performance, which are obtained by spatial saliency map based on Gaussian mixture model (GMM) in luma space and temporal saliency map obtained by background subtraction. In addition, we design a two-pass background estimation framework, in which the initial estimation is used for temporal saliency estimation, and the other is to detect foreground and update model parameters. The proposed algorithm consists of the following three steps.

3.1 Spatial Saliency Analysis

In this section, we calculate the spatial saliency based on Gaussian mixture model. Firstly, we separate a frame into several clusters based on GMM [9], and then generate spatial saliency map by calculating the weighted distance from each point to the cluster centers. We use algorithm 1 to obtain the image clusters.

Algorithm 1: **Image frame cluster based on GMM**

Input: the current video frame

1. Obtain luma histogram of the current frame, and find the local maximums p_i of the histogram
2. Remove small peaks of the histogram. If $\|p_i - p_j\|_2 < \varepsilon$, the bigger peak will be preserved.
3. Use the remaining peaks as initial mean $\mu_i = p_i$ $i = 1,...,m$, m is the number of remaining peaks.
4. Use the midpoint between two peaks which can identify interval of an initial Gaussian cluster, denoted as $[bl\ br]$, to calculate initial covariance matrices of each Gaussian cluster.
$$\sigma_i^2 = \sum_{q=bl}^{q=br} \left(h_q - \mu_i\right)^2 * c(h_q) \Big/ \sum_{q=bl}^{q=br} c(h_q)$$
where $c(h_q)$ is the histogram value at h_q
5. Normalize the heights of peaks and set them as the initial weights of Gaussian clusters, denoted as w_i
$$w_i = c(\mu_i) \Big/ \sum_{i=1}^{m} c(\mu_i)$$
6. Run EM algorithm to find more accurate w_i, μ_i, σ_i^2.

Output: w_i, μ_i, σ_i^2.

Then we use the Gaussian clusters to calculate spatial saliency. We suppose the luminance intensities of background centralize around several peaks in the histogram, and the outstanding objects in picture have different luminance intensities from background. Furthermore, we believe that the areas covering the outstanding objects are much smaller than the background area. Based on the hypothesis, we divide Gaussian clusters into background clusters and foreground clusters. We sort

w_i by descending order, find the first k weights which satisfy $\sum_{j=1}^{k} w_j > \eta$, and set the corresponding Gaussian clusters to background clusters. The left clusters are regarded as foreground clusters.

Based on the labels of cluster, the spatial saliency map is obtained by calculating weighted distance from each point to the cluster centers, which is shown as follow:

$$ss_{t,x} = \rho * \sum_{i=1}^{k} w_i^2 * (I_{t,x} - \mu_i^B)^2 + (1-\rho) \sum_{j=1}^{m-k} w_j^2 * (I_{t,x} - \mu_j^F)^2 \quad (5)$$

where μ_i^B is the mean of background cluster, μ_j^F is the mean of foreground cluster, and ρ is set to 0.6 in this paper. Because background is much bigger than the moving objects, the mixing weights of background clusters are usually larger. So the outstanding objects which are much different from background clusters can be marked salient.

3.2 Spatio-temporal Saliency Map Generation

Since spatial saliency may mark some outstanding objects which are part of background, in this section, we combine spatial saliency and temporal saliency to get more accurate saliency information. For moving object detection, the moving foreground can be regarded as temporal saliency, so we use GMM to generate temporal saliency map.

First, we use the recent GMM based background subtraction method [6] to obtain a preliminary binary object map $fg_{t,x}$, where $fg_{t,x}$ equals 0 in background region and equals 1 in object region.

After obtaining the binary foreground map, temporal saliency map can be obtained by (6) and we find θ equals 0.3 can get good performance.

$$st_{t,x} = \theta * (1 - fg_{t,x}) + (1-\theta) * fg_{t,x} \quad (6)$$

Because the salient areas in spatial saliency map include some other static objects, we do not use the sum of temporal and spatial saliency maps, and we find the product of temporal and spatial saliency maps can eliminate some error. So a single saliency map is generated by.

$$STs_{t,x} = ss_{t,x} * st_{t,x} \quad (7)$$

It is observed that the spatio-temporal saliency contains the preliminary semantic information of the frame which describe the probability of a pixel belong to foreground.

(a) (b) (c)

Fig. 1 An example of Spatio-temporal saliency and learning rate scale function (a) Frame 90 from"Walk1" (b) Spatio-temporal saliency extraction result (c) learning rate scale function

3.3 Learning Rate Control Scheme

In this section, we use GMM again to detect moving objects. After obtaining the spatio-temporal saliency map, the learning rate of GMM model can be adaptively adjusted according to the saliency semantic information. We choose logistic function as learning rate scale function, as shown as Fig 1. Supposed that μ_{st} is the mean of saliency map, and σ_{st} is the variance of saliency map. We use the threshold $\tau = \mu_{st} + \sigma_{st}$ to distinguish salient region and un-salient region. The learning rate scale function can be defined as follow:

$$SF_{t,x} = a / \left(1 + \exp(b * (STs_{t,x} - \tau))\right) + c \quad (8)$$

where a is 999, and c is 1.0 in this paper. b is a parameter to control the size of distinguish belt, which is set to 0.1 in this paper. So the learning rate of GMM can be adjusted as follow:

$$\alpha_{t,x} = \gamma * SF_{t,x} \quad (9)$$

where γ is 0.0001. It can be found that the range of $\alpha_{t,x}$ is [0.0001, 0.1], and if the regions have higher saliency, the lower learning rate $\alpha_{t,x}$ will be assigned to the regions.

In addition, the overall proposal can be described as follow:

Algorithm 2: **Foreground Detection Scheme**

Input: video sequence,
For $t = 1, ..., L$

1. Compute $ss_{t,x}$ by algorithm 1 and $fg_{t,x}$ by GMM [6]
2. Compute $STs_{t,x}$ by (6) and (7)
3. Calculate $\alpha_{t,x}$ using (7) and (8)
4. Based new learning rate $\alpha_{t,x}$, perform GMM [6] again to get the final foreground map. At meanwhile, update $w_{t-1,x,n}$, $\mu_{t-1,x,n}$, $\sigma^2_{t-1,x,n}$ for next frame as follow:

   If $I_{t,x}$ matches the n th Gaussian model

   Use (2), (3), (4) to update the model parameters

   else

   A model replacement is performed to

incorporate $I_{t,x}$ into the GMM.

$k = \arg\min_{n=1,.....N} w_{t-1,x,n}$

$\mu_{t,x,k} = I_{t,x}, \sigma^2_{t,x,k} = \sigma^2_0, w_{t,x,k} = w_0$

$w_{t,x,n} = w_{t-1,x,n} - \alpha w_{t-1,x,n}$ *when* $n \neq k$

end

Output: the final foreground map of each frame.

4. Experiment and Result

To evaluate the performance of our proposed model, we applied it to extract moving objects on CAVIAR's datasets [10]. In the experiment, Zivkovic's model [6] was used as an anchor, and the constant learning rate α in [6] was set to 0.001. For a more comprehensive comparison, background subtraction with a higher learning rate (α = 0.01) was also compared with proposal.

Fig 2 illustrates the subject results of sequence "Walk1". Fig 2 (a) includes the 60th, 90th and 120th frame from "Walk1". Pictures in Fig 2 (b) are the results of GMM with α = 0.001. It can be seen that a lot of background noises are classified into foreground, although the man can be detected. When we increase the learning rate, the noise sensitivity is reduced significantly as shown in Fig 2 (c). But when the man stop and keep static for a while, GMM with higher learning rate may classify some parts of the man into background. However, the proposed spatio-temporal saliency can help us use different model parameters in different regions. Because spatial saliency is robust to illumination changes, the saliency values of background areas are low, while the area covering the person shows a high degree of saliency no matter the person is moving or not as shown in Fig 2 (d). So the higher learning rate α is set to the background region, while the lower learning rate is assigned to foreground. From Fig 2(e), we can see that the proposal can detect the man, and eliminate the background noise.

For quantitative analysis of our proposed method, ROC measure was applied, where true positive (TP), false positive (FP) were used. We firstly used image segmentation to split images into objects, and then marked the foregrounds manually. Let Rgt and Rd be the ground truth region and the detected region respectively. The region $Rgt \cap Rd$ is defined as TP, and the region $\overline{Rgt} \cap Rd$ is considered as FP.

So detection rate and false alarm rate can be obtained as follow:

$$DR = n(TP)/n(Rgt) \quad (9)$$

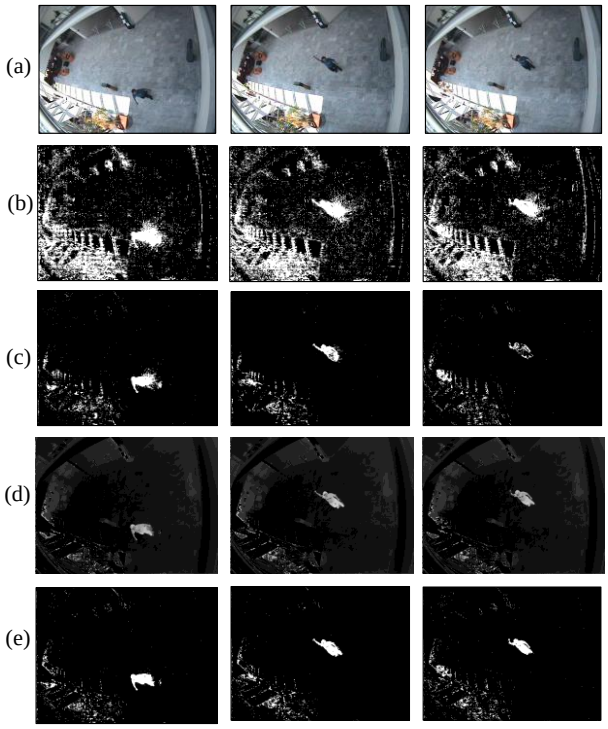
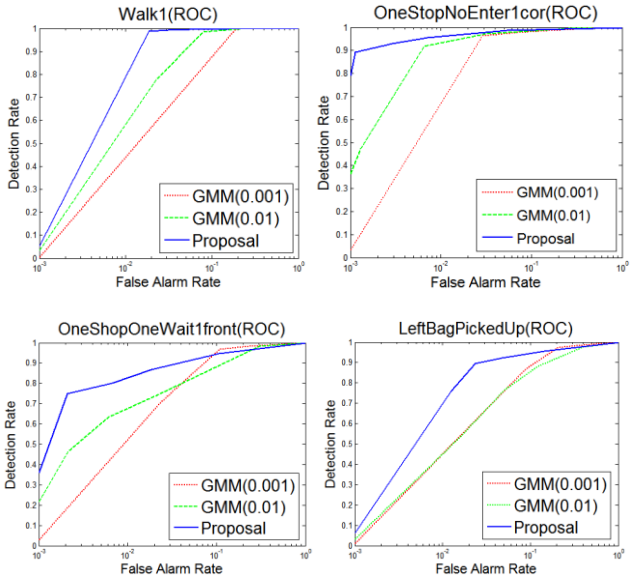


Fig 2 (a) Test frames;(b) GMM(α = 0.001); (c) GMM(α = 0.01); (d) Spatio-temporal saliency; (e) Proposal

$$FAR = n(FP)/n(\overline{Rgt}) \quad (10)$$

where $n(.)$ is an operator to count the pixel number in a region. Using detection rate and false alarm rate, we can obtain the ROC curves.



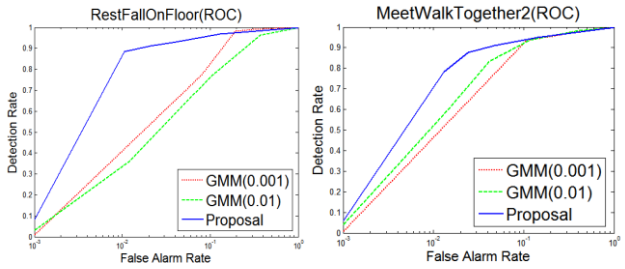


Fig3 ROC curve of six videos in CAVIAR's datasets.

The ROC curves in Fig 3 illustrates that the proposal achieves better ROC performance than anchors for six videos in the CAVIAR's datasets. Due to space limitation, we don't provide all the results here. Compared with anchors, the proposal can maintain a high detect rate while the false alarm rate decreases.

The above experimental results show that the proposed method is robust to illumination changes and movement with different moving speeds, which can achieve good balance between detecting moving objects and eliminate noises

5. Conclusion

In this paper, a more robust object detection algorithm based on spatio-temporal saliency is proposed. Spatial saliency based on Gaussian mixture model in luma space and temporal saliency obtained by background subtraction, are calculated as an auxiliary semantic knowledge to adjust pixel-wise learning rate of object detection model adaptively, which achieves good balance between detecting moving objects and eliminate noises. Experiment results show our proposal can achieve better performance than the existing background subtraction method based on GMM.

Acknowledgments

This work is supported by the major national science and technology special projects (2010ZX03004-003-03); the National Natural Science Foundation of China under Grant No. 61172173, 60970160, 61070080, 61003184, 60832002; the National Grand Fundamental Research 973 Program of China under Grant No.2009CB320906;

References

[1] C. Lakshmi Devasena, R. Revathi, M. Hemalatha,"Video Sur-veillance Systems - A Survey", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011, pp:635-642

[2] Soumaya Ghorbel, Maher Ben Jemaa and Mohamed Chtourou, "Object-based Video compression using neural networks", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011,pp:139-148

[3] C. Stauffer and W. Grimson. "Adaptive background mixture models for real-time tracking". In IEEE Conference on Computer Vision and Pattern Recognition, 1999, pp. 246-252.

[4] A. Elgammal, D. Harwood, L. Davis, "Non-parametric model for background subtraction", in: Proceedings of the 6th European Conference on Computer Vision, 2000, pp. 751–767.

[5] A. Mittal, N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation", in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, pp. 302–309.

[4] Z. Zivkovic, and Ferdinand van der Heijden "Efficient adaptive density estimation per image pixel for the task of background subtraction", Pattern Recognition Letters, 2006, pp 773–780.

[6] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Anal. Machine Intell, vol. 20, no. 11, Nov. 1998, pp. 1254-1259

[8] L. Itti, N. Dhavale, F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in: Proceedings of SPIE 48th Annual International Symposium on Optical Science and Technology, vol. 5200, Aug 2003, pp. 64-78.

[9] Heng-Da Cheng , and Ying Sun, "A Hierarchical Approach to Color Image Segmentation Using Homogeneity", IEEE Trans on Image Processing, 2000, pp. 2071-2082.

[10]http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

Yang Xia received the B.S degrees from Wuhan University of technology in 2005, Wuhan, China. He is currently pursuing the Ph.D. degree in National Engineering Research Center For Multimedia Software, Wuhan University, Wuhan, China.His research interests include image/video processing, video coding and artificial intelligence.

Ruimin Hu received the B.S and M.S degrees from Nanjing.University of Posts and Telecommunications, Nanjing China,in 1984 and in 1990 respectively. and Ph.D degree in Communication and Electronic System from Huazhong University of Science and Technology, Wuhan, China in 1994. Dr. Hu is the director of National Engineering Research Center For Multimedia Software, Wuhan University and Key Laboratory of Multimedia Network Communication Engineering in Hubei province. He is Executive Chairman of the Audio Video coding Standard (AVS) workgroup of China in Audio Section. He has published two books and over 100 scientific papers. His research interests include audio/video coding and decoding, video surveillance and multimedia data processing.

Zhongyuan Wang received the B.S. degree and M.S degree in computer science from Wuhan University, Wuhan, China, in 1995 and 2001, and he received the Ph.D. degree in Communication and Information System in Wuhan University in 2008. From 2001, he worked as a Member of Research Staff in National Multimedia Software Engineering Research Center of Wuhan University. His

research interests include video compression, multimedia communications.

Tao Lu received the B.S and M.S degrees from Computer Science and Engineering Department, Wuhan institute of technology, Wuhan, China. He is currently pursuing the Ph.D. degree in National Engineering Research Center For Multimedia Software, Wuhan University, Wuhan, China. His research interests include image/video processing, computer vision and artificial intelligence.

Recognition and Tracing Scheme Study of Moving Objects by Video Monitoring System

Peilong XU[1]

[1] **The Growing Base for State Key Laboratory, Qingdao University,**
No. 308, Ningxia Road, Qingdao 266071, China.

Abstract

Objective: In this paper a recognition and tracing scheme for moving objects by video monitoring system was studied. Methods: During moving objects recognition, Multi frame sampling method was used to establish the initial background. Edges of the moving objects were drawn according to the changes of the images, and the influence factors for edges drawing were eliminated. In the end, tracing for moving objects could be realized by recognition of morphological characteristics. Results: The experiments indicate that, for field environment, the number of collected frames between 120 to 180 could get better image background. The shadows of the moving objects are the main factor which influence detection of object edges, and they could be eliminated by Shadow edge detection operator. For vehicles tracing, adjacent frame matching method could be used to reflect the time-space transformation of the vehicles. Conclusion: This scheme could realize the recognition and tracing of moving objects by video monitoring system effectively.

Keywords: *Image Recognition, Image Tracing, Video Monitoring System.*

1. Introduction

Video monitoring systems are widely used in fields like industry, transport, security and military, and are playing more and more important roles. With time progressing, the video monitoring systems have even reached scales of hundreds and thousands ways. It is impossible to rely entirely on human to monitor so many systems. In this case, all kinds of intelligent monitoring system emerge as the times require. Nowadays, the monitoring systems could realize auto alarming of the invasion objects in certain distances through all kinds of sensors and image recognition software[1]. But this is far from enough. How to realize recognition and tracing of the moving objects by video monitoring system has become an important problem to solve.

For recent years, the objects recognition and tracing techniques have made great progress. In many brands of digital cameras, face recognition function has been developed, and the face could be positioned and be focused automatically. But these techniques could just realize recognition and comparison of some preset specific shapes, and then track them. They techniques still could not realize recognition and tracing of the multi type and multi angle objects.

This study developed a intelligent monitoring system scheme according to society video monitor needs, which could recognize and real timely trace the targets. This scheme realized classification statistics and tracing of the moving objects with high recognition rate.

2. Method for Realizing the System

Traditional target segmentation algorithms are mainly based on Iteration threshold segmentation algorithm. The general progress is to detect edges and acquire difference images, then two value the data through threshold segmentation so as to highlight the parts in the images we are interested in. But this method usually needs large amounts of computation, and demands highly time complexity[2]. So it is not suitable for real time analysis for images of the moving objects. In order to ensure the real time of the system reaction, this study used background subtraction algorithm to realize quick and effective segmentation of moving objects. The key of the algorithm is that how to establish the background models.

2.1 Background subtraction algorithm

Usually, there are two kinds of background development scheme for using background subtraction algorithm. First, select two frames of adjacent images, and put the former image as background of the latter one. Detect changes of the two images using differential operation. Second, preset an unified background, and detect image changes through comparing all acquired real time images with the preset background image by differential operation. For the first scheme, it is no need to preset background, so that it is more suitable for moving monitoring facilities. But the targets volume it could monitor is related to the targets movement speeds. If the target stopped moving after coming into monitoring views, then it could not be detected by the monitoring system. This is called "Hole" phenomenon in video monitoring. For this reason, the first

scheme is not suitable for video monitoring systems with fixed camera. The second scheme could customer the problem of "Hole" phenomenon caused by the algorithm in the first scheme and is better for tracing moving objects, but this scheme need monitoring systems to establish and real time update backgrounds by themselves.

Through analysis of the above two schemes, our study try to establish a background subtraction algorithm with an uniformed background, and based on this algorithm to set up a monitoring system. Since the background quality has great influence on the targets recognition and tracing, this study used a background estimation method by multi frame difference image to set up the initial background. And through a background changing estimate strategy, the background updates only on certain conditions. To ensure the system could work normally in different light conditions and monitoring scene changes accidentally, this study used a filtering algorithm based on differential two value image processing, and this increased robustness of the algorithm effectively.

2.2 Method for background development and update

This scheme used improved background differential method to separate moving objects with background. The scheme includes two parts which are development and update of background.

2.2.1 Method for acquire background images

The background images are developed by a multi-frame subtraction images dynamic evaluation method, and the images are collected in accordance with certain time intervals. It is supposed that 3 continues images named a, b, c, and $\{B_{i,j}^t\}$, $\{O_{i,j}^t\}$ are background and moving object of the t frame. The processing procedure should be:

1) Separate the 3 frame continues images into 2 groups. The first group include frame a and b, and the second group include frame b and c. Gray scale difference subtractions between frames are operated to each pixel of the two groups, then the absolute values were preserved in $\{N_{i,j}^{-1}\}$ and $\{N_{i,j}^{+1}\}$. See in formula (1):

$$N_{i,j}^{-1} = \left| I_{i,j}^a - I_{i,j}^b \right| \quad , \quad N_{i,j}^{+1} = \left| I_{i,j}^b - I_{i,j}^c \right| \qquad (1)$$

2) Because $\{N_{i,j}^{-1}\}$ and $\{N_{i,j}^{+1}\}$ are difference values of adjacent two frames, they are highly similar and the histogram shows double peaks. The threshold T_0 of $\{N_{i,j}^{-1}\}$ calculated through OTSU method could be used as the best separation threshold of foreground and background in $\{N_{i,j}^{-1}\}$ and $\{N_{i,j}^{+1}\}$. Compare values in $\{N_{i,j}^{-1}\}$ and $\{N_{i,j}^{+1}\}$ with T_0 separately. If for any point X (i, j), the corresponding values in $\{N_{i,j}^{-1}\}$ and $\{N_{i,j}^{+1}\}$ are all bigger than T0, then it could be judged that this point are moving in all the 3 frames, so that the point could be included into moving objects $\{O_{i,j}^t\}$ in foreground. See in formula (2):

$$O_{i,j}^t = \begin{cases} 255 & if\ D_{i,j}^{-1} > T_0\ AND\ D_{i,j}^{+1} > T_0 \\ 0 & else \end{cases} \qquad (2)$$

3) According to moving images $\{O_{i,j}^t\}$, all pixels valued 255 are eliminated from input b frames, and the rests are background images selected from b frames$\{B_{i,j}^t\}$. See in formula (3):

$$B_{i,j}^t = \begin{cases} I_{i,j}^t & if\ O_{i,j}^t = 0 \\ 0 & else\ \ O_{i,j}^t = 255 \end{cases} \qquad (3)$$

4) Deal all collected frames with method above, and supplement lost parts in background, and then a entire initial background image could be acquired.

2.2.2 Method for background image update.

After the monitoring system is used for a certain periods, the background would have some changes inevitably, for example influence of weather, light and displacement because of other factors. If the background is not updated throughout working, it must influence accuracy of recognition and tracing for moving objects[3].

This study could realize background update through method of establishing statistical models for each pixels, and calculating probabilities to judge background changes. When a certain pixel values were changing constantly during a certain times, then the grey value of the previous background would be replaced by the value of this point. Or else, the background would not be changed. In actual operation, the background judging model is established by calculating two parameters of each images—the mean value μ and variance σ. For the newly collected sample value S for a certain point (x, y), if formula (4) were satisfied, the point would be regarded to be the new background pixel.

$$f(s) = \frac{1}{\sqrt{2\pi\sigma}} Exp\left(-\frac{(s-\mu)^2}{2\sigma^2}\right) \geq T \qquad (4)$$

T is the probability threshold value in this formula, and could be set dynamically.

2.3 Recognition and tracing of moving objects

After background was established, the background difference method was used to collect shapes of moving objects in the background. The method is to use the grey values of each pixels of the current image to subtract with those of the background. See in formula (5):

$$N_{i,j}^{t} = \left| I_{i,j}^{t} - B_{i,j}^{t} \right| \quad (5)$$

Because this system need to recognize moving objects, but through experiments we got to know that the moving objects produced shadows are important factors to influence system to recognize targets accurately. So that the first step to recognize objects is to eliminate influences of the shadows. Through studies of moving objects, two characters were found: one is that the difference values of the moving object shadows to the background are smaller than the those of objects themselves to the background. And the other is that the grey values of the shadows are usually smaller than those of the surrounding areas. Then the shadows could be eliminated by threshold setting. But only use threshold to eliminate moving targets would cause lost of details of moving objects. So edge detecting becomes a key problem for moving objects recognition.

2.3.1 Edge detection of the moving objects and their shadows

In detection of the objects edges, the Prewitt operator and the Sobel gradient operator are usually used. The Prewitt operator finish the calculation by counting grey differences of each pixels with their adjacent points, and then by neighbor convolution of the model and images using the horizontal and vertical vectors. This method has better effect on smoothing noises, but also has shortcoming of lower positioning precision. The Sobel gradient operator also detects image edges using the horizontal and vertical vectors, which is similar with the Prewitt operator. But the Sobel gradient operator did weighted processing on position influences of the pixels, so that it is more accurate in detection. This study used the Sobel gradient operator to detect image edges, and the calculation method is shown in formula (6).

$$E_{i,j}^{'} = \max\left(\frac{1}{4} \sum_{y=j-1}^{j+1} \gamma_y \left| I_{i-1}^{t} - I_{i+1}^{t} \right|, \frac{1}{4} \sum_{x=i-1}^{i+1} \gamma_x \left| I_{x,j+1}^{t} - I_{x,j-1}^{t} \right| \right) \quad (6)$$

Through the above calculation, and after image binaryzation, we could get the entire binary image of the whole moving object area. After this, the edges of the shadows could be get through method below: A window of 5×5 moves on the binary image, and if the central pixel grey value reached 255 in the window, then count convolution of the window subgraph with 4 sensitive one dimensional Laplasse operator. The maximum value of the 4 convolution absolute value is regarded as the basis of judging whether the central pixel is the moving object edge pixel. See in formula (7):

$$EC_{i,j}^{t} = \max\left\{ \left| BE_{i,j}^{t} \times K_p \right| : p = 1,2,3,4 \right\} \quad (7)$$

K_p is the p[th] convolution operator in the formula.

Through the above process, shadow areas were subtracted from the motion edge depicted binary image, influences of shadows on moving objects recognition could be eliminated, and the texture features of the moving objects are entirely conserved. This is the basic of realizing moving objects recognition and tracing.

2.3.2 Recognition method for moving objects

The study testing targets for this study is the monitoring cameras set on a school gate area. In this circumstance, most popular moving objects are vehicles and pedestrians[5]. Usually, the target recognition work through recognizing shapes, colors and textures features of the targets. But in this complicated circumstance, high misjudging rate will happen if this recognition method were used. So our scheme used targets recognition method based on morphology.

The first step is to set up a three dimensional feature library for vehicles and pedestrians. In feature depicting for vehicles and pedestrians, their length, width and shape features could be shown by morphological parameters. Because morphological features could reflect quantitative difference of shapes of vehicles and pedestrians, it is effective for recognition of the targets. According to study needs, our monitoring system selected morphological features include the following:

- Size ratio: Deal with the targets as a rectangular and get their ratio of length, width and height.
- Rectangle filling degree: Each object has its own shape filling degree. We could recognize the shape of the targets according to this.
- Projection rate: It is ratio of A size to convex polygon area size, and this feature could depict the irregular of the object edges.
- Eccentricity: This is a parameter for detecting whether the target is concentrated or not, and through calculating the eccentricity, compact degrees of the targets could be depicted.

To solve problems of highly recognition time complexity caused by vast number of data in feature database, fast linear classifier for self adaption feature selection is used to realize targets fast recognition in this study. The processing procedure of the fast linear classifier is that, select features with big differences to start initial classification for the targets, for example, large cars, small cars, pedestrians and other, and then self adaption select features fit for further classification and finish targets recognition (See in Fig 1). In occasion of too many moving targets causes high presser for the monitoring system, the system could lower its classification level to insure system recognition speed.

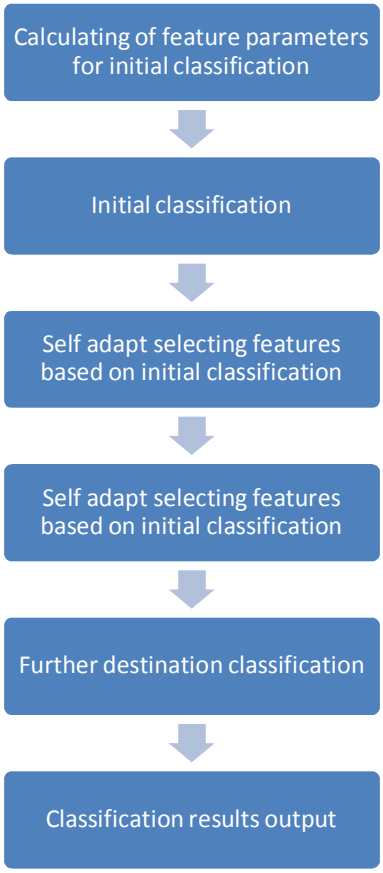


Fig 1. Target recognition chart

2.3.3 Tracing for the moving objects

Method for tracing moving objects relies on calculating matching degrees of moving targets between former and latter frames. In this study, moving objects in adjacent two frames are matched according to object position, size, average color grey value etc., to realize consistency labeling for the same targets. And the action path for the objects are recorded according to the matched position. The method is shown in formula (8):

$$\begin{aligned} C_{a,b} &= \alpha DIS_{a,b} + \beta AVE_{a,b} + \gamma AREA_{a,b} \\ DIS_{a,b} &= \sqrt{\left(\frac{X_{c,a} - X_{c,b}}{R_x}\right)^2 + \left(\frac{Y_{c,a} - Y_{c,b}}{R_y}\right)^2} \\ AVE_{a,b} &= \left[\begin{array}{l} \left|M_{R,a} - M_{R,b}\right| + \left|M_{G,a} - M_{G,b}\right| \\ + \left|M_{B,a} - M_{B,b}\right| \end{array}\right] / 256 \\ AREA_{a,b} &= \left|(S_a - S_b) / S_a\right| \end{aligned} \quad (8)$$

In the formula, (R_x, R_y) symbols image resolution, α, β, γ show target position (X_c, Y_c), average color value (MR, MG, MB), and weight parameters for relative changing of target size S in function calculation. And $\alpha+\beta+\gamma=1$. The system used feature relative variation to increase adaption of target matching.

3. Result and analysis

In the experiment, IK-HD1 type 3CCD camera produced by Toshiba company is used as image collection facility. This camera owns following features: 1) Output pixel: 1920×1080; 2) Output port: digital HD-SDI (SMPTE 292M), DVI output; 3) Manual / automatic mode white balance settings are available. Digital collection facility is MV9300HD video capture card from WOSHI company. Its mainly parameters are : Output quality: 10 bit; Compress mode: H.264; 8 video collection channels and 4 voice collection channels. is Precision T7500 type graphics workstation from Dell company is used to be data collection platform. This workstation has 4 channels memory system and NVIDIA Quadro FX3800 display chip, which has relatively high speed for graphics processing.

In software development, Microsoft Visual C++ (VC++) is used and Matlab numerical calculation software is used in programming of formulas in this study, these programs are eventually compiled into VC++ procedure for call. There are 3 compiling methods used in this study: 1) Use keil compiler MCC of Matlab; 2) Use Matcom compiler; 3) Use COM Build tool of Matlab. Among these methods, method 1 is the simplest, but could not call powerful Matlab image toolbox. Method 2 is high efficient, but imperfect in supporting graphics and image functions. Method 3 is fast in program working, could be used out from the Matlab environment, and supports almost all Matlab functions. It is also perfect in supporting graphics functions, and is a MathWorks company recommended Matlab mixed programming method. Therefore, The 3rd method was used in this study to develop image processing program.

3.1 Establishment of background image

The calculation results from the multi-frame subtraction method indicate that, under the same frame sampling frequency, background quality is positively related to frame sampling time. Let t be sampling time, and n be sampling frames, the background collection effect is shown in Fig 2. When t equals to 2 sec, the background image C came from image A and B showed large blanks; and when t equals to 6 sec, the background image F came from D and E could basically fulfill background establishment requirements.

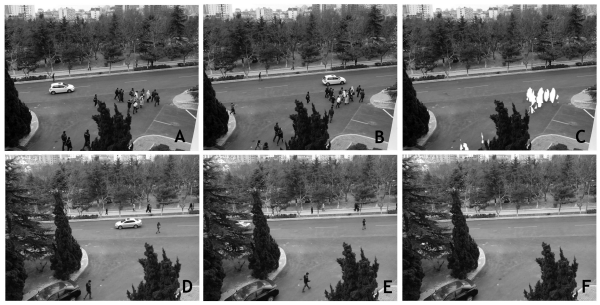


Fig 2. Initial background established under different t and n values (In the figure, C is the background established from A and B, t= 2; F is the background established from D and E, t=6)

In general transportation and traffic condition, when sampling rate is 5 per second, the bigger t value is, the more clear the background will be. When t equals to12, the background is imperfect, and when t equals to 36, the background is clear, and when t equals to 72, the background definition shows almost no difference comparing with t equals to 36. After many experiments, the results indicates that, in general transportation and traffic condition, a background clear enough for monitoring system could be established when t value reach 36. But in conditions of too many pedestrians or traffic jam, the t value needs to become bigger to establish a better background for the monitoring system to recognize and track targets effectively.

3.2 Targets recognition and tracing

In target edges detection aspects, in order to decrease pressers for calculation system, firstly the collected RGB images are converted into grayscale images, and then Sobel gradient operator is used. The target edges are detected through Matlab programming. Then the images are processed with edge function, and the input parameter of the function is the two dimensional matrix, indicator string and some numerical parameters with restricted method after imread. The edge detection procedures used Matlab are as follows:

```
f=imread('1.jpg');
f=rgb2gray(f);%  convert to Gray scale map
f=im2double(f);% The funtion im2double, the value is normalized to 0 ～ 1
%  Use vertical Sobcl operator, and select threshold grates automatically
[VSFAT Threshold]=edge(f, 'sobel','vertical'); % Edge detection
figure,imshow(f),title(' initial image,  ');% Show the initial image
figure,imshow(VSFAT),title( ' Vertical image edge detection ');
% Show the edge detection image
% Using the horizontal and vertical Sobel operator to select threshold automatically.
SFST=edge(f,'sobel',Threshold);
figure,imshow(SFST),title(' Horizontal and vertical image edge detection ');
% Show the edge detection image
% Use specified angles of 45 degrees Sobel operator filter to specified the threshold
s45=[-2 -1 0;-1 0 1;0 1 2];
SFST45=imfilter(f,s45,'replicate');% Function : to filter arrays of arbitrary types or multidimensional image
SFST45=SFST45>=Threshold;
figure,imshow(SFST45),title(' angles of 45 degrees edge detection') ;
% Show the edge detection image
```

The background elimination effect for moving object after edge detection is shown in fig 3.

Fig 3. The background elimination effect for moving object after edge detection

During target recognition and tracing, the system frame rate is 15/ second, and image resolution is 1920×1080 pixels. In the actual experiment, videos stored with local hard disk were used to test the function of recognition, and the experiment results were calculated artificial counting and statistical classification results comparison. The results got from testing 1200 second images recognition is shown in table 1.

Table 1: Results of recognition rate experiments

Name of moving object	*Recognition rate (%)*	*Average recognition time (ms)*
Pedestrians	91.6	27.11
Small cars	87.1	29.52
Big buses	93.9	25.70
Small buses	92.8	25.87
Trucks	89.8	26.64
Bicycles	90.5	27.46

The experiment shows that this system could finish targets recognition under frame rate of 15 / second, and recognition rates are higher than 87%. It also could finish target tracing according to front and rear frames matching

degree, but the recognition of small cars are relatively lower, because that the rectangular degree of the cars are usually difficult to control, and the system often mistake them with small buses. In future studies, this system could be further developed by improving the target morphological database.

4. Conclusion

This study designed a set of moving objects recognition and tracing methods according to unique requests of monitoring system. And researchers also developed an effective scheme in background establishment and moving objects shadows elimination regarding to the actual conditions. This study proved through experiments that this scheme could fulfill video monitoring system needs for moving objects recognition and tracing. And the recognition rate and tracing speed both reached the design standards.

References

[1] Weihua Liu. An image restoration algorithm based on image fusion, International Review on Computers and Software, 2012,Vol.7, n.3,pp. 1245-1249.
[2] Abo-Eleneen Z.A., Abdel-Azim Gamil. An improved image segmentation algorithm based on MET method, International Journal of Computer Science Issues, Vol.9 n.5-3, 2012, pp.346-351.
[3] Jun Sun, Yan Wang, Xiaohong Wu.A New Image Segmentation Algorithm and Its Application in Lettuce Object Segmentation, TELKOMNIKA, Vol.10, n.3,2012,pp. 227-563.
[4] Kondapalli Varaprasad S., Chiranjeevi Manike, Mishra Kundan Kumar, Tanuja Kdbs. Image processing and analysis for DTMRI, International Journal of Computer Science Issues, Vol.9 No. 1-1,2012, pp.266-272.
[5] Ren Mingwu, Yang Jingyu, Sun Han. Tracing boundary contours in a binary image, Image and Vision Computing, Vol.20 No.2,2002,pp:125-131.
[6] Kang Lie, Zhong Sheng, Wang Fang. A new contour tracing method in a binary image, 2011 International Conference on Multimedia Technology, ICMT2011,July 26, 2011 - July 28, 2011,Hangzhou,China,2011,pp.6183-6186.

Peilong XU, born in 1977. The author has achieved Master degree from Tongji University of Shanghai in 2007. He is currently an engineer of computer science in the State Key Laboratory, Qingdao University. The author's research interests include software engineering, image processing, and spectral analysis. Recently, he have published paper named Design and Implementation of Landscape System for East and West Huashi Street in Beijing Based on Virtual Reality Technology on journal Applied Mechanics and Materials（Trans Tech Publications Inc, Switzerland）etc.

The steady-state solution analysis for the degenerate nonlocal parabolic equation

Miaochao Chen[1], Peilong Xu[2] and Dexin Dong[3]

[1] Department of Mathematics, Chaohu College
Bantang Road, Chaohu, 238000, P. R. China

[2] The Growing Base for State Key Laboratory, Qingdao University
No. 308, Ningxia Road, Qingdao 266071, P. R. China

[3] Guangxi Beibu Gulf Marine Research Center, Guangxi Academy of Sciences
Nanning, 530007, P. R. China

Abstract

In this paper, we investigate the steady-state solution for the degenerate nonlocal parabolic equation.we prove that the equation corresponds a unique steady-state solution under certain conditions.

Keywords: *Parabolic Equation, The Steady-State Solution, Ohmic Heating, Nonlocal Parabolic Equation.*

1. Introduction

In this short paper,we investigate the steady-state solution for the following parabolic equation with nonlocal and degenerate source, i.e.,

$$u_t - \nabla\cdot(u^3\nabla u) = \frac{\lambda\exp(-u^4)}{(\int_\Omega \exp(-u^4)dx)^2},$$

(1)

where $x\in\Omega\subset R^2$ and $t>0$.

With the homogeneous Dirichlet boundary conditions as

$$u(x,t)=0, \qquad x\in\partial\Omega, t>0$$

(2)

and

$$u(x,0)=u_0(x)>0. \qquad x\in\Omega$$

(3)

where $\lambda>0$ and $\Omega=\{x\in R^2 : 0<\rho<|x|<R\}$.

In the past several decades, many physical phenomena have been formulated into nonlocal mathematical models.Let us mention,for instance,Lacey [1,2] has obtained the nonlocal parabolic equations

$$\begin{cases} u_t - \nabla u = \dfrac{\lambda f(u)}{(\int_\Omega f(u)dx)^2}, x\in\Omega, t>0, \\ u=0, x\in\partial\Omega, t>0, \\ u(x,0)=u_0(x), x\in\Omega \end{cases}$$

(4)

Where u is the temperature of the heated object.

Eq.(4), as a kind of Ohmic heating model,which comes from the more general parabolic-elliptic equations

$$\begin{cases} u_t - \nabla\cdot(\kappa(u)\nabla u) = \sigma(u)|\nabla\phi|^2, x\in\Omega, t>0, \\ \nabla\cdot(\sigma(u)\nabla\phi)=0, x\in\Omega, t>0. \end{cases}$$

(5)

Where ϕ is the voltage at the ends of the conductor.

These two equations were studied in [1,2,3] and [4-9] respectively.

Investigation on Eq.(1-3) Mainly include three problems: the existence and uniqueness of the steady-state solution , the rate of blow-up and asymptotic analysis for the equations.

The work of this paper is motivated by the steady-state source problem

$$\nabla\cdot(w^3\nabla w) + \frac{\lambda\exp(-w^4)}{(\int_\Omega \exp(-w^4)dx)^2} = 0,$$

(6)

where $x\in\Omega$, $w=0$ and $x\in\partial\Omega$.

The existence of solution to the problem (6) has closely Relationship with the following problem

$$\nabla\cdot(w^3\nabla w) + \mu\exp(-w^4) = 0,$$

(7)

where $x \in \Omega$, $w = 0$ and $x \in \partial\Omega$.

Here we set $\mu \geq 0$ and $\lambda(\mu) = \mu(\int_{\Omega} \exp(-w^4)dx)^2$.

2. Main Results

The main result of this paper reads as follows:

Theorem 2.1

Assume that $\Omega = \{x \in R^2 : 0 < \rho < |x| < R\}$ and that $\lambda^* = |\partial\Omega|^2 / 2$,we have

(i) If $0 < \lambda < \lambda^*$,the problem (6) corresponds a solution at least .

(ii) If $\lambda \geq \lambda^*$,the problem (6) have no solution.

Theorem 2.2

Assume that $\Omega = \{x \in R^2 : 0 < \rho < |x| < R\}$ and that $\lambda^* = |\partial\Omega|^2 / 2$ .If $0 < \lambda < \lambda^*$,then we have the problem (6) corresponds a unique steady-state solution.

3. Proof of Theorem 2.1 and Theorem 2.2

First of all,we prepare some definitions, notations which will be needed in the proof of our results.

We assume $w(r;\mu)$ is radially symmetric, Let $w(r;\mu)$ be a solution of (7).By the maximum principle,form (7),we have

$$(w^3 w_r)_r + \frac{1}{r} w^3 w_r + \mu \exp(-w^4) = 0, \rho < r < R;$$
$$w(\rho) = w(R) = 0 \quad (8)$$

Which implies

$$-(rw^3 w_r)_r = \mu r \exp(-w^4), \rho < r < R, \quad (9)$$

and

$$-((rw^3 w_r)^2)_r = \mu r^2 w^3 w_r \exp(-w^4), \rho < r < R, \quad (10)$$

From (9),we obtain a unique solution

$$r_0 = r_0(\mu) \in (\rho, R)$$

Such that

$$w(r_0;\mu) = \max_{[\upsilon,R]} w(r;\mu) = M(\mu).$$

Integrating both sides of (9) and (10) over (r, r_0) ,we have,for $\rho < r < R$,

$$\frac{1}{2}(rw^3 w_r)^2 = \frac{1}{4}\mu(r^2 e^{-w^4} - r_0^2 e^{-M^4}) + \frac{1}{2} rw^3 w_r,$$

This equality infers that

$$(\frac{1}{2} - rw^3 w_r)^2 = \frac{1}{4} + \frac{1}{2}\mu(r^2 e^{-w^4} - r_0^2 e^{-M^4}), \quad (11)$$

Set $L_1(\mu) = \lim_{r \to \rho+} rw^3 w_r$ and $L_2(\mu) = \lim_{r \to R-} rw^3 w_r$.

From (11),we have

$$L_1(\mu) = \begin{cases} \frac{1}{2} - \sqrt{\frac{1}{4} + \frac{1}{2}\mu(\rho^2 - r_0^2 e^{-M^4})}, L_1(\mu) \leq \frac{1}{2}, \\ \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{1}{2}\mu(\rho^2 - r_0^2 e^{-M^4})}, L_1(\mu) > \frac{1}{2}, \end{cases} \quad (12)$$

and

$$L_2(\mu) = \frac{1}{2} - \sqrt{\frac{1}{4} + \frac{1}{2}\mu(R^2 - r_0^2 e^{-M^4})} \quad (13)$$

By the definition of $\lambda(\mu)$,it holds that

$$\lambda(\mu) = \frac{4\pi^2}{\mu}(L_1(\mu) - L_2(\mu))$$

Set $\Gamma(\mu) = \frac{1}{2\pi}\sqrt{\lambda(\mu)} = \frac{1}{\sqrt{\mu}}(L_1(\mu) - L_2(\mu))$,

(14)

Combining (12) and (13),we have

$$\Gamma(\mu) = \begin{cases} \sqrt{\frac{1}{4\mu} + \frac{1}{2}y_1} - \sqrt{\frac{1}{4\mu} + \frac{1}{2}y_2}, L_1(\mu) \leq \frac{1}{2} \\ \sqrt{\frac{1}{4\mu} + \frac{1}{2}y_1} + \sqrt{\frac{1}{4\mu} + \frac{1}{2}y_2}), L_1(\mu) > \frac{1}{2} \end{cases} \quad (15)$$

where $y_1 = R^2 - r_0^2 e^{-M^4}$ and $y_2 = \rho^2 - r_0^2 e^{-M^4}$.

Through a series of preparations,we derive a fact of $\Gamma(\mu)$.

Lemma 1

(i) If $L_1(\mu) \leq \frac{1}{2}$, hence $\Gamma(\mu) < (R + \rho)/\sqrt{2}$.

(ii) If $L_1(\mu) > \frac{1}{2}$,

hence $\Gamma(\mu) < (R + \rho)/\sqrt{2} \Leftrightarrow \mu r_0^2 e^{-M^4} > \frac{1}{2}$,

and $\Gamma(\mu) = (R + \rho)/\sqrt{2} \Leftrightarrow \mu r_0^2 e^{-M^4} = \frac{1}{2}$.

Through a series of calculation yields, we can prove the lemma 1.Here we omit the proof of lemma 1 because of the length of the article.

Proof of Theorem 2.1

Proof.Set $y=\frac{1}{2}-rw^3w_r$. Combining (9) and (11),we have

$$\frac{1}{2}r\frac{dy}{dr}=y^2+\frac{1}{2}\mu r_0^2e^{-M^4}-\frac{1}{4}.$$

(16)

In the case of $\mu=\mu_1$,we then obtain

$$\frac{1}{2}r\frac{dy}{dr}=y^2$$

(17)

Now according to (9),(13) and Lemma 1,we see that there exists $r_1>\rho$,such that

$$r_1w^3(r_1;\mu_1)w_r(r_1;\mu_1)=\frac{1}{2}.$$

Integrating both sides of (17) over (r,r_0), we have,for $r_1<r<r_0$,

$$\frac{1}{1/2-rw^3w_r}=2+2(\ln r_0-\ln r),$$

This is a contradiction of the equation for $r\to r_1$ and we then complete the Proof of Theorem 2.1.

In order to prove Theorem 2.2, We need to derive a fact of the following two problems.

Lemma2

Denote $\mu r_0^2(\mu)e^{-M^4(\mu)}=1/2 \quad , \quad \mu>0,$

(18)

We then have a unique solution

$$\mu_I=\frac{(R-\rho)^2}{2\rho^2R^2(\ln R-\ln\rho)^2}$$

(19)

Proof.From (12),we have $\lim_{\mu\to\infty}L_1(\mu)=\infty$. Now according to theorem 2.1 and Lemma 1(ii),we obtain $\mu r_0^2(\mu)e^{-M^4(\mu)}>1/2$ and $L_1(\mu)<1/2.$ which implies that there exists μ_I satisfies (18), Integrating both sides of (17) over (ρ,r_0) and (r_0,R) respectively,we have

$$\frac{1}{1/2-L_1(\mu_I)}-2=2(\ln r_0-\ln\rho),$$

(20)

and $$2-\frac{1}{1/2-L_2(\mu_I)}=2(\ln R-\ln r_0).$$

(21)

Using (12) and (13),we infer that

$$\begin{cases}1/2-L_1(\mu_I)=\sqrt{\frac{1}{2}\mu_I\rho},\\ 1/2-L_2(\mu_I)=\sqrt{\frac{1}{2}\mu_IR}.\end{cases}$$

(22)

Combining (20) to (22),we obtain a unique solution

$$\mu_I=\frac{(R-\rho)^2}{2\rho^2R^2(\ln R-\ln\rho)^2}.$$

Lemma3

Denote $$L_1(\mu)=\frac{1}{2},\mu>0,$$

(23)

We then have a unique solution

$$\mu_{II}=\frac{(\arctan\frac{\sqrt{R^2-\rho^2}}{\rho})^2}{2\rho^2(\ln R-\ln\rho)^2}.$$

(24)

Proof.Similar to the proof of Lemma 2,we have

$$\frac{1}{2}r\frac{dy}{dr}=y^2+\frac{1}{2}\mu_{II}\rho^2.$$

(25)

Integrating both sides of (17) over (ρ,r_0) and (r_0,R) respectively,we have

$$\frac{1}{\sqrt{2\mu_{II}\rho^2}}\arctan\frac{1}{\sqrt{2\mu_{II}\rho^2}}=\ln r_0-\ln\rho \quad (26)$$

and $$\frac{1}{\sqrt{2\mu_{II}\rho^2}}\arctan\frac{1-2L_2(\mu_{II})}{\sqrt{2\mu_{II}\rho^2}}+\frac{1}{\sqrt{2\mu_{II}\rho^2}}\arctan\frac{1}{\sqrt{2\mu_{II}\rho^2}}=\ln R-\ln r_0 \quad .$$

(27)

From (13),we have

$$1/2-L_2(\mu_{II})=\sqrt{\frac{1}{2}\mu_{II}(R^2-\rho^2)}.$$

(28)

Combining (26) to (28),we obtain a unique solution

$$\mu_{II} = \frac{(\arctan\frac{\sqrt{R^2-\rho^2}}{\rho})^2}{2\rho^2(\ln R - \ln\rho)^2}.$$

Proof of Theorem 2.2

Proof.Set $$G(\mu) = \frac{1}{4\mu} - \frac{1}{2}r_0^2 e^{-M^4}$$

(29)

in view of(16),we observe

$$\frac{1}{2}r\frac{dy}{dr} = y^2 - \mu G(\mu)$$

(30)

We have three steps to prove Theorem 2.2.

Step 1 If $0 < \mu < \mu_I$,Using lemma 2,it holds that

$$\mu r_0^2(\mu)e^{-M^4(\mu)} < 1/2,$$

which implies $G(\mu) > 0$. Using lemma 1 and Theorem 2.1,we obtain $L_1(\mu) < 1/2$.

Integrating both sides of (30) over (ρ, r_0) and (r_0, R) respectively,we have

$$\frac{1}{\sqrt{\mu G(\mu)}}(\ln\frac{1-2\sqrt{\mu G(\mu)}}{1+2\sqrt{\mu G(\mu)}} - \ln\frac{1-2L_1(\mu)-2\sqrt{\mu G(\mu)}}{1-2L_1(\mu)+2\sqrt{\mu G(\mu)}}) = 4(\ln r_0 - \ln\rho)$$

(31)

and

$$\frac{1}{\sqrt{\mu G(\mu)}}(\ln\frac{1-2L_2(\mu)-2\sqrt{\mu G(\mu)}}{1-2L_2(\mu)+2\sqrt{\mu G(\mu)}} - \ln\frac{1-2\sqrt{\mu G(\mu)}}{1+2\sqrt{\mu G(\mu)}}) = 4(\ln R - \ln r_0)$$

(32)

From (12) and (13),we obtain

$$\begin{cases} 1/2 - L_1(\mu_I) = \sqrt{\frac{1}{2}\mu\rho^2 + \mu G(\mu)}, \\ 1/2 - L_2(\mu_I) = \sqrt{\frac{1}{2}\mu\rho^2 + \mu G(\mu)}. \end{cases}$$

(33)

Combining (31) to (33),we then have

$$\frac{1}{\sqrt{G(\mu)}}(\ln\frac{\sqrt{R^2/2+G(\mu)}-\sqrt{G(\mu)}}{\sqrt{R^2/2+G(\mu)}+\sqrt{G(\mu)}} - \ln\frac{\sqrt{\rho^2/2+G(\mu)}-\sqrt{G(\mu)}}{\sqrt{\rho^2/2+G(\mu)}+\sqrt{G(\mu)}}) = 4\sqrt{\mu}(\ln R - \ln r_0),$$

which implies $G'(\mu) \neq 0$.

According to the definition of $\Gamma(\mu)$,we have

$$\Gamma(\mu) = \sqrt{R^2/2+G(\mu)} - \sqrt{\rho^2/2+G(\mu)}.$$

(34)

Hence, we have $\Gamma'(\mu) > 0$ in the case of $0 < \mu < \mu_I$.

Step 2 If $\mu_I < \mu < \mu_{II}$, Using lemma 2 and lemma 3 , it holds that

$$\mu r_0^2(\mu)e^{-M^4(\mu)} > \frac{1}{2} \text{ and } L_1(\mu) < 1/2,$$

which implies $G(\mu) < 0$.

Integrating both sides of (30) over (ρ, r_0) and (r_0, R) respectively,we have

$$\frac{1}{\sqrt{-\mu G(\mu)}}(\arctan\frac{1}{2\sqrt{-\mu G(\mu)}} - \arctan\frac{1-2L_1(\mu)}{2\sqrt{-\mu G(\mu)}}) = 2(\ln r_0 - \ln\rho)$$

(35)

and

$$\frac{1}{\sqrt{-\mu G(\mu)}}(\arctan\frac{1-2L_2(\mu)}{2\sqrt{-\mu G(\mu)}} - \arctan\frac{1}{2\sqrt{-\mu G(\mu)}}) = 2(\ln R - \ln r_0)$$

(36)

Combining (33)to(36),we obtain

$$\frac{1}{\sqrt{-G(\mu)}}(\arctan\frac{\sqrt{R^2/2+G(\mu)}}{\sqrt{-G(\mu)}} - \arctan\frac{\sqrt{\rho^2/2+G(\mu)}}{\sqrt{-G(\mu)}}) = 2\sqrt{\mu}(\ln R - \ln r_0)$$

Which implies $G'(\mu) \neq 0,\ \mu_I < \mu < \mu_{II}$,Thus

$$\Gamma'(\mu) = G'(\mu)(\frac{1}{2\sqrt{R^2/2+G(\mu)}} - \frac{1}{2\sqrt{\rho^2/2+G(\mu)}}) > 0.$$

Step 3 If $\mu > \mu_{II}$, Using lemma 2 and lemma 3 , it holds that

$$\mu r_0^2(\mu)e^{-M^4(\mu)} > \frac{1}{2} \text{ and } L_1(\mu) < 1/2,$$

which implies $G(\mu) < 0$. Integrating both sides of (30) over (ρ, r_0) and (r_0, R) respectively,we also obtain (35) and (36). Combining (12)to(13),we obtain

$$\begin{cases} 1/2 - L_1(\mu_I) = -\sqrt{\frac{1}{2}\mu\rho^2 + \mu G(\mu)}, \\ 1/2 - L_2(\mu_I) = \sqrt{\frac{1}{2}\mu\rho^2 + \mu G(\mu)}. \end{cases}$$

(37)

Combining (35)to(37),we obtain

$$\frac{1}{\sqrt{-G(\mu)}}(\arctan\frac{\sqrt{R^2/2 + G(\mu)}}{\sqrt{-G(\mu)}} + \arctan\frac{\sqrt{\rho^2/2 + G(\mu)}}{\sqrt{-G(\mu)}}) = 2\sqrt{\mu}(\ln R - \ln\rho),$$

which implies $G'(\mu) \neq 0$.

According to the definition of $\Gamma(\mu)$,we have

$$\Gamma(\mu) = \sqrt{R^2/2 + G(\mu)} + \sqrt{\rho^2/2 + G(\mu)}.$$

Hence, we have $\Gamma'(\mu) > 0$ in the case of $\mu > \mu_{II}$

We then complete the proof of Theorem 2.2.

4. Conclusions

In this paper,we consider the degenerate nonlocal parabolic equation

$$u_t - \nabla \cdot (u^3 \nabla u) = \frac{\lambda \exp(-u^4)}{(\int_\Omega \exp(-u^4)dx)^2},$$

with homogeneous Dirichlet boundary condition, where $\lambda > 0$, $\Omega = \{x \in R^2 : 0 < \rho < |x| < R\}$.

We prove that in the case of $0 < \lambda < |\partial\Omega|^2/2$,the equation corresponds a unique steady-state solution.

References

[1] A.A.Lacey,Thermal runaway in a non-local problem modelling Ohmic heating.I.Model derivation and some special cases,European J.Appl.Math,Vol.6, 1995, pp. 127-144.

[2] A.A.Lacey,Thermal runaway in a non-local problem modelling Ohmic heating.II.General proof of blow-up and asymptotics of runaway, European J.Appl.Math, Vol.6, 1995, pp. 201-224.

[3] J.W.Beberbes,A.A.Lacey,Global existence and finite-time blow-up for a class of nonlocal parabolic problems, Adv. Differential Differential Equations, Vol.2,1997, pp. 927-953.

[4] S.N.Antontsev,M.Chipot,The Analysis of blow-up for the thermistor problem,Sb.Math.J,Vol.38,1997, pp. 827-841.

[5] A.Barabanova,The blow-up of solutions of a non-local thermistor problem, Appl.Math.Lett, Vol.9,1996, pp. 59-63.

[6] W.Allegretto,H.Xie,A non-local thermistor problem Eur. J.Appl. Math, Vol.6,1995, pp. 83-94.

[7] D.E.Tzanetis, Blow-up of radially symmetric solutions of a nonlocal problem modelling Ohmic heating,Electron.J.Diff. Eqns, Vol. 11,2002, pp. 1-26.

[8] N.I.Kavallaris, D.E.Tzanetis,On the blow-up of a non-local parabolic problem,Appl.Math.lett, Vol. 19,2006, pp. 921-925.

[9] N.I.Kavallaris, D.E.Tzanetis,On the blow-up the nonlocal thermistor problem, Proc. Edinb.Math.Soc, Vol. 50,2007, pp. 389-409.

[10] Basma Zahra, Anis Sakly and Mohamed Benrejeb. Stability Study of Fuzzy Control Processes Application to a Nonlinear Second Order System, International Journal of Computer Science Issues, Vol. 9, No.2-2,(2012) pp. 97-106.

Corresponding author：Miaochao Chen,born in 1981. The author has achieved bachelor's degree from Capital Normal University of Beijing in 2004.Then,He has received the Master degree in Mathematics and applied mathematics from Southeast University of Nanjing in 2012. Currently, he works in the department of Mathematics at chaohu college. He has published three papers.The papers respectively were published in Journal of Chaohu college(Vol.12,2009,pp.28-30.),Proceedings of the 2011 3rd International Conference on Computer Technology and Development(Vol.1, 2011,pp.55-59, ASME PRESS.) and Journal of Daqing Normal University(.(Vol.32, 2012,pp.56-59.).His research interests include Partial differential equations and Functional analysis.

Peilong Xu,born in 1977.The author has achieved Master degree from Tongji University of Shanghai in 2007. He is currently an engineer of computer science in the State Key Laboratory, Qingdao University. His research interests include software engineering, image processing, and spectral analysis.

Dexin Dong,born in 1980. The author research interests include Partial differential equations and oceanographic physics.

Energy-Aware Scheme used in Multi-level Heterogeneous Wireless Sensor Networks

Mostafa SAADI*§, Moulay Lahcen HASNAOUI‡, Abderrahim BENI HSSANE§,
Said BENKIRANE§, Mohamed LAGHDIR§

§MATIC Laboratory, Mathematics and Computer Science Department, Faculty of Sciences,
Chouab Doukkali University, El Jadida, Morocco.

‡Computer Science Department, Faculty of Sciences Dhar el Mahraz,
Sidi Mohammed Ben Abdellah University, Fez, Marocco.

saadi_mo@yahoo.fr, mlhnet2002@yahoo.ca, abenihssane@yahoo.fr, sabenk1@hotmail.com, laghdirm@yahoo.fr

*Corresponding Author

***Abstract*—The wireless sensor networks (WSNs) is a power constrained system, since nodes run on limited power batteries which shorten its lifespan.The main challenge facing us in the design and conception of Wireless Sensor Networks (WSNs) is to find the best way to extend their life span. The clustering algorithm is a key technique used to increase the scalability and life span of the network in general. In this paper, we propose and evaluate a distributed energy-efficient clustering algorithm for WSNs. This heterogeneous-energy protocol is a new clustering algorithm to decrease probability of failure nodes and in which we introduce the node's remaining energy so as to determine the cluster heads. We study the impact of heterogeneity of nodes on WSNs that are hierarchically clustered. Finally, simulation results show that the proposed algorithm increases the life span of the whole network and performs better than LEACH and EEHC according to the metric:first node dies.**

***Keywords*—Wireless Sensor Networks; Clustering Algorithm; Heterogeneous Environment; Energy-Efficient**

I. Introduction

Continued enhancement of Micro-Electro-Mechanical Systems (MEMS) and wireless communication technologies have enabled the deployment of large scale wireless sensor networks (WSNs). It comprises a big number of sensor nodes deployed in ad hoc manner in an unreachable field to give the end-user the ability to instrument, observe, and react to events and phenomena in a specified environment. WSNs provide unforeseen applications: ranging from military applications such as battlefield mapping and target surveillance, to creating context-aware homes; the number of applications is endless [1], [2], [3].

Since they are exposed to atrocious and dynamic environments and limited in their energy level, processing power and sensing ability, WSNs must deliver only processed and concise data. Therefore, any inefficient use of these WSNs leads to a poor performance and consequently a short life cycle. Routing techniques are the most important issue for networks where resources are limited.[4], [5], [6]

In most of the applications, sensors are supposed to spot the events and then send the collected data to the Base Station (BS) where parameters characterizing these events are evaluated. Since the cost of forwarding data is higher than computation, [1], [2], [5], clustering sensors into groups so as to communicate information only to cluster heads which communicate information to the processing center (BS), is a kind of key technique used to reduce energy consumption and then increase the life span of the network [6], [7].

In this respect, there are two types of schemes that operate differently. The conventional centralized algorithms operate with a global knowledge of the whole network and any error in transmission or a failure of a critical node will potentially bring about a serious protocol failure; whereas the distributed algorithms are executed locally with partial nodes, which can prevent any failure caused by a single node [6], [7], [8], [9], [10].

In this paper, we propose a new energy-efficient cluster head selection algorithm to reduce energy consumption dubbed EASM. This heterogeneous-energy protocol decreases the probability of failure nodes and in which we introduce the node's remaining energy so as to determine the future cluster heads.

The operation of this algorithm is divided into rounds. Each of these rounds consists of a set-up and a steady-state phase. During the set-up phase cluster-heads are determined and the clusters are organized. During the steady-state phase data transfers to the base station occur. This protocol is proposed to increase the whole network life span on a heterogeneous network with a BS located far away from

the sensor area.

The remainder of this paper is organized as follows. Section **II** presents the related work and describes the heterogeneous WSN model. Section **III** exhibits the details and analyzes the properties of the newest one. Section **IV** evaluates the performance of our protocol by simulations and compares it with other existing protocols. Finally, Section **V** gives concluding remarks.

II. Problem out line

A. Related Work

In most WSN applications the power supply is limited, so preserving the consumed energy of the network is a challenge that must be considered when developing a routing protocol for WSNs.

A comprehensive survey of the routing protocols for WSNs can be found in [11]. In general, these protocols can be categorized into two classes according to the node's participating style: flat protocols and clustering protocols. Those in [12], [13], [14], [15] belong to the first class.The second class can be also categorized into two subclasses: the clustering algorithms applied in homogeneous networks are called homogeneous schemes, where all nodes have the same initial energy and the clustering algorithms applied in heterogeneous networks are referred to as heterogeneous clustering schemes, where all the nodes of the sensor network are equipped with different amounts of energy.

Many homogeneous clustering algorithms exist in literature such as LEACH [6], PEGASIS [16], HEED [17] and RE-LEACH [18]. Low-Energy Adaptive Clustering Hierarchy (LEACH), which is one of the most fundamental protocol frameworks in the literature, utilizes randomized rotation of the Cluster-Heads (CHs) to uniformly distribute the energy budget across the network. The sensor nodes are grouped into several clusters and in each cluster, one of the sensor nodes is selected to be CH. Each node will transmit its data to its own CH which forwards the sensed data to the BS finally. Both communications between sensor nodes and CH and that between CHs and the BS are direct, single-hop transmission. Based on the framework of LEACH, several protocols are proposed in the open literature. In [16], a scheme called Power-Efficient GAthering in Sensor Information System (PEGASIS) is proposed. In this system, each node communicates only with a close neighbor and takes turns transmitting to the BS, thus reducing the amount of energy spent per round. In [16], nodes will be organized to form a chain, which can be computed by each node or by the base station. The requirement of global knowledge of the network topology makes this method difficult to implement. In[17], HEED is a distributed clustering algorithm, which selects the cluster-heads stochastically. The election probability of each node is correlative to the residual energy. But in heterogeneous environments, the low-energy nodes could own larger election probability than the high-energy nodes in HEED.

WSNs are more possibly heterogeneous networks than homogeneous ones. Thus, the protocols should be fit for the characteristic of heterogeneous WSNs. Many heterogeneous clustering algorithms exist in literature such as SEP[8], M-LEACH[19] , EECS[20], LEACH-B[21], DEEC[9] and SDEEC[22]. The EECS[20] protocol elects the cluster-heads with more residual energy through local radio communication. In cluster formation phase, EECS considers the tradeoff of energy expenditure between nodes to the cluster-heads and the cluster-heads to the base station. But on the other hand, it increases the requirement of global knowledge about the distances between the cluster-heads and the base station. The EEHC [24] protocol is developed for the 3-level heterogeneous networks, which include three types of nodes according to the initial energy, i.e., the super nodes, the advance nodes and the normal nodes. The rotating epoch and election probability is directly correlated with only the initial energy of nodes. EEHC performs poorly when heterogeneity is a result of operation of the sensor network.

In this paper, we also focus on the design of power efficient network layer solutions. Our work is inspired by the previous approaches, but it differs by designing the protocol with the integration of the cross-layer design principle which is proven to be a pertinent method to meet the challenges of power-constrained WSNs. The EAMS protocol assigns different epoch of being a cluster-head to each node according to the initial and residual energy. A novel clustering-based routing protocol proposed in this paper improve the effective life span of the WSNs with a limited energy supply.

EASM is an energy-aware scheme clustering used in heterogeneous wireless sensor networks. In witch, every sensor node independently elects itself as a cluster-head based on its initial energy and residual energy. To control the energy expenditure of nodes by means of adaptive approach, our algorithm use the orientation to BS to transmit the sensing data, and doesn't require any global knowledge of energy at every election round.

B. Heterogeneous WSN model

In this study, we describe the network model. Assume that there are N sensor nodes, which are uniformly dispersed within an M $\times$ M square region (Fig. 1).

The nodes always have data to transmit to a base station, which is often far from the sensing area. The network is organized into a clustering hierarchy, and the cluster-heads execute data aggregation to reduce redundant data produced by the sensor nodes within the clusters.

We consider the heterogeneous networks with nodes heterogeneous in their initial amount of energy. We assume there are three types of sensor nodes, i.e.,the super nodes, the advanced nodes and the normal nodes[24]. Note E_0 the initial energy of the normal nodes, and m the fraction of

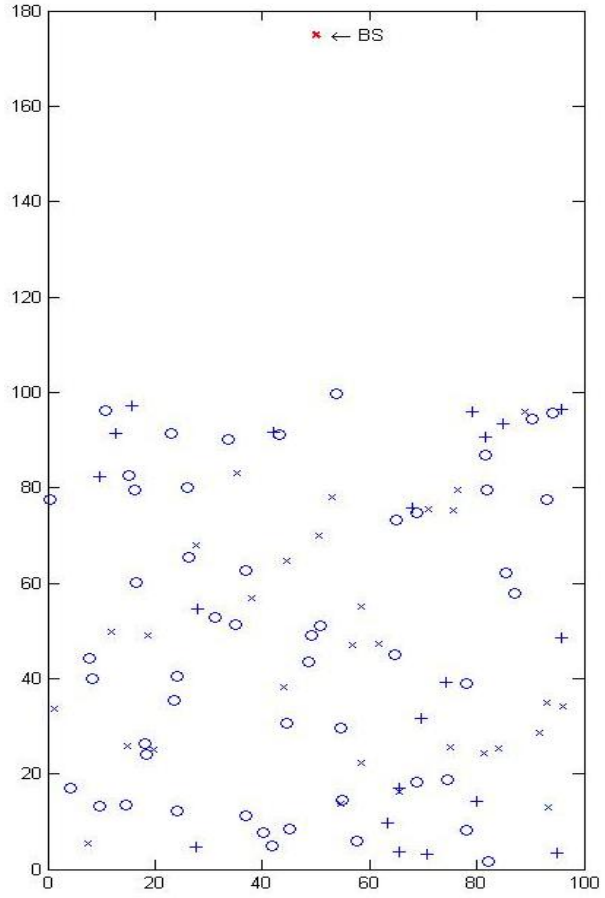


Fig. 1. 100 nodes randomly deployed in the network
o: normal node; x: advanced node; +: super node

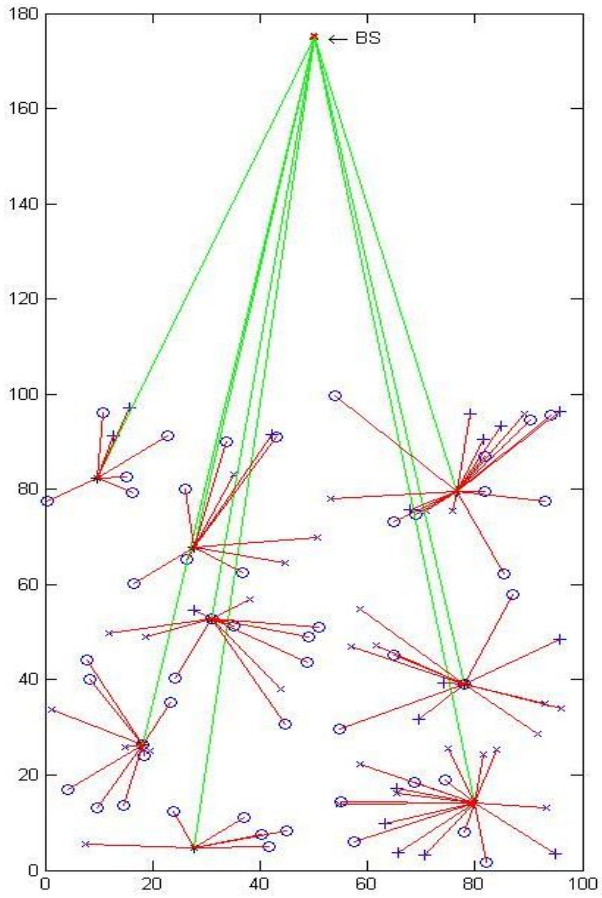


Fig. 2. Dynamic cluster structure by EASM algorithm

the total nodes N, and m_0 is the percentage of the total number of nodes m which are equipped with β times more energy than the normal nodes, we call these nodes as super nodes. The rest $N \times (1-m_0)$ nodes are equipped with α times more energy than the normal nodes, we call these nodes as advanced nodes and the remaining $N \times (1-m)$ as normal nodes.

Thus there are $N \times m \times (1-m_0)$ advanced nodes equipped with initial energy of $E_0 \times (1+\alpha)$, $N \times m \times m_0$ super nodes equipped with initial energy of $E_0 \times (1+\beta)$ and $(1-m) \times N$ normal nodes equipped with initial energy of E_0.

The total initial energy of the three-level heterogeneous networks is given by:

$$\begin{aligned} E_{tot} &= N \times m \times (1-m_0) \times E_0 \times (1+\alpha) \\ &+ N \times m \times m_0 \times E_0 \times (1+\beta) + (1-m) \times N \times E_0 \\ &= N \times E_0 \times (1+m \times (\alpha + m_0 \times (\beta - \alpha))) \end{aligned} \quad (1)$$

The cluster-heads (Fig. 2) transmit the aggregated data to the BS directly. We assume that the nodes are stationary as supposed in [7]. More interestingly, a similar energy model as proposed in [7]is used in this study. According to the radio energy dissipation model illustrated in (Fig. 3), and in order to achieve an acceptable Signal-to-Noise Ratio (SNR) in transmitting an L-bit message over a distance d, the energy expended by the radio is given by :

$$E_{Tx}(l,d) = \begin{cases} lE_{elec} + l\varepsilon_{fs}d^2, d < d_0 \\ lE_{elec} + l\varepsilon_{mp}d^4, d \geq d_0 \end{cases} \quad (2)$$

Where E_{elec} is the energy dissipated per bit to run the transmitter E_{Tx} or the receiver E_{Rx} circuit, and ε_{fs} and ε_{mp} depend on the transmitter amplifier model used and d is the distance between the sender and the receiver.

In most WSN applications the power supply is limited, so preserving the consumed energy of the network is a challenge that must be considered when developing a routing protocol for WSNs. In the next section, we describe the EASM algorithm in details.

III. Explanation of the Proposed Protocol: (EASM)

We assume a network with N nodes uniformly deployed within M× M square region, the network topology remains stagnant over time and the BS location is known. In EASM, a

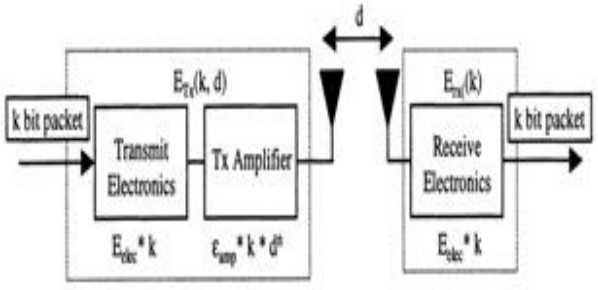


Fig. 3. Radio Energy Dissipation Model

TABLE I
RADIO PARAMETERS USED IN OUR SIMULATIONS

Parameter	Value
E_{elec}	$5nJ/bit$
ε_{fs}	$10pJ/bit/m^2$
ε_{mp}	$0.0013pJ/bit/m^4$
E_0	$0.5J$
E_{DA}	$5nJ/bit/message$
d_0	$70m$
Message size	4000 bits
p_{opt}	0.1

new optimal probability threshold is introduced, where each node i uses to determine whether itself to become a cluster-head in each round r , given as follows:

$$T(s_i) = \begin{cases} \frac{p_i}{1-p_i(rmod\frac{1}{p_i})} \times \frac{E_{residual}(r_i)}{E_{initial}(i)} & , \ if \ s_i \ \varepsilon \ G \\ 0 & , \ otherwise \end{cases} \quad (3)$$

Where $E_{residual}(r_i)$, $E_{initial}(i)$ are the residual and the initial energy respectively. p_i is the cluster-head probability and r_i is the number of consecutive rounds in which a node has not been cluster-head within an epoch.

When r_i reaches the value $1/p_{opt}$ the threshold $T(i)_{opt}$ is reset to the value it had before the inclusion of the remaining energy into the threshold-equation (3).

Also, the probabilities for normal, advanced and super nodes are defined as follow:

$$p_n = \frac{p_{opt}}{(1+m\times(\alpha+m_0\times(\beta-\alpha)))} \quad (4)$$

$$p_a = \frac{p_{opt}}{(1+m\times(\alpha+m_0\times(\beta-\alpha)))}\times(1+\alpha) \quad (5)$$

$$p_s = \frac{p_{opt}}{(1+m\times(\alpha+m_0\times(\beta-\alpha)))}\times(1+\beta) \quad (6)$$

where m is the fraction of the total nodes N, and m_0 is the percentage of the total number of nodes m which are equipped with β times more energy than the normal nodes; The rest nodes are equipped with α times more energy than the normal nodes.

In each round r, when node i finds it is eligible to be a cluster head, it will choose a random number between 0 and 1. If the number is less than threshold $T_{opt}(i)$, the node i becomes a cluster head during the current round.

Each node that has elected itself a cluster-head for the current round broadcasts an advertisement message to the rest of the nodes. For this "cluster-head-advertisement" phase, the cluster-heads use a CSMA MAC protocol, and all cluster-heads transmit their advertisement using the same transmitted energy. The non-cluster-head nodes must keep their receivers on during this phase of set-up to hear the advertisements of all the cluster-head nodes. After this phase is complete, each non-cluster-head node decides the cluster to which it will belong for this round. This decision is based on the received signal strength of the advertisement. Assuming symmetric propagation channels, the cluster-head advertisement heard with the largest signal strength is the cluster-head to whom the minimum amount of transmitted energy is needed for communication. In the case of ties, a random cluster-head is chosen [6].

Each non cluster head node communicates its data during its allocated transmission time (TDMA) to its own cluster head. After that, each non cluster head can turn on the sleep mode. The cluster head node must keep its receiver on in order to receive all the data from the nodes in the cluster.

When all the data is received, the cluster head node performs signal processing functions to compress the data into a single signal. When this phase is completed, each cluster head can send the aggregated data to the BS.

IV. SIMULATION

In this section, we evaluate the performance of EASM protocol. We consider a WSN with N = 100 nodes randomly distributed in a 100m × 100m sensing area. We assume the BS is far away from the sensing region and placed at location$(x = 50; y = 175)$. The nodes in the network are divided in three heterogenous energy levels and are energy-constrained.

To compare the performance of EASM with other protocols, The radio parameters used in our simulations are shown in TABLE I. We assume that all nodes know their location coordinates. We will consider the following scenarios and examine several performance measures.

After deployment of WSN, the nodes consume energy during the course of the WSN life span. In fact, energy is removed whenever a node transmits or receives data and whenever it performs data aggregation. Once a node runs out of energy, it is considered dead and can no longer transmit or receive data.

Firstly, we run simulation for our proposed protocol EASM to detect the round when the first node dies and compare the results to LEACH and EEHC protocols under two kinds of 3-level heterogeneous networks . Figure Fig. 4

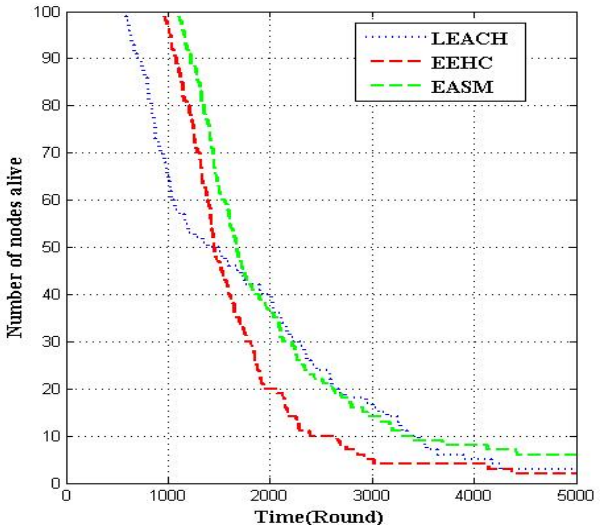


Fig. 4. Number of nodes alive over time.($\alpha = 1.5$,$m = 0.5$,$\beta = 3$ and $m_0 = 0.4$)

shows the results of the case with $\alpha = 1.5$, $m = 0.5$, $\beta = 3$ and $m_0 = 0.4$. It is obvious that the stable time of EASM is prolonged compared to that of LEACH and EEHC.

Second, we run simulation for our proposed protocol EASM to compute the round when the first node dies when $\alpha = 2$,$m = 0.3$,$\beta = 5$ and $m_0 = 0.6$, and compare the results to LEACH and EEHC protocols. Fig.5 shows the number of rounds when the first node dies.

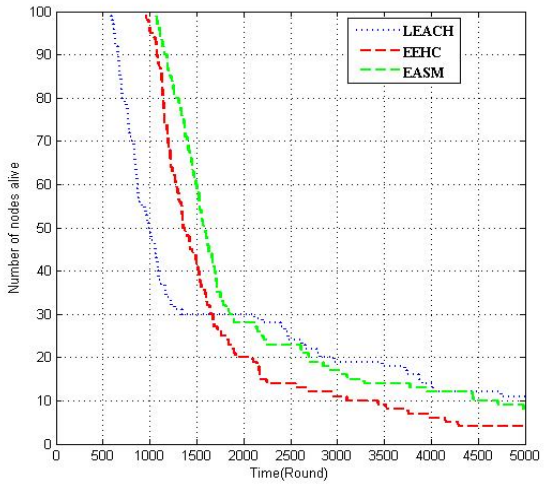


Fig. 5. Number of nodes alive over time.($\alpha = 2$,$m = 0.3$,$\beta = 5$ and $m_0 = 0.6$)

For EEHC, the stability period of EEHC is much longer than that of LEACH. Though achieves the stability period longer by about 37% than LEACH (see Fig.4 and 5). This is because EEHC is an energy-aware protocol, which elects the cluster-heads according to the residual energy of nodes. Being also an energy-aware protocol, EASM outperforms other clustering protocols. In fact, EASM obtains 19% more rounds than EEHC.

Fig. 6 shows the comparison between all nodes in terms of FND and HNA. Obviously, we can remark that our protocol EASM contains a larger period of stability time than LEACH and EEHC, which increases the efficiency of the network. We notice the same results for HNA.

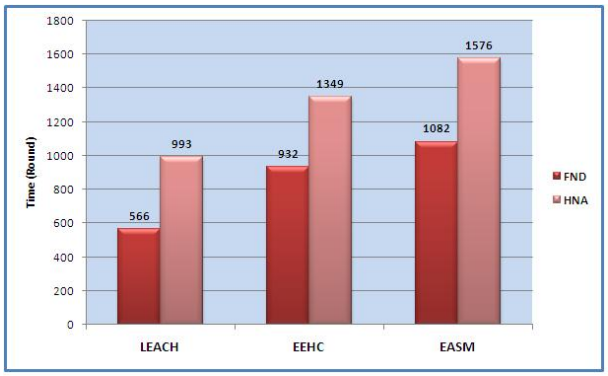


Fig. 6. FND and HNA

A longer stable time metric is important because it gives the end user reliable information of the sensing area, which extend the network lifetime. This reliability is vital for sensitive applications like tracking fire in forests.

Third, we run simulation for our proposed protocol EASM to compute the number of received messages at the BS over energy dissipation and compare the results to LEACH and EEHC protocols. Fig.7 shows that the messages delivered by EASM to the BS are better than the others ones; this means that EASM is an energy-aware adaptive clustering protocol.

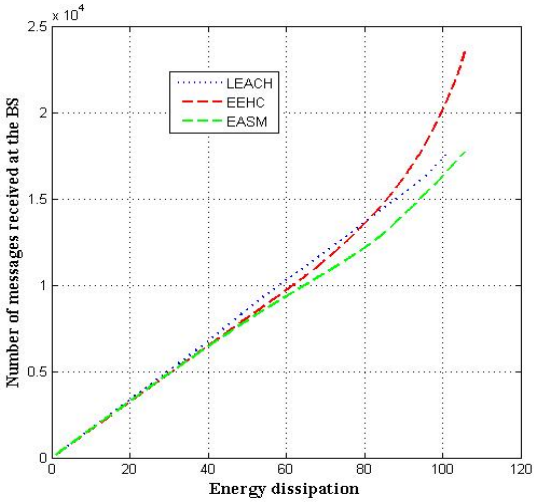


Fig. 7. Number of message received at the BS over energy spent

Fig. 8 shows the remaining energy over time for all simulated protocols and it reveals that EASM consumes less energy in comparison to the others, which helps to extend the network life span.

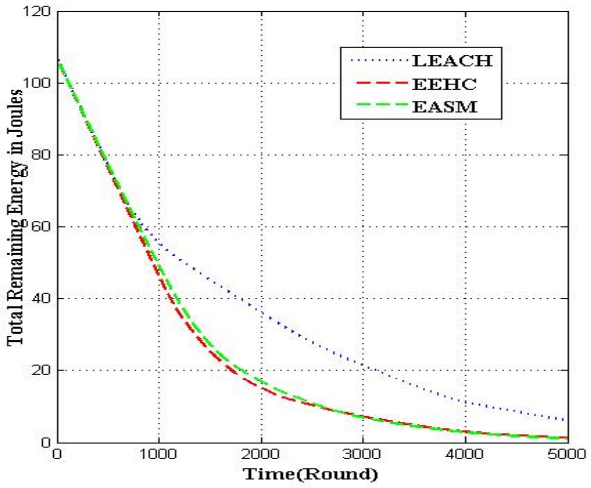


Fig. 8. Total remaining energy over rounds under 3-level heterogeneity of LEACH, EEhC and EASM

According to the simulation results, we can obviously state that EASM is a more efficient protocol than LEACH and EEHC, and consequently can be considered as an energy-aware protocol.

V. Conclusion

It has been explained in details that EASM is an energy-aware adaptive clustering protocol used in Multi-level heterogenous WSNs. To control the energy expenditure of nodes by means of adaptive approach, EASM uses new optimal probability threshold which takes the ratio of residual energy and initial energy into account. In order to increase more the EASM protocol performances, we implemented a dynamic way to distribute the spent energy more equitably between nodes. Thus, saves energy in a better way and consequently increases the life span of the WSNs

To sum up, we can say that the proposed algorithm EASM extends and outperforms better the performances of EEHC protocol.

References

[1] F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirici, A survey on sensor networks, IEEE communications magazine 40 (8) (2002)102-114.
[2] V. Mhatre, C. Rosenberg, D. Kofman, R. Mazumdar, N. Shroff, Design of surveillance sensor grids with a lifetime constraint, in: 1st European Workshop on Wireless Sensor Networks (EWSN), Berlin, January 2004.
[3] S.Taruna, Kusum Jain, G.N. Purohit Application Domain of Wireless Sensor Network: A Paradigm in Developed and Developing Countries, International Journal of Computer Science Issues, 8(4)2 : 611-617, 2011.
[4] D. Estrin, L. Girod, G. Pottie, and M. Srivastava, Instrumenting the world with wireless sensor networks, In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, (ICASSP 2001), Salt Lake City, Utah, USA, vol. 4, pp. 2033- 2036, May 2001.
[5] R. Min, M. Bhardwaj, S. Cho, E. Shih, A. Sinha, A. Wang, A.Chandrakasan, "Low-power wireless sensor networks", VLSI Design 2001, Invited Paper, Bangalore, January 2001.
[6] W.R. Heinzelman, A.P. Chandrakasan, H. Balakrishnan, Energyefficient communication protocol for wireless microsensor networks, in: Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS-33), January 2000.
[7] W.R. Heinzelman, A.P. Chandrakasan, H. Balakrishnan, An application-specific protocol architecture for wireless microsensor networks, IEEE Transactions on Wireless Communications 1 (4) (2002) 660-670.
[8] G. Smaragdakis, I. Matta, A. Bestavros, SEP: A Stable Election Protocol for clustered heterogeneous wireless sensor networks, in: Second International Workshop on Sensor and Actor Network Protocols and Applications (SANPA 2004), 2004.
[9] L. Qing, Q. Zhu, M. Wang, "Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks". ELSEVIER, Computer Communications 29, pp 2230-2237,2006.
[10] M.A. KOULALI, M. EL KOUTBI, A. KOBBANE, and M. AZIZI, QGRP : A No-vel QoS-Geographic Routing Protocol for Multimedia Wireless Sensor Networks, International Journal of Computer Science Issues, 8(6) : 51-66, 2011.
[11] Jiang, Q. and Manivannan, D. (2004) 'Routing protocols for sensor networks', IEEE Consumer Communications and Networking Conference, January, pp.93-98.
[12] Hedetniemi, S. and Liestman, A. (1988) 'A survey of gossiping and broadcasting in communication networks', Networks,Vol. 18, pp.319-349.
[13] Heinzelman, W.R., Kulik, J. and Balakrishnan, H. (1999) 'Adaptive protocols for information dissemination in wireless sensor networks', ACM MobiCom, Seattle, pp.174-185.
[14] Sohrabi, K., Gao, J., Ailawadhi, V. and Pottie, D.J. (2002) 'Protocols for self-organization of a wireless sensor network', IEEE Personal Communications, October, pp.16-27.
[15] Intanagonwiwat, C., Govindan, R. and Estrin, D. (2000) 'Directed diffusion: a scalable and robust communication paradigm for sensor networks', ACM MobiCom, pp.56-67.
[16] S. Lindsey, C.S. Raghavenda, PEGASIS: power efficient gathering in sensor information systems, in: Proceeding of the IEEE Aerospace Conference, Big Sky, Montana, March 2002.
[17] O. Younis, S. Fahmy, HEED: A hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks, IEEE Transactions on Mobile Computing 3 (4) (2004) 660-669.
[18] Mostafa SAADI, M. L. HASNAOUI, A. BENI HSSANE, S. BENKIRANE, M. LAGHDIR, "Review and Extension of LEACH Protocol for Wireless Sensor Networks",in: Proceeding of "International Conference on Complex Systems" (ICCS12) November 2012, Agadir, Morocco.
[19] V. Mhatre, C. Rosenberg, Design guidelines for wireless sensor networks: communication, clustering and aggregation, Ad Hoc Network Journal 2 (1) (2004) 45-63.
[20] M. Ye, C. Li, G. Chen, J. Wu, EECS: an energy efficient cluster scheme in wireless sensor networks, in: IEEE International Workshop on Strategies for Energy Efficiency in Ad Hoc and Sensor Networks (IEEE IWSEEASN2005), Phoenix, Arizona, April 7-9, 2005.
[21] A. Depedri, A. Zanella, R. Verdone, An energy efficient protocol for wireless sensor networks, in: Autonomous Intelligent Networks and Systems (AINS 2003), Menlo Park, CA, June 30-July 1, 2003.
[22] Brahim Elbhiri and Saadane Rachid and Driss Aboutajdine","Stochastic Distributed Energy-Efficient Clustering (SDEEC) for Heterogeneous Wireless Sensor Networks", "Computer Networks and Internet Research CNIR",November 2009,9(2),pp.11-17.
[23] S. Bandyopadhyay, E.J. Coyle, An energy efficient hierarchical clustering algorithm for wireless sensor networks, in: Proceeding of INFOCOM 2003, April 2003.
[24] Dilip Kumar, Trilok C. Aseri, R.B. Patel, "EEHC: Energy efficient heterogeneous clustered scheme for WSNs", ELSEVIER, Computer Communications, 32 (2009) 662667
[25] Yingchi Mao, Zhen Liu, Lili Zhang, Xiaofang Li, "An Effective Data Gathering Scheme in Heterogeneous Energy WSNs" , International Conference on Computational Science and Engineering,2009.

Authors:

Mostafa SAADI : Received the B.Sc. degree in Computer Sciences at the University Hassan 2^{nd} , Faculty of Sciences Ain-Chook, Casablanca, Morocco, in 2003, and a M.Sc. degree in Mathematical and Computer engineering at the University Chouaib Doukkali, Faculty of Sciences, El Jadida (FSJ), Morocco, in 2009. He has been working as a professor of Computer Sciences in high school since 2003, in Sidi Rahal Beach, Morocco. Currently, he is working toward his Ph.D. at FSJ. His current research interests performance evaluation, analysis and simulation of Wireless Sensor networks.

Dr. Moulay Lahcen HASNAOUI: Received his Ph.D in modelling and simulation of semiconductor devices at the Paris-Sud University, France (1991-1995). He worked as research associate in developing fuel cell at Department of Engineering Physics, Polytechnic School, Montreal, Canada (1996-1996). He earned his bachelor's degree in Computer Science from University of Montreal, Canada (1998- 2002). Self-employed as a software developer (2002-2004). He worked as research assistant professor at Mathematics and Computer Science Department at the Faculty of Sciences, MATIC Laboratory, El Jadida, Morocco, between 2004-2011. He is working as research assistant professor at Computer Sciences Department at the Faculty of Sciences Dhar Al Mahraz, Fez (2011).

Abderrahim BENI HSSANE: Is a research and an assistant professor at Science Faculty, Chouab Doukkali University, El Jadida, Morocco, since September 1994. He got his B.Sc. degree in applied mathematics and his Doctorate of High Study Degree in computer science, respectively, in 1992 and 1997 from Mohamed V University, Rabat, Morocco. His research interests focus on performance evaluation in wireless networks.

Said BENKIRANE: Obtained his Certificate in telecommunications engineering at the National Institute of Posts and Telecommunications, Rabat, Morocco, in 2004, and his M.Sc. degree in computer engineering and network from the University of Sidi Mohammed Ben Abdellah Fez, Morocco in 2006. He has been working as professor of Computer Sciences in high school since 2007, in El Jadida, Morocco. He is a member of a research group e-NGN (e-Next Generation Networks) for Africa and Middle East. Currently, he is pursuing his Ph.D at the Faculty of Sciences, Chouaib Doukkali University, El Jadida, Morocco. His main research areas include wireless and mobile computing and mobile telecommunications systems.

Comprehensive evaluation on housing market supply & demand based on principal component analysis: the case of Xi'an, China

Jianping Yang[1], Yanwei Zhang[2]

[1] Department of Management, Xi'an University of Architecture Technology
Xi'an, Shanxi Province, China

[2] Department of Management, Xi'an University of Architecture Technology
Xi'an, Shanxi Province, China

Abstract

At present, China's housing prices and structures have been severe distorted, which have been done serious harm on the development of national economy and people's lives, the government has implemented the most strictest macro-control. This paper applies the correlation and principal component analysis on the supply & demand of housing market. Aims to points out clear direction for the implementation of macro-control policies in the future.

Keywords: *Macro-control; Correlation (Statistics); Principal Component Analysis; Comprehensive Evaluation; Supply & Demand Coordination Degree; Tourism Real Estate*

1 Introduction

In recent years, China's housing market has entered a white-hot stage, the prices and structures of housing are irrational, and the housing market distortions do serious harm to the social and economic development. After 2010, the government has taken the most severe macro-control policies to promote the healthy and orderly development of the housing market. The supply & demand relationship development is the culprit of many problems and most intuitive response to market conditions, studying the current housing market supply & demand is imminent. The purpose of research is to study the supply & demand analysis indicators to provide accurate information on the country's macro-control of intensity and direction in the future.

2 Some Related References Background

Until now, there have been many references studying supply & demand of housing, and correlation and principal component analysis (PCA) have been applied widely in various domain including science, commerce, agriculture, medicine, and industry. In [1], housing supply is affected by construction costs, demand is determined by the renting/buying considerations of the public, adopting three price indices of the Taiwan housing market, the CCI, the RPI and the HPI, and examines long-term and short-term correlations among the three indices [1]. The reference [2] focuses on the housing demand increase in Belgium. An overview of the housing market is presented wherein several are identified such as rental, private, new-build and secondhand. Forty-three percent of households looking for housing in 2009 bought an existing house, 43 percent rented, 11 percent opted for new self-build, and three percent bought a new house [2]. The reference [3] hypothesizes that the increase in money supply accosted by rapid economic growth leads to strong investment demand in the Taiwanese housing market. When the growth rate of money supply is below the model's estimated threshold value, household number, income, and user cost of housing capital are significant variables. Results suggest that non-linear movement of housing prices is primarily driven by investment demand [3]. The reference [4] discusses briefly the methodology used, the housing situation, the dynamics of the housing sector, and etc. the main parts of the research is to analyze the policies and practices, supply factors and demand factors. Showing the housing policy adopted by the government aims to diversify housing types according to household incomes. And analyzing two supply factors which are land development regulations and ownership, and housing production and ownership, and two demand factors which are housing finance and housing subsidies [4]. The reference [5] aims to solve the current high price's dilemma, according to the effective housing demands, and combining the local real income levels then build up three high, middle and low real estate markets with independent operation, and mutual connection and conversion [5], and etc [6-8].

In [9], principal component analysis (PCA) was employed to optimize an LUR model for PM2.5. An optimized surrogate of vehicle emissions was produced by PCA and employed as the predictor variable in the model, resulted that the method used can contribute to LUR techniques in two major ways: 1) by improving the predictive power of

the input variable, by substituting a principal component for a single variable and 2) by creating an orthogonal set of predictor variables, and thus fulfilling the no colinearity assumption of the linear regression methods. The proposed PCA method should be universally applicable to LUR methods and will expand their economical attractiveness [9]. The reference [10] applies the principal component analysis and parallel analysis to smoothed tetrachoric correlation matrices were investigated in a simulation study. To evaluate the effect of several smoothing algorithms, 360 different types of data sets were simulated [10], and etc [11-12].

3 China's Housing Market Performances under the Macro-control Background

3.1 A slump in China's housing market

At the beginning of 2011, the country has implemented series of forceful measures, such as purchase limit, credit limit, price limit, and etc. the government used widely of the land, finance, taxation and other means to return to the reasonable prices and structures of housing. According to the concerned statistic data of 2011 shown, real estate enterprises invested 4430.8 billion in the construction of housing, it grown 30.2% than 2010, increasing ratio decreased 2.6%. The sales of commercial housing area and the transactions of second-hand housing area also declined sharply. The commodity housing sales area is 970.3 million square meters, increasing ratio declined 4.4%; second-hand housing transaction area is 93.38million square meters, declined 33.0% than 2010. The boom index of real estate has dropped fall since in June 2011, it has fallen to 99.87 and 98.89 in November and December respectively, which dropped below the boom line 100 for the first time since 2009.

3.2 Real estate industry structure confronted adjustment

At present, under the background of the draconian control policies and the inflation, the real estate industry is confronted with the adjustment of industrial structure. The real estate industry structure ratio will be transformed from the domination of the housing real estate in the world into the commercial real estate as its core with a collection of the industry chain. Along with the policy changes in recent years, the financing costs, transaction costs and holding costs of the real estate have increased gradually, and the housing real estate investment function has been disappeared step by step, thus housing estate will be gradually transformed into a product against inflation. Along with the Chinese enterprises in the industrial restructuring, consumer consumptive habits transformation, the reform of financial system and the adjustment of industrial structure, China's commercial real estate development model will be changed from simple property rental into commercial estate as the core with the integration of the industry chain and capital chain.

National implementation of stringent regulation policy aims to adjust market supply & demand, and from the performance of the housing market and the real estate industry we can see that, the implementation of the policy also played a certain role, thus analysis of housing market supply & demand coordination degree seems much more necessary.

4 Housing Market Supply & Demand Analysis: The Case of Xian

4.1 The selection of Indexes and analysis tools

1) In this paper, the analysis of data processing, correlation, principal component and parameter interval estimation, and the calculation of comprehensive evaluation value, all use the software of Excel 2007 and SPSS Statistics 20.

2) According to the related references, there have selected five indexes which can synthetically reflect the housing market supply & demand: housing capital fund / the sales of housing X_1, housing pre-sale area / housing area sales X_2, the completed housing area / housing vacancy X_3, housing average price growth rate / urban per capita disposable income growth X_4, housing sales price index / housing sales price index X_5, show as Table 1.

Table 1 2004-2011 the data of different indexes

	X_1	X_2	X_3	X_4	X_5
2004	1.362	0.568	0.236	3.165	1.035
2005	0.924	0.603	0.313	2.387	1.046
2006	1.26	0.699	0.251	-0.556	1.023
2007	1.219	0.584	0.091	0.288	1.024
2008	1.566	0.756	0.086	0.838	1.027
2009	1.261	0.911	0.063	-0.008	1.002
2010	1.014	0.953	0.064	0.913	1.091
2011	0.856	0.667	0.426	1.9614	1.012

Notes: the above data derived from 2003-2012 the statistical year books of Shaanxi Province

4.2 The analysis of index correlation

The data in table 1 is processed by the correlation analysis, and then obtained correlation coefficients and significance test of the indexes, show as Table 2.

Table 2 Pearson correlation coefficients and significance test of the indexes

Indexes		X_1	X_2	X_3	X_4	X_5
correlation coefficients	X_1	1.000				
	X_2	0.014	1.000			
	X_3	-0.566	-0.568	1.000		
	X_4	-0.264	-0.480	0.517	1.000	
	X_5	-0.254	0.291	-0.228	0.216	1.000
Sig.(2-tailed)	X_1					
	X_2	0.974				
	X_3	0.143	0.142			
	X_4	0.527	0.229	0.189		
	X_5	0.544	0.484	0.588	0.607	

* *Correlation is significant at the 0.05 level (2-tailed).

According to table 2, it can be known that 1) the significance of the t-test between the correlation coefficient among the five index is significantly greater than 0.05, thus the null hypothesis is accepted, and there is no significant linear correlation among the five index.2) There is a correlation between different index. Results X_1 is negatively correlated with X_3, X_4 and X_5, X_2 is negatively correlated with X_3 and X_4, negative correlation is existed between X_3 and X_5 ,other are positively correlated, so that they provide information about the overlap. Therefore, the direct usage of these indexes is difficult to accurately evaluate the housing market supply & demand, require the application of principal component analysis method.

5 Comprehensive Evaluation of Housing Market Supply & Demand Coordination Degree Based on Principal Component Analysis

5.1 The principal component analysis basic model

The principal component analysis method is a multivariate statistical method aims to use the idea of dimension reduction, by researching the inner structure relation of the index system, to transform a lot of indices into a few independent indices which are comprehensive indices contain most information of the original index (80% ~ 85%).

If there are m indicators, the observed value of each indicator is n, so the principal component analysis model is established as follows:

$$\begin{cases} Z_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{n1}X_n \\ Z_2 = a_{12}X_1 + a_{22}X_2 + \cdots + a_{n2}X_n \\ \vdots \\ Z_m = a_{1m}X_1 + a_{2m}X_2 + \cdots + a_{nm}X_n \end{cases} \quad (1)$$

In (1), $a_{1i}, a_{2i}, \ldots a_{ni} (i = 1,2,\ldots m)$ is the characteristic vector of the characteristic value of the covariance matrix of X, and X_1, $X_2, \cdots X_n$ is the standardized variable of X, $a = (a_{ij})_{m\times n} = (a_1, a_2, \cdots a_m)$, $R_{ai} = \lambda_i a_i$, R is the correlation coefficient matrix, λ_i is the corresponding characteristic value, a_i is the unit orthogonal vector, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$.

5.2 The process of the principal component analysis

The data in table 1 is analyzed by the principal component analysis method, and then obtained total variance explained, show as Table 3.

Table 3 Total Variance Explained

Component	Eigen-vector of different indexes				
	X_1	X_2	X_3	X_4	X_5
1	-0.248	-0.323	0.404	0.335	-0.040
2	-0.422	0.339	-0.050	0.150	0.631
3	0.625	-0.345	-0.332	0.583	0.438

Component	Extraction sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	2.255	45.098	45.098
2	1.396	27.915	73.013
3	0.869	17.376	90.389

The data in table 3 shows that: in the first three indices of cumulative contribution rate of 90.389%, indicating that the first three index reflects the original index mostly, can be substituted for the original five index of the Xi'an housing market supply & demand for comprehensive evaluation, according to the standardization eigenvector, there can get the expressions of the three new comprehensive index:

$$CI(1) = -0.248X_1 - 0.323X_2 + 0.404X_3 + 0.335X_4 - 0.040X_5 \quad (2)$$

$$CI(2) = -0.422X_1 + 0.339X_2 - 0.050X_3 + 0.150X_4 - 0.631X_5 \quad (3)$$

$$CI(3) = -0.625X_1 + 0.345X_2 - 0.332X_3 + 0.583X_4 - 0.438X_5 \quad (4)$$

From the above expression can be clearly seen that 1) by the reducing dimension, can make the 5 original interrelated supply & demand indexes into three new independent comprehensive indexes, and retain the information of the original index 2) most of the information of different indexes on CI(x) effect is also different. In the CI(1), X_2 and X_4 have the greatest impact on the evaluation value; In CI(2), the impact of X_5 and X_1 is the biggest, but the former has positive effect, the latter produces negative effect; in the CI(3), X_1 and X_4 is the maximum impacts 3) the three new comprehensive index contribution different rates will exercised different influence on CI(x), so there should goes further in the comprehensive evaluation.

5.3 Comprehensive evaluation on the housing market supply & demand coordination degree

1) The comprehensive evaluation expression
According to the principal component expressions, there can calculate 2004-2011 Xi'an housing market supply & demand comprehensive evaluation value of CI(x); according to the contribution rate and the contribution rate of (Table 3), there can calculate the comprehensive indicator weight K_1=0.4989, K_2=0.219 3088, K_3=0.1922, respectively, and then establish Xi'an housing market supply & demand coordination degree comprehensive evaluation expression CE:

$$CE=CI(1)*K_1+CI(2)*K_2+CI(3)*K_3 \quad (5)$$

2) The comprehensive evaluation grade
According to the parameter interval estimation principle, CE is analyzed by the method of confidence intervals, therefore, Xi'an housing market supply & demand comprehensive evaluation grade can be calculated, show as Table 4.

Table 4 Comprehensive evaluation grade of the housing market supply & demand coordination grade

Grade	Perfect coordinated	Well coordinated	Coordinated
Range	≥ 1.293	(0.848,1.293)	(-0.041,0.848)

Grade	Not well coordinated	Not coordinated
Range	(-0.485,-0.041)	≤ -0.485

3) The comprehensive evaluation results
According to the comprehensive evaluation expression and grade, Xi'an housing market supply & demand comprehensive evaluation results can be obtained, show as Table 5.

Table 5 the comprehensive index value CI(x) and comprehensive evaluation value CE of the housing market supply & demand

	CI(1)	**CI(2)**	**CI(3)**	**CE**	**Grade**
2004	0.593	0.734	2.875	1.075	Well coordinated
2005	0.461	0.817	2.115	0.889	Well coordinated
2006	-0.664	0.254	0.588	-0.140	Not well coordinated
2007	-0.398	0.369	1.146	0.135	Coordinated
2008	-0.358	0.365	1.628	0.247	Coordinated
2009	-0.624	0.405	0.887	-0.016	Not well coordinated
2010	-0.272	0.717	1.293	0.335	Coordinated
2011	0.361	0.776	1.750	0.756	Coordinated

From table5 and Figure 1, it can be known: 1) In 2004 and 2005 Xi'an housing market supply & demand coordination grades both are the Well coordinated, 2004 Xi'an housing market supply & demand coordination degree reached 1.075, it decline from 0.889 of 2005 to minimum value -0.140 of 2006 in the 8 years, the grade of coordination degree is down to coordinated from not well coordinated. 2006-2008 supply & demand improved slightly, rising the grade of coordinated. because of the 2008 financial crisis, 2008-2009 the comprehensive evaluation value decline to -0.016, nearly the minimum value of 2006, return to the grade of not well coordinated; In 2009-2011, because the country introduced a series of incentives, Xi'an housing supply & demand coordination degree high speed increased by 0.772 points; In 2011, owing to the implementation of the most stringent regulatory policies by far, Xi'an housing market supply & demand coordination degree increased by 0.071 comparing 2011 to 2010, and up to the grade of coordinated.2) At present, Xi'an housing market supply & demand is at the grade of coordinated. The implementation of the macro-control policies played a significant effect on curbing the speculative demand, reducing price and adjusting housing structure. Therefore, the government of housing market ironhanded macro-control can't relax, and must guide effective demand at the same time.

6 Recommendations

Under the strict macro-control background, curbing the speculative demand and guiding the effective demand at the same time is the only key, and there some recommendations on the housing market and real estate industry development.

6.1 Focus on the housing market regulation, meanwhile deepen comprehensive housing policies

Under the market economic system, the housing real estate, like other commodity markets, run under the interaction of the supply & demand mechanism, price mechanism and competition mechanism, but because of the particularity and complexity of the supply & demand, the supply & demand balance of the housing market is a long and difficult task under normal circumstances. So there should give priority to the market adjustment, meanwhile continue to use the visible hand to make up for market failure, to correct market irregularities, by the comprehensive use of economic, legal and requisite administrative means to achieve the equilibrium of supply & demand. There should strengthen the macro-control policies, the banking and credit system, and actively and effectively play the regulatory role of the property tax, to deepen the policy of regulation, in order to curb the speculative demand and guide the effective demand, prompting prices return to reasonable.

6.2 Strengthen the construction of the indemnificatory housing, establish multifunctional housing supply system

There should strengthen the government oversight and enforcement, meanwhile improve the laws and regulations for indemnificatory housing. There should accelerate the implementation of the "Twelfth Five-Year Plan" of the indemnificatory housing industry, further enhance the construction of the affordable housing, capped-price housing, public rental housing, low-rent housing as well as the shed housing vigorously. To steadily solve the housing problem of all income groups, such as city employees (mainly college graduates newly participate in the work) as well as the migrant workers' quickly. All of this is in order to establish indemnificatory housing supply system step by step, which is based on protection of basic needs and guidance of rational consumption, is the government provide basic protection mainly and the market meet the multi-level needs mainly.

6.3 Coordinate at all levels of real estate markets, balance supply & demand

There should establish the monitoring and evaluation system of the land market and perfect the land market information communication mechanism, improve the differentiation of the land supply policy to ensure the land transaction is fair and reasonable moderate, further stabilize market expectations. There should coordinate the development of secondary real estate market, guide he real estate developer to invest reasonably and the buyers consume rationally, and improve the real estate market information system to make the information transparent, in order to avoid blind investment and consumption. There should promote the reform of state-owned real estate enterprises, improve enterprise and personal credit file and evaluation system, guide the healthy development of the real estate market. We know that China's large population, poor little rich, so less indemnificatory housing in solving the housing problems of lower-middle class is just a drop in the bucket, it's necessary to promote reasonable housing concept vigorously, encourage low-income and middle-income groups to purchase second-hand and rental housing, and guide the second-hand rental housing market actively, that play the linkage effects of the real estate markets at all levels to promote the equilibrium of supply & demand.

6.4 Conformed and combined with the tourism, accelerate the sustainable development of tourism real estate

Tourism real estate is a new convergence industry which is combined tourist industry with real estate industry. A successful tourism real estate projects always have itself cultures, which are not just rely merely on the tourism resources. Therefore, the real estate the real estate developers should find their strength and make the best use of opportunities instead of using housing development ideas on the development of tourism real estate, they need to make innovations and open up different development concept. 1) housing real estate should be transformed from low-end to high-end, from a single housing real estate to the diversification of the tourism real estate, from one-time development to sustainable development, and increase the diversification of the tourist hotels, vigorously develop the service industry, and then strengthen the real estate industry 2) whereby the quality of tourism industry to promote the development of the real estate market, which will be the long-term way of Xi'an development. So there should establish the orderly development of high-end tourism real estate goals, adhere to high standard, high-quality and high-end development principles, to construct a diversified tourism real estate supply system which includes high-star hotel, the hotel property, high-end leisure houses, and so on.

6.5 Accelerate the adjustment of the real estate industry structure, promote the housing industrialization gradually

There should deepen the policy control, promote the structural adjustment of the real estate industry, and establish a diversified real estate product supply system which take commodity housing as a basic, affordable housing as a protection, travel health real estate as a characteristic. Gradually form the housing supply system which has the protection in the low-end, has the support in

the medium-end, has the market in the high-end, and has the reputation in the characteristics. There should gradually promote the housing industrialization. First, there should develop a variety of housing standards, and narrow the gap between the standard; Second, there should focus on building energy efficiency, and reduce the construction cost; Third, there should adapt to the needs of the residents, and strengthen the demonstration and driving effect of the government; Fourth, there should take the concept of green and sustainable development, healthy and comfortable living environment as the prerequisite to accelerate the process of housing industrialization.

7 Conclusions

In this paper, the application of correlation analysis and principal component analysis method on the comprehensive evaluation of Xi'an housing market supply & demand under the macro control, the results show that under the sustaining implementation of the housing policies in the last two years Xi'an housing market has obtained satisfaction effect, so the study verifies the effect of the government regulation and the efficiency of the principal component analysis method in the application of the housing market supply & demand comprehensive evaluation. The results provide the meaningful guidance for the intensity and direction of the macro-control in the future, and what great important is the meaningful reference and valuable theoretical basis for the future research.

References:

[1]Tsai, I-chun, Housing Supply, Demand and Price: Construction Cost, Rental Price and House Price Indices Housing Supply, Demand and Price: Construction Cost, Rental Price and House Price Indices, Asian Economic Journal, Vol. 26 Issue 4, p381-396, Dec 2012

[2]Loosveldt, Filiep, Investment in housing-the case for demand and supply-side subsidies; the example of the Belgian housing market, Housing Finance International; Vol. 26 Issue 3, p28-30, Mar2012

[3]Ming-Chi Chen, Chin-Oh Chang, Chih-Yuan Yang, Bor-Ming Hsieh, Investment Demand and Housing Prices in an Emerging Economy, Journal of Real Estate Research, Vol. 34 Issue 3, p345-373, Jul-Sep2012

[4]Bellal, Tahar, Housing supply in Algeria: affordability matters rather than availability, Theoretical and Empirical Researches in Urban Management 12, p 97-114, Aug 2009

[5]Qingquan, Li, Guohua, She, Building Cascaded Hierarchical Real Estate Market Based on the Housing Consumption Theory and the Security Theory, Management Science and Engineering 5.3, p 68-71, 2011

[6]Fingleton, Bernard, Housing Supply, Housing Demand, and Affordability, Urban Studies, v 45, n 8, p 1545-1564,2008

[7]Dol, Kees, Kleinhans, Reinout, Going too far in the battle against concentration? On the balance between supply and demand of social housing in Dutch cities, Urban Research & Practice, Vol. 5 Issue 2, p273-283, Jul2012

[8]Kemp, Brian, The Housing Market--A Case for Supply and Demand Analysis, Economics, 13, 58, 40-5, Sum 77.

[9]Olvera, Hector A., Garcia, Mario, Li, Wen-Whai, Yang, Hongling, Amaya, Maria, Myers, Orrin, Burchiel, Scott W., Berwick, Marianne Pingitore, Nicholas E. Principal component analysis optimization of a PM2.5 land use regression model with small monitoring network, Science of the Total Environment; Vol. 425, p27-34, May2012

[10]Debelak, Rudolf, Tran, Ulrich S.,Principal Component Analysis of Smoothed Tetrachoric Correlation Matrices as a Measure of Dimensionality, Educational & Psychological Measurement; Vol. 73 Issue 1, p63-77, Feb2013

[11]Khatun, Tahmina, Measuring environmental degradation by using principal component analysis. Environment, Development & Sustainability; Vol. 11 Issue 2, p439-457, Apr2009

[12]Cadima, Jorge, Calheiros, Francisco Lage, Preto, Isabel P., The eigenstructure of block-structured correlation matrices and its implications for principal component analysis, Journal of Applied Statistics; Vol. 37 Issue 4, p577-589, Apr2010

First Author Jianping Yang: male, born in 1969, Ph. D, Professor, master's tutor, national registered real estate valuer, National Registered Consulting Engineer (investment), is the committee of the Architectural Society of China economic housing construction real estate economy professional, Shaanxi Province Civil Society Construction Economy, and the executive director of Xi'an real estate Appraisal Association, the main research directions are construction and real estate economics and management of the city, investment decision and project evaluation, the management of the construction and real estate enterprises, building energy saving and other relevant teaching and research work. Hosted and participated in the National Natural Science Foundation of China, natural science foundation, longitudinal and transverse scientific research and teaching research for more than 30 items, is the editor in chief, deputy editor of more than 10 textbooks and other books, published teaching and academic papers nearly 20 pieces, and won the award of provincial government science and technology second prize, science and technology progress province college first prize, college scientific and technological progress first prize, and the school project award, the paper award and the teaching achievement award nearly 20 items.

Second Author Yanwei Zhang: female, born in 1987, is a postgraduate student of the Xi'an university of architecture & technology, School of management, the main research direction is the tourism project development and management.

The Study of The Bay of Mount Saint-Michel by Using Graph Theory in The Analysis of Satellite Images

Seyfallah BOURAOUI

Dynamique Globale et Déformation Active (UMR 7516-CNRS), University of Strasbourg, Strasbourg, France

Abstract

In this paper, a new approach for mapping based on the concept of objects and relationships between these objects is proposed to take advantage from both supervised and unsupervased classification methods. On the one hand, objects obtained after a supervised classification are represented by an adjacency graph model. On the other hand, objects obtained after unsupervised classification are represented by an adjacency graph data, and the goal is to measure the matching between this two graphs in order to improve the results of unsupervised classification in association with those obtained from supervised classification. This study concerned the coastal Bay of Mont Saint-Michel, the data used are from SPOT 5 optical satellite images.

Keywords: *Clustering, graph theory, Classification, graph matching, spatial relations, mapping.*

1. Introduction

Remote sensing techniques allow the extraction and the analysis of large and different type of information provided from height resolution satellite images. Analyzing with relevant the content of this height resolution satellite imagery following a complex process and needs advanced techniques of data mining to provide a solution to the automatic map generation problem [1]. Several algorithms and methods are developed in this purpose to detect complex object from satellite images [1].

The bay of mount Saint-Michel is a coastal area subject to large environmental and anthropogenic pressures. In fact, the natural and human activities increases the pressures facing the coast, need quick solutions for the management of the territory. In this concern, the analysis of remot sensing images play an important role, and specially the study of the coastal environment from satellite imagery can address a variety of fields ranging from simple mapping of the land to the study of the foreshore and hydro-sedimentary dynamics.

Bay of Mont Saint-Michel was chosen as a study site to share its scientific interest which is worn for several decades. This area has also a global reach with environmental and heritage very important. On the other hand, the bay is one of the most complex and most dynamic coast area in the world. The multiple challenges posed by the management of the bay requires a thorough knowledge of this area and its current and future developments.

The bay is a very dynamic environment that is changing rapidly. Satellite images are very well suited for monitoring regularly updated and they allow a multiscale study, in addition to traditional data used by managers. More accurate images is finer, closer now than aerial photographs. Finally satellite images permettes many additional treatments because of their spectral range with infrared. The advantage of using satellite images to study the nearshore been the subject of several works [2, 3, 4].

One of the classical techniques using THR images or called traditional techniques of image processing are based on the pixels or regions are no longer applicable, the community is currently interested in the technical processing of objects and relations between these objects. The interest in such approach is argumented by the possibility to extract the maximum information and have intelligent processing of information in the same way as in the human résonnement.

This paper is organized as follows: section 2 presents the description of the classification system. Section 3 presents the supervised classification. In Section 4, we present the method used in unsupervised classification. In Section 5, we present the construction of adjacency graphs and matching algorithm. Finally, in Section 6, we give the conclusion and some perspectives.

2. The System Description

System description is given in Fig. 1. Two different classification methods can be run in parallel in tow differents processus for the same image mapping, a supervised and unsupervised algorithms are used for this purpose. We will detail each of these classification method following the various stages that make up the system architecture. The supervised classification is performed from samples selection and from ROI (Region of Interest) of each sampled image consiste the critical step in the process of supervised classification. The quality and the suitability of the results will depend on this sampling

phase. That is why it is important to select the samples that cover all classes of objects. It is not very clear for a non-expert in the field to distinguish between different objects especially on coastal environment or objects are very similar and in overlapping. So to get a good set of sampling data, we requires the help of an expert (knowledge of the geography of the field), while the second method of unsupervised classification does not require domain knowledge for the processing.

This two classification methods are independent in the system architecture and will be run in parallel (needs multi-threding system). From the classification results of each of these two methods, we will build an adjacency graph. In the case of supervised classification, the system creates a model graph where each node represents a label that corresponds to an object classified by this method and each edge represents the adjacency relation (adjacency graph). For unsupervised classification, the system creates a graph or data each node represents a form found and each arc represents the adjacency relation between these objetcs. Was used in this study the adjacency graph to represent the relationship between the diffents objects (Next-To) that can implicitly express other relationships such as the relationship (surrounded-by) can be expressed by (all the 4 or 8 Next-To an object are the same object).

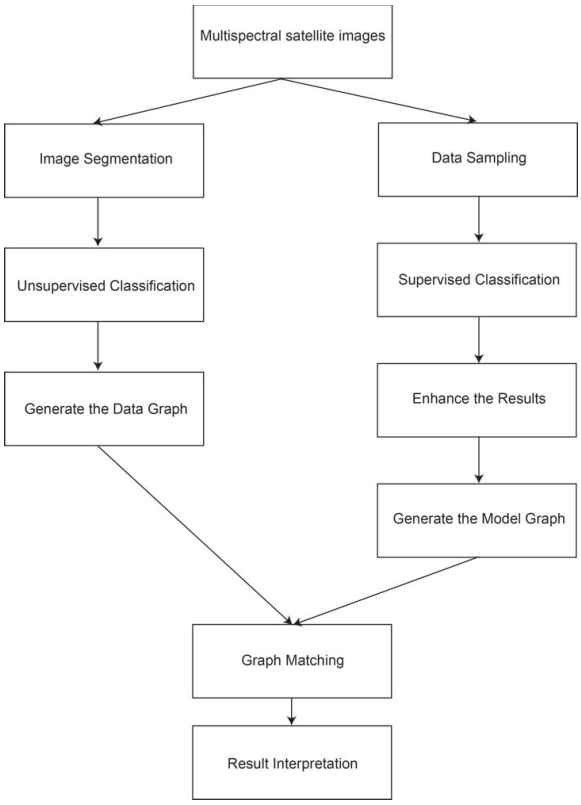


Fig. 1: The system architecture.

3. Supervised classification:

This classification method needs a set of ROI or samples representing the diffirent classes of objects presented in at least one of the sampling images. We applied the standard algorithm of maximum likelihood classification, the major advantage of this method of classification is there short runtime, howevere it can gives an incorrect results, especially in our study area (coastline) where the major objects have a very similar signature spectral and situated in overlapping. For these reasons, if we use only the spectral information in the classification process, we can not distinguished between objects "sea", "water body" and "channel and shallow sea." (Fig. 2)

After finishing the supervised classification, we can apply an algorithm for the correction of the classification based on a set of rules taking into account the nature of the possible neighbors of each classe object found. We apply different rules taking into account the number of neighbors (4 or 8 neighbors) of all pixel situated in the border of two different classes. We present an example and the syntaxe of one rule used in this processing:

- if (label (i, j) == "Wed") and (label (i-1, j) == agricultural areas with low vegetation, bare soil ") or (label (i + 1, j) == "agricultural areas with low vegetation, bare soil") or (label (i, j-1) == "agricultural areas with low vegetation, bare soil") or (label (i, j + 1) == "agricultural areas with low vegetation, bare soil") then label (i, j) == "water channel".

The classe "water" is easily identifiable with this system, and the problem is how to isolate the other classes involved in this main classe (water) . It is possible to isolate this class of objects by performing a numerical thresholding values applied in a single spectral band (monospectral thresholding). Water surfaces are observable by a sensor sensitive to near infrared: all pixels whose numerical value is below a threshold value can be assigned to the class "water" because water absorbs radiation. But it is difficult to distinguish clearly between these subclass (Sea, Channel and shallow sea, river, body of water). To solve this problem, we apply rules based on the notion of object to enhancing the final results. Below, we gives an example of this used rules.

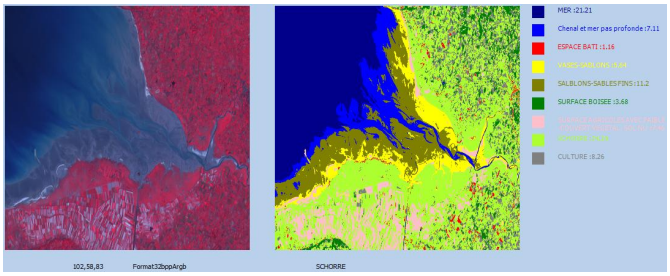


Fig. 2: Results of supervised classification for nine classes.
Is generally the format rules for classification step to improve results take into account the aspect of spatial objects. In other words, it refines the first results of the first classification by introducing the constraint direct vicinity.

- ***If (object classified channel) and (not (Next-To (object, sea))) then becomes streams, and bodies of water.***

Noticed that these rules can solve the problem of water classes, but these do not help to know other informations : the extracting of the salt marsh, the direct neighborhood information of the object. We use neo-canal information to separate between two objects and to limit the salt marsh on the map of the neo-gradient. We use the NDVI (Normalized Difference Vegetation Index) neo-Canal introduced by Rouse et al. [5].

$$NDVI = \frac{NIR - RED}{NIR + RED} \qquad (1)$$

It varies between -1 and 1. This index is very efficient for the detection of active vegetation, We use also another index to highlight the mineral surfaces, bare soil, called the index of Brilliance (IB) given by the following formula:

$$IB = \sqrt{(PIR^2 + R^2)} \qquad (2)$$

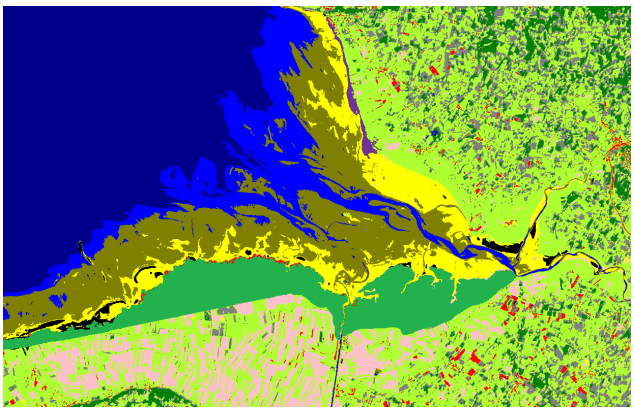


Fig. 3: The final result of supervised classification.

4. Unsupervised classification:

This type of classification is guided by the segmentation phase. The growth areas segmentation algorithm was used with a small changes in the settings to better adapt the algorithm to this area study (coastline area). The idea is to group pixels with the same gray level into the same group (cluster), the major problem faced here is that a small change in grayscale value of a pixel will assign the pixel to a new group. For this, we used a parameter k = 5, adapted to group the pixels which have the same gray level (+ / - k) and respecting to initial grayscale of the starting pixel. We recompute the mean grayscale of the for each iteration using the formula (4- neighbors):

NewNg = (OldNg * 2 + ¼ (Ngg + Ngd + Ngh + Ngb))/ 3 (3)

With: NewNg: the new gray level of the pixel.
OldNg: the old gray level of the pixel.
NGG: the gray level of the left neighbor.
Ngd: the gray level of the right-hand neighbor.
Ngh: the gray level of the neighbor above.
Ngb: the gray level of the neighbor below.

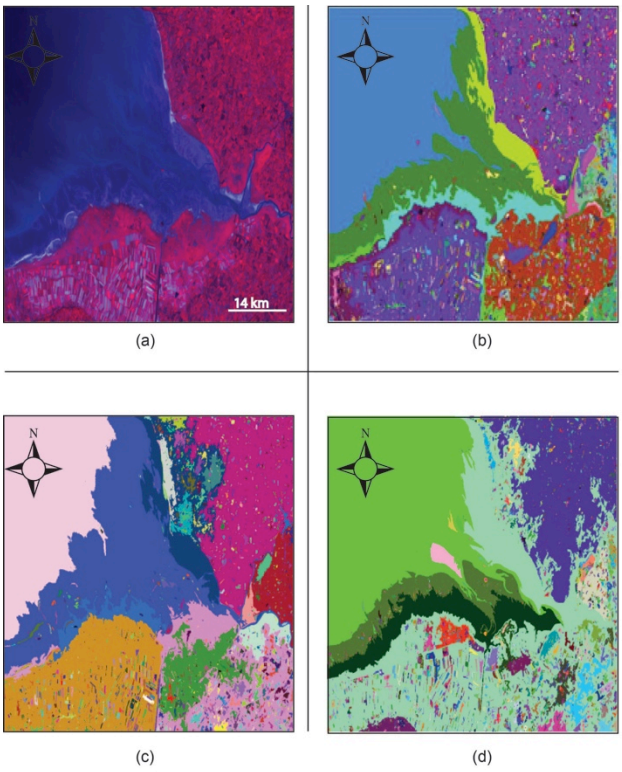


Table 1: (a) original image, (b) segmenting results with k = 5, (c) segmentation results with k = 10, (d) segmentation results of NDVI neo-canal and using k = 5;

In a second step, we applied the segmentation algorithm on a the gradient map to determine the border of the different object, then we calculate for each region the index of consistency between each region and its neighbors. According to this index, we will grouped the similar region into a single region. The consistency index between two regions is given by:

$$IC = \frac{1}{i*j} * \sum_i \sum_j (|PiR1 - PjR2|) \quad (4)$$

With IC: the consistency index.
PIR1: pixel i of region 1.
PjR2: pixel j of region 2.

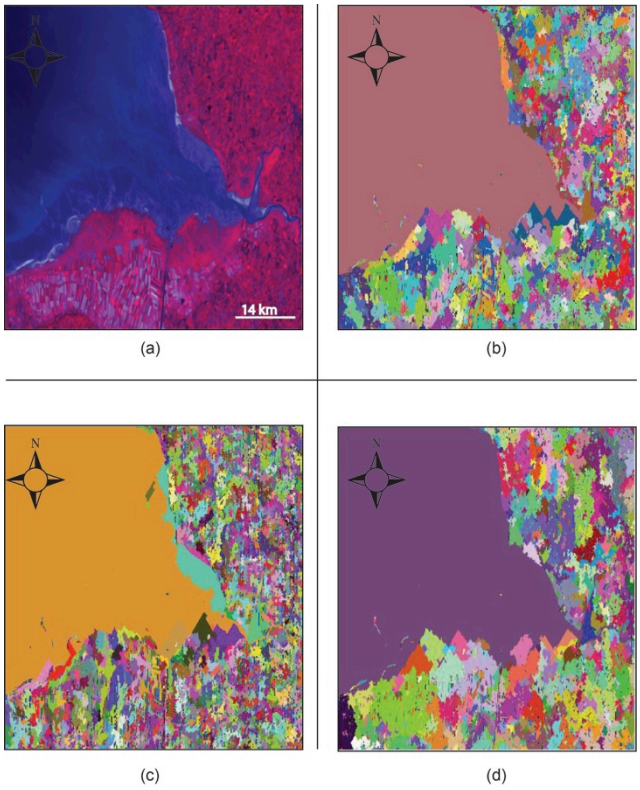


Table 2: (a) original image, (b) segmenting with k = 5, CI = 15, (c) segmentation with k = 10, IC 20 =, (d) segmentation of NDVI from k = 5, IC = 20;

5. The construction of the two graphs and their matching

This part contains three subparts, in the first subpart, we construct a model graph from the supervised classification: nodes represent labels and the average gray level and the size of each class (number of pixels) in each object and the arcs represent the adjacency relationship (direct neighbors).

In the secound subpart, we built a data graph from unsupervised segmentation respects (k = 10, CI = 25). Nodes represent the average gray levels for each region and size of each class. In general, the size of data graph is greater than that size of the model graph.

In the second part, we group the nodes of each graph separately using the given gray level value close to the gray level of the node in the graph model and the regions include the turn of nodes in the graph model and each cluster is recalculated at average gray regions by the formula (3). Spatial information used here is the relationship of Next-To as bijective relation.

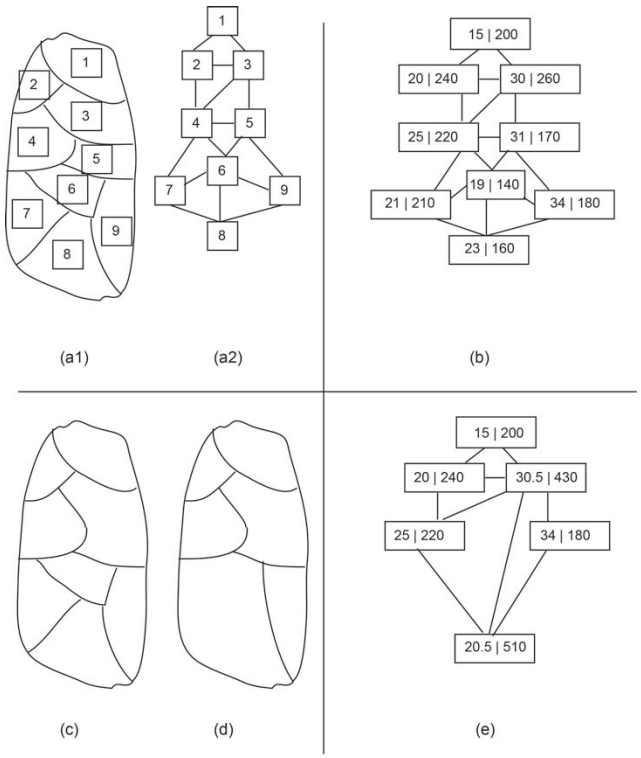


Fig. 4: (a1) the result of the unsupervised segmentation, (a2) its adjacency graph, (b) the composition of the value nodes left the gray level and the right size of the region, (c) the result of the first iteration, (d) the final result with Nb-node = 6, (e) the final result graph.

However, an ambiguity in the decision phase can appeared when the system faces the situation where there is more than one neighbor node checked the condition (same gray level and both are neighbor of the initial node). In this case, the decision will be for the largest region (number of pixels). In other words, this node will be added to the critical node that represents the largest region. At the end

of this stage, the system left two graphs having the same number of nodes and labeling graph data is as following:

For each node in the model graph model do:

Find the node in the data graph that minimizes *distance*, this distance is calculated by the formula:

$$\text{distence} = \sqrt{(\text{NGM} - \text{NGD}) * (\text{sizeM} - \text{sizeD})} \quad (5)$$

if size (data node) is less than size (model node) then:

data node ← model node.

Delete all the pixels belonging to this node in the model graph.

- Give to this node of data graph the same label as the corresponding node of model graph.

repeat until have covered all model graph

Return the data graph.

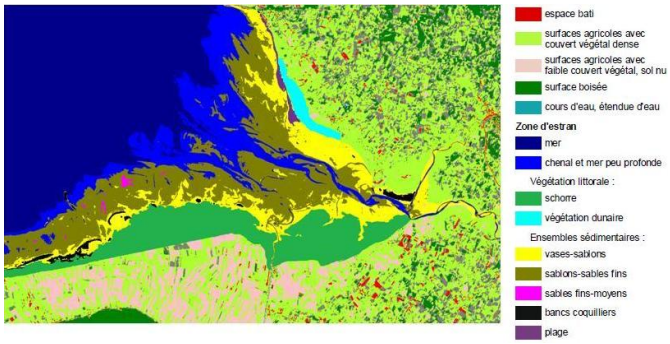


Fig. 5: The final map after the application of the graph matching algorithm and the labeling.

The table below gives a description of the entire geographic objects found and their neighbor, after applying the graph matching algorithm for the SPOT 5 satellite image of the bay of Mount Saint-Michel.

Object	***Neighbors***
Sea	Channel and shallow sea
Channel and shallow sea	Sea fine-sand sand fine sand middleweight Vases shellfish beds rivage
Fine sand	Channel sands and deep sea middleweight sand Vases shellfish beds
Means sand	Channel and deep sea fine sand Vases shellfish beds
Vases and fine sand	Channel and deep sea fine sand dune beach means shellfish beds Shore Woody Bare soil surface Agricultural dense built
Shellfish beds	Channel and deep sea fine sand middleweight Vases and fine sand Shore
Beach	Shellfish beds sand middleweight Vases and fine sand Shore Build
Dune	Vases range fine sand shore Bare soil Agricultural dense built
Shore	Channel and deep sea vases fine sand dune shellfish beds built Sol naked Agricultural dense
Watercourse	Bare soil surface beach Agricultural dense
Surface Woody	Vases fine sand rivers bare soil Agricultural dense built
Bare soil	Vases fine sand dune shore rivers Woody Agricultural dense built area
Dense agricultural	Vases fine sand dune shore rivers Woody Bare soil surface mount
Built	Vases range fine sand shore dune woody sol dense agricultural

Table 3: Table of knowledge of the entire geographic components with their neighborhoods of the bay of Mount Saint-Michel. according to the results obtained in Fig. 5.

6. Conclusion

In this work, we deal with the detection and the representation of differents geographic patterns of the bay of Mount Saint-Michel. Two different segmentation methods are used: the first one based on a supervised algorithm using set of geographic patterns for the learning phase and the secound is unsupervised algorithm. Both of this algorithms can be used separately or in parallel for mapping the coastal area. All the patterns are represented in the form of adjacency graph for helping the clustering with the introduciotn of spatial information in the classification process. Mapping and the labelling of all geographic patterns are performed from the model graph built from the results of supervised classification and the data graph obtained from the unsupervised classification. Our algorithm was tested with set of optical SPOT 5 images with different resolutions. This study provide more than 14 differents geographic objetc with their neighbor and can be used for the monitor of the bay of Mount Saint-Michel from multitemporal images.

References

[1] S. Bouraoui. A system to detect residential area in multispectral satellite images. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 2, November 2011.

[2] D. Guo, H. Xiong, V. Atluri, and N.R. Adam. Object discovery in high-resolution remote sensing images : a sementic perspective. Knowledge and Information Systems, 19(2) : 211-233, 2009.
[3] J. Inglada and J. Michel. Qualitative spatial reasoning for high-resolution remote sensing images analysis. Geoscience and remote sensing, IEEE Transactions, 47(2) :599-612, Feb. 2009.
[4] E. Guray and L. Nicola. Automatic Learning of Structural Models of Cartographic Objects. GbRPR 2005, LNCS 3434, Poitiers,France, pp. 273-280,April 2005.
[5] J.W. Rouse,R.H.Haas,J.A.Schell, D.W. Deering,J.C Harlan, Monitoring the vernal advancement of natural vegetation, Final report, NASA/GCSFC, Greenbelt, MD, 1974. [6] J.W. Rouse,R.H.Haas,J.A.Schell, D.W. Deering,J.C Harlan, Monitoring the vernal advancement of natural vegetation, Final report, NASA/GCSFC, Greenbelt, MD, 1974.

A Novel Malicious Web Crawler Detector: Performance and Evaluation

DeXiang Zhang[1], DiFan Zhang[2] and Xun Liu[3]

[1] Information and Network Center, Qingdao University
Qingdao, Shandong 266071, China

[2] Information and Network Center, Qingdao University
Qingdao, Shandong 266071, China

[3] Library, Qingdao University
Qingdao, Shandong 266071, China

Abstract

Internet demands a robust and resilient protected communication and computing environment to enable information flows flawlessly with no down time. However, the Internet is exposed to the general public which will lead to loss of sensitive information as well as copyright protected content. To address this issue, in this paper, we proposed two schemes to fight against unwanted automatic web crawlers, TSSNBS (Too Simple Sometimes Naive Blocking Schema) and ABS (Adaptive Blocking Schema). We validated the effectiveness of the two schemes by implementing an advanced integrated crawler detection system and applied on a high trafficked site in the real world, exposed with real attackers.

Keywords: *Crawler, Internet, Network, Algorithm.*

1. Introduction

Internet demands a robust and resilient protected communication and computing environment to enable information flows flawlessly with no down time. Nevertheless, the openness nature of Internet causes security risk of information leakage because the information are accessible to anyone without proper access control. In recent years, much research has been devoted to the construction of web technologies; contrarily, few have investigated the construction of architecture. Fortunately, several access control schemes were introduced including WebDAV [1] and The PLAIN Simple Authentication and Security Layer (SASL) Mechanism [2].

Information flows in the Internet. To better organize the world's information and make it universally available and accessible, crawlers, or known as web spiders, are invented to traverse against the Internet to fetch information [3].

To keep information confidential and prevent automatic crawling programs not behaving normally, we proposed a novel method of detecting unwanted automatic crawlers. However, considering user anonymity and local law requirements, raw logs was never processed without stripping out user information.

To future help minimizing the effectiveness of this technique, we proposed a dynamic blocker to block the malicious request in real time.

2. System Design

2.1 Overview

Crawlers behave significantly different from normal users since they are automated programs with pre-defined routines, thus allowing researchers to use fingerprint based techniques to classify them. Per analysis of the behaviors of several commonly seen crawlers and robots, we concluded several commonly seen patterns. By detecting those patterns, we can figure out malicious traffic effectively.

By utilizing known HTTP and TCP features, active and passive network sensors can be put in the system to monitor those traffic and with HTTP features as well as TCP features, those traffic can be got rid of from the entire system with little computational resource consumption.

2.2 Crawler Pattern Analysis

Most crawlers are not script awareness and are simply traversing against all links found in a page with a fixed

interval. For those crawlers, we found the following patterns and are surprisingly high performing in detection.

(1) Continuous Requests: Many crawlers are programmed to parse an entry page, extract links in the entry page and visit each link immediately or after a fixed or random interval. For robots, in order to fetch the whole site as fast as possible, the interval is likely to be short. Regardless of the interval, in the access log, we can observe consequent and continuous requests. By defining an adequate threshold of visiting the site, we can figure out possible crawlers [4,5].

(2) Not Accepting Cookies: Since HTTP is stateless, to keep state of the user, cookies are used. However, due to the nature of crawlers which is stateless, it does not keep cookies sent from the server. Thus, requests from the same or similar (in the same C class) IP address which never send cookies information can be very suspicious.

(3) Bogus User Agents: Users cannot access the Internet directly. Instead, users use User Agents. Most commonly seen user agent is web browser. All user agents use an user agent string to identify itself. All browsers will send out User Agent information. However, many crawlers are omitting user agents; others are simply identifying themselves as crawlers or very old browsers including Internet Explorer 3.0 running on Windows 95 or Netscape4.78 on Solaris. Since those old browsers are not capable for the current Internet, we can safely define a blacklist of user agent or even use machine learning algorithms to automatically generate a white-list.

(4) Not Loading/Executing Scripts: Opposed to web browsers which has integrated scripting engine (mostly ECMAScript interpretation engine, whether fully functional and complying with standards or not), spiders are not equipped with scripting engines in most cases for simpler implementation and faster execution. Thus, by putting pitfalls and triggers in the source code, we may be able to implement traps for web spiders and automated bots. However, considering the instability nature of the Internet, thresholds should be set and timeouts should be available [6].

(5) High Fetch Rates: Another common approach in implementing web spiders and crawlers is to fetch pages as fast as possible. However, normal users tend to load several pages at a time, read the pages and load another batch of pages after a relatively long period.

2.3 Blockage of Requests

After detecting malicious traffic, traditionally, we implement firewall rules or system configuration rules to block the traffic. However, this requires human involvement and thus is offline. Also, it requires much human invocation. To address such issues, according to RFC2616 [7], we first use standard HTTP error responses, and then use connection based blockages.

(1) HTTP Error Message based Blockage: To prevent requests being made successfully, we use HTTP error messages. By returning a 40x error, the client won't be able to receive any useful message and the application server won't even receive such request, saving much system resources and traffic. We used a customized HTTP error message, code 444, as a respond. Such respond totally ignores the request and returns nothing but the HTTP header.

(2) Connection based Blockage: For most malicious requests, HTTP Error Message based blockage should be enough. However, more some DDoS aimed malicious requests, even returning HTTP Error Messages help to take down the server. Hence, we have connection based blockage. In such blockage mode, we maintain a table of blocked internet protocol addresses and scan this table every time accepting a new connection.

2.4 Strategies against Malicious Crawler

We proposed two different strategies in blocking malicious crawlers, TSSNBS (Too Simple Sometimes Naive Blocking Schema) and ABS (Adaptive Blocking Schema) to block malicious crawlers. We use TSSNBS as the primitive decision making algorithm, and for unclear crawlers, we use ABS which is powerful yet resources consuming[8].

(1) TSSNBS (Too Simple Sometimes Naive Blocking Schema): The TSSNBS algorithm aims to provide a quick and relatively inaccurate method to detect malicious web spiders. Thus, we choose to be stricter that ambiguous requests will be marked as malicious. This will increase false positive ratio, however, will block almost all real bots.

For user agents, we have an extended list of known bad keywords and another list of known-to-be-good client list which was represented in Regular Expressions to provide better compatibility and scalability. To provide better performance, we used JIT technologies and cached byte-codes generated from PCRE library.

(2) ABS (Adaptive Blocking Schema): The adaptive blocking schema is a combination of machine learning and advanced metrics mentioned in features of malicious crawlers above. Specifically, it is stateful. For each suspicious connection passed from the TSSNBS, it will create a session and keep record of it for an period of time. The session will be served a specially designed trap in JavaScript. Normal user agents will try to execute this script and respond accordingly.

For clients not accepting cookies and not responding the traps in JavaScript correctly, we accumulate a counter which is cleared after a certain period of time. If the counter reached a carefully designed threshold before it expires, we identify them as crawlers.

(3) Blocking Strategies: Initially, when a connection is identified as malicious, we stop it immediately by sending a HTTP error message, and keep a volatile counter with a fixed timeout value. When there's more connecting coming from the same requester, we increase this counter by one. When it reaches a certain threshold, we mark such requester as abuser and pass it to the kernel space driver. This driver will then drop the packet from such requester when it reaches the Ethernet buffer. When it's dropped in the buffer, the traffic will not be noticed in any user land applications.

2.5 Implementation

After several weeks of onerous coding, we finally have a working implementation of our system. The implementation was delivered as three decoupled parts – a web server module to dynamically load block list from shared memory which runs in nginx, a web service module which analyzes and serves designated requests to differentiate bots from users, and a backend server to generate reports and make final decisions.

The implementation has two modes – TSSNBS mode and dual mode. Under TSSNBS mode, only basic rules are applied. Under dual mode, TSSNBS are applied first and for ambiguous traffic, ABS is applied.

The HTTP Error Message based blockage module was implemented as a web server module, and connection based blockage was implemented with UNIX shared memory and net filter. Shared memory was used to provide a simple interface to communicate from user land to kernel space in a reasonably fast fashion.

3. Evaluation

Evaluating complex systems is difficult. Only with precise measurements might we convince the reader that performance matters. We use two metrics to evaluate the malicious crawlers, detection ratio P_d and false positive ratio P_f.

$$P_d = \frac{Count(crawlers\ \det ected)}{Count(crawlers)} \times 100\% \quad (1)$$

$$P_f = \frac{Count(false\ \ positive\ crawlers)}{Count(crawlers\ \det ected)} \times 100\% \quad (2)$$

3.1 Testing Environment

To evaluate the algorithm, we run the evaluation program on a testing platform. The testing platform is a famous technology media focused on mobile applications and in-depth analysis of relevant news. The site has two servers with a load balancing configuration and is running nginx [9], an open source light weight web server as front-end server.

The testing environment runs Debian Linux 6.0.6 with up-to-date patches, Nginx 1.2.4 and Redis 2.0. Redis was chosen as memory-cached temporary storage engine. The version of kernel is *Linux 2.6.32-5-amd64 #1 SMP Sun May 6 04:00:17 UTC 2012 x86_64 GNU/Linux*. The test was conducted on December, 2012.

3.2 Experiment Steps

(1) Deploy Crawler Blocking Plugin: To deploy the implementation, we chose clang compiler *clang version 4.0 (tags/clang-421.0.60) (based on LLVM 3.1svn)*, llvm-gcc *gcc version 4.2.1 (LLVM build 2336.11.00)* and gcc *gcc version 4.2.1 20070831 patched*. Standard UNIX tools including *sed,awk, autotools* and *m4* are also used in the deployment.

(2) Deploying the Blockage Plugin: We deploy the blockage plugin in two places – HTTP Error Message based blockage plugin on the web server, and connection based blockage in the kernel as a kernel extension. The kernel extension was compiled with Sun Studio compiler to provide the best performance.

(3) Running the Plugin: To minimize the impact of the experiment to the site availability, we used DNS rotation, and send about 20% traffic to the testing environment. We run the experiment for about 48 hours, collecting gigabytes of HTTP access log.

3.3 Data Metrics and Procession

We use several metrics to measure the performance and effectiveness of our system.

(1) False Positive: We define false positives as known-to-be-good clients identified as malicious traffic. Those traffic will be blocked once it reaches a certain threshold. Lower false positive ratio indicates better accuracy.

(2) Missed Crawlers: Those are malicious traffic labeled as good by our filter. They are considered harmful.

(3) Load Average: Load average is the metric to measure system utilization. We measure the load average against traffic volume and compare load average for different configurations. Lower load average indicates better performance. In our particular system, load average is calculated by counting context switching count:

$$L = \frac{N_{context\ switch}}{T} \quad (3)$$

We take T = 15 seconds to provide better accuracy and prevent bias resulting from accidental incidents such as garbage collection or memory reclaim.

(4) Data Visualization: We collect raw HTTP logs with several customized fields. However, raw data are never meant to be processed. Hence, we first normalize data and associate it with results from the detector. We use GNUPlot to visualize the data into EPS format. We use GNU GhostScript To distill the visualized results.

4. Discussion

Our evaluation strategy represents a valuable research contribution in and of itself and evaluation strives to make these points clear.

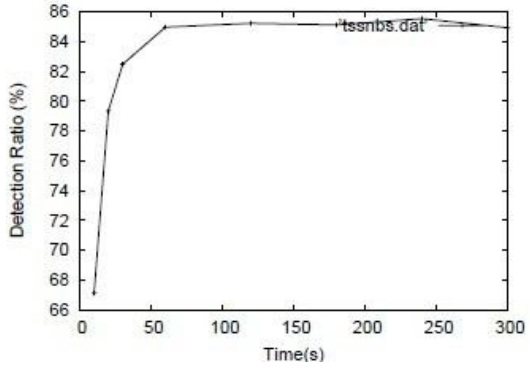


Fig.1 TSSNB Performance – Detection Ratio

4.1 Detection Algorithms

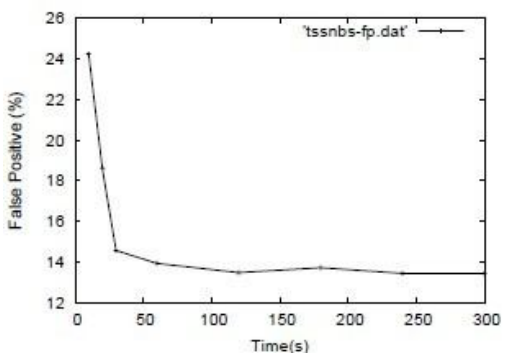


Fig.2 TSSNB Performance – False Positive Ratio

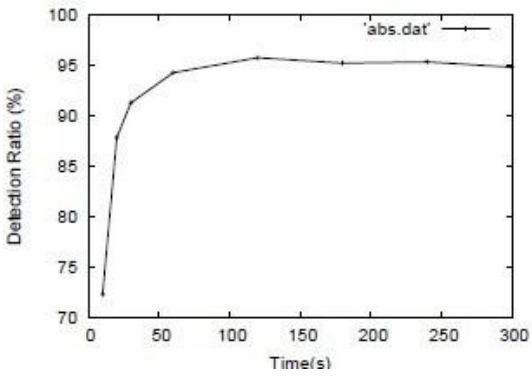


Fig.3 ABS Performance – Detection Ratio

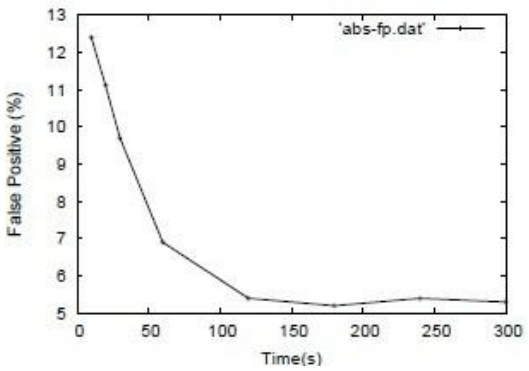


Fig.4 ABS Performance – False Positive Ratio

From Fig.1, we can see clearly that after about 60 seconds, the detection ratio is stabled at about 86%. While a detection ratio of 86% will be sufficient for most applications, for mission critical applications, it is not acceptable. Also, we may lost up to 15% legitimate traffic according to Fig.2.

However, from Fig.3 and Fig.4, we see a great performance improvement which is almost 97% detection ratio and 3% false positive ratio, which means after more optimization and white-listing functionality; it could be applied in production systems.

4.2 Blockage Algorithms

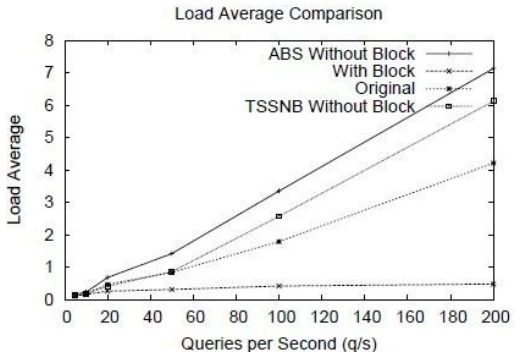


Fig.5 Load Average Comparison

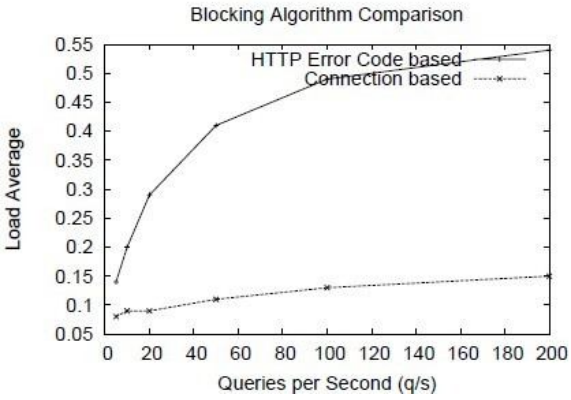


Fig.6 Blocking Algorithm Load Average Comparison

Fig.5 was plotted with data specially collected from known malicious bots. The load average is almost linear to QPS value. Since TSSNB consumes less computational resources, it provides better performance compared with ABS algorithm. However, both of them consumes considerable CPU time and will affect performance of the entire system significantly. However, with the blockage module deployed, the overhead is about 0:2 in load average, which does not affect the performance at all. Also, on peak hours, it will reduce the load average efficiently.

From Fig.6, we can draw two conclusions: 1) both algorithm has an excellent job in controlling load average, 2) HTTP Error Code based blockage still consumes much more computational resources.

Since the HTTP Error Code based blockage requires more memory to run by nature, and since the web server must be invoked to process, it's not surprising that HTTP based algorithm has bad performance. However, as the kernel module requires proprietary software to compile and has to be upgraded on every kernel upgrade, also considering the preliminary implementation lacks security audit, it's possible that buffer overflow may occur, corrupting the whole kernel space and causing downtime. Hence, on not quite heavily loaded sites, HTTP Error Code based blockage should be sufficient.

5.Conclusion

In this paper, we discussed about several approaches of implementing a situation aware anti bot system. After combining several different techniques and algorithms, we reached a good performance.

Even with the most simple algorithm described in this paper, we are able to get fairly acceptable performance. However, with the highly hand crafted and optimized algorithm and implementation, we are able to reduce server load significantly.

To further improve the detection ratio and reduce the false positive rate, we can use a white-list feature to exclude several known suspicious-thus-legitimate clients. Also, we may use SVM and machine learning algorithms to future detect unknown crawlers. In this way, we can archive better performance and will be production ready. We plan to explore more issues related to these issues in future work.

For large scale systems, we may consider implementing the algorithm in FPGA chipsets and deploy it as hardware to have even better performance and easier deployment. Also, we may consider utilizing watermarking techniques to trace the root of the attacks and cooperate with Internet Service Providers and local law enforcements to take down those servers.

Acknowledgments

The testing system was deployed on a famous Chinese IT media, ifanr.com.

References

[1] G. Clemm, J. Reschke, E. Sedlar, and J. Whitehead, "Web distributed authoring and versioning (webdav) access control protocol," 2004.

[2] K. Zeilenga, "The plain simple authentication and security layer (sasl) mechanism," 2006.

[3] Kyle Zeeuwen,Matei Ripeanu and Konstantin Beznosov,"Improving malicious URL re-evaluation scheduling through an empirical study of malware download centers",Proceeding of the 2011 Joint WICOW/AIRWEB Workshop on Web Quality,pages 42-49.

[4] Pang-Ning Tan and Vipin Kumar,"Discovery of Web Robot Sessions Based on their Navigational Patterns",Data Mining and Knowledge Discovery,vol.6,No.1,january 2012,pages 9-35.

[5] Shinil Kwon,Young-Gab Kim and Sungdeok Cha,"Web robot detection based on pattern-matching technique",Journal of Information Science,vol.38,No.2,April 2012,pages 118-126.

[6] Jingyu Zhou and Yu Ding,"An Analysis of URLs Generated from JavaScript Code",Proceedings of the 2012 IEEE/ACIS 11th International Conference on Computer and Information Science,page 688-693.

[7] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, "Rfc 2616, hypertext transfer protocol – http/1.1," 1999. [Online]. Available: http://www.rfc.net/rfc2616.html

[8] Ashif S.Harji,Peter A.Buhr and Tim Brecht,"Comparing high-performance multi-core web-server architectures",Proceeding of the 5th Annual International Systems and Storage Conference.

[9] I. Sysoev. [Online]. Available: http://www.nginx.org.

[10] Basma Zahra, Anis Sakly and Mohamed Benrejeb. Stability Study of Fuzzy Control Processes Application to a Nonlinear Second Order System, International Journal of Computer Science Issues, Vol. 9, No.2-2,(2012) pp. 97-106.

Dexiang Zhang woks in the Information and Network Center of Qingdao University since 2000.He graduated from Qingdao University with bachelor's degree at 2000 and with a master's degree at 2006.His current research interest includes internet security and information security. He published six papers and involved in the preparation of one book.

Difan Zhang is a researcher on Internet Security at Information and Network Center, Qingdao University. Graduated from Towson University with a Master of Science in Information Technology, his current research interest includes Distributed Intrustion Detection Systems and Mobile Device Security.

Xun Liu is a researcher on Internet Security at library of Qingdao University. He graduated from Qingdao University with a master of MPM.

Convergent Projective Non-negative Matrix Factorization

Lirui Hu[1,2,3], Jianguo Wu[1,3] and Lei Wang[1,3]

[1] Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University
Hefei, 230039, China

[2] School of Computer Science and Technology, Nantong University
Nantong, 226019, China

[3] School of Computer Science and Technology, Anhui University
Hefei, 230039, China

Abstract

In order to solve the problem of algorithm convergence in projective non-negative matrix factorization (P-NMF), a method, called convergent projective non-negative matrix factorization (CP-NMF), is proposed. In CP-NMF, an objective function of Frobenius norm is defined. The Taylor series expansion and the Newton iteration formula of solving root are used. An iterative algorithm for basis matrix is derived, and a proof of algorithm convergence is provided. Experimental results show that the convergence speed of the algorithm is higher, however it is affected by the initial value of the basis matrix; relative to non-negative matrix factorization (NMF), the orthogonality and the sparseness of the basis matrix are better, however the reconstructed results of data show that the basis matrix is still approximately orthogonal; in face recognition, there is higher recognition accuracy. The method for CP-NMF is effective.

Keywords: *Non-negative Matrix Factorization, Projective, Convergence, Face Recognition.*

1. Introduction

According to the point of view which perception of the whole is based on perception of its parts, a data technology, called non-negative matrix factorization (NMF) $X \approx WH$ [1], was constructed. The method had revealed the essence of describing data, and it had been widely applied to the fields of data dimension reduction, image analysis, pattern recognition [1, 2], text mining, spectral data analysis [3], and so on. NMF is a current research focus.

Projective non-negative matrix factorization (P-NMF) $X \approx WW^T X$ [4] was proposed based on NMF. Since it was constructed from the projection angle, the basis matrix W was only computed in the algorithm for P-NMF. The computational complexity was lower for one iteration step for P-NMF, as only one matrix had to be computed instead of two for NMF. On the basis of optimization rule $\arg\min_{W\geq 0} \frac{1}{2}\left\|X - WW^T X\right\|_F^2$, the basis matrix W for P-NMF was forced to tend to be orthogonal. So, the orthogonality and the sparseness of the basis matrix were better in P-NMF than in NMF, and then the method for P-NMF was more beneficial to the applications of data dimension reduction, pattern recognition, and so on.

However, the proof of algorithm convergence was not given in the paper [4]. Now, we use the objective function $F = \frac{1}{2}\left\|X - WW^T X\right\|_F^2 (W \geq 0, X \geq 0)$ and the iterative formula

$$W_{ij} = W_{ij} \frac{2(XX^T W)_{ij}}{(WW^T XX^T W)_{ij} + (XX^T WW^T W)_{ij}} \quad (1)$$

in the paper [4] to do an experiment. In Eq. (1), X consists of the first five images of each person in the ORL facial image database, a total of 200 data. We set the rank of the basis matrix W 80 and initialize it with non-negative data. In order to reduce the amount of computation and improve the speed of operation, each image is reduced to a quarter of the original. After 10000 iteration steps, we will see that the varied curve of the objective function values versus iteration steps is severely concussive, and the algorithm does not converge. Here, in order to make the graphics clearly seen, we give the varied curve of objective function values versus iteration steps after 100 iteration steps, and the curve is shown in Fig. 1.

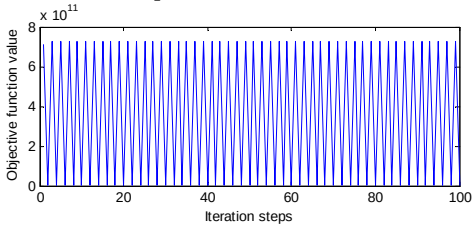


Fig. 1 Objective function values versus iteration steps

In order to solve the problem of algorithm convergence in P-NMF, a method is proposed based on P-NMF $X \approx WW^T X$ in this paper. We call it convergent projective non-negative matrix factorization (CP-NMF). In this method, another iterative algorithm for the basis matrix W is constructed, and strict proof of algorithm convergence is provided, and the convergence speed of the algorithm is higher. Like P-NMF, the orthogonality and the sparseness of the basis matrix are still better in CP-NMF. We compare this method with the methods of NMF, LNMF [5], and NMFOS [6], and the experimental results show that this method has higher recognition accuracy in face recognition.

The rest of this paper is organized as follows. In Section 2, our method is introduced in detail. Firstly, an iterative formula for basis matrix W is derived strictly based on $X \approx WW^T X$. Secondly, a proof of algorithm convergence is provided. Finally, the algorithm steps are given. In Section 3, the convergence of the algorithm is validated by numerical experiments, and it is emphasized that the convergence speed of the algorithm is affected by the initial value of the basis matrix W and the basis matrix W is still approximately orthogonal. Moreover, by numerical experiments in face recognition, we compare this method with NMF and some extended methods in Section 4, and explain the effectiveness of the method. In the end, conclusions are drawn in Section 5.

2. Convergent Projective Non-negative Matrix Factorization (CP-NMF)

We consider an objective function [4]

$$F = \frac{1}{2}\left\| X - WW^T X \right\|_F^2 \qquad (2)$$

where $X \geq 0, W \geq 0.$ Obviously, F is a function defined in P-NMF. We may minimize F to get W .

2.1 The Iterative Rule for Basis Matrix W

For any element w_{ab} of W , let $F_{w_{ab}}$ stand for the part of F relevant to w_{ab} in Eq. (2). So, writing w instead of w_{ab} in the expression of $F_{w_{ab}}$, we may get a function $F_{w_{ab}}(w).$ Obviously, the first order derivative of $F_{w_{ab}}(w)$ at w_{ab} is the first order partial derivative of F with respect to w_{ab} . That is

$$F'_{w_{ab}}(w_{ab}) = \frac{\partial F}{\partial w_{ab}} = \frac{\partial(\frac{1}{2}\sum_{ij}[X_{ij} - (WW^T X)_{ij}]^2}{\partial w_{ab}}$$

$$= \sum_{ij} -(X_{ij} - (WW^T X)_{ij})\frac{\partial (WW^T X)_{ij}}{\partial w_{ab}}$$

$$= \sum_{ij}(-X_{ij} + (WW^T X)_{ij})\frac{\partial(\sum_k (WW^T)_{ik} X_{kj})}{\partial w_{ab}}$$

$$= \sum_{ij}(-X_{ij} + (WW^T X)_{ij})(\sum_k \frac{\partial (WW^T)_{ik}}{\partial w_{ab}} X_{kj})$$

$$= \sum_{ij}(-X_{ij} + (WW^T X)_{ij})(\sum_k \frac{\partial(\sum_l W_{il}W_{kl})}{\partial w_{ab}} X_{kj})$$

$$= \sum_{ij}(-X_{ij} + (WW^T X)_{ij})(\sum_k \frac{\partial (W_{ib}W_{kb})}{\partial w_{ab}} X_{kj})$$

$$= \sum_j \sum_i (-X_{ij} + (WW^T X)_{ij})(\sum_k \frac{\partial (W_{ib}W_{kb})}{\partial w_{ab}} X_{kj})$$

$$= \sum_j [(-X_{aj} + (WW^T X)_{aj})(\sum_k \frac{\partial (W_{ab}W_{kb})}{\partial w_{ab}} X_{kj}) +$$
$$\sum_{i \neq a}(-X_{ij} + (WW^T X)_{ij})(\sum_k \frac{\partial (W_{ib}W_{kb})}{\partial w_{ab}} X_{kj})]$$

$$= \sum_j [(-X_{aj} + (WW^T X)_{aj})(\sum_{k \neq a} \frac{\partial (W_{ab}W_{kb})}{\partial w_{ab}} X_{kj} +$$
$$\frac{\partial (W_{ab}W_{ab})}{\partial w_{ab}} X_{aj}) + \sum_{i \neq a}(-X_{ij} + (WW^T X)_{ij}) W_{ib} X_{aj}]$$

$$= \sum_j [(-X_{aj} + (WW^T X)_{aj})(\sum_{k \neq a} W_{kb} X_{kj} + 2W_{ab} X_{aj}) +$$
$$\sum_i (-X_{ij} + (WW^T X)_{ij}) W_{ib} X_{aj} -$$
$$(-X_{aj} + (WW^T X)_{aj}) W_{ab} X_{aj}]$$

$$= \sum_j [(-X_{aj} + (WW^T X)_{aj})(\sum_{k \neq a} W_{kb} X_{kj} + W_{ab} X_{aj}) +$$
$$\sum_i (-X_{ij} + (WW^T X)_{ij}) W_{ib} X_{aj}]$$

$$= \sum_j [(-X_{aj} + (WW^T X)_{aj})(W^T X)_{bj} +$$
$$\sum_i (-X_{ij} W_{ib} X_{aj}) + \sum_i (WW^T X)_{ij} W_{ib} X_{aj}]$$

$$= -\sum_j X_{aj}(W^T X)_{bj} + \sum_j (WW^T X)_{aj}(W^T X)_{bj}$$

$$-\sum_j X_{aj}(W^T X)_{bj} + \sum_j X_{aj}(W^T W W^T X)_{bj}$$
$$= -(XX^T W)_{ab} + (WW^T XX^T W)_{ab} - (XX^T W)_{ab} + (XX^T WW^T W)_{ab}$$
$$= -2(XX^T W)_{ab} + (WW^T XX^T W)_{ab} + (XX^T WW^T W)_{ab}. \quad (3)$$

Similarly, in order to get the second order derivative of $F_{w_{ab}}(w)$ at w_{ab}, we can get

$$\frac{\partial(-2(XX^T W)_{ab})}{\partial w_{ab}} = -2(XX^T)_{aa},$$
$$\frac{\partial(WW^T XX^T W)_{ab}}{\partial w_{ab}} = (WW^T XX^T)_{aa} + (W^T XX^T W)_{bb} + (XX^T W)_{ab} W_{ab},$$

and

$$\frac{\partial(XX^T WW^T W)_{ab}}{\partial w_{ab}} = (WW^T XX^T)_{aa} + (XX^T W)_{ab} W_{ab} + (XX^T)_{aa}\sum_k W_{kb}^2.$$

So, the second order derivative of $F_{w_{ab}}(w)$ at w_{ab} is

$$F''_{w_{ab}}(w_{ab}) = \frac{\partial(-2(XX^T W)_{ab})}{\partial w_{ab}} + \frac{\partial(WW^T XX^T W)_{ab}}{\partial w_{ab}} + \frac{\partial(XX^T WW^T W)_{ab}}{\partial w_{ab}}$$
$$= -2(XX^T)_{aa} + 2(WW^T XX^T)_{aa} + (W^T XX^T W)_{bb} + 2(XX^T W)_{ab} W_{ab} + (XX^T)_{aa}\sum_k W_{kb}^2. \quad (4)$$

Similarly, in order to get the third order derivative of $F_{w_{ab}}(w)$ at w_{ab}, we can get

$$\frac{\partial(XX^T)_{aa}}{\partial w_{ab}} = 0,$$
$$\frac{\partial(WW^T XX^T)_{aa}}{\partial w_{ab}} = (XX^T)_{aa} W_{ab} + (XX^T W)_{ab},$$
$$\frac{\partial(W^T XX^T W)_{bb}}{\partial w_{ab}} = 2(XX^T W)_{ab},$$
$$\frac{\partial(XX^T W)_{ab} W_{ab}}{\partial w_{ab}} = (XX^T W)_{ab} + (XX^T)_{aa} W_{ab},$$

and

$$\frac{\partial((XX^T)_{aa}\sum_k W_{kb}^2)}{\partial w_{ab}} = 2(XX^T)_{aa} W_{ab}.$$

So, the third order derivative of $F_{w_{ab}}(w)$ at w_{ab} is

$$F'''_{w_{ab}}(w_{ab}) = 6(XX^T W)_{ab} + 6(XX^T)_{aa} W_{ab}. \quad (5)$$

Similarly, in order to get the fourth order derivative of $F_{w_{ab}}(w)$ at w_{ab}, we can get

$$\frac{\partial(XX^T W)_{ab}}{\partial w_{ab}} = \frac{\partial(\sum_k (XX^T)_{ak} W_{kb})}{\partial w_{ab}} = (XX^T)_{aa}$$

and

$$\frac{\partial((XX^T)_{aa} W_{ab})}{\partial w_{ab}} = (XX^T)_{aa}.$$

So, the fourth order derivative of $F_{w_{ab}}(w)$ at w_{ab} is

$$F^{(4)}_{w_{ab}}(w_{ab}) = 12(XX^T)_{aa}, \quad (6)$$

and other order derivatives of $F_{w_{ab}}(w)$ with respect to w are

$$F^{(n)}_{w_{ab}}(w) = 0 \quad (7)$$

where $n \geq 5$.

Thus, the Taylor series expansion of $F_{w_{ab}}(w)$ at w_{ab} is

$$F_{w_{ab}}(w) = F_{w_{ab}}(w_{ab}) + F'_{w_{ab}}(w_{ab})(w - w_{ab}) + \frac{1}{2}F''_{w_{ab}}(w_{ab})(w - w_{ab})^2 + \frac{1}{6}F'''_{w_{ab}}(w_{ab})(w - w_{ab})^3 + \frac{1}{24}F^{(4)}_{w_{ab}}(w_{ab})(w - w_{ab})^4. \quad (8)$$

Meantime, to emphasize time of w_{ab} in numerical calculation, we write $w_{ab}^{(t)}$ instead of w_{ab} in the brackets of $F_{w_{ab}}(w)$. So, equation

$$F_{w_{ab}}(w) = F_{w_{ab}}(w_{ab}^{(t)}) + F'_{w_{ab}}(w_{ab}^{(t)})(w - w_{ab}^{(t)}) + \frac{1}{2}F''_{w_{ab}}(w_{ab}^{(t)})(w - w_{ab}^{(t)})^2 + \frac{1}{6}F'''_{w_{ab}}(w_{ab}^{(t)})(w - w_{ab}^{(t)})^3 + \frac{1}{24}F^{(4)}_{w_{ab}}(w_{ab}^{(t)})(w - w_{ab}^{(t)})^4 \quad (9)$$

is gotten from Eq. (8).

Now, we define a function

$$G_{w_{ab}}(w, w_{ab}^{(t)}) = F_{w_{ab}}(w_{ab}^{(t)}) + F'_{w_{ab}}(w_{ab}^{(t)})(w - w_{ab}^{(t)}) + \frac{1}{2}[\frac{(WW^T XX^T W)_{ab} + (XX^T WW^T W)_{ab}}{w_{ab}^{(t)}} + \frac{(W^T XX^T W)_{bb} W_{ab} + 2(XX^T W)_{ab} W_{ab}^2}{w_{ab}^{(t)}} + \frac{((XX^T)_{aa} \sum_k W_{kb}^2) W_{ab}}{w_{ab}^{(t)}}](w - w_{ab}^{(t)})^2 + \frac{1}{6} F'''_{w_{ab}}(w_{ab}^{(t)})(w - w_{ab}^{(t)})^3 + \frac{1}{24} F^{(4)}_{w_{ab}}(w_{ab}^{(t)})(w - w_{ab}^{(t)})^4. \quad (10)$$

Theorem 1. $G_{w_{ab}}(w, w_{ab}^{(t)})$ is an auxiliary function for $F_{w_{ab}}(w)$.

Proof: $G_{w_{ab}}(w, w_{ab}^{(t)}) = F_{w_{ab}}(w)$ is obvious when $w_{ab}^{(t)} = w$. We need show that $G_{w_{ab}}(w, w_{ab}^{(t)}) \geq F_{w_{ab}}(w)$ when $w_{ab}^{(t)} \neq w$.

Because $W \geq 0, X \geq 0,$

$$(WW^T XX^T W)_{ab} = \sum_k (WW^T XX^T)_{ak} W_{kb}^{(t)} \geq (WW^T XX^T)_{aa} W_{ab}^{(t)}$$

and

$$(XX^T WW^T W)_{ab} = \sum_k (XX^T WW^T)_{ak} W_{kb}^{(t)} \geq (XX^T WW^T)_{aa} W_{ab}^{(t)} = (WW^T XX^T)_{aa} W_{ab}^{(t)}.$$

So,

$$(WW^T XX^T W)_{ab} + (XX^T WW^T W)_{ab} \geq 2(WW^T XX^T)_{aa} W_{ab}^{(t)}.$$

When $W_{ab}^{(t)} > 0,$

$$\frac{(WW^T XX^T W)_{ab} + (XX^T WW^T W)_{ab}}{W_{ab}^{(t)}} \geq 2(WW^T XX^T)_{aa}.$$

In fact, $W_{ab} = W_{ab}^{(t)} = w_{ab}^{(t)}$. So

$$\frac{(WW^T XX^T W)_{ab} + (XX^T WW^T W)_{ab}}{w_{ab}^{(t)}} + \frac{(W^T XX^T W)_{bb} W_{ab} + 2(XX^T W)_{ab} W_{ab}^2}{w_{ab}^{(t)}} + \frac{((XX^T)_{aa} \sum_k W_{kb}^2) W_{ab}}{w_{ab}^{(t)}} \geq -2(XX^T)_{aa} + 2(WW^T XX^T)_{aa} + (W^T XX^T W)_{bb} + 2(XX^T W)_{ab} W_{ab} + (XX^T)_{aa} \sum_k W_{kb}^2 = F''_{w_{ab}}(w_{ab}^{(t)}),$$

and then

$$G_{w_{ab}}(w, w_{ab}^{(t)}) \geq F_{w_{ab}}(w).$$

Thus, $G_{w_{ab}}(w, w_{ab}^{(t)})$ is an auxiliary function for $F_{w_{ab}}(w)$ according to the definition 1 of reference [7].

Theorem 2. $F_{w_{ab}}(w)$ is nonincreasing under the update

$$w_{ab}^{(t+1)} = \arg\min_w G_{w_{ab}}(w, w_{ab}^{(t)})$$

Proof: Because

$$F_{w_{ab}}(w_{ab}^{(t+1)}) \leq G_{w_{ab}}(w_{ab}^{(t+1)}, w_{ab}^{(t)}) \leq G_{w_{ab}}(w_{ab}^{(t)}, w_{ab}^{(t)}) = F_{w_{ab}}(w_{ab}^{(t)}),$$

$F_{w_{ab}}(w)$ is nonincreasing.

Using the definition of auxiliary function and Theorem 2, we can get the local minimum of $F_{w_{ab}}(w)$ if only the local minimum of $G_{w_{ab}}(w, w_{ab}^{(t)})$ is gotten. To get a local minimum of $F_{w_{ab}}(w)$, we may calculate the first order derivative of $G_{w_{ab}}(w, w_{ab}^{(t)})$ with respect to w, and have

$$G'_{w_{ab}}(w, w_{ab}^{(t)}) = F'_{w_{ab}}(w_{ab}^{(t)}) + [\frac{(WW^T XX^T W)_{ab} + (XX^T WW^T W)_{ab}}{w_{ab}^{(t)}} + \frac{(W^T XX^T W)_{bb} W_{ab} + 2(XX^T W)_{ab} W_{ab}^2}{w_{ab}^{(t)}} + \frac{((XX^T)_{aa} \sum_k W_{kb}^2) W_{ab}}{w_{ab}^{(t)}}](w - w_{ab}^{(t)}) + \frac{1}{2} F'''_{w_{ab}}(w_{ab}^{(t)})(w - w_{ab}^{(t)})^2 + \frac{1}{6} F^{(4)}_{w_{ab}}(w_{ab}^{(t)})(w - w_{ab}^{(t)})^3. \quad (11)$$

In order to get the root of the equation

$$G'_{w_{ab}}(w, w_{ab}^{(t)}) = 0, \quad (12)$$

we have known that the function $G'_{w_{ab}}(w, w_{ab}^{(t)})$ is a Taylor series expansion with respect to w from Eq. (11), and then may use the Newton iteration formula of solving root to get the root of Eq. (12). That is

$$w = w_{ab}^{(t)} - \frac{G'_{w_{ab}}(w_{ab}^{(t)}, w_{ab}^{(t)})}{G''_{w_{ab}}(w_{ab}^{(t)}, w_{ab}^{(t)})}. \quad (13)$$

In Eq. (13),

$$G'_{w_{ab}}(w_{ab}^{(t)}, w_{ab}^{(t)}) = F'_{w_{ab}}(w_{ab}^{(t)})$$

and

$$G''_{w_{ab}}(w_{ab}^{(t)}, w_{ab}^{(t)}) = \frac{(WW^T XX^T W)_{ab}}{w_{ab}^{(t)}} + \frac{(XX^T WW^T W)_{ab} + (W^T XX^T W)_{bb} W_{ab}}{w_{ab}^{(t)}} + \frac{2(XX^T W)_{ab} W_{ab}^2 + ((XX^T)_{aa} \sum_k W_{kb}^2) W_{ab}}{w_{ab}^{(t)}}.$$

Using Eq. (3), we simplify the Eq. (13) to

$$w = \frac{M}{D} \quad (14)$$

where

$$M = [2(XX^T W)_{ab} + (W^T XX^T W)_{bb} W_{ab} + 2(XX^T W)_{ab} W_{ab}^2 + ((XX^T)_{aa} \sum_k W_{kb}^2) W_{ab}] w_{ab}^{(t)} \quad (15)$$

and

$$D = (WW^T XX^T W)_{ab} + (XX^T WW^T W)_{ab} + (W^T XX^T W)_{bb} W_{ab} + 2(XX^T W)_{ab} W_{ab}^2 + ((XX^T)_{aa} \sum_k W_{kb}^2) W_{ab}. \quad (16)$$

So, the iterative rule of w_{ab} is

$$w_{ab}^{(t+1)} = \frac{M}{D}. \quad (17)$$

Because the Newton iteration formula of solving root is convergent, we may use this iterative rule Eq. (17) and make the auxiliary function $G_{w_{ab}}(w, w_{ab}^{(t)})$ local minimum, and thus make the objective function $F_{w_{ab}}(w)$ local minimum. If all elements of W are updated by Eq. (17), the local minimum of the objective function F may be gotten. Therefore, the algorithm converges.

The Eq. (17) is the iterative update rule for the basis matrix W.

2.2 Algorithm Steps

Using Eq. (17), we may get an algorithm to compute the basis matrix W. As follows:

Step1: initialize W and X with non-negative data;
Step2: update W by Eq. (15), Eq. (16) and Eq. (17);
Step3: repeat step2 until algorithm converges.

W_n and W_{n+1} are respectively used to denote the *n*th and *n*+1th iterative result of the basis matrix W. The condition of algorithm convergence is that there is

$$\|W_{n+1} - W_n\|_F^2 < \varepsilon, \forall \varepsilon > 0 \quad (18)$$

for an arbitrarily small positive number ε. In inequality (18), *F* stands for Frobenius norm.

3. Experiments and Analysis

In the following experiments, X consists of the first five images of each person in the ORL facial image database, a total of 200 data. We set the rank of the basis matrix W 80. In order to reduce the amount of computation and speed up the operating speed, every image is reduced to half.

3.1 Algorithm Convergence

In order to make the process of algorithm convergence seen more clearly in the graph, we set the larger ε 0.001, and randomly initialize W with non-negative data. The set precision of algorithm convergence is obtained after 60 iterations. In this case, the varied curve of the objective function values versus iteration steps is shown in Fig. 2. We can see that the convergence speed of the algorithm is higher. In addition, we initialize W with the first two images of each person in the ORL, and do an experiment again. The same precision of algorithm convergence is obtained after 39 iterations. The varied curve is shown in Fig. 3. We can see faster convergence of the algorithm. This shows the initial value of the basis matrix affects the convergence speed of the algorithm. The reason is that the convergence speed of Newton iteration formula is dependent on initial value, and the initial value is close to the root convergence faster.

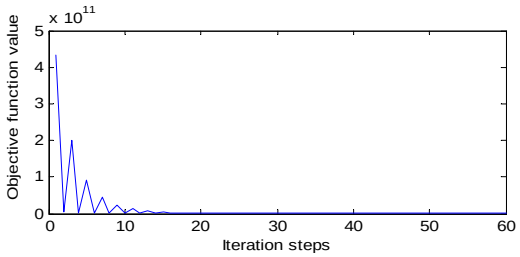


Fig. 2 Objective function values versus iteration steps when the basis matrix is initialized randomly with non-negative data

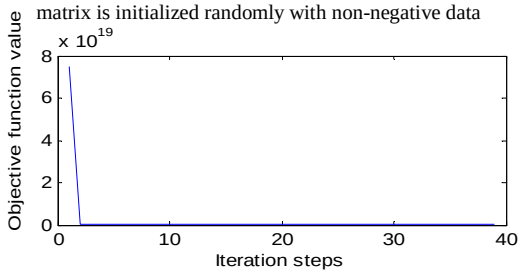


Fig. 3 Objective function values versus iteration steps when the basis matrix is initialized with the known non-negative data

3.2 Analysis of the basis matrix

In the experiment of obtaining Fig. 2, we set the convergent precision ε 0.00001 for the base matrix W while the other set data are unchanged, and do an experiment again. After the algorithm converges, the basis matrix image is shown in Fig. 4. We respectively take the vector $W^T x$ and $(W^T W)^{-1} W^T x$ as the feature vector of the data x and reconstruct x, and reconstructed results are respectively shown in Fig. 6 and Fig. 7.

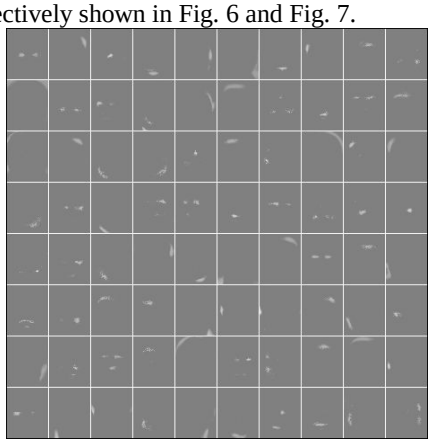


Fig. 4 Basis matrix image

Fig. 5 Original image x; Fig. 6 $W(W^T x)$; Fig. 7 $W(W^T W)^{-1} W^T x$

From the basis matrix image, we can see that the basis matrix is very sparse. This shows that the basis matrix W is forced to tend to be orthogonal by optimizing the objective function F.

From the reconstructed images, we can see that two reconstructed images are all effective, and this shows that the basis matrix W is effective; getting the reconstructed image of x is better by $W(W^T W)^{-1} W^T x$ than by $W(W^T x)$, and this shows that the basis matrix W is still approximately orthogonal, therefore getting the feature vector of data x is better using $(W^T W)^{-1} W^T x$ than using $W^T x$.

The orthogonality and the sparseness of the basis matrix may be computed quantificationally [8, 9]. Without doubt, because this method is still based on the objective function in Eq. (2) for optimization, the orthogonality and the sparseness of the basis matrix are still better. Here, we don't repeat them.

4. Results of Face Recognition and Analysis

In learning phase, X consists of the first five images of each person in the ORL facial image database, a total of 200 data. In order to reduce the amount of computation, and speed up the operating speed, each image is reduced to a quarter of the original. We set ε 0.00002, and initialize randomly the basis matrix W with non-negative data. After the algorithm converges, we get the basis matrix W and feature matrix $(W^T W)^{-1} W^T X$, and take the feature matrix as a template library.

In the pattern recognition test phase, we take the after five images of each person in the ORL facial image database, a total of 200 data, as test data, and reduce every image to a quarter of the original, use $(W^T W)^{-1} W^T x$ to compute the feature vector of test image x by the basis matrix W obtained in the learning phase, and use the nearest neighbor rule for face recognition. We compare this method with the methods of NMF, LNMF, and NMFOS. When the ranks (i.e., the feature subspace dimensions) of the basis matrix are set different values, the results of the face recognition are shown in Fig. 8.

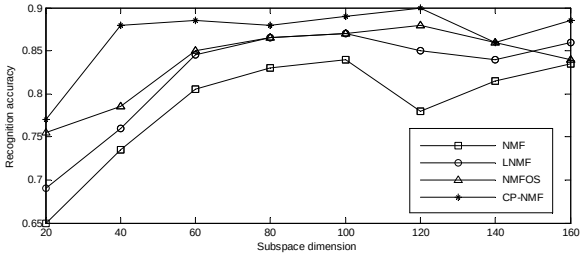


Fig. 8 Comparison of the results of face recognition in the ORL

As can be seen from the Fig. 8, the recognition accuracy is obviously higher using CP-NMF than using NMF. The cause is that the basis matrix W is forced to tend to be orthogonal by the objective function for CP-NMF in Eq. (2) so that the basis matrix is more orthogonal in CP-NMF than in NMF. So the discriminative power of the feature vector $(W^TW)^{-1}W^Tx$ for CP-NMF is better. Meantime, when the rank of the basis matrix is greater than or equal to 60, the recognition accuracy is slightly higher using CP-NMF than using LNMF or NMFOS. This is because there are also approximately orthogonal constraints for the basis matrixes in the objective functions for LNMF and NMFOS so that the discriminative power of the feature vectors is also good. But the discriminative power of the feature vector $(W^TW)^{-1}W^Tx$ for CP-NMF is better.

In addition, when the rank of the basis matrix for CP-NMF is between 40 and 160, the recognition accuracy becomes more stable. This is because the orthogonality and the sparseness of the basis matrix for CP-NMF are always better so that the recognition accuracy is less affected by the number of the rank of basis matrix.

5. Conclusion

In this paper, we propose a method, called convergent projective non-negative matrix factorization (CP-NMF). In CP-NMF, the algorithm steps are given. The convergence speed of the algorithm is higher. Relative to NMF, the orthogonality and the sparseness of the basis matrix are better. Relative to NMF and some extended NMF methods with orthogonal constraints for the basis matrixes in the objective functions, there is higher recognition accuracy in face recognition.

Acknowledgment

This work was supported by Key Technologies Research and Development Program of Chinese Anhui Province under grant No. 07010202057.

References

[1] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, 1999, vol. 401, pp. 788−791.
[2] L. Y. Ma, N. Z. Feng and Q. Wang, “Non-negative matrix factorization and support vector data description based one class classification,” International Journal of Computer Science Issues, 2012, Vol. 9, No. 5, pp. 36−42.
[3] M. W. Berry, M. Browne, A. N. Langville, et al., “Algorithms and applications for approximate non-negative matrix factorization,” Computational Statistics & Data Analysis, 2007, vol. 52, pp. 155−173.
[4] Z. J. Yuan and E. Oja, “Projective nonnegative matrix factorization for image compression and feature extraction,” In: Proceedings of the fourteenth Scandinavian Conference on Image Analysis, 2005, pp. 333−342.
[5] S. Z. Li, X. W. Hou, H. J. Zhang, et al., “Learning spatially localized, parts-based representation,” In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2001, pp. 1−6.
[6] Z. Li, X. Wu and H. Peng, “Non-negative matrix factorization on orthogonal subspace,” Pattern Recognition Letters, 2010, vol. 31, pp. 905−911.
[7] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” In Advances in Neural Information Processing Systems 13, MIT Press, 2001.
[8] L. Li and Y. J. Zhang, “Linear projection-based non-negative matrix factorization,” Acta Automatica Sinica, 2010, vol. 36, pp. 23−39.
[9] Z. R. Yang, Z. J. Yuan and J. Laaksonen, “Projective nonnegative matrix factorization with applications to facial image processing,” Intenational Journal of Pattern Recognition and Artificial Intelligence, 2007, vol. 21, pp. 1353−1362.

Lirui Hu was born in Xiangtan, China, in November 1966. He received his Master degree in 2000 in mathematics from Guizhou University, Guizhou China. Presently, he is a doctoral candidate in computer application technology at the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei China. He is an associate professor at School of Computer Science and Technology at Nantong University, Nantong China. His research interests include image processing, pattern recognition and machine learning.

Jianguo Wu was born in Suzhou, China, in August 1954. He received his Doctor degree in 1998 from Beijing Institute of Technology, Beijing China. He is a professor at School of Computer Science and Technology at Anhui University, Anhui China. His research interests include Chinese information processing, image processing and pattern recognition.

Lei Wang was born in Xuancheng, China, in March 1987. Presently, he is a master’s candidate in computer application technology at the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei China. His research interests include image processing and pattern recognition.

Job Scheduling Model for Cloud Computing Based on Multi-Objective Genetic Algorithm

Jing Liu[*1], Xing-Guo Luo[2], Xing-Ming Zhang[3], Fan Zhang[4] and Bai-Nan Li[5]

[1,2,3,4,5] **National Digital Switching System Engineering & Technology Research Center, Zhengzhou 450002, China**

Abstract

Cloud computing is an emerging high performance computing environment with a large scale, heterogeneous collection of autonomous systems and flexible computational architecture. To improve the overall performance of cloud computing, with the deadline constraint, a task scheduling model is established for reducing the system power consumption of cloud computing and improving the profit of service providers. For the scheduling model, a solving method based on multi-objective genetic algorithm (MO-GA) is designed and the research is focused on encoding rules, crossover operators, selection operators and the method of sorting Pareto solutions. Based on open source cloud computing simulation platform CloudSim, compared to existing scheduling algorithms, the results show that the proposed algorithm can obtain a better solution, and it provides a balance for the performance of multiple objects.

Keywords: *Task Scheduling, Cloud Computing, Multi-Objective Genetic Algorithm, CloudSim.*

1. Introduction

Cloud computing, the long-held dream of "computing as a utility", is emerging as a new paradigm of large-scale distributed computing driven by economies of scale, in which a pool of highly scalable, heterogeneous, virtualized, and configurable and reconfigurable computing resources (e.g., networks, storage, computing units, applications, data) can be rapidly provisioned and released with minimal management effort in the data centers [1-6]. Economically, the main appeal of cloud computing is that customers only use what they need, and only pay for what they actually use. Resources are available to be accessed from the cloud at any given time, and from any location via the internet [7]. However, data centers use a significant and growing portion of energy, an average data center consumes as much energy as 25,000 households. Therefore, energy-aware computing is crucial for cloud computing systems that consume considerable amount of energy.

The resource demands for different jobs fluctuate over time. Job scheduling system, which efficiently allocates resources to required tasks under the constraint of the Service Level Agreements (SLAs), is a fundamental issue in achieving high performance in cloud computing and of great significance for improving resource load balance, security, reliability and reducing energy consumption of the whole system. However, it is a big challenging problem for efficient scheduling algorithm design and implementation in cloud computing environment.

To reduce the energy consumption, Pinheiro et al. propose a model for minimization of power consumption in a heterogeneous cluster of computing nodes serving multiple web-applications, which periodically monitors the load of resources and makes decisions on switching nodes on/off to minimize the overall power consumption [8]; Raghavendra et al. combine five different power management policies and explore the problem in terms of control theory, but the system fails to support variable SLAs for different applications [9]; Lee et al. propose two algorithms based on pricing model, using processor sharing in order to balance between profit and resource utilization [10]; Gang et al. propose a linear programming driven genetic algorithm, aiming to establish the best scheduler in a utility grid by minimizing the combined costs of all users in a coordinated way [11].

All of the above mentioned methods consider the profit or the energy in their study, but do not the relationship between them. To overcome the deficiencies of the above algorithms, in this paper, we first establish a macroscopic scheduling model with cognition and decision components for the cloud computing, which considers both the requirements of different jobs and the circumstances of computing infrastructure, then propose a job scheduling algorithm based on Multi-Objective Genetic Algorithm (MO-GA), taking into account of the energy consumption and the profits of the service providers, and providing a dynamic selection mechanism of the most suitable scheduling scheme for users according to the real-time requirements; at last, we take some experiments to validate our design and compare our MO-GA based scheduling model to the traditional ones.

2. Job Scheduling Model

In cloud computing, service requests have heterogeneous resource demands because some services may be CPU-intensive whereas others are I/O-intensive. Cloud resources need to be allocated not only to satisfy Quality of Service (QoS) requirements specified by users via SLAs, but also to reduce energy usage and improve the profits of the service providers.

2.1 Model Architecture

Fig 1 shows the functional architecture of the scheduling model we have established, the detail functions of the main components are introduced as follows:

Request cognition component should be fully aware of the special needs for different businesses, which may include the computing, storage and communication requirements for computing, arrival law and concurrent conditions, security and privacy requirements, QoS of the service and so on;

Service decomposition component decomposes the service request into different level of granularities with different processor preferences. In the next procedure, the task manager will analyze the resource requirements of each granularity, and mapping it on to optimal processors to reach a effective solution.

Task manager is responsible for task status management (start, stop, cancel…), determining the scheduling sequence and resource assignment for the requests and allocating suitable resources to each job under the help of the scheduling algorithm.

Resource cognition component plays the role of managing the available resources, monitoring the performances of resources, dynamic optimization of scheduling strategy and error notification.

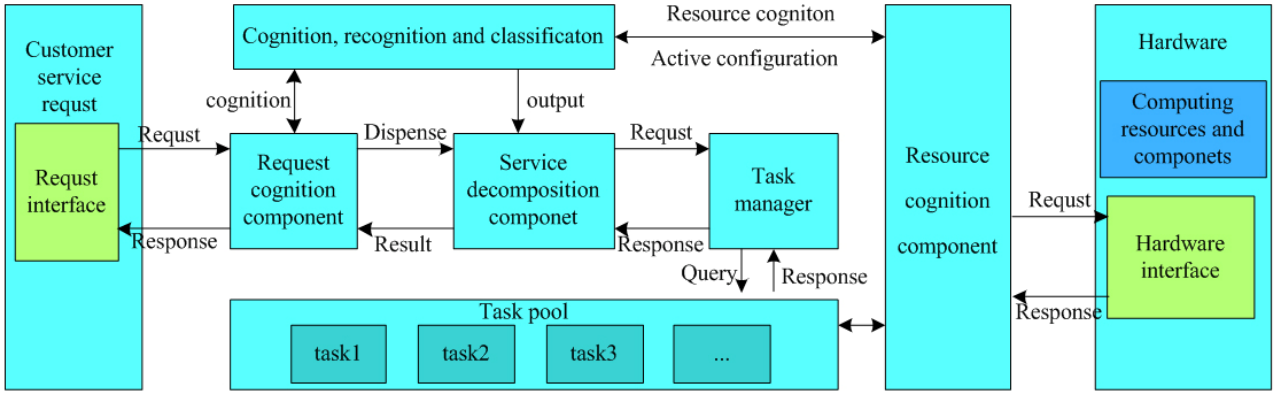


Fig. 1. Functional Architecture of Job Scheduling Model for Cloud Computing

2.2 Problem Formulation

In our model, a cloud application is considered as a collection of work items or jobs that carry out a complex computing task by using cloud resources, and the set $A = (a_1, a_2, \ldots, a_M)$ is a batch of applications arrived in a period. During the scheduling process, the client submits a service request for application $a_i (1 \le i \le M)$, with the resource requirements represented as a triplet (t_i, n_i, d_i), where, t_i represents the reservation time of the application for virtual machines (VMs), which are the virtualized computing elements in cloud computing by means of virtualization technology, n_i for the number of VMs needed for a_i and d_i for the deadline after what the application will be considered to be failed. The problem need to solve for this algorithm is how to schedule these M applications to the given N clouds under the constraints and make the objective function optimal. Where, the N clouds distributed in different geographical areas around the world are usually heterogeneous, while in a cloud, all the VMs are considered homogeneous with the virtualization techniques.

2.3 Objective Function

Suppose application a_i is scheduled to execute on cloud C_j, and p_j represents the Power of each VM in C_j, then, the energy consumption for execution of a_i is given by:

$$E_{ij} = p_j n_i t_i \quad (1)$$

And the profit of the service provider is:

$$R_{ij} = n_i t_i pr - co_{ij} \quad (2)$$

where, the pr is the price unit charged by provider for application a_i, and co_{ij} is the cost of the provider for executing the application a_i.

Combing Eq. (1) and (2), the objective functions can be written as follows:

$$\min E = \min(\sum_{i=1}^{M}\sum_{j=1}^{N} E_{ij}) \quad (3)$$

$$\max R = \max(\sum_{i=1}^{M}\sum_{j=1}^{N} R_{ij}) \quad (4)$$

where, E and R is the total energy consumption and profit for the execution of M application on N clouds respectively.

2.4 Constraints

The constraints are listed as follows:

(1) The application a_i has to be finished before the deadline d_i, otherwise, the schedule is considered to be failed;

(2) Each application can be allocated to only one cloud.

3. MO-GA Scheduling Algorithm

3.1 Encoding Rule

Each schedule is expressed as a 2 by M matrix, where, M is the length of the chromosome. The first row of the matrix represents the requested applications, and second of the matrix is the corresponding number of the cloud where the application is executed. Fig. 2 shows an example of scheduling result, in which, application 2 is assigned to cloud 0, and application 1 is allocated to cloud 5.

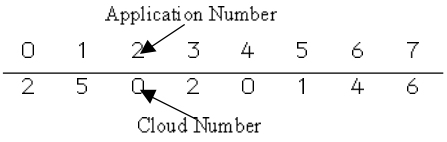


Fig. 2 Encoding example of a Scheduling

According to the above rule, we can see that each application can only be assigned to one cloud, while a cloud may be able to process several applications.

3.2 Population Initialization

The population initialization affects the quality of the future generations, and is an important step in the whole algorithm. In this paper, this step is conducted by combing the random and greedy initialization methods. Owning to the greedy initiation method, the scheduler rejects the applications not meeting the deadline constraint which may cause the whole scheduling failed. This kind of initialization method helps add variety to the initial population and avoid biasing the search of MO-GA.

3.2 Genetic Algorithm

Genetic algorithm is a search heuristic that mimics the process of natural evolution based on a population of candidate solutions. It is routinely used to generate useful solutions to optimization and problems. In the process of evolution, a modification is performed by those operators on each individual. Each chromosome represents a scheduling result, and an evaluation operator (fitness) is called to evaluate the offspring.

(1) Individual Evaluation

In this paper, the fitness is deduced from the energy consumption and profits of the service providers. Only the solutions with the best rank after the evaluation of the fitness function are stored in the Pareto archive which contains the different non-dominated solutions generated through the generations.

(2) Selection operation

The selection operation is based on tournament operator of k individuals, with two strategies: elitism and crowding. The elitism strategy makes use of the individuals in Pareto archive and selects the best ones according to the non-dominated concept to the next generations, allowing the convergence of the evolution process. Crowding strategy takes advantage of crowding distance to estimate the intensity of surrounding solutions and remove the solutions which were too crowded by ranking the crowding distance of each individual. The crowding distance is defined as the circumference of the rectangle defined by its left and right neighbors, and infinity if there is no neighbor.

(3) Crossover Operation

The crossover operator uses two individuals s_1, s_2 to generate two new individuals s_1', s_2'. For individual s_1, first, the operator randomly generates two integers i, j, where, $1 \le i \le j \le N$; then, copies the tasks in s_1 before i and after j to s_1', and maps the tasks between i and j to a temporary individual s_1^m according to the tasks allocation result in s_2; finally, copies the tasks in s_1^m to corresponding place in s_1', as shown in Fig. 3. The individual s_2' is generated using the same method.

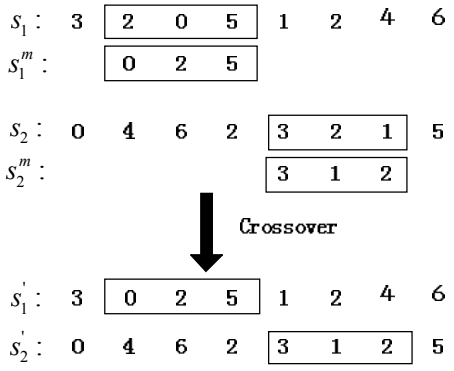


Fig. 3 The crossover operation mechanism

(4) Mutation Operation
The mutation operation chooses two tasks in a individual randomly, and swaps their allocation position to generate a new individual.

3.3 Optimal Selection in Pareto Archive

The results of MO-GA algorithm are a set of Pareto solutions, providing a wide range of possible options, while reducing the efficiency of scheduling process. In practice, users sometimes need to adjust the degree of preference for a particular objective dynamically. This step provides an approach to pick up an optimal solution among the external Pareto archive according to the current requirement. A two dimensional vector is introduced to represent the weighting for a particular objective, whose direction points to the most favorable solution.

Fig. 4 shows an example with 3 two-dimensional vectors, where, p_1 - p_5 represent the external archive of after the MO-GA algorithm, $v_1 = (0,1)$, $v_2 = (\sqrt{2}/2, \sqrt{2}/2)$ and $v_3 = (1,0)$ represents three kind of requirements respectively. For example, p_1 is the optimal solution for vector v_1, and p_3 for v_1, p_5 for v_1.

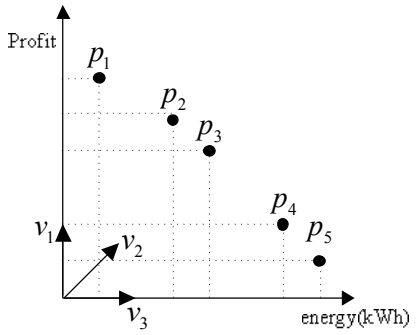


Fig. 4. The schematic diagram of optimal selection

3.4 Implementation Steps

Based on the above ,the implementation steps of this algorithm is listed following:
(1) Initial the population by greedy and random methods;
(2) Modify the individual during the evolution process of the MO-GA algorithm according to the operators indicated in Sec. 3.2, and store the results to external Pareto archive;
(3)Select the optimal solution according to the vector and implement the scheduling result to distributed cloud federation.

4. Simulations and analysis

In order to validate the effectiveness of the proposed algorithm, simulation experiments based on the platform of CloudSim were performed.

4.1 Experimental Settings

The parameter settings in this experiment are as follows

Table 1: Characteristics of the VMs in Each Cloud

NO.	*Power (kW)*	*Frequency(GHz)*	*Memory (GB)*	*Amount*
1	0.28	1.6	1	250
2	0.44	1.8	1	200
3	0.54	2.0	2	150
4	1.04	2.4	2	100
5	1.59	2.6	4	80
6	2.31	2.8	4	60
7	2.43	3.0	4	30
8	3.45	3.2	8	20

(1) Cloud federation parameter. Table 1 shows the characteristics of the VMs which compose the cloud federation in CloudSim. There are eight clouds in this experiment, and the VMs in the same cloud are homogeneous.
(2) Application settings. The request consists of 10000 applications, and the task arrival rates in our experiments are low, medium and high(the high rate has ten times more applications arrival during the same period of time than the medium, so does the medium to low). The length of each task obeys [500, 8000] bimodal distribution, and each has a crest at 2000 and 6000 and near, which is in MI(mega-instructions).
(3) The population size is 20, number of generations is 1000, crossover and mutation rate is 0.95 and 0.1

respectively. The users should pay ¥2 for each VM per hour, and the provider should pay ¥0.5 per kWh for electricity.

4.2 Performance Evaluation

We conduct several experiments and with different parameters of our algorithm. Comparison between our algorithm, maximum applications scheduling algorithm and random scheduling algorithm is listed as following. The maximum applications scheduling algorithm aims to maximize the number of scheduled applications, while the random scheduling algorithm randomly assigns the applications to the cloud. The results of each experiments of the MO-GA have been deduced from 30 independent runs. Table 2 shows the comparison of the MO-GA algorithm with the selection vector $v = (\sqrt{2}/2, \sqrt{2}/2)$ to the other two algorithms, according to the different arrival rates.

Table 2: Experimental comparison for three algorithms

Algorithm	*Arrival Rate*	*Energy (kWh)*	*Profit (¥)*	*Failed Applications*
MO-GA	Low	2340	2988	91
	Medium	2506	2927	130
	High	3875	2846.5	214
Maximum applicatio -ns	Low	4213	2826	105
	Medium	3950	2811.5	159
	High	3904	2796	281
Random	Low	1979	893	2703
	Medium	624.8	251.7	8211
	High	8.2	5.6	9635

We can conclude from the Table 2 that:

(1) Compared to the other algorithms, the MO-GA based scheduling method can obtain a higher profit, while consume lower energy. When the arrival rate is low, the MO-GA methods consumes 44.46% less energy, and obtains 5.73% higher profit, but the improvement reduces with the growth of the arrival rate;
(2) The more the application rate is high, the worse are the results and the higher the number of failed applications is;
(3) There is little difference between the profit and failed applications in MO-GA and maximum applications scheduling algorithm, while the failed applications varies a lot between the two algorithm;
(4) The random scheduling algorithm considers nothing of the resource requirements of the applications and the computing ability of the VMs, thus, poor results are obtained.
(5) With the increase of the arrival rate, more applications are rejected. This is because that all the cloud becomes saturated and busy at the high arrival rate, with no ability to accept the new arrival applications.

To investigate the influence of the selection vector to the MO-GA scheduling result, we conduct an experiment and compare the results of the MO-GA algorithm with three different vectors. The results are listed in Table 3

Table 3: Experimental comparison for three selection vectors

Vector	*Arrival Rate*	*Energy (kWh)*	*Profit (¥)*	*Failed Applications*
$v_1 = (0,1)$	Low	2647	3036.7	93
	Medium	2730	2964.8	131
	High	4070	2856.2	249
$v_2 = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$	Low	2340	2988	91
	Medium	2506	2927	130
	High	3875	2846.5	214
$v_3 = (1,0)$	Low	2334	2980	91
	Medium	2481	2931.4	125
	High	3819	2854	213

From the results, we can see that:
(1) The vector orientations that favor a specific objective obtain a significant improvement on this objective, especially at the low and medium arrival rates.
(2) v_2 corresponds to the scheduling favorite no objective, and the result is also average compared to the other two situations.

4. Conclusions

In this paper, we have established a scheduling model for cloud computing based on MO-GA algorithm to minimize energy consumption and maximize the profit of service provides under the constraint of deadlines. We first propose a job scheduling architecture under the environment of cloud computing, which contains several components to analyze the application, and allocate the suitable resources to the applications to improve the effectiveness and efficiency of the computing; then, the MO-GA based scheduling algorithm is proposed, at last, several experiments are conducted to validate our scheduling models

Acknowledgments

This work was financially supported by the National High-Tech Research and Development Plan of China (2009AA012201), Key Project of Shanghai Science and Technology Commission (08dz501600).

References

[1] Armbrust M, Fox A, Griffith R, Joseph A D, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A and Stoica I, "A view of cloud computing", Communications of the ACM, Vol. 53, No. 4, 2010, pp. 50-58.

[2] Nidhi Jain Kansal and Inderveer Chana, "Cloud. Load Balancing Techniques : A Step Towards Green Computing", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No. 1, 2012, pp. 238-246.

[3] Iosup, A., Ostermann, S., Yigitbasi, M.N., Prodan, R., Fahringer, T. and Epema, D.H.J, "Performance Analysis of Cloud Computing Services for Many-Tasks Scientific Computing", IEEE Transactions on Parallel and Distributed Systems, Vol. 22, No. 6, 2011, pp. 931-945.

[4] Arunadevi.M and R.S.D Wahidabanub, "Design of Power Efficient Schema for Energy Optimization in Data Center With Massive Task Execution Using DVFS", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, 2012, pp. 407-414.

[5] Almutairi, A., Sarfraz M., Basalamah S., Aref W. and Ghafoor A, "A Distributed Access Control Architecture for Cloud Computing", IEEE Software Vol. 29, No. 2, 2012, pp. 36-44.

[6] Junaid Qayyum, Faheem Khan, Muhammad LaL, Fayyaz Gul, Muhammad Sohaib and Fahad Masood, "Implementing and Managing framework for PaaS in Cloud Computing", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No. 3, 2011, pp. 474-479.

[7] Sanjeev Narayan Bal, "Clouds for Different Services", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, 2012, pp. 273-277.

[8] E. Pinheiro, R. Bianchini, E.V. Carrera and T. Heath, "Load balancing and unbalancing for power and performance in cluster-based systems", in Proceedings of the Workshop on Compilers and Operating Systems for Low Power, 2001, pp. 182-195.

[9] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang and X. Zhu, "No ''power'' struggles: coordinated multi-level power management for the data center", SIGARCH Computer Architecture News, Vol. 36, No. 1, 2008, pp. 48-59.

[10] Lee Y.C., Wang, C., Zomaya, A.Y. and Zhou B.B., "Profit-driven service request scheduling in clouds", In: Cluster, Cloud and Grid Computing (CCGRID), 2010, pp. 15-24.

[11] Garg, S.K., Konugurthi P. and Buyya R, "A linear programming driven genetic algorithm for meta-scheduling on utility grids", International Journal of Parallel Emergent and Distributed Systems, Vol. 26, No. 6, 2011, pp. 493-517.

Jing Liu is a Ph.D. student at National Digital Switching System Engineering & Technology Research Center, China. He has completed his Bachelor's and Master's degrees in Information Engineering University, Zhengzhou, Henan province. He is actively involved in research on resource management in virtualized data centers for Cloud computing, job scheduling and PSS.

Xing-guo Luo is Professor of National Digital Switching System Engineering & Technology Research Center, China. His interests include cloud computing and wireless communication.

Xing-ming Zhang is Professor of National Digital Switching System Engineering & Technology Research Center, China. His interests include cloud computing and Network on a Chip.

Fan Zhang is Lecturer of National Digital Switching System Engineering & Technology Research Center, China. His interests include cloud computing and Network on a Chip.

Bai-nan Li is a Ph.D. student at National Digital Switching System Engineering & Technology Research Center, China. His interests include cloud computing and resource allocation.

Implementation of Data Mining in Estimating The Growth Of Local Sheep

Aan Kardiana[1], Lilis Khotijah[2]

[1] Faculty of Information Technology, YARSI University
Jakarta, 10510, Indonesia

[2] Faculty of Animal Science, Bogor Agricultural University
Bogor, 16680, Indonesia

Abstract

Data mining is a process to use statistical technique, mathematics, artificial intelligence, and learning machine to extract, identify beneficial information and discovery knowledge from database. In this research, the authors apply this method to estimate the growth of local sheep. Research method consists of several phases, namely: Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evolution and Knowledge Presentation. Data as amount of 4357 samples, processed by using CART (Classification and Regression Tree) and Correlation Analysis method. The Average Daily Gain is target variable is and indicator variable consist of dry matter intake from : Grass; Corn; Cassava Meal; Coconut Meal; $CaCO_3$; Salt; Premix; Urea; Corn Oil; Corncob; Soybean Meal; Fish Meal and Sunflower Oil. The knowledge presentation gotten is Coconut Meal as dominant indicator variable. The optimal regression trees that has 41 terminal nodes with relative error of 0,659, can be used to determine composition ingredient base on daily gain expected.

Keywords: Data mining, regression tree, estimation, average daily gain

1. Introduction

Beef Self Sufficient Program 2010 is a government programs to supply animal protein in order to feed security. Until now, beef production ability is just able to give contribution around 70-75% from national needs, whereas government launch beef production role can give contribution around 90-95% from national needs. Mutton and lamb in Indonesia only reaches 0.24 g. It is still very low than in several other countries, such as German of 3.33 g, Russia of 3.36 g, and China of 6.36 g. Those numbers will increase continuously in line with the increase of population and awareness level upon the importance of animal protein for nation intelligence [1].

Efforts to increase sheep role as contributor of qualified animal protein source is significantly determined by its productivity level. A lot of researches in livestock field have been done to discover sheep potency and increase that productivity. Therefore, there are many research data collected, but those are not yet utilized optimally now. Current processing method is tabulation and parametric statistic (regression, correlation, and variance analysis). Publication of processing result is still limited only in environmental science farm, whereas many data generated from many researches can be information source not only for livestock field, but can be useful for other related knowledges either directly or indirectly.

Several problems faced are: data generated from researches is quite big so it needs big database; Research results in this field are still partially connected, not yet comprehensively integrated to use for developing livestock sector.

To respond above mentioned problems, new processing methods that can process big data and integrate research results are needed. Another approach that can be used is data mining. This research will apply Data Mining method in collected research data to find valuable hidden information which can be used in developing livestock sector.

2. Literature Study

2.1 Data mining

Turban *et al.* [2] defines data mining as process to use statistical technique, mathematics, artificial intelligence, and learning machine to extract and identify related beneficial information and knowledge from any big database.

Data mining is an essential step in the process of knowledge discovery, consists of an iterative sequence of the following steps [3]:

1. Data cleaning, to remove noise and inconsistent data.
2. Data integration, where multiple data sources may be combined.

3. Data selection, where data relevant to the analysis task are retrieved from the database.
4. Data transformation, where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. Data mining, an essential process where intelligent methods are applied in order to extract data patterns.
6. Pattern evaluation, to identify the truly interesting patterns representing knowledge based on some interestingness measures.
7. Knowledge presentation, where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Larose [4] expresses that the task of data mining are:
1. Description. Simply researcher want to find ways to describe pattern and trend existing in data.
2. Estimation. Estimation model is developed by using complete data that contains value from target variable as prediction value. Then, based on value substitution of prediction variable, it is known that estimation model resulted can known target variabel value. Target variable as numerical.
3. Classification. Classification has categorical target variable.
4. Prediction. Prediction is almost the same with estimation and classification, except in prediction, value from result variable will be exist in the future.
5. Cluster. Constitute data group that has similarity.
6. Association. Find attribute that appear simultaneously.

Data mining task that will be done in this research are description and prediction by using regression tree method.

Breiman *et al.* [5] expresses that regression tree is partitioned by a sequence of binary splits into terminal nodes. In each terminal node t, the predicted response value y(t) is constant. Regression tree formation phases are:
1. Growing the initial tree
   The initial tree is grown through phase:
   a. Select root node.
   b. Determine all splits that might be formed from all indicator variable and calculate homogeneity level.
   c. Select the best indicator variable that has the highest homogeneity level.
   d. Do changing on other branch node.
   e. Stop growing the tree if there is no change on homogeneity level significantly.
2. Determine optimal tree
   The initial tree that has been formed has big size, as a result of using tree formation stop criteria. It is difficult to present the knowledge. To avoid estimation of overfitting, the prunning process is done use the 10-cross validation sample therefore optimal tree is generated.

2.2 Livestock Productivity

Livestock productivity is determined by consumption value of food substance, the increase of body weight and effectiveness to use feed. The increase of weight constitutes ability from animal to change food substances contained in feed to form muscle tissue (meat) that can be known by repeated weighing every day, week or month [6]. Food consumption value is total food consumed by animal if they are given adlib. efficiency to give feed constitutes ration between total feed consumed with total increase of body weight generated [7].

3. Result and Discussion

Data collected is 4357 data [8] used to develop regression tree that use CART (Classification and Regression Tree) method supported by Salford Predictive Modelling (SPM) software issued by Salford System [9].

The Average Daily Gain is target variable, while indicator variables consist of dry matter intake from: Grass; Corn; Cassava Meal; Coconut Meal; CaCO3; Salt; Premix; Urea; Corn Oil; Corncob; Soybean Meal; Fish Meal and Sunflower Oil.

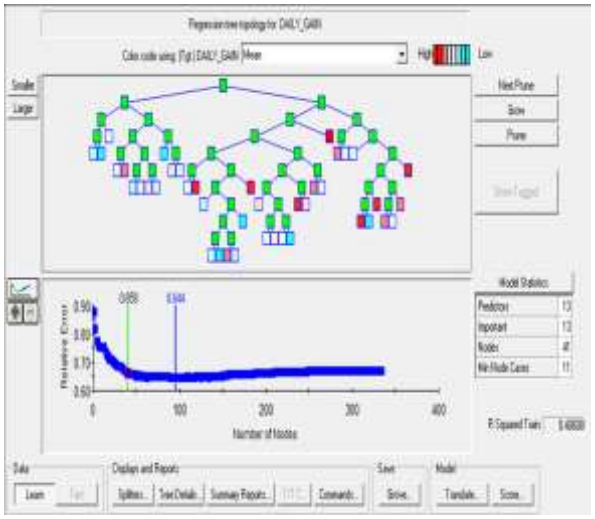


Fig. 1 Regression Tree Toplogi

Figure 1 shows that optimal regression tree has 41 terminal nodes with relative error is 0,659 and involves 10 indicator variables.

The dominant indicator variable is dry matter intake from Coconut Meal. This variable becomes the best split on root node, with the highest Variable Importance and Improvement value among 12 other variables (Figure 2 and Figure 3).

Fig. 2 The Variable Importance

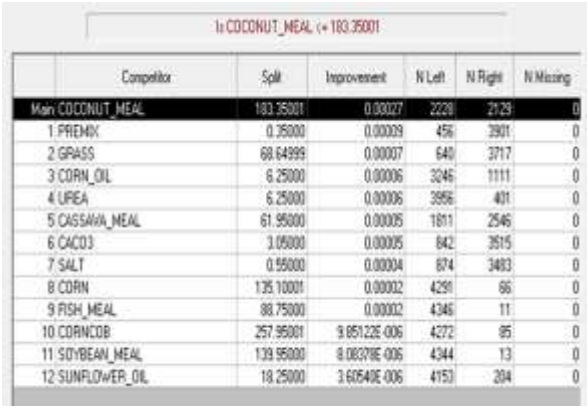


Fig. 3 Root Splits

This result is in line with Pearson Correlation Coefficient between indicator variable with target variable as stated in Table 1. Coconut Meal variable has the highest Pearson Correlation Coefficient value with Average Daily Gain variables (0,405) among other indicator variables.

This is in line also with analysis result that indicates Coconut Meal has the highest protein content among other food material sources, where protein constitutes main food to form tissue in the infancy. This is also in accordance with NRC [10] that one of factors influences average daily gain is total protein consumed everyday.

Root node is splitted by dry matter intake from Coconut Meal variable, if less than or the same with 183,35001 g split to node 2 and if more than 183,35001 g is split to node 3 (Figure 4).

Table 1: Correlation Coeffisien with Indicator Variable and Average Daily Gain

Correlation Coeffisien	*Average Daily Gain*
Coconut Meal	0,405
Grass	0,152
Corn Oil	0,144
Cassava Meal	0,059
Salt	0,054
Fish Meal	0,024
Soybean Meal	0,020
Corn	0,015
Premix	0,014
Sunflower Oil	0,011
Corncob	0,006
CaCO3	0,005
Urea	-0,073

Node 2 and node 3 are developed become next nodes based on splits that has the highest Variable Importance and Improvement value on those nodes.

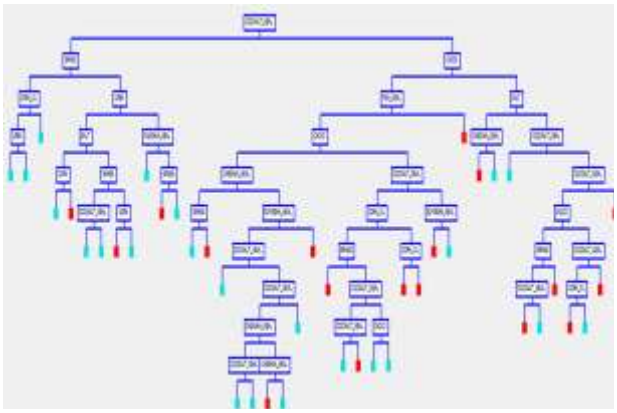


Fig. 4 Optimal Tree

The biggest tree has 335 terminal nodes and after pruning by using 10-Cross Validation, optimal tree is gotten with 41 terminal nodes. Figure 4 indicates whatever indicator variable contained in optimal regression tree that also involved in estimation of body weight increase value as target variable. This tree gives knowledge representation concerning combination of whatever indicator variables that can be used to determine composition ingredient base on daily gain expected.

Recommendation concerning estimation of food material content that can be used in accordance with the increase of

average daily gain expected based on rule on terminal nodes are if dry matter intake value from ingredient:

```
/*Terminal Node 14*/
if
(
   COCONUT_MEAL > 183.35 &&
   FISH_MEAL <= 88.75 &&
   CACO3 <= 3.15 &&
   CASSAVA_MEAL <= 104.5 &&
   GRASS > 151.05
)

{
   terminalNode = -14;
   mean = 0.152231
}

/*Terminal Node 18*/
if
(
   FISH_MEAL <= 88.75 &&
   CACO3 <= 3.15 &&
   SOYBEAN_MEAL <= 139.95 &&
   COCONUT_MEAL > 227.8 &&
   COCONUT_MEAL <= 400.7 &&
   CASSAVA_MEAL > 172 &&
   CASSAVA_MEAL <= 185.05
)

{
   terminalNode = -18;
   mean = 0.122449
}

/*Terminal Node 21*/
if
(
   COCONUT_MEAL > 183.35 &&
   FISH_MEAL <= 88.75 &&
   CACO3 <= 3.15 &&
   CASSAVA_MEAL > 104.5 &&
   SOYBEAN_MEAL > 139.95
)

{
   terminalNode = -21;
   mean = 0.138462
}

/*Terminal Node 27*/
if
(
   FISH_MEAL <= 88.75 &&
   CACO3 > 3.15 &&
   CACO3 <= 6.35 &&
   COCONUT_MEAL > 183.35 &&
   COCONUT_MEAL <= 258.45 &&
   CORN_OIL > 6.55 &&
   CORN_OIL <= 8.3
)

{
   terminalNode = -27;
   mean = 0.1449
}

/*Terminal Node 29*/
if
(
   FISH_MEAL <= 88.75 &&
   CACO3 > 3.15 &&
   CACO3 <= 6.35 &&
   COCONUT_MEAL > 258.45 &&
   SOYBEAN_MEAL <= 29.6
)

{
   terminalNode = -29;
   mean = 0.12377
}

/*Terminal Node 31*/
if
(
   COCONUT_MEAL > 183.35 &&
   CACO3 <= 6.35 &&
   FISH_MEAL > 88.75
)

{
   terminalNode = -31;
   mean = 0.163633
}

/*Terminal Node 35*/
if
(
   SALT > 0.95 &&
   CACO3 > 6.35 &&
   CACO3 <= 6.85 &&
   GRASS <= 133.45 &&
   COCONUT_MEAL > 207.4 &&
   COCONUT_MEAL <= 220.5
)

{
   terminalNode = -35;
   mean = 0.149837
}
```

```
/*Terminal Node 37*/
if
(
   SALT > 0.95 &&
   COCONUT_MEAL > 207.4 &&
   COCONUT_MEAL <= 387.45 &&
   CACO3 > 6.35 &&
   CACO3 <= 6.85 &&
   GRASS > 133.45
)

{
   terminalNode = -37;
   mean = 0.153565
}

/*Terminal Node 40*/
if
(
   SALT > 0.95 &&
   CACO3 > 6.85 &&
   COCONUT_MEAL > 314.6 &&
   COCONUT_MEAL <= 387.45
)

{
   terminalNode = -40;
   mean = 0.130885
}
/*Terminal Node 41*/
if
(
   CACO3 > 6.35 &&
   SALT > 0.95 &&
   COCONUT_MEAL > 387.45
)

{
   terminalNode = -41;
   mean = 0.149706
}
```

That means, if CACO3 > 6.35 g and SALT > 0.95 g and COCONUT_MEAL > 387.45g, so AVERAGE DAILY GAIN will be 149.706 g.

4. Conclusions

From this research we conclude that :

1. Implementation of data mining in estimating the growth of local sheep generates maximum size regression tree that contains 335 terminal nodes.
2. The dominant indicator variable is dry matter intake from Coconut Meal.
3. The optimal regression tree that has 41 terminal nodes with relative error of 0,659 can be used to determine composition ingredient base on average daily gain expected.

Acknowledgments

This study is one of research roadmap of the Faculty of Information Technology YARSI University, and funded by the Directorate General of Higher DIPA Education Ministry of National Education through Grant named "Hibah Unggulan Perguruan Tinggi".

References

[1] Heriyadi D. Domba dan kambing di Indonesia: Potensi, Masalah dan Solusi, Majalah TROBOS No. 101 Februari 2008 Tahun VIII, 2008.

[2] Turban E, Aronson JE, Liang TP. Decision Support Systems and Intelligent Systems, Seventh 7/E, Prentice Hall, 2005.

[3] Han J, Kamber M. Data Mining: Concepts and Techniques, Second Edition, Morgan Kaufmann Publiser, Elsevier, San Francisco, 2006.

[4] Larose DT. Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley & Sons Inc, New Jersey, 2005.

[5] Breiman L, Friedman JF, Olshen RA, Stone CJ. Classification and Regression Tree, Chapman & Hall Inc, New York, 1993.

[6] Tillman AD, Reksohadiprodjo S, Prawirokusumo S, Hartadi H, Lebdosoekojo S. Ilmu Makanan Ternak Dasar, Gadjah Mada University Press, Yogyakarta, 1998.

[7] Parakkasi A. Ilmu Nutrisi Ternak Ruminansia, UI Press, Jakarta, 1995.

[8] INTDK. Kumpulan Data Hasil Penelitian Nutrisi Ternak Domba 2010-2012, Laboratorium Ilmu Nutrisi Ternak Daging dan Kerja, Departemen Ilmu Nutrisi dan Teknologi Pakan, Fakultas Peternakan, IPB, 2012.

[9] Steinberg D, Golovnya M. CART® 6.0 User's Manual, San Diego, CA: Salford Systems, 2006.

[10] NRC. Nutrient Requirements of Small Ruminant. The National Academies Press, Washington Dc, 2006.

Aan Kardiana holds BSc from Bandung Institute of Technology and MSc from IPB (Bogor, Indonesia) in 2000. He is currently an academic, research staff and Head of Computational Intelligent Research Group of Faculty of Information Technology, YARSI University. His research interests are data mining, statistics and e-Health. He is also a member of YARSI E-Health Research Center (YEHRC); has won some research grants; published a number of papers in national proceeding and international journal.

Lilis Khotijah holds BSc and MSc from Faculty of Animal Science from from Bogor Agricultural University (IPB) in 1999. She is currently an academic, lecturer of Animal Nutrition and research staff of Faculty of Animal Science, Bogor Agricultural University. Her research interest is Nutrition Reproduction of Animal.

An Optimal Scheduling Algorithm for Real Time Applications in Grid System

S.Baghavathi Priya[1], T.Ravichandran[2]

[1] Research Scholar, Jawaharlal Nehru Technological University, Hyderabad, Andhra Pradesh, India.
Associate Prof. /IT. Rajalakshmi Engineering College Chennai, Tamil Nadu India.

[2] Principal, Hindustan Institute of Technology, Coimbatore, Tamil Nadu, India.

Abstract

The objective of the proposed work is to use an optimal scheduling algorithm for real-time application. A grid is considered to be an infrastructure that bonds and unifies globally remote and diverse resources in order to provide computing support for a wide range of applications. Real time applications in an industrialized technological infrastructure such as telecommunication systems, factories, defense systems, aircraft and space stations pose relatively rigid requirements on their performance. Aircraft scheduling represents the best example of real-time applications. The main focus of this work is to check the time taken for turn-around activities which comprises of taxi in, load/unload baggage, deboarding, water fueling, cleaning, catering, boarding, de-icing, take off processes, thus relating in the lowest flight delays and shortest waiting time. The optimal scheduling algorithm is used for aircraft take-offs. The penalties are associated with proper scheduling but delayed turn around activities, improper scheduling and early/late takeoffs.

Keywords: *Grid computing, Real-time systems, Task-Scheduling, Turn-around activities, Penalty.*

1. Introduction

The Grid computing refers to a new technology infrastructure paradigm which is based on the Web and the Internet. Grids provide scalable, secure, and reliable mechanisms for discovering and negotiating access to remote resources including clusters. Grid applications involve large amounts of data and/or computing and often require secure resource sharing across the organizational boundaries. Grid computing comprises of a combination of a decentralized architecture for resource management on one hand and a layered architecture of a specific hierarchy for the implementation of various services of the grid. In a grid environment, heterogeneous computer systems located over a large area or around the globe can be integrated and made to appear as a single computational resource, to be optimally or maximally utilized by the user community without any loss or wastage of time, investment or resources.

The consistent decrease in the cost of hardware has led to the employment of computers in many applications. As a consequence, the complexity of modern computer systems has increased. Proportionally, and more effort is needed to maintain the dependability of these computer systems. Systems in which the complexity exists in the dimension of time are called real-time systems. Examples of current real-time systems range from very simple micro controllers in embedded systems to highly sophisticated and complex systems such as air traffic control and avionics. A typical real-time system consists of a controlling system, a controlled system, and the environment. The environment in which a real-time system is to operate plays an important role in the design of the system. Many environments are well defined and deterministic. Many applications will be large, complex, distributed and dynamic; contain many types of timing constraints; need to operate in fault-prone highly nondeterministic environments, and evolve over a long system life time. Issues in real-time systems are resource management, architecture, software. Resource management issues deal with scheduling, resource reclaiming, fault tolerance and communication.

Task scheduling meeting task deadlines is of great importance in real-time systems, because failure to meet task deadlines may result in severe consequences possibly loss of human life. Scheduling of tasks involves the allocation of processes (including resources) and time to tasks in such a way that certain performance requirements are met. The scheduling algorithms have to satisfy not only the timing constraints of tasks, but also the resource constraints and/or precedence constraints among tasks. Similarly, providing predictable inter task communication is also of great significance in real-time systems, because unpredictable delays in the delivery messages can affect the completion time of tasks participating in the message communication. Classical task-scheduling theory typically employs metrics such as minimizing schedule length and the sum of completion times. In static scheduling, the performance requirement is that every task in the system

meets its deadline. In dynamic scheduling, however, the system may be unable to meet the deadlines of all tasks because it lacks knowledge about future task arrivals.
As the history of the field suggest there are many different variants of evolutionary algorithms. The common underlying idea behind all these techniques is the same: given a population of individuals the environmental pressure causes natural selection and this causes a rise in the fitness of the population. Given a quality function to be maximized, a set of candidate solutions should be randomly created. Based on this fitness, some of the better candidates are chosen to seed the next generation by applying recombination and/or mutation to them. Recombination is an operator applied to two or more selected candidates and results one or more new candidates. Mutation is applied to one candidate and results in one new candidate. Executing recombination and mutation leads to a set of new candidates that compete-based on their fitness with the old ones for a place in the next generation. This process can be iterated until a candidate with sufficient quality is found or a previously set computational limit is reached.

The rest of the paper is organized as follows; Section 2 illustrates the related work on real-time applications and new approach. Section 3 describes proposed architecture for aircraft take-off system. Section 4 describes optimal scheduling algorithm for real-time applications. The experimental results are discussed in section 5. Section 6 concludes the paper and presents the future work.

2. Related work and Our New approach

We first review related work on real-time applications in computational Grid. Then, we use evolutionary algorithm for providing optimal solution to complex problem.

2.1 Related Previous Work

Grid Computing concepts are clearly explained in Grid and Cluster computing. Online scheduling approach was proposed for multiple mixed-parallel workflows in grid environments [1]. Resource management issues deal with scheduling, resource reclaiming, fault tolerance and communication, architecture issues involve processor architecture, network architecture and i/o architecture, software issues encompass specification and verification of real-time systems, programming languages and databases. These are the major issues in real-time systems [2]. Task scheduling, preemptive and non-preemptive scheduling, resource reclaiming, static scheduling algorithms and dynamic scheduling algorithms are dealt in deadline scheduling for real-time systems [3]. Resource management in a dynamic real-time system involves both scheduling and reclaiming of resources are explained resource management in real-time systems and networks [4]. Airport logistics, which is a framework of resource management in the air transportation system and focuses on processes supporting turn-around [5]. The future traffic management is the use of planned four-dimensional trajectories (4DTs) for airborne and surface operations at busy airports were explained [6]. The optimization of airport operations is recognized as a challenge that aims at finding the best trade-off solution in order to maximize the airport capacity and minimize both pollution and noise [7]. Evolutionary algorithms (EAs) are stochastic optimization techniques based on the principles of natural evolution. Numerous applications of these techniques were explained [8]. Efficient scheduling of the aircraft for take-off can reduce the total separations and increase throughput. A runway controller is responsible for take-off scheduling. These concepts were explained [9].

2.2 Our New Approach

The scheduling of tasks in grid real-time systems has recently attracted many researchers. The demand for more and more complex real-time applications, which have high computational needs with timing constraint. The first objective is to check the performance of the turn-around activities by calculating the delay and waiting time for completing that particular activity. In most real-time applications, computational complexity is a challenging factor. Priority based scheduling algorithm is used to check the time taken for turn-around activities. Evolutionary algorithms can solve often complex problems. Evolutionary algorithms, such as shuffled frog-leaping optimization algorithm has been used to arrive at near-optimum solutions to complex and large-scale problems. The third objective is to report penalty by using objective function. This function has been used for providing penalty for sub-optimal solutions and invalid results.

3. Grid Scheduling

3.1 The Grid Scheduling Architecture

The Fig.1 shows the architecture and functionalities supported by various units of the grid scheduling system. One of the most complex airport processes is the turn-around process. The turnaround is the collective name for all those activities that affect an aircraft while it is on the ground. One process where most of the actors are involved is the turn-around process. The turn-around process starts when an aircraft touches down and is going on until the aircraft takes off again. Grid agent is to check the

performance of the turn-around activities by calculating the delay and waiting time for completing that particular activity. Global scheduler receives the information about the status of turn-around activity. Scheduled has been prepared according to the information received from Grid Information Service (GIS).GIS provides the information about the workstation. Local scheduler provides the information of the resources (runways) availability to the GIS. If there were no delay in completing the turn-around activity then prepared schedule will be executed. Otherwise the schedule should be revised. If there were no delay in completing the schedule then the aircraft will take-off as per schedule. Otherwise the scheduling delay will be informed to the GIS, in turn it will be reported to the scheduling agent to pay the penalty.

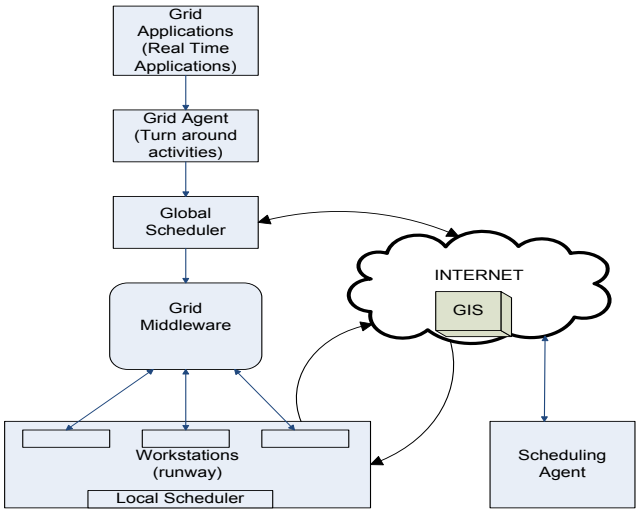


Fig. 1 The Grid Scheduling Architecture.

3.2 Priority Based Scheduling

In priority based scheduling algorithm, each process is assigned and the higher priority processes are scheduled before the lower priority processes. At any point of time, the process having the highest priority among all the ready processes is scheduled first. In case two processes are having the same priority, they are executed in the first come first serve order. The priority scheduling may be either preemptive or non-preemptive. The choice is made whenever a new process enters the ready queue while some process is executing. If the newly arrived process has the higher priority than the currently running process, the preemptive priority scheduling algorithm preempts the currently running process and allocates CPU to the new process. On the other hand, non-preemptive scheduling algorithm allows the currently running process to complete its execution and the new process has to wait for the CPU. In the case of scheduling turn-around activities in air traffic management system, non-preemptive scheduling algorithm is used.

Consider four aircrafts A1, A2, A3, A4 with their required processing time (in milliseconds) and priorities as shown in the following table.

Table 1: Priority Scheduling

Aircrafts	A1	A2	A3	A4
Processing time(ms)	7	4	3	2
Priority	4	3	1	2

The aircrafts will be scheduled for completing turn-around activities as A3, A4, A2, and A1.

A3	A4	A2	A1
0 3	3 2	5 4	9 7 16

Waiting time of A1 = 9 ms Turn around time for A1 = 16ms
Waiting time of A2 = 5 msTurn around time for A2 = 9ms
Waiting time of A3 = 0 ms Turn around time for A3 = 3ms
Waiting time of A4 = 3 ms Turn around time for A4 = 5ms

4. Evolutionary Algorithm

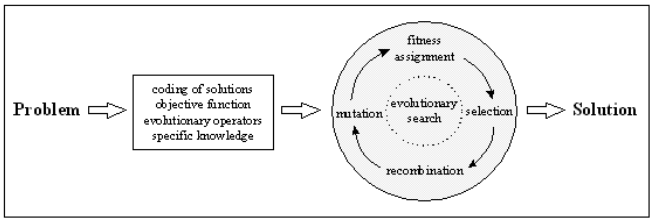


Fig. 2 Problem solution using evolutionary algorithm.

The Fig. 2 shows problem solution using evolutionary algorithm. This section presents the Shuffled Frog-leaping algorithm used in this paper for providing near-optimum solutions.

The Shuffled Frog-Leaping (SFL) algorithm is a memetic metaheuristic that is designed to seek a global optimal solution by performing a heuristic search. It is based on the evolution of memes carried by individuals and a global exchange of information among the population (Eusuff and Lansey 2003). In essence, it combines the benefits of the local search tool of the particle swarm optimization (Kennedy and Eberhart 1995), and the idea of mixing information from parallel local searches to move toward a global solution (Duan et al. 1993). The SFL algorithm has

been tested on several combinatorial problems and found to be efficient in finding global solutions (Eusuff and Lansey 2003).

The SFL algorithm involves a population of possible solutions defined by a set of frogs (i.e. solutions) that is partitioned into subsets referred to as memeplexes. The different memeplexes are considered as different cultures of frogs, each performing a local search. Within each memeplex, the individual frogs hold ideas, that can be influenced by the ideas of other frogs, and evolve through a process of memetic evolution. After a number of memetic evolution steps, ideas are passed among memeplexes in a shuffling process (Liong and Atiquzzaman 2004). The local search and the shuffling processes continue until convergence criteria are satisfied (Eusuff and Lansey 2003). The SFL algorithm is described by the corresponding flowchart in figure 3.

First, an initial population of 'P' frogs is created randomly. For S-dimensional problems, each frog i is represented by S variables as Xi=(xi1, xi 2, ... , xiS). The frogs are sorted in a descending order according to their fitness. Then, the entire population is divided into m memeplexes, each containing n frogs (i.e. P=m6n). In this process, the first frog goes to the first memeplex, the second frog goes to the second memeplex, frog m goes to the mth memeplex, and frog m+1 goes to the first memeplex, and so on.

Within each memeplex (figure 1b), the frogs with the best and the worst fitness are identified as X_b and X_w, respectively. Also, the frog with the global best fitness is identified as Xg. Then, an evolution process is applied to improve only the frog with the worst fitness (i.e. not all frogs) in each cycle. Accordingly, the position of the frog with the worst fitness is adjusted as follows:

$$\text{Change in frog position } (D_i) = \text{rand}()*(X_b - X_w) \quad (1)$$

$$\text{New position } X_w = \text{Current position } X_w + D_i; \quad (2)$$
$$(D_{max} >= D_i >= - D_{max})$$

Where rand ( ) is a random number between 0 and 1; and D_{max} is the maximum allowed change in a frog's position. If this process produces a better frog (solution), it replaces the worst frog. Otherwise, the calculations in equations (1) and (2) are repeated with respect to the global best frog (i.e. X_g replaces X_b). If no improvement becomes possible in this latter case, then a new solution is randomly generated to replace the worst frog with another frog having any arbitrary fitness. The calculations then continue for a specific number of evolutionary iterations within each memeplex (Eusuff and Lansey 2003). The main parameters of the SFL algorithm are: number of frogs P, number of memeplexes, and number of evolutionary iterations for each memeplex before shuffling.

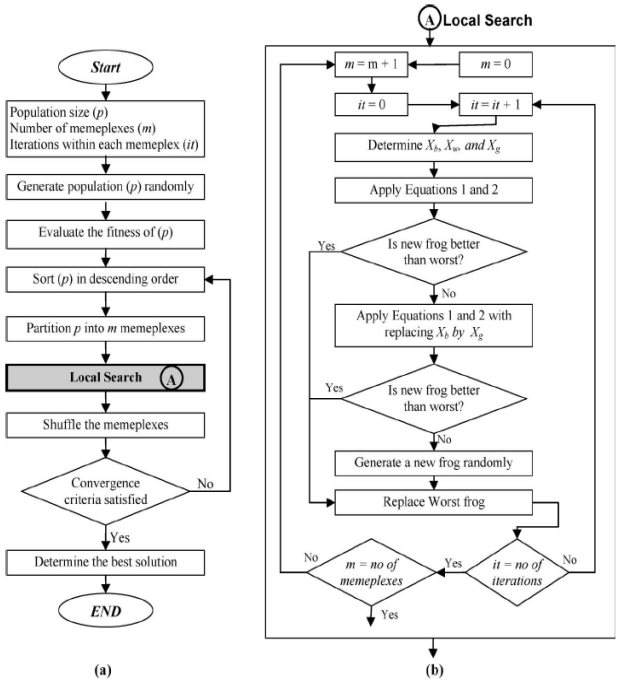


Fig 3. Flowchart of the shuffled frog-leaping algorithm

Penalties help to limit the turn-around activities delay and improper scheduling of an aircraft. Penalties considered in this paper include aircraft delay penalty (P_A), turn-around activities delay penalty (P_{TA}), improper schedule penalty (P_S) of an aircraft. The trade-off situation occurs as the system schedules turn-around activities to maximize proper scheduling while doing fastest turn-around activities for aircraft. The objective function of this model is to minimize total system penalties (P_T) while doing proper scheduling through optimal scheduling algorithm.

The objective function can be expressed as :

$P_T = \alpha P_E + (1-\alpha) P_S$ ---------- (1)

Where

P_E ----- Expected aircraft delay penalty and turn-around activities.

5. Results and Discussions

Gridsim Toolkit 4.0 which allows modeling and simulation entries in grid system.

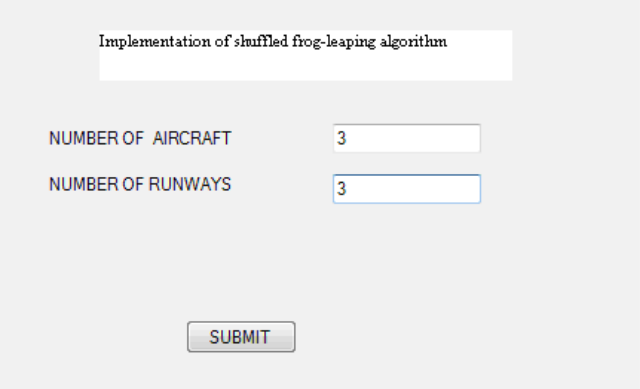


Fig 4. Implementation of an algorithm

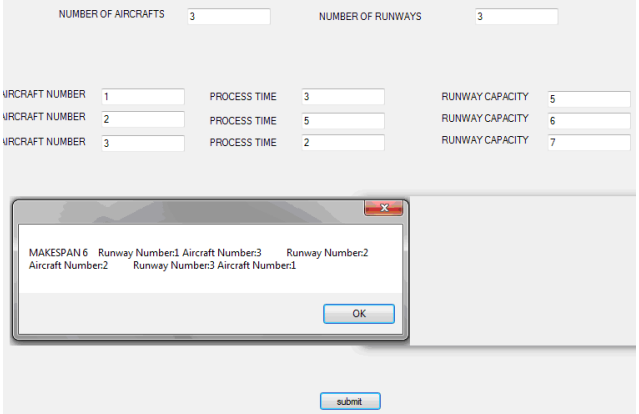


Fig 5. Initial order of assignment

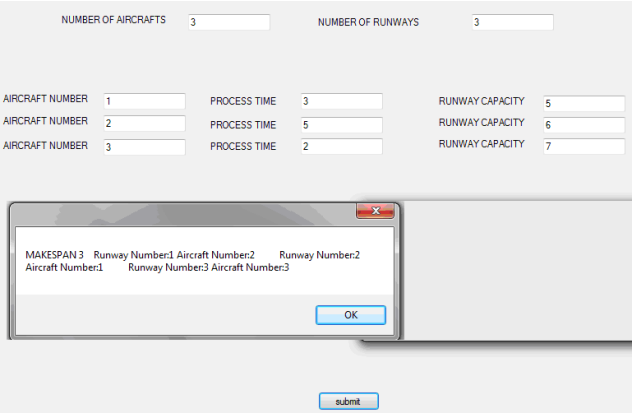


Fig 6. Final order of assignment

6. Conclusions and Future Work

In this paper, we have checked delay and waiting time of turn-around activity. Based on the calculation, scheduling of aircraft will be planned in suitable runway for achieving minimum completion time. The shuffled frog-leaping algorithm was used and implemented for aircraft take-off. This algorithm gives the results of order of aircraft and its suitable runway in a optimized way. Penalty will be calculated and reported to a scheduling agent if any delay in turn-around activity or delay in scheduling. In the future, we plan to bring fault tolerant system in grid real-time applications.

References

[1] C.S.R.Prabhu, "Grid and cluster computing", (Deputy Director General, National Informatics Centre, A.P, Hyderabad) Prentice Hall of India Private Limited 2008.

[2] Jane W.S.Liu, " Real-Time Systems, Pearson, 2000.

[3] John A. Stankovic, Marco Spuri, Krithi Ramamritham, Giorgio C.Buttazzo, "Deadline Scheduling for Real-time systems, Kluwer Academic Publishers, 1998.

[4] C. Siva Ram Murthy and G.Manimaran, "Resource management in real-time systems and networks", PHI Learning Private Limited, 2009.

[5] Anna Norin, Tabias Anderson Granberg, Peter Varbrand&Di yuan, "Integrating Optimization and Simulation to gain more efficient airport logistics", Eighth USA/Europe Air Traffic Management Research and Development Seminar(ATM 2009).

[6] Stephen Atkins, Christipher Brinton, Yoon Jung, "Implication of variability in Airport Surface operation on 4-D Trajectory Planning".

[7] S.Bagassi, D.Francia, F.Persiani, "Simulating airport operations in a synthetic environment", International Conference on Innovative methods in product design. June 2011.

[8] Enrique Alba and Carlos cotta, " Evolutionary Algorithms", February 19, 2004.

[9] Jason A.D. Atkin, Edmund K.Burke, John S.Greenwood, Dale Reeson, "Hybrid Metaheuristics to Aid Runway Scheduling at London Heathrow Airport", Transportation Science, Vol.41, No.1, February 2007, PP.90-106.

S.Baghavathi Priya received the B.E Degree in Computer Science and Engineering Sundranar University and M.Tech Degree from Dr.M.G.R Educational and Research Institute, Chennai, India. She is pursuing PhD in Faculty of Computer Science Engineering, Jawaharlal Nehru Technological University, Hyderabad, Andhra Pradesh, India. Presently she is working as an Associate Professor in the department of Information Technology, Rajalakshmi Engineering College, Chennai, India. Her research area includes Grid Computing and Distributed Computing. She has published around 10 papers in National and International conferences and journals.

Dr.T.Ravichandran received B.E degree from Bharathiar University, Tamilnadu, India and M.E degree from Madurai Kamaraj University, Tamilnadu, India in 1994 and 1997, respectively and PhD degree from the Periyar University, Salem, Tamilnadu, India, in 2007. He is currently the Principal of Hindustan Institute of Technology, Coimbatore, Tamilnadu, India. His research interests include Distributed Computing, Image Processing, Internet Computing and Security, Mobile Computing, Performance Evaluation and Fault Tolerant Computing. He has published more than 40 papers in National and International Conferences and Journals. He is the life member in professional society of IEEE, CSI and ISTE.

A Group Decision Making Methodology for Emergency Decision

Tiejun CHENG[1], Fengping WU[2] and Yanping CHEN[3]

[1]Business School, Hohai University
Nanjing 211100, China

[2]Institute of Planning and Decision-making, Business School, Hohai University
Nanjing 211100, China

[3]Business School, Hohai University
Nanjing 211100, China

Abstract

As the emergency is always unconventional, sudden and complex, it is necessary to invite experts from different fields to make decisions. However, the decision makers are usually hesitant and cannot get hold of the emergency because of the lack of information and knowledge. In this paper, a group decision-making methodology based on intuitionistic fuzzy sets is proposed to solve the emergency group decision-making problem. The intuitionistic fuzzy set that was introduced by Atanassov can consider the degree of membership, the degree of non-membership and hesitant degree. As the preferences of emergency decision makers are usually hesitating and incomplete, the incomplete intuitionistic judgment matrix can be constructed to convey the preferences of decision makers. Considering the known elements of the incomplete intuitionistic judgment matrix, the incomplete preference is estimated according to some principles. Then, the individual's preference is aggregated into the group preference through IFWG operators. According to the results of the proposed method, the best emergency plan can be figured out. Finally, a case in emergency decision making in Jiangsu coastal development is introduced to demonstrate the feasibility and efficiency of the proposed method.

Keywords: *Incomplete intuitionistic judgment matrix, Group decision making, Emergency management.*

1. Introduction

Emergency events often lead to casualties, economic losses, destructions to the ecological environment and other unexpected catastrophic consequences [1-3]. In China, the emergency events have caused 200 thousand people died, 2 million people disabled, and the economic loss that was about 5 percent of the GDP every year [4]. In the emergency planning and management, how to choose the best from many emergency plans to minimize the losses of the destructive events is a valuable research topic [5-6].

As the emergency is always complex and involves many aspects, it needs the consensus decision that is made by experts, government workers, the public and other relevant departments. Accordingly, using group decision support systems (GDSS) to handle emergency decision problems could be extremely valuable. Yu and Lai proposed a distance-based group decision-making (GDM) methodology to solve unconventional multi-person multi-criteria emergency decision-making problems. The results demonstrated that the proposed distance-based multi-criteria GDM methodology can improve decision-making objectivity and emergency management effectiveness [7]. Mendonca et al. designed and used of a gaming simulation as a means of assessing one group decision support system (GDSS) for emergency response [8]. Levy and Taji proposed a GANP multi-criteria Decision Support System (DSS) that used quadratic mathematical programming and interval preference information [9]. Nils and Giampiero developed a participatory methodology that helps infrastructure providers, spatial planners and emergency responders converge their views on safety in infrastructure planning[10]. Jutta et al. proposed the multi-criteria decision support and evaluation of strategies for nuclear remediation management [11]. Selcuk and Cengiz developed a decision support system (DSS) based on fuzzy information axiom (FIA) in order to make the decision procedure easy [12]. Liu puted forward a Multiple Attribute Decision Making (MADM) based on water bloom emergency management decision-making methods, and applied to the lake reservoir water bloom emergency management program's selection [13].

The present studies have shown that GDSS can improve emergency management effectiveness and decision transparency because it can integrate group wisdom of multiple decision-makers into one group wisdom. In the process of emergency decision-making, how to express the preference of each decision-maker in the group realistically is a key issue for group decision making method. As emergency is always complex and uncertainty, the decision makers are usually hesitant and can't get hold of enough

knowledge of the emergency. The emergency decision makers from different fields may be familiar with some aspects of the emergency, but not all. It is important to consider the incomplete and hesitating complements of the decision language when the decision makers express their preference. So, the paper tries to convey the information of decision makers in emergency management based on intuitionistic fuzzy sets.

Intuitionistic fuzzy set was proposed by Atanassov. It is commonly used because that it can consider the degree of membership, the degree of non-membership and the hesitancy degree[14]. Yu and Lai utilized fuzzy QFD method as a tool that makes the subjective judgment of the problem [7]. Dursun et al. used the ordered weighted averaging (OWA) operator to aggregate decision makers' opinions [15]. Chen et al. presented a new method to deal with fuzzy multiple attributes group decision-making problems based on ranking interval type-2 fuzzy sets [16]. Ye proposed an extended technique for order preference by similarity to ideal solution (TOPSIS) method for group decision making with interval-valued intuitionistic fuzzy numbers to solve the partner selection problem under incomplete and uncertain information envioronment[17]. Malekly and Meysam described the rating values regarding to each alternative and criteria throughout the phases in a fuzzy environment by means of linguistic variables [18]. Ben combined fuzzy logic with case-based reasoning to identify useful cases that can support the decision making [19].

The main purpose of the proposed multi-criteria GDM methodology is to improve decision accuracy, and to enhance decision transparency and thus to increase decision effectiveness. The rest of this paper is organized as follows. In Section 2, the general framework for the methodology is decribed. In Section 3, the multi-criteria GDM methodology based on intuitionistic fuzzy sets Theory is described in detail. For illustration and verification purposes, Section 4 presents a practical emergency decision case to illustrate the implementation process, and to verify the effectiveness of theproposed methodology. Finally, some concluding remarks are drawn in Section 5.

2. Preliminaries

In this section, the description of the emergency decision problem is given. Then, a general framework for the multi-criteria GDM methodology is presented. Finally, the basic knowledge of intuitionistic fuzzy sets is given.

2.1 Description of the emergency decision problem

As the emergency is always unconventional, sudden and complex, it is necessary to invite experts from different fields to make decisions. It is impossible to make an emergency plan considering all aspects of the emergency. The realistic choice is that we should have many emergency plans and let the decision makers to choose a best one. So, the emergency decision is a group decision-making problem. As the emergency decision-making must be made in a short time using partial or incomplete information, the decision makers may be hesitant and unfamiliar with some aspects of the emergency. The paper tries to introduce intuitionistic fuzzy sets to solve the problem. The description of the emergency group decision-making problem is as the following:

$Y=(Y_1,\ Y_2,\cdots,Y_n)$: the emergency plans that are made by emergency department to deal with the emergency . Y_i stands for the i th emergency plan, $i=1,2,\cdots,n$.

$E=(e_1,\ e_2,\cdots,e_l)^T$:the decision makers from different field to deal with the emergency, e_k stands for the k th decision maker, .

$\mu_{ij}^{(k)}$:the certain degree to which Y_i is preferred to that is assessed by emergency decision maker e_k .

$v_{ij}^{(k)}$: the certain degree to which Y_j is preferred to Y_i that is assessed by emergency decision maker e_k .

$1-\mu_{ij}^{(k)}-v_{ij}^{(k)}$:the uncertain degree to which Y_i is preferred to Y_j that is assessed by emergency decision maker.

$\xi=(\xi_1,\ \xi_2,\cdots,\xi_l)^T$: the weight vector of the emergency decision makers.

2.2 The general framework for the GDM methodology

The general framework for the GDM methodology is given as Fig.1. First, the emergency group decision making problem is described. As the emergency is always complex, the decision maker is usually hesitant and cannot get hold of the emergency because of the lack of information. So the incomplete intuitionistic judgment matrix is proposed when the decision makers express their preference for the emergency plan. Based on intuitionistic fuzzy set, we can get the average intuitionistic preference value and the comprehensive intuitionistic preference value. Finally, choose the best emergency plan to deal with the emergency.

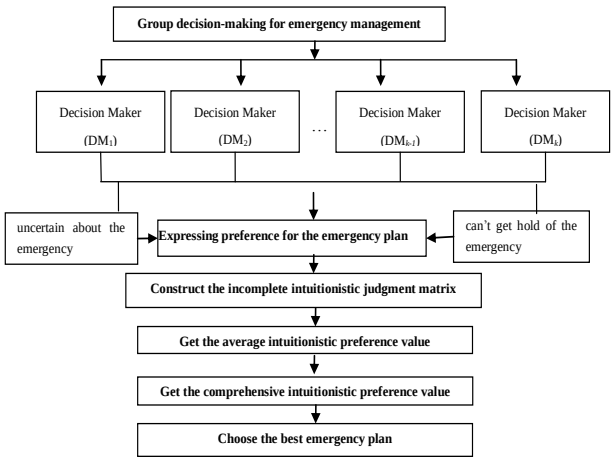


Fig. 1 General framework for the GDM methodology.

2.3 Basic knowledge of intuitionistic fuzzy sets

Definition1. Let $Q=(q_{ij})_{n\times n}$ be the intuitionistic judgment matrix[20], where $q_{ij}=(\mu_{ij},v_{ij})(i,j=1,2,\cdots,n)$, μ_{ij} stands for the decision maker's preference to Y_i when he or she compare Y_i with Y_j , v_{ij} stands for the decision maker's preference

$$\mu_{ij}\in[0,1],v_{ij}\in[0,1],0\le\mu_{ij}+v_{ij}\le 1,\mu_{ji}=v_{ij},v_{ji}=\mu_{ij},\mu_{ii}=v_{ii}=0.5\ \ (i,j=1,2,\cdots,n) \quad (1)$$

then we call Q the intuitionistic judgment matrix.

Definition2. Let $Q=(q_{ij})_{n\times n}$ be the intuitionistic $Q=(q_{ij})_{n\times n}$ judgment matrix, if it contains incomplete elements and complete elements, be the incomplete elements, if

$$0\le\mu_{ij}+v_{ij}\le 1,\mu_{ji}=v_{ij},v_{ji}=\mu_{ij},\mu_{ii}=v_{ii}=0.5 \quad (2)$$

then we call Q the intuitionistic judgment matrix.

Definition3. If $q_{ij}=(\mu_{ij},v_{ij})$ and $q_{kl}=(\mu_{kl},v_{kl})$ are two intuitionistic fuzzy values, then

$(1)\ \bar{q}_{ij}=(v_{ij},\mu_{ij}).$

$(2)q_{ij}+q_{kl}=(\mu_{ij}+\mu_{kl}-\mu_{ij}\cdot\mu_{kl},\ \ v_{ij}\cdot v_{ij}).$

$(3)q_{ij}\cdot q_{kl}=(\mu_{ij}\cdot\mu_{ij},\ \ v_{ij}+v_{kl}-v_{ij}\cdot v_{kl}).$

$(4)\lambda q_{ij}=(1-(1-\mu_{ij})^{\lambda},v_{ij}^{\lambda}),\lambda>0.$

$(5)q_{ij}^{\lambda}=(\mu_{ij}^{\lambda},1-(1-v_{ij})^{\lambda}),\lambda>0.$

Definition4. Let $Q=(q_{ij})_{n\times n}$ be the incomplete intuitionistic judgment matrix, if $q_{ij}=q_{ik}\otimes q_{kj},q_{ij},q_{ik},q\in\Omega$, then we call Q the consistency incomplete intuitionistic judgment matrix.

Definition5. Let $Q=(q_{ij})_{n\times n}$ be the incomplete intuitionistic judgment matrix, if $(i,j)\cap(k,l)\neq\varnothing$, then we call the element q_{ij} and q_{kl} are adjacent.

Definition6. Let $Q=(q_{ij})_{n\times n}$ be the incomplete intuitionistic judgment matrix, if each unknown element can be got from its adjacent elements, Q is acceptable, or Q is unacceptable.

In the face of the emergency, the decision maker ($e_k\in E$) is usually hesitant and uncertain, he or she gives the preference after compare two contingency plans, and we can get $q_{ij}^{(k)}=(\mu_{ij}^{(k)},v_{ij}^{(k)})$, where μ_{ij} stands for the decision maker's preference to Y_i when he or she compare Y_i with Y_j , v_{ij} stands for the decision maker's preference .

Theorem1. Let $q_{ij}^{(1)},q_{ij}^{(2)},...,q_{ij}^{(m)}$ be m intuitionistic fuzzy values, where $q_{ij}^{(c)}=(\mu_{ij}^{(c)},v_{ij}^{(c)})$, $c=1,2,...,m$, and let $w=(w_1,w_2,...,w_m)^T$ be the weight vector of $q_{ij}^{(1)},q_{ij}^{(2)},...,q_{ij}^{(m)}$, then the aggregated value $\bar{q}_{ij}$ of $q_{ij}^{(1)},q_{ij}^{(2)},...,q_{ij}^{(m)}$ is also an intuitionisitic fuzzy value, where $\bar{q}_{ij}$ is obtained by using the intuitionisitic fuzzy weighted arithmetic averaging operator:

$$\bar{q}_{ij}=\sum_{c=1}^{m}w_c q_{ij}^{(c)},\ i,j=1,2,...,n \quad (3)$$

or by using the intuitionist fuzzy weighted geometric averaging operator:

$$\bar{q}_{ij}=\prod_{c=1}^{m}(q_{ij}^{(c)})^{w_c},\ i,j=1,2,...,n \quad (4)$$

In particular, if $w=(1/m,1/m,...,1/m)^T$, then(3)and(4)are, respectively, reduced to the intuitionistic fuzzy arithmetic averaging operator:

$$\bar{q}_{ij}=\frac{1}{c}\sum_{c=1}^{c} q_{ij}^{(c)},\ i,j=1,2,...,n \quad (5)$$

and the intuitionisitic fyzzy geometric averaging operator:

$$\bar{q}_{ij}=(\prod_{c=1}^{m}(q_{ij}^{(c)}))^{\frac{1}{m}},\ i,j=1,2,...,n \quad (6)$$

3. Group decision making model base on intuitionistic fuzzy sets

As the emergency is always complex, the decision maker is usually hesitant and cannot get hold of the emergency because of the lack of knowledge, the paper introduces the incomplete intuitional judgment matrix to express the preference of the decision maker. The decision makers express their preference according the knowledge about the emergency, then the paper aggregates individual preference to group preference, and finally get the best emergency plan.

3.1 Step1: Construct the incomplete intuitionistic judgment matrix

As the emergency is complex and sudden, the decision maker may be hesitant and can't get enough knowledge, he or she can make space when express the preference, then we can get the incomplete intuitional judgment matrix $Q_k=(q_{ij}^{(k)})_{n\times n}$, where $q_{ij}^{(k)}=(\mu_{ij}^{(k)},\nu_{ij}^{(k)}),0\le\mu_{ij}^{(k)}+\nu_{ij}^{(k)}\le 1,\mu_{ji}^{(k)}=\nu_{ij}^{(k)},\nu_{ji}=\mu_{ij},\mu_{ii}^{(k)}=\nu_{ii}^{(k)}=0.5(i,j\in\Omega)$.

As defined in 2.3, Q_k should be acceptable. If Q_k is unacceptable, the decision maker needs to construct a new one until it is acceptable.

3.2 Step2: Construct the improved incomplete intuitionistic judgment matrix

As described in 3.1, we can get the acceptable incomplete intuitionistic judgment matrix from each emergency decision maker. As there are incomplete and unknown elements in the intuitionistic judgment matrix, we should estimate them through other known elements.

Let $Q=(q_{ij})_{n\times n}$ be the acceptable incomplete intuitionistic judgment matrix, if each unknown element can be got through

$$\dot{q}_{ij}=(\bigotimes_{k\in N_{ij}}(q_{ik}\otimes q_{kj}))^{\frac{1}{n_{ij}}} \quad (7)$$

where $N_{ij}=\{k|q_{ik},q_{kj}\in\Delta\}$, then we get the improved $\dot{Q}=(\dot{q}_{ij})_{n\times n}$.

$$\dot{q}_{ij}=\begin{cases} q_{ij}, & q_{ij}\in\Omega \\ \dot{q}_{ij}, & q_{ij}\notin\Omega \end{cases} \quad (8)$$

The improved intuitionistic judgment matrix $\dot{Q}=(\dot{q}_{ij})_{n\times n}$ contains both the direct intuitionistic preference information given by the emergency decision maker and the indirect intuitionistic preference information derived from the known intuitionistic preference information.

3.3 Step3: Get the average intuitionistic preference value through IFWA operators

Through institutionistic fuzzy weighted aggregation (IFWA) operators:

$$\dot{q}_i^{(k)}=\frac{1}{n}(q_{i1}^{\cdot(k)}\oplus q_{i2}^{\cdot(k)}\oplus\cdots\oplus q_{in}^{\cdot(k)}) \quad (9)$$

we can aggregate the intuitionistic preference value of emergency plan ,then get the average intuitionistic preference value.

3.4 Step4: Get the comprehensive intuitionistic preference value through IFWG operators

Through intuitionistic fuzzy weighted geometric (IFWG) operator:

$$\dot{q}_i=(\xi_1\dot{q}_i^{(1)}\oplus\xi_2\dot{q}_i^{(2)}\oplus\cdots\oplus\xi_l\dot{q}_i^{(l)}) \quad (10)$$

We can aggregate the intuitionistic preference value of emergency plan, and then get the comprehensive intuitionistic preference value.

3.5 Step5: Choose the best emergency plan

Definition6. For any intuitionistic fuzzy number $q_{ij}=(\mu_{ij},\nu_{ij})$, we can asses it through the score function $s(q_{ij})$:

$$s(q_{ij})=\mu_{ij}-\nu_{ij} \quad (11)$$

Where $s(q_{ij})$ is the score value, $s(q_{ij}) \in [-1,1]$ .The larger the score $s(q_{ij})$, the greater the intutionistic fuzzy value q_{ij} .

Definition7. For any intuitionistic fuzzy number , we can asses it through the accuracy function:

$$h(q_{ij})=\mu_{ij}+\nu_{ij} \quad (12)$$

to evaluate the degree of accuracy of the intuitionistic fuzzy value q_{ij} , where $h(q_{ij}) \in [-1,1]$. The larger the value of $h(q_{ij})$, the more the degree of accuracy of the intuitionistic fuzzy value q_{ij} .

Normally, we use score function to judge the intuitionistic fuzzy Numbers, in some special circumstances, such as the score value of two groups of intuitionistic fuzzy number is the same and it cannot through the score function to judge, then we can use the accuracy function to judge.

Definition8. Let $q_{ij}=(\mu_{ij},\nu_{ij})$ and $q_{kl}=(\mu_{kl},\nu_{kl})$ be two intuitionistic fuzzy values, $s(q_{ij})=\mu_{ij}-\nu_{ij}$ and $s(q_{kl})=\mu_{kl}-\nu_{kl}$ be the scores of q_{ij} and q_{kl} , respectively, and let $h(q_{ij})=\mu_{ij}+\nu_{ij}$ and $h(q_{kl})=\mu_{kl}+\nu_{kl}$ be the accuracy degrees of q_{ij} and q_{kl} , respectively, then

If $s(q_{ij})<s(q_{kl})$, then q_{ij} is smaller than , denoted by $q_{ij}<q_{kl}$.

If $s(q_{ij})=s(q_{kl})$, then

(1)If $h(q_{ij})=h(q_{kl})$, then q_{ij} and q_{kl} represent the same information, denoted by $q_{ij}=q_{kl}$.

(2)If $h(q_{ij})<h(q_{kl})$, then q_{ij} is smaller than q_{kl} , denoted by $q_{ij}<q_{kl}$.

According formula (5) and (6), we can sort the comprehensive intuitionistic preference value $q_i^{\cdot}(i=1,2,\cdots,n)$, then we can sort the emergency plans $Y_i(i=1,2,\cdots,n)$ and choose the best one.

4. Application

In June 2009, the State Council of China reviewed the Jiangsu Coastal Area Development Plan. The Jiangsu Coastal Area has brought fast development of economy since 2009. However , the coastal areas is also easy to happen emergency in its development ,such as safe production, land expropriation demolition, traffic accident, natural disaster and so on. The safety of coastal needs our attention. The paper takes the emergency in Jiangsu coastal area development for example for simulation analysis. To assess the emergency plans, we consider the following four aspects: economic loss, personnel losses, environmental impact and social influence. We suppose that there are four decision makers to choose the best plan from four emergency plans of Jiangsu coastal area development. In order to deal with the emergency, the emergency department has made four emergency plans considering with different situations. The committee comprise of four decision makers $e_k(k=1,2,3,4)$ (whose weight vector is $\xi=(0.22,0.25,0.3,0.23)^T$) has been set up to provide assessment information on the emergency plans.

Step1.The decision makers $e_k(k=1,2,3,4)$ provide their preference information by incomplete intuitionistic judgment matrix $Q^{(k)}=(q_{ij}^k)_{4\times4}(k=1,2,3,4)$ as follows, respectively:

$$Q_1=\begin{pmatrix}(0.5,0.5) & (0.4,0.5) & (x_1,x_2) & (0.3,0.5)\\ (0.5,0.4) & (0.5,0.5) & (0.5,0.3) & (0.4,0.5)\\ (x_2,x_1) & (0.3,0.5) & (0.5,0.5) & (0.3,0.6)\\ (0.5,0.3) & (0.5,0.4) & (0.6,0.3) & (0.5,0.5)\end{pmatrix}$$

$$Q_2=\begin{pmatrix}(0.5,0.5) & (0.3,0.6) & (0.5,0.3) & (x_3,x_4)\\ (0.6,0.3) & (0.5,0.5) & (0.6,0.3) & (0.5,0.4)\\ (0.3,0.5) & (0.3,0.6) & (0.5,0.5) & (0.4,0.5)\\ (x_4,x_3) & (0.4,0.5) & (0.5,0.4) & (0.5,0.5)\end{pmatrix}$$

$$Q_3=\begin{pmatrix}(0.5,0.5) & (x_5,x_6) & (0.5,0.4) & (0.3,0.6)\\ (x_6,x_5) & (0.5,0.5) & (0.3,0.5) & (0.6,0.3)\\ (0.4,0.5) & (0.4,0.5) & (0.5,0.5) & (0.4,0.3)\\ (0.6,0.3) & (0.3,0.6) & (0.3,0.4) & (0.5,0.5)\end{pmatrix}$$

$$Q_4=\begin{pmatrix}(0.5,0.5) & (0.3,0.5) & (0.5,0.3) & (0.3,0.6)\\ (0.5,0.3) & (0.5,0.5) & (x_7,x_8) & (0.5,0.4)\\ (0.3,0.5) & (x_8,x_7) & (0.5,0.5) & (0.3,0.6)\\ (0.6,0.3) & (0.4,0.5) & (0.6,0.3) & (0.5,0.5)\end{pmatrix}$$

Step2. Use (7) to construct the improved intuitionistic judgment matrix $\dot{Q}^{(k)}=(\dot{q}_{ij}^{(k)})_{4\times4}(k=1,2,3,4)$ of $Q^{(k)}=(q_{ij}^{k})_{4\times4}(k=1,2,3,4)$.

$$\dot{Q}_1=\begin{pmatrix}(0.5,0.5) & (0.4,0.5) & (0.44,0.29) & (0.3,0.5)\\ (0.5,0.4) & (0.5,0.5) & (0.5,0.3) & (0.4,0.5)\\ (0.39,0.44) & (0.3,0.5) & (0.5,0.5) & (0.3,0.6)\\ (0.5,0.3) & (0.5,0.4) & (0.6,0.3) & (0.5,0.5)\end{pmatrix}$$

$$\dot{Q}_2=\begin{pmatrix}(0.5,0.5) & (0.3,0.6) & (0.5,0.3) & (0.42,0.44)\\ (0.6,0.3) & (0.5,0.5) & (0.6,0.3) & (0.5,0.4)\\ (0.3,0.5) & (0.3,0.6) & (0.5,0.5) & (0.4,0.5)\\ (0.44,0.42) & (0.4,0.5) & (0.5,0.4) & (0.5,0.5)\end{pmatrix}$$

$$\dot{Q}_3=\begin{pmatrix}(0.5,0.5) & (0.37,0.52) & (0.5,0.4) & (0.3,0.6)\\ (0.52,0.37) & (0.5,0.5) & (0.3,0.5) & (0.6,0.3)\\ (0.4,0.5) & (0.4,0.5) & (0.5,0.5) & (0.4,0.3)\\ (0.6,0.3) & (0.3,0.6) & (0.3,0.4) & (0.5,0.5)\end{pmatrix}$$

$$\dot{Q}_4=\begin{pmatrix}(0.5,0.5) & (0.3,0.5) & (0.5,0.3) & (0.3,0.6)\\ (0.5,0.3) & (0.5,0.5) & (0.52,0.32) & (0.5,0.4)\\ (0.3,0.5) & (0.32,0.52) & (0.5,0.5) & (0.3,0.6)\\ (0.6,0.3) & (0.4,0.5) & (0.6,0.3) & (0.5,0.5)\end{pmatrix}$$

Step3. Use (3) to aggregate all corresponding to the emergency plan Y_i, and then get the averaged intuitionistic fuzzy value of the emergency plan over all the other emergency plans.

$$q_1^{(1)}=\frac{1}{4}(q_{11}^{(1)}\oplus q_{12}^{(1)}\oplus q_{13}^{(1)}\oplus q_{14}^{(1)})$$
$$=\frac{1}{4}((0.5+0.4+0.44+0.3),(0.5+0.5+0.39+0.5))$$
$$=(0.41,0.47)$$

$q_1^{(2)}=(0.43,0.46)$, $q_1^{(3)}=(0.42,0.51)$,$q_1^{(4)}=(0.4,0.48)$

$q_2^{(1)}=(0.48,0.45)$, $q_2^{(2)}=(0.55,0.38)$, $q_2^{(3)}=(0.48,0.42)$,$q_2^{(4)}=(0.51,0.38)$

$q_3^{(1)}=(0.37,0.51)$, $q_3^{(2)}=(0.3,0.53)$, $q_3^{(3)}=(0.43,0.45)$,$q_3^{(4)}=(0.36,0.53)$

$q_4^{(1)}=(0.53,0.33)$, $q_4^{(2)}=(0.46,0.46)$, $q_4^{(3)}=(0.53,0.45)$,$q_4^{(4)}=(0.48,0.4)$

Step4. Use (4) to aggregate all into a collective intuitionistic fuzzy value of the emergency plan over all the other emergency plans:

$$\dot{q}_1=(\xi_1\dot{q}_1^{(1)}\oplus\xi_2\dot{q}_1^{(2)}\oplus\xi_3\dot{q}_1^{(3)}\oplus\xi_4\dot{q}_1^{(4)})$$
$$=(0.42,0.47)$$

$$\dot{q}_2=(0.50,0.41)$$

$$\dot{q}_3=(0.37,0.50)$$

$$\dot{q}_4=(0.50,0.42)$$

Finally, choose the best emergency plan. Through formula (11), we can get:

$s(\dot{q}_1)=-0.05$, $s(\dot{q}_2)=0.09$, $s(\dot{q}_3)=-0.13$, $s(\dot{q}_4)=0.08$.

Then

$$q_2>q_4>q_1>q_3$$

and hence

$$Y_2\succ Y_4\succ Y_1\succ Y_3.$$

The emergency plan 2 is the best.

5. Conclusions

In emergency decision making, the decision makers may be hesitated and lack of knowledge. To solve this group decision making problem, a method that based on incomplete intuitionistic judgment matrix is proposed for emergency management. In this paper, the incomplete intuitionistic judgment matrix is constructed to convey the information of experts in group decision making. Finally, a case in emergency decision making in Jiangsu coastal development is introduced to demonstrate the feasibility and efficiency of the proposed method.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (41271537), Ministry of Education, Humanities and Social Science Projects (11YJC630025).

References

[1] Fozel, "Time pressure and stress as a factor during emergency egress", Safety Science, vol.38, No.2, pp.95-107,2001.

[2] Henry Jackson, "First responders: problems and solutions: Office of Emergency Management", Technology in Society, vol. 25, No.4, pp. 539-541,2003.

[3] Luh, Peter B, B Xong, S.C Chang, "Group elevator scheduling with advance information for normal and emergency modes", IEEE Transactions on Automation Science and Engineering, Vol.5, No.2, pp. 245-258,2008.

[4] Z.P Fan, Y Liu, R.J Shen ,"Risk decision analysis method for emergency response based on prospect theory", System Engineering Theory and Practice, Vol.32, No.5, pp. 977-984, 2012.

[5] Jennifer Wilson, Arthur Oyola-Yemaiel, "The evolution of emergency management and the advancement towards a profession in the United States and Florida", Safety Science, Vol.32, No.1-2, pp. 117–131, 2001.
[6] Bertsch Valentin, Geldermann Jutta, "Preference elicitation and sensitivity analysis in multicriteria group decision support for industrial risk and emergency management",International Journal of Emergency Management, Vol.5,No.1-2, pp.7-24, 2008.
[7] L Yu, Lai K K, "A distance-based group decision-making methodology for multi-person multi-criteria emergency decision support", Decision Support Systems, Vol.51, No.2,pp. 307-315,2011.
[8] Mendonca D, Beroggi G E G, van Gent D, "Designing gaming simulations for the assessment of group decision support systems in emergency response",Safety Science.Vol.44, No.6, pp.523-535,2006.
[9] Levy J K, Taji K," Group decision support for hazards planning and emergency management: A Group Analytic Network Process (GANP) approach", Mathematical and Computer Modelling,Vol.46, No.7, pp. 906-917,2007.
[10] Nils Rosmullera, Giampiero E.G. Beroggib, "Group decision making in infrastructure safety planning", Safety Science, Vol.42, No.4, pp.325–349,2004.
[11] Jutta Geldermann, Valentin Bertsch, Martin Treitz, et al, "Multi-criteria decision support and evaluation of strategies for nuclear remediation management", Omega, Vol.37, No.1, pp.238-251,2009.
[12] Selcuk Cebi , Cengiz Kahraman, "Developing a group decision support system based on fuzzy information axiom", Knowledge-Based Systems , Vol.23, No.1, pp. 3–16,. 2010
[13] Liu Zaiwen, Li Lin, Wang Xiaoyi,Chen Chen." Researches of water bloom emergency management decisionmaking method and system based on fuzzy multiple attribute decision making".International Journal of Computer Science Issues, Vol9, No.5, pp. 48-53, 2012.
[14] Atanassov K, "Intuitionistic fuzzy sets", Fuzzy Sets and Systems, Vol.20, No.1, pp. 87-96,1986.
[15] Dursun, M., Karsak, E. E. , Karadayi, M. A.. "A fuzzy multi-criteria group decision making framework for evaluating health-care waste disposal alternatives". Expert Systems with Applications, Vol.38, No.9, pp.11453-11462,2011.
[16] S.M Chen , M.W Yang, L.W Lee, W Szu, "Fuzzy multiple attributes group decision-making based on ranking interval type-2 fuzzy sets", Expert Systems with Applications, Vol.39, No.5, pp. 5295-5308,2012.
[17] F Ye. "An extended TOPSIS method with interval-valued intuitionistic fuzzy numbers for virtual enterprise partner selection". Expert Systems with Applications, Vol.37, No.10, pp. 7050-7055, 2010.
[18] Malekly H, Meysam M. S, Hashemi H,"A fuzzy integrated methodology for evaluating conceptual bridge design", Expert Systems with Applications. Vol.37, No.7, pp. 4910-4920, 2010.
[19] Ben Y. N.,Bellamine Narjès, Ben G. H.. "Integrating fuzzy case-based reasoning and particle swarm optimization to support decision making". International Journal of Computer Science Issues, Vol. 9, No.33-3, pp. 117-124, 2012.
[20] Xu Z S. "Intuitionistic Fuzzy Information Aggregation: Theory and Applications". Beijing. Sciences Press, Apr 2008, pp:134-140.

Tiejun CHENG received the M.S. degree in Business school of Hohai University, Nanjing, China, in 2009. She is currently working toward the Ph.D. degree at Hohai University. Her research interests include decision making methods and emergency management.

Fengping WU received the Ph.D. degree at Hohai University in 1998. He is a director, full professor in School of Business , Hohai University.

Yanping CHEN received the Ph.D. degree at Hohai University in 2009. Her research interests include decision making methods and water resource management.

Analysis of the impact of parameters values on the Genetic Algorithm for TSP

Avni Rexhepi[1], Adnan Maxhuni[2], Agni Dika[3]

[1,2,3] Faculty of Electrical and Computer Engineering, University of Pristina, Kosovo

Abstract

Genetic algorithms (GAs) are multi-dimensional and stochastic search methods, involving complex interactions among their parameters. Researchers have been trying to understand the mechanics of GA parameter interactions by using various techniques. It still remains an open question for practitioners as to what values of GA parameters (such as population size, choice of GA operators, operator probabilities, and others) to use in an arbitrary problem.

Genetic algorithm (GA) parameters are explored to minimize the time needed to find a solution. The basic GA code stays the same throughout the entire system. Variable parameters include mutation rate, crossover rate, crossover operator, number of generations, population size, etc. When an optimization problem is encoded using genetic algorithms, one must address issues of population size, crossover and mutation operators and probabilities, stopping criteria, selection operator and pressure, and fitness function to be used in order to solve the problem. This paper tests a relationship between (1) size of initial population, (2) mutation probability, and (3) number of generations in runs of Genetic Algorithm for solving the TSP.

TSP is an NP hard problem, so using Genetic Algorithm we can find a solution on reasonable amount of time. In this paper we describe the results of the solution for the Traveling Salesman Problem (TSP) for Kosovo municipalities, using the genetic algorithm, with different settings for the parameters of the Genetic Algorithm [12].

Keywords: *Genetic Algorithms, TSP, Parameter Selection, Initial Population, Crossover, Mutation*

1. Introduction

TSP is a problem, where traveling salesman wants to visit each of a set of cities exactly once, starting from hometown and returning to his hometown. His problem is to find the shortest route for such a trip. TSP has a model character in many branches of Mathematics, Computer Science, Operations Research, etc. Linear programing, heuristics and branch and bound which are main components for the most successful approaches to hard combinatorial optimization problems, were first formulated for the TSP and used to solve practical problem instances in 1954 by Dantzig, Fulkerson and Johnson.

When the theory of NP-completeness developed, the TSP was one of the first problems to be proven NP-hard by Karp in 1972. New algorithmic techniques have first been developed for or at least have been applied to the TSP to show their effectiveness. Such examples are branch and bound, Lagrangean relaxation, Lin-Kernighan type methods, simulated annealing, etc. [3].

Representation model is: Let $Kn=(V_n, E_n)$ be the complete undirected graph with $n=|V_n|$ nodes and $m=|En|=\binom{n}{2}$ edges. An edge e with endpoints *i* and *j* is also denoted by *ij*, or by (*i*,*j*). We denote by R^{En} the space of real vectors whose components are indexed by the elements of E_n. The component of any vector $z \in R^{En}$ indexed by the edge $e=ij$ is denoted by z_e, z_{ij}, or $z(i,j)$.

Given an objective function $c \in R^{En}$, that associates a "length" c_e with every edge e of K_n, the symmetric traveling salesman problem consists of finding a Hamiltonian cycle such that its *c*-length (the sum of the lengths of its edges) is as small as possible.

Of special interest are the Euclidean instances of the traveling salesman problem. In these instances the nodes defining the problem correspond to points in the two-dimensional plane and the distance between two nodes is the Euclidean distance between their corresponding points.

More generally, instances that satisfy the triangle inequality, i.e., $c_{ij} + c_{jk} \geq c_{ik}$ for all the three distinct i,j and k, are of particular interest.

For our case, we consider the locations of the cities/municipalities in Kosovo map as nodes of the graph. For to do this, we take their geographic coordinates and then based on that, we calculate their position in our map scaled to a smaller size, for to calculate the real positions and distances, by using the real life values for distances in kilometers between the cities.

2. Genetic Algorithms

Genetic Algorithms (GA) [1,2] are computer algorithms that search for good solutions to a problem within a large number of possible solutions. They were proposed and developed in the 1960s by John Holland, his students, and his colleagues at the University of Michigan. These computational paradigms were inspired by the mechanics of natural evolution, including survival of the fittest, reproduction, and mutation. These mechanics are well suited to resolve a variety of practical problems, including computational problems, in many fields. Some applications of GAs are optimization, automatic programming, machine learning, economics, immune systems, population genetic, and social system.

GAs have been successfully applied to many problems of business, engineering, and science. Because of their operational simplicity and wide applicability, GAs play an important role in computational optimization and operations research [6].

The genetic algorithm transforms a population (set) of individual objects, each with an associated fitness value, into a new generation of the population using the Darwinian principle of reproduction and survival of the fittest and analogs of naturally occurring genetic operations such as crossover (sexual recombination) and mutation. Each individual in the population represents a possible solution to a given problem. The genetic algorithm attempts to find a very good (or best) solution to the problem by genetically breeding the population of individuals over a series of generations.

2.1 Basic elements of GAs

Most GAs methods are based on the following elements: populations of chromosomes, selection according to fitness, crossover to produce new offspring, and random mutation of new offspring. The chromosomes in GAs represent the space of candidate solutions. Possible chromosomes encodings are binary, permutation, value, and tree encodings. GAs require a fitness function which allocates a score to each chromosome in the current population. Thus, it can calculate how well the solutions are coded and how well they solve the problem [2].

The selection process is based on fitness. Chromosomes that are evaluated with higher values (fitter) will most likely be selected to reproduce, whereas, those with low values will be discarded. The fittest chromosomes may be selected several times, however, the number of chromosomes selected to reproduce is equal to the population size, therefore, keeping the size constant for every generation. This phase has an element of randomness just like the survival of organisms in nature. The most used selection methods, are roulette-wheel, rank selection, steady-state selection, and some others. Moreover, to increase the performance of GAs, the selection methods are enhanced by elitism. Elitism is a method, which first copies a few of the top scored chromosomes to the new population and then continues generating the rest of the population. Thus, it prevents losing the few best found solutions.

Crossover is the process of combining the bits of one chromosome with those of another to create an offspring for the next generation that inherits traits of both parents.

Mutation is performed after crossover to prevent falling all solutions in the population into a local optimum of solved problem.

So, general outline of basic GA is:

1. Start: Randomly generate a population of N chromosomes.
2. Fitness: Calculate the fitness of all chromosomes.
3. Create a new population:
   a. Selection: According to the selection method select 2 chromosomes from the population.
   b. Crossover: Perform crossover on the 2 chromosomes selected.
   c. Mutation: Perform mutation on the chromosomes obtained.
4. Replace: Replace the current population with the new population.
5. Test: Test whether the end condition is satisfied. If so, stop. If not, return the best solution in current population and go to Step 2.

Each iteration of this process is called generation.

The genetic algorithm object determines which individuals should survive, which should reproduce, and which should die. It also records statistics and decides how long the evolution should continue. A typical genetic algorithm will run forever, so we must build functions for specifying when the algorithm should terminate. These include terminate-upon generation, in which you specify a certain number of generations for which the algorithm should run, and terminate-upon-convergence, in which you specify a value to which the best-of-generation score should converge. One can customize the termination function to use own stopping criterion and must tell the algorithm

when to stop. Often the number-of generations is used as a stopping measure, but you can use goodness-of-best-solution, convergence-of-population, or any problem-specific criterion if you prefer.

There are some flavors of genetic algorithms. For example, the first is the standard 'simple genetic algorithm' described by Goldberg in his book [2]. This algorithm uses non-overlapping populations and optional elitism. Each generation the algorithm creates an entirely new population of individuals. The second is a 'steady-state genetic algorithm' that uses overlapping populations. In this variation, you can specify how much of the population should be replaced in each generation. The third variation is the 'incremental genetic algorithm', in which each generation consists of only one or two children. The incremental genetic algorithms allow custom replacement methods to define how the new generation should be integrated into the population. So, for example, a newly generated child could replace its parent, replace a random individual in the population, or replace an individual that is most like it. The fourth type is the 'deme' genetic algorithm. This algorithm evolves multiple populations in parallel using a steady-state algorithm. Each generation the algorithm migrates some of the individuals from each population to one of the other populations.

The base genetic algorithm class contains operators and data common to most genetic algorithms.

The genetic algorithm contains the statistics, replacement strategy, and parameters for running the algorithm. The population object, a container for genomes, also contains some statistics as well as selection and scaling operators.

The number of function evaluations is a good way to compare different genetic algorithms with various other search methods[3]. The basic algorithm is as follows:

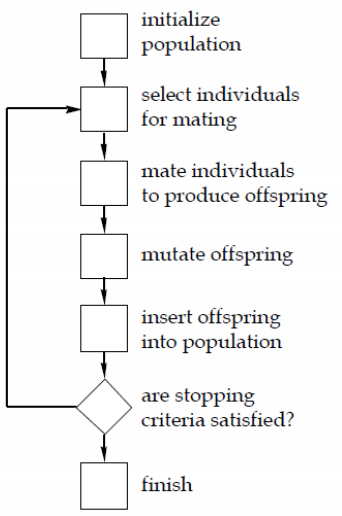


Figure 1 – Genetic Algorithm

2.2 TSP

In order to calculate the shortest traveling distance from an initial city, by visiting each one only once and returning to the initial one, we consider the locations as nodes of the graph, in the graph model. In the meantime, the distances between the cities are edges of the graph.

For length of the edges we take inter-city distances in kilometers. We have created a matrix of distances between all the municipalities, where the matrix elements c_{ij} are elements of the square symmetrical matrix, since distance ij is equal to the distance in the other side ji. The diagonal of the matrix will contain zero values, since diagonal elements of the matrix c_{ij}, where $i=j$, will represent the distance of the city to itself, so in fact it will be the traveling distance of zero kilometers, therefore these elements will be equal to $c_{ij}=0$.

By using complete graph in the definition of the TSP, the existence of a feasible solution is guaranteed, while for general graphs deciding the existence of a Hamiltonian cycle is an NP-complete problem. The number of Hamiltonian cycles in K_n, i.e. the size of the set of feasible solutions of the TSP is (n-1)!/2.

The algorithmic treatment of the TSP ensures an approximation algorithm that cannot guarantee to find the optimum, but which is the only available technique to find a good solution to a large problem instances. To assess the quality of a solution, one has to be able to compute a lower bound on the value of the shortest Hamiltonian cycle.

We have built an application in C#, with an image with the small-scaled size of the Kosovo map, with depicted municipality boundaries and locations (Figure 2).

It is possible to select a particular city by clicking on the map and we have also added buttons that make it possible to create locations for the biggest cities and locations for the all municipalities.

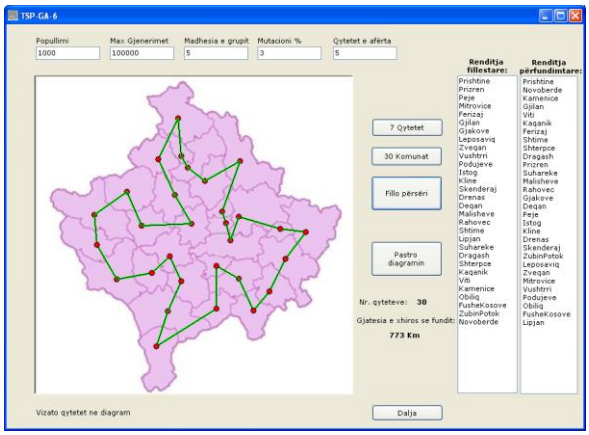


Figure 2 – Screenshot of the application

We used the geographical coordinates of the cities, to calculate their positions. By running the application, we can calculate the shortest traveling distance between the selected cities, by finding a solution of the TSP using a genetic algorithm.

User can set up the values for initial population, maximal number of generations, size of the group, percentage of mutations and number of close cities/locations (used by the algorithm, while finding closest locations).

3. Simulation and Results

We analyzed the results of different cases with different values for the parameters of initial population, probability of mutation and number of generations.

Firstly, we have set the value of the maximal number of generations to 10,000 and we calculated the fitness for the cases where the size of the initial population is: 1000, 5000 and 10000 and for each of these cases, we calculated the results for the mutation rate of: 1%, 3%, 5% and 10% (as in Figures: 3, 4 and 5).

Then, we compared the results for each value of the mutation rate (1%, 3%, 5% and 10%), with different values of the initial population parameter values (1000, 5000 and 10000, as in Figures: 6, 7, 8 and 9).

We repeated this for the values of maximal number of generations of 50,000 and 100,000, too (see Appendix).

In each figure, the y-axis is the value of the fitness and the x-axis is the maximal number of generations, which serves as the interruption parameter for the genetic algorithm.

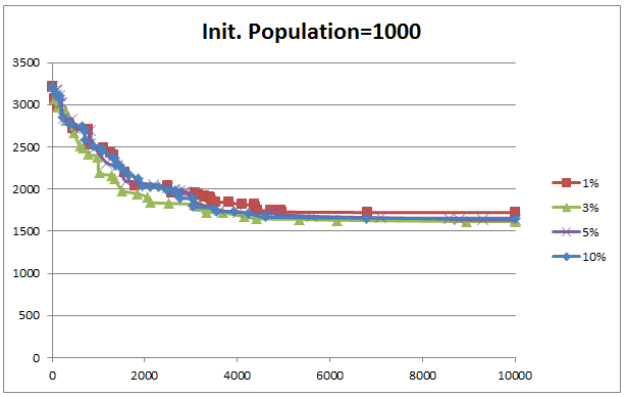


Figure 3 – Initial population size: 1000; pm=1%, 3%, 5% and 10%.

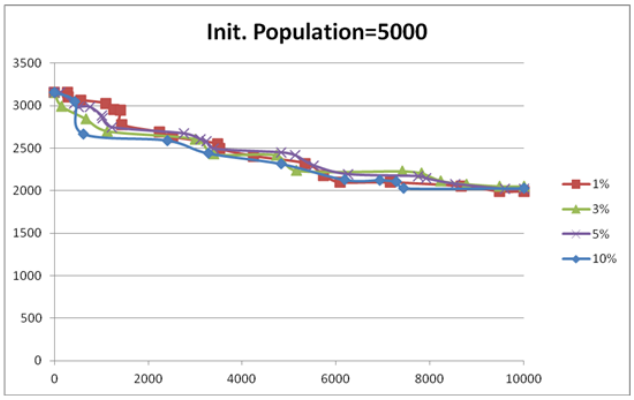


Figure 4 – Initial population size: 5000; pm=1%, 3%, 5% and 10%.

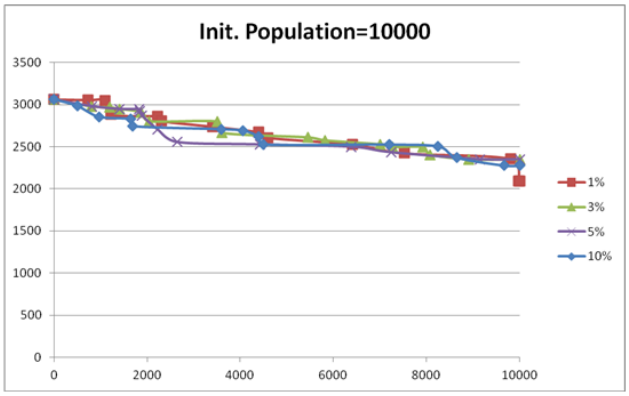


Figure 5 – Initial population size: 10000; pm=1%, 3%, 5% and 10%.

Now we present the results for the cases when, for some particular probability of mutation, we take different values for the parameter of the Initial Population. In Figure 6 we show the changes in fitness for different values of the number of initial population, having the same probability of mutation (pm=1%). Similarly, for other pm values (3%, 5% and 10%), in Figure 7, 8 and 9.

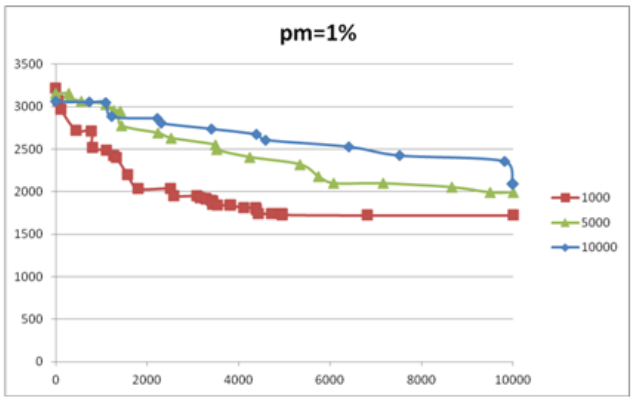


Figure 6 – pm=1%,; Initial population sizes: 1000, 5000 and 10000.

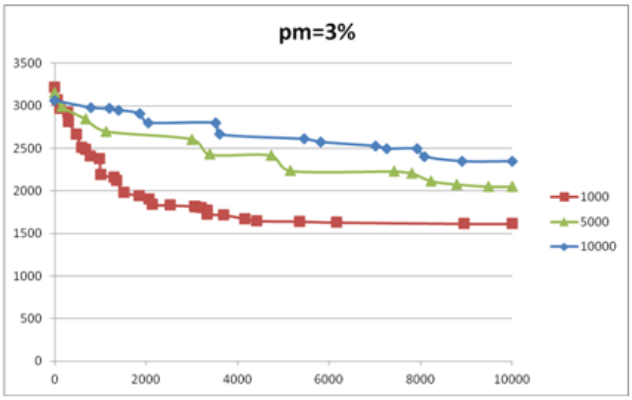


Figure 7 - pm=3%,; Initial population sizes: 1000, 5000 and 10000

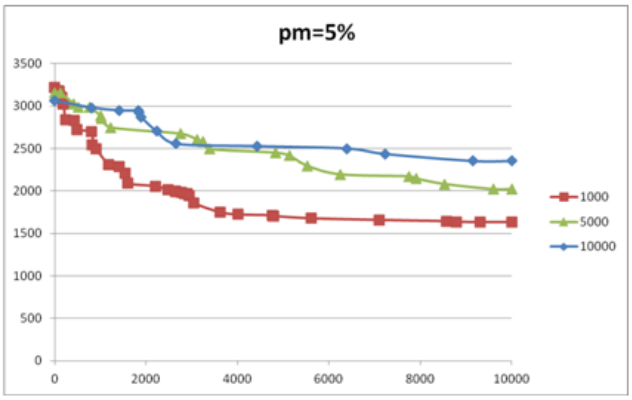


Figure 8 - pm=5%,; Initial population sizes: 1000, 5000 and 10000

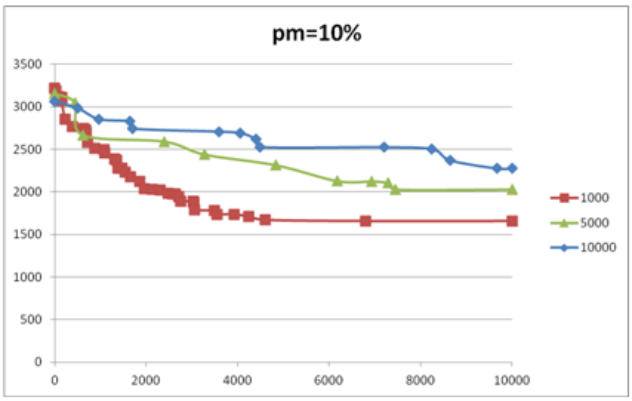


Figure 9 - pm=10%,; Initial population sizes: 1000, 5000 and 10000

4. Conclusions

Since the use of correct population size is a crucial factor for successful GA applications, more efforts need to be spent in finding correct population sizing estimates for particular problems. Our results show that there is a slight or no difference in fitness results for the case of TSP with GA, for different values of the initial population size, with all tested values of mutation parameter. So increasing the mutation rate doesn't contribute to a better results.

There is no use of increasing the value of the probability of mutation, since it doesn't pay-off as it increases the execution time/processor time and doesn't contribute to a much better fitness result.

Considering the effect of changing the initial population size for fixed mutation rate, we see that the effect of the mutation is important for small initial populations, since it contributes to promoting new solutions in the solution space.

For each percentage of the mutation, results show that the bigger the value of the initial population, the less is the fitness value improved.

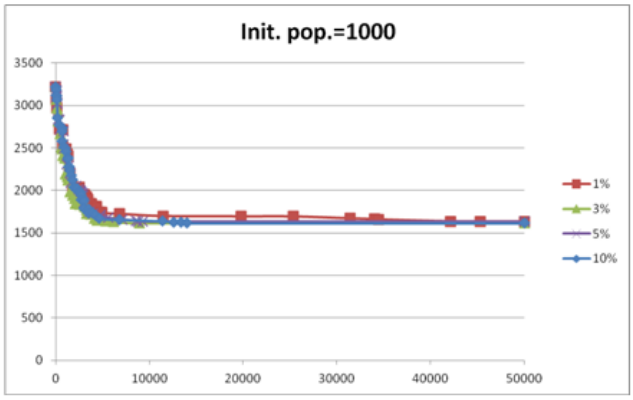
The effect of mutation is especially noticed in the first generations of the cases with small value of the intimal population (when there is no enough number of good solutions). Otherwise, when the initial population is of higher value, it means that there is a higher diversity of solutions and so the mutation doesn't bring much new and better solutions.

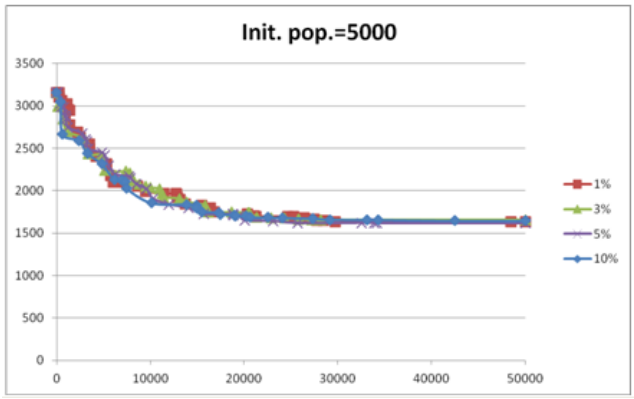
When the population is large, the diversity in the initial random population is large and the best solution in the population is expected to be close to the optimal solution.

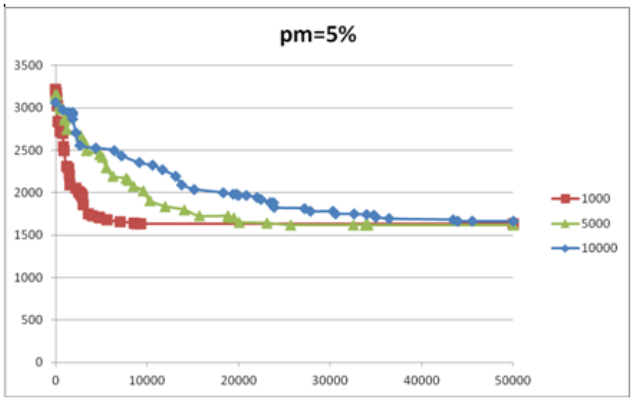
Appendix

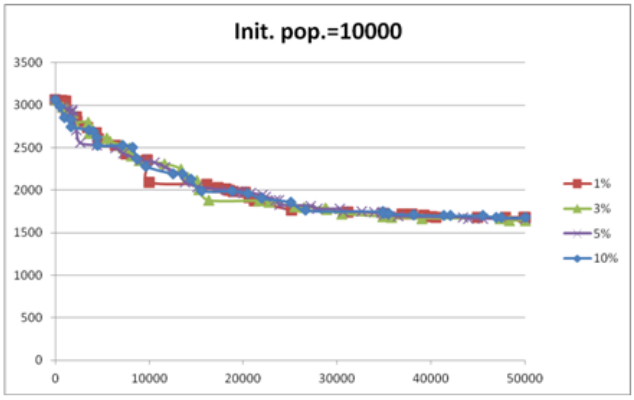
Here we will present the results for the cases when the value of the parameter of maximal number of generations is: 50,000 and 100,000.

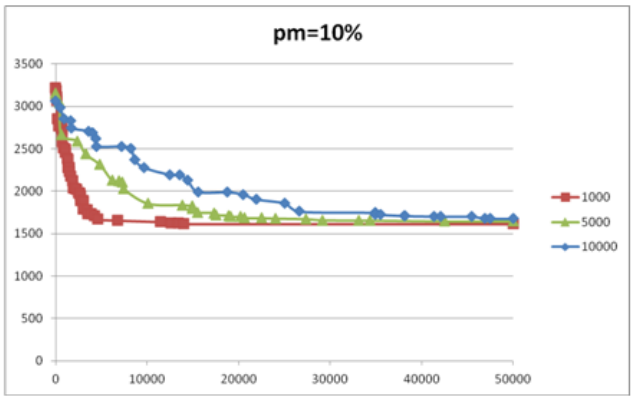
For each value of the number of generations, we present the results for cases with values of the initial population (1000, 5000 and 10000) for different values of the probability of mutation (1%, 3%, 5% and 10%) and then cases with same probabilities of mutation for different values of initial population (1000, 5000 and 10000) .

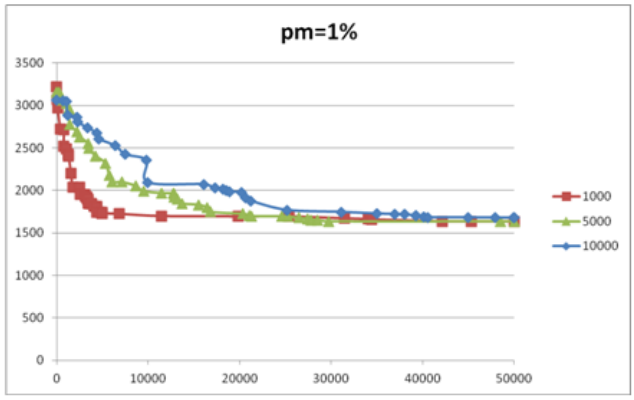


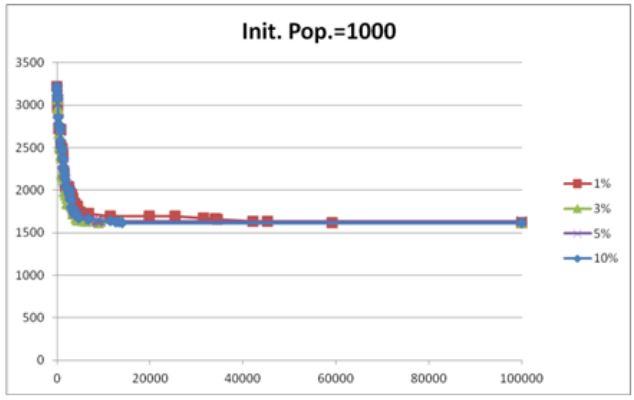


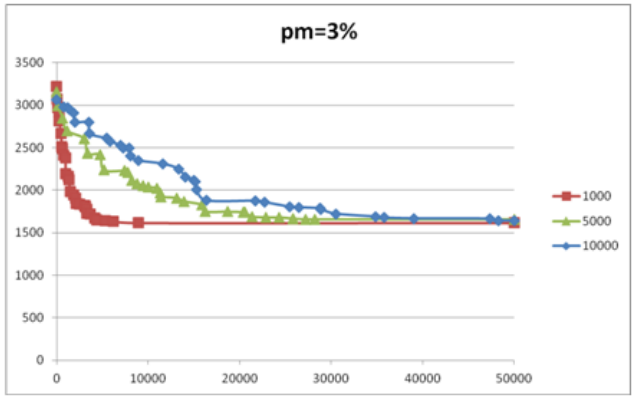


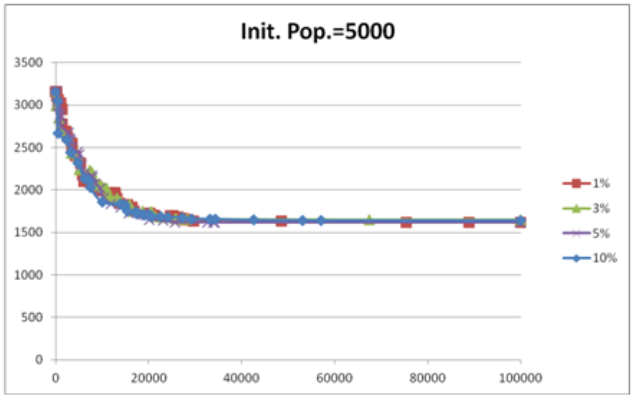


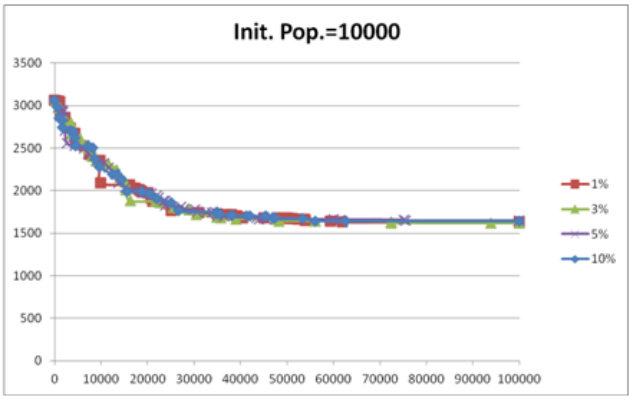


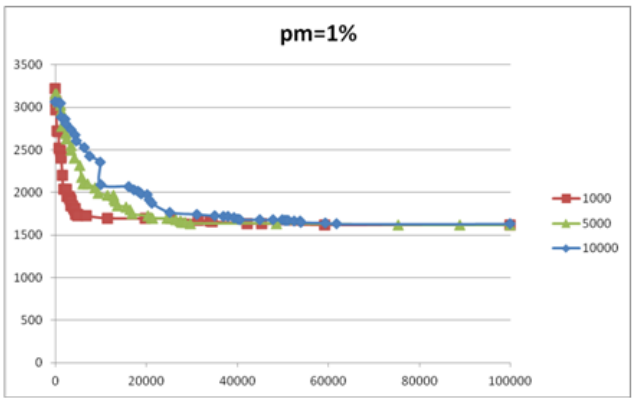


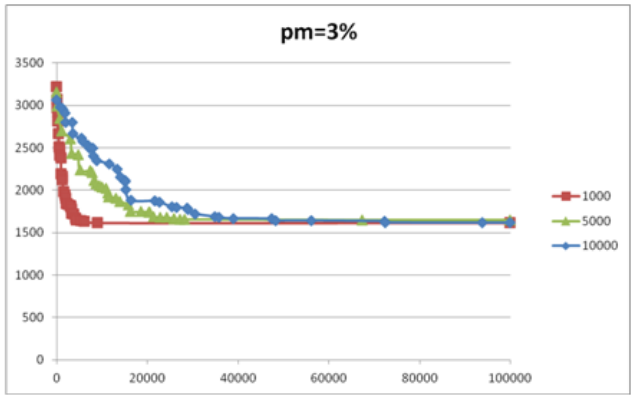


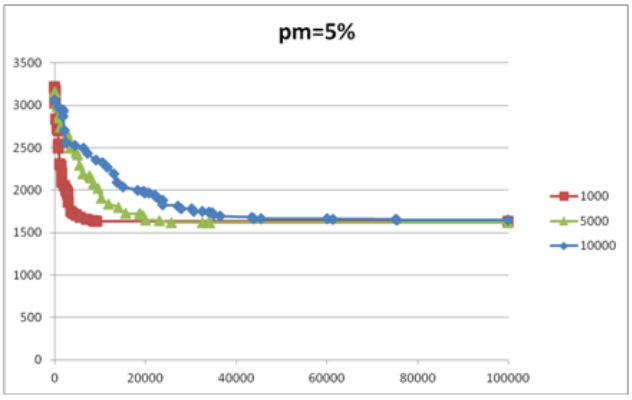












References

[1] J. H. Holland, *Adaptation in natural and artificial systems*. Ann Arbor, MI: University of Michigan Press, 1975.

[2] D. E. Goldberg, *Genetic algorithms in search, optimization, and machine learning*. Reading, MA: Addison-Wesley, 1989.

[3] Michael Junger, Gerhard Reinelt, Giovanni Rinaldi, *The Traveling Salesman Problem*, M.O. Ball et all, Eds. Handbooks in OR & Ms, Vol. 7, Elsevier Science, B.V. 1997

[4] T.-L. Yu, D. E. Goldberg, and Y.-P. Chen, "A genetic algorithm design inspired by organizational theory: A pilot study of a dependency structure matrix driven genetic algorithm," IlliGAL Report No. 2003007, University of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory, Urbana, IL, 2003.

[5] K. Sastry, D. E. Goldberg, and G. Kendall, "Genetic algorithms: A tutorial," in Introductory Tutorials in Optimization, Search and Decision Support Methodologies, ch. 4, pp. 97–125, Springer, 2005.

[6] Martin Pelikan, Genetic Algorithms, MEDAL Report No. 2010007, 2010.

[7] R.Sivaraj, T.Ravichandran Computer Engineering and Intelligent Systems, www.iiste.org, ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online) Vol. 3, No.1, 2012.

[8] K. Deb, S. Agrawal, Understanding Interactions among Genetic Algorithm Parameters, KanGAL Report Number 1999003.

[9] E.M. Khalilzad , S.Hosseini, ISSN (Online): Recovery of Faulty Cluster Head Sensors by Using Genetic Algorithm (RFGA), 1694-0814, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012.

[10] D. Anand, Feature Extraction for Collaborative Filtering: A Genetic Programming Approach, ISSN (Online): 1694-0814, www.IJCSI.org, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 1, September 2012.

[11] Ahmed Azouaoui, Ahlam Berkani and Pr. Mostafa Belkasmi, An Efficient Soft Decoder of Block Codes Based on Compact Genetic Algorithm, ISSN (Online): 1694-0814 www.IJCSI.org. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 2, September 2012.

[12] A. Rexhepi, A. Dika, A. Maxhuni, Solving TSP using Genetic Algorithm – Case of Kosova, WSEAS 6th WSEAS European Computing Conference (ECC '12), Prague, Czech Republic, 2012.

Avni Rexhepi, MSc. (DB: 12/12/1969). Master of Science in computer science (2004) - University of Prishtina. Teaching and research Assistant in the University of Prishtina, since 1996.

Adnan Maxhuni, MSc. (DB: 31/12/1967). Master of Science in computer science (2005) - University of Prishtina. Teaching and research Assistant in the University of Prishtina, since 1993.

Agni Dika, Prof. Dr. (DB: 21/02/1950). PhD in computer sciences (1989) - University of Zagreb, Croatia. Ordinary Professor at the Faculty of Electrical and Computer Engineering, in the University of Prishtina.

Interoperability between .Net framework and Python in Component way

M. K. Pawar[1], Dr. Ravindra Patel[2] and Dr. N. S. Chaudhari[3]

[1] Assistant Professor,
UIT, RGPV, Bhopal

[2] Associate Professor,
UIT, RGPV, Bhopal

[3] Professor, Deptt. Of CSE
IIT, Indore

Abstract

The objective of this work is to make interoperability of the distributed object based on CORBA middleware technology and standards. The distributed objects for the client-server technology are implemented in C#.Net framework and the Python language. The interoperability result shows the possibilities of application in which objects can communicate in different environment and different languages. It is also analyzing that how to achieve client-server communication in heterogeneous environment using the OmniORBpy IDL compiler and IIOP.NET IDLtoCLS mapping. The results were obtained that demonstrate the interoperability between .Net Framework and Python language. This paper also summarizes a set of fairly simple examples using some reasonably complex software tools**.**

Keywords: *Component Interoperability, Component objects, cross communication among .NET and Python.*

.

1. Introduction and Background

There is an increasing demand of development of component based technology in software industries [1]. For the reason that in another engineering discipline in which, components have successfully developed and also have adapted to build the systems. (For example, Civil engineering, Mechanical engineering Electronics engineering etc.). As a result we can also think the same concept in software engineering. CORBA is continuously progressing in the research area of Component based software engineering. Since, CORBA middleware makes available the common platform [2] for various oops based language, some of the languages are very powerful in terms of compatibility of CORBA and some languages are less supportive the CORBA middleware. In the past few years, component-based software's have been well developed and motivated, for example Enterprise Java beans EJB of Sun Microsystems, CORBA Component Model of the OMG (Object Management Group) and COM (Component object model), DCOM and COM+ of Microsoft. Still there is a need to lot of development in CORBA standards and services for language compatibility in component based technology.

The overviews of well developed components are as follows:

Microsoft's COM, DCOM and COM+: Microsoft has implemented a COM component to develop the desktop applications [3]. DCOM is being implemented to operate remote applications, and COM+ is a higher version of COM. The limitations of the above components are running under the windows operating system. The awareness of Microsoft system based tools is required to implement the above domain specific components.

Java Beans Components: SUN Microsystems are required the familiarity of enterprise Java beans and Remote method invocation (RMI) to develop the Java Beans component [4]. Components of Java beans are platform independent, which overcome some of the limitations of Microsoft's component. RMI is used to invoke the component of one Java program into another Java program within the network boundary. The limitation of RMI is also that, it runs only for Java based applications.

CORBA of OMG Group: The domain specific limitations of Microsoft's and SUN Microsystems, the OMG has launched common object request broker architecture (OMG/CORBA). The application developer

has to use CORBA component model [5], which cross the boundaries of domain specific applications. An ORB provides different services that enable the one component to communicate other component in a transparent manner. The CORBA supports the architecture of various programming language developed by different vendors. For the language and environmental interoperability [6], CORBA provides Interface Definition Language (IDL), which is used to implement the component in any programming language.

The ORB services are used for component communication as shown in figure: 1. The Stubs and Skeletons are generated for each component by using their IDL compiler, for example (C++ to IDL ACE+TAO, omniORB etc, Java to IDL, idlj, python to IDL, omniORBpy, and .NET to IDL, IDLtoCLS etc). Stubs and skeletons file play crucial role for client server communication.

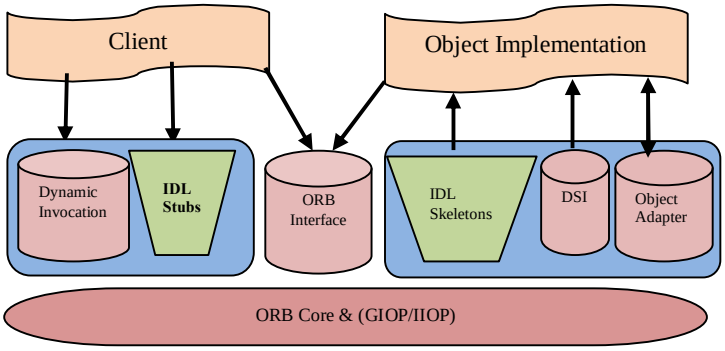


Fig1: ORB Architecture

To communicate with the ORB [7], the application uses a static IDL stub on the client end and static IDL skeleton at server end, which invokes the implementation of an IDL file that contains the interface definition.

2. Overview of IDL Compiler Tools

There are many IDL compiler tools were developed by different vendors and they successfully achieve the adaptive environment for most of the languages [8]. By using CORBA we can achieve various object-oriented languages interoperable in any environment and successfully build the component-based application [9] [3]. IDL compilers that support the CORBA [10] standard such as: **IIOP.NET**, interoperation between .NET, and CORBA or J2EE, Jacob wrote in Java IDL-to-Java Compiler, **R2CORBA,** a CORBA implementation of the Ruby Programming Language, VBOrb**,** CORBA Visual Basic clients and servers, MICO**,** IDL to C++ mapping, ACE ORB (TAO), IDL to C++ mapping, **omniORB**, ORB with C++ and Python bindings, ORBit**,** C and Perl bindings, *idlj - The IDL-to-Java Compiler* etc.
In our example we have used the omniORB IDL to python language mapping and **IIOP.NET** channel for C#.Net mapping, to make interoperable using CORBA middleware. Here we summarize the IDL compiler tools omniORB and IIOP.net channel:

2.1 OmniORB

The OmniORB [11] [12] is an Object Request Broker (ORB) that develop the specification of the Common Object Request Broker Architecture (CORBA). **OmniORB** is a robust high performance CORBA ORB for C++ and Python. It is an open source implementation and freely accessible under the terms of the GNU Lesser General Public License (for the libraries), and GNU General Public License (for the tools). OmniORB has always been designed to be portable. It runs on many versions of UNIX, Windows etc, It is designed to be easy to port to the new environment. The IDL to C++ mapping for all target platforms is similar. The main features of OmniORB are Multithreading and Portability. The major limitations of OmniORB are that it does not have its own interface repository and standard Portable Interceptor API.

2.2 IIOP Channel

The main requirement to communicate with the Common Object Request Broker Architecture (CORBA), a channel [10] for the IIOP protocol was implemented by Dominic Ullmann and Patrik Reali. The IIOP.NET does not provide interoperability with python ORB in different environments. The major importance of the IIOP.NET project has been to maintain the IIOP protocol between Java and .NET. However, IIOP.NET can also support the compatibility with C++ client and server through ACE+TAO.

3. Problem Definition

The main advantage of CORBA technology to achieve interoperability of component objects. We have implemented the two very simple client server model based on CORBA standards. IIOP.NET allows interoperation between .NET, CORBA and other distributed objects. This is done by incorporating CORBA/IIOP support into .NET, influencing the Remoting framework [10]. Since, python IOR does not support by .NET framework due to limitation of standards and lack of development in this area. We have implemented the python server that accepts the request of C#.Net Client. To test the efficiency of interoperable

objects based on CORBA an example has implemented with the aim to calculate the multiplication and division operation. C#.Net and python seems important in terms of efficiency of communication between component objects, to test interoperability of these objects in different environments. In our approach, First server has implemented in IIOP C#.NET for multiplication and second server has implemented in Python and integrated mixed C#.NET and Python. The proposed model is demonstrated in figure: 2 as follows:

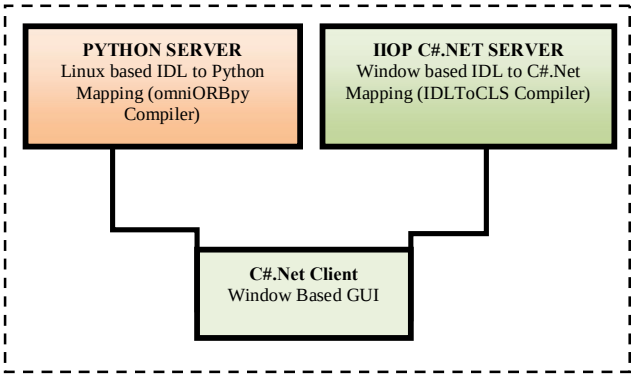


Fig2: Communication between Python Server, IIOP C#.Net Server and C#.Net Client

In the Client-Server model, python server generates an IOR (Interoperable Object Reference), this IOR is copied into the client implementation file to invoke the server object, if the client and server running on the same machine then it can easily copy from server to client. If the client and server running on different machines, then, we need to remotely copy the IOR from server to client. It is a little difficult for the developer. So, we have implemented an approach that overcomes to the remotely copy the IOR, and extend the interoperability between C#.net and python.

4. Example

In our example, the work carried out on a network of two different machines. IIOP C#.NET Server based on Microsoft's windows 7 Operating System, and **IIOP.NET-1.9.3**, IDLToCLS compiler, IDL to C#.NET mapping. C#.NET server computes the result of the multiplication operation by using input parameters. Python Server based on Ubuntu10.10 and **omniORBpy-2.7** IDL compiler, IDL to Python mapping. The python server computes the result of a division operation by using the same parameters. GUI based client has implemented in C#.Net, which passes the input parameters to the distributed objects and receives the result of multiplication and division by using the same input parameters The process diagram of the communication is shown in fig: 3

Procedure:

1. Launch the C#.Net Server and Python Server on Different Machine by using different port no.
2. Launch the client application that receives the input parameter and choice for multiplication and division operation
3. Client side we have two choices to choose the operation. Choose one for multiplication and another for division.
4. C#.Net client simply sends the input parameters to the C#.Net Server and python server.
5. These input parameters are passed into a python object that computes the result.
6. The Result is sent back to the client.

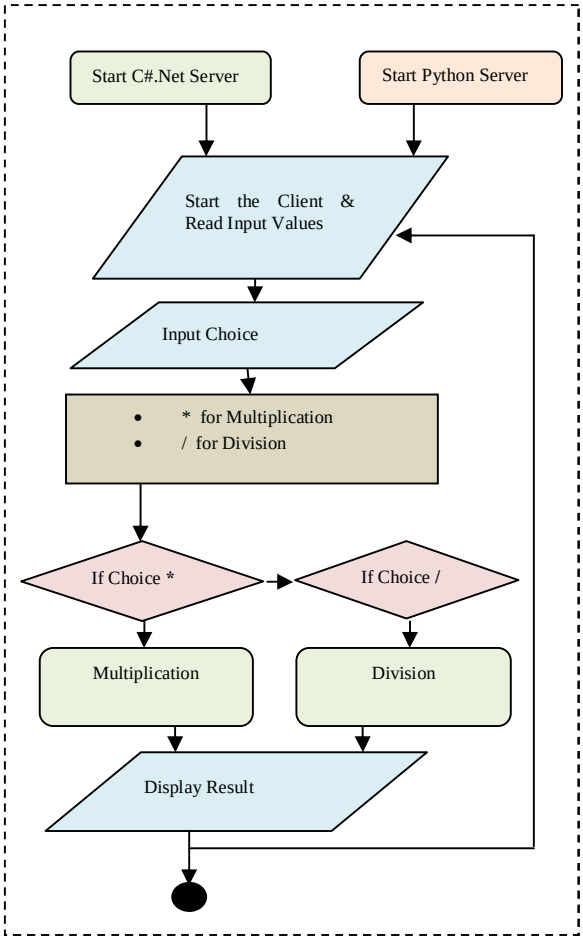


Fig: 3 Client- Server Communication Process

IDL Interface: straightforwardness of the IDL file is the proposed action [13][14]; this makes possible testing and directs to the transparency code. The IDL file content is

very simple as shown in fig: 4. The IDL contains the definition of multiplication and division and an interface containing the method to calculate the operation. After initialization the C#.net server, python server and C#.net client, the client receives two input parameters, to compute the result of multiplication and division operation.

```
//Division.idl
interface Division
{
double div(in double a, in double b);
};
```

```
//Multiplication.idl
using System;
using System.Runtime.Remoting;

namespace CalciClient
{
    public interface Multiply

    {
        double mul(double a, double b);
    }
}
```

Fig: 4 The IDL Interface for C#.Net and Python

The distributed object recognizes the assignment of finding an implementation repository for the input parameters and forwards the calculated result to the client. Python server code as shown in figure: 5, receives the input parameter and establish the connection between C#.net client and python server. This input parameter is passed to the python object. These input parameters are used by a python object to evaluate the result of division operation and the result is given to the C#.net client.

5. Results

There are two cases, in which we have evaluated the results:

Case1: **Communication between C#.Net server and C#.net client:**

Initially we start running the C#.net server and a Python server on a different machine by using different port no. as shown in figure: 6 and 8, and then we launch the C#.net client application. In this case, client-server communication using CORBA services is excellent due to the same environment of client and server.

```
#!/usr/bin/env python
import sys
import socket
from omniORB import CORBA, PortableServer

# Import the stubs and skeletons for the Example module

import _GlobalIDL, _GlobalIDL__POA

class division (_GlobalIDL__POA.Division):
   def div(self, a,b):
      return a/b

 # Initialise the ORB
orb = CORBA.ORB_init(sys.argv, CORBA.ORB_ID)
 # Find the root POA
poa = orb.resolve_initial_references("RootPOA")
 # Create an instance of Div
ei = division()
 # Create an object reference, and implicitly activate the object
eo = ei._this()

# calling python object using input parameter form C#.Net Client

x=ei.div(data[0],data[1])
print orb.object_to_string(eo)
conn.sendall(str(x))
print "Result of Division send to the client....."

poaManager = poa._get_the_POAManager()
poaManager.activate()
orb.run()
```

Fig: 5 Python Server code

The multiplication result, which is computed by C #.net server by using the input parameter, is shown in figure: 7

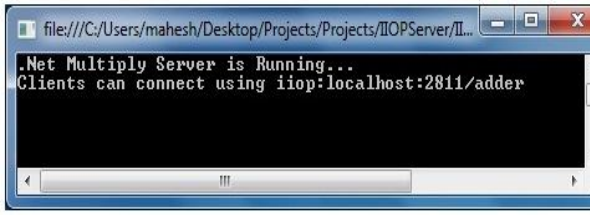


Fig: 6 Running process of C#.Net Server on port no.2811

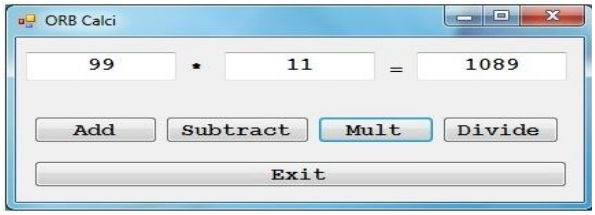


Fig: 7 Running process of C#.Net Client on port no.2811

Case2: **Communication between Python server and C#.net client:**

In case1, client server communication is very strong using the CORBA standard and services. But, in case2, shown in figure: 8, for client server communication in different environment and different language (e.g. .Net framework and python), python ORB & IOR does not support directly. As a result, we ultimately employ the python servant object all the way through communication. In this case, the copy of IOR is not requisite to the client for servant object invocation. As an alternative, we use the python object services on the server side.

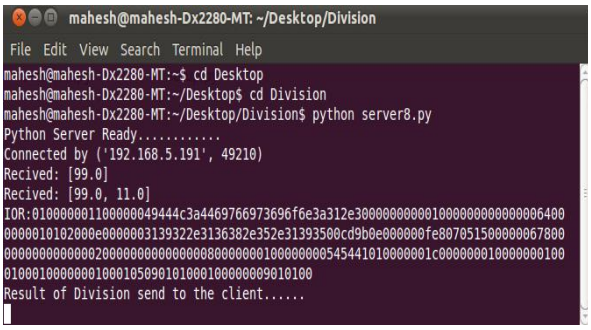


Fig:8 Running process of Python Server

The result of a division operation, which is computed by the Python server by using the same input parameter and CORBA, is shown in figure: 9

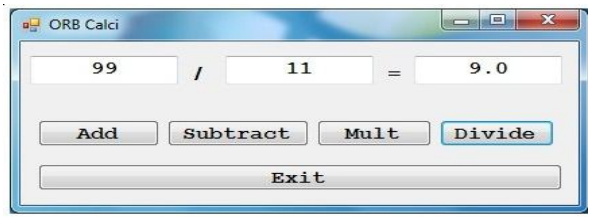


Fig: 9 Running process of C#.Net Client on port no.2809

6. Conclusion

In this paper we are presented opportunities of usage CORBA middleware standard and services for the component object for different programming languages .Net framework and python, with challenging importance on cross communication implementation. We have tested that how a C#. Net client written in C#, and running on Windows, can communicate directly with a .NET server, and python server by using the IIOP protocol and omniORBpy. Additionally, we have also seen how to write C#.Net client, using IIOP.NET, which can communicate with python server, running on Linux.

CORBA is a benchmark which supports the architecture of various programming languages, which makes it very reasonable means to implement the component based application. But, sometimes it may be challenging, As CORBA is used with various external tools such as ACE+TAO, omniORB, idlj, IIOP.net, omniORBpy etc.

Some component technology already exists, in which development of the application is not complicated because of their domain specific nature, but when we cross the domain, then there is need to such kind of standard and technology in which we can develop component based application. After the execution of implementation, it can be concluded that server and client in python and .NET framework is the most effective way for component object communication.

References

[1] F Bronsard, D Bryan, W Kozaczynski; Toward software plug-and-play SSR'97 Proceedings of the 1997 symposium on Software reusability Pages 19 – 29, ACM New York, NY, USA ©1997.

[2] Hall, L.; Hung, C.; Hwang, C.; Oyake, A.; Yin, J.; , "COTS-based OO-component approach for software interoperability and reuse (software systems engineering methodology)," Aerospace Conference, 2001, IEEE Proceedings. , Vol. 6, no., pp. 2871-2878 Vol. 6, 2001.

[3] Onderka Z., Cichy M.; The Comparison of the Communication Eciency for the CORBA and DCOM Standards in the Client Server Systems, Computer Networks, 2011. Will be published in Studia Informatica.

[4] Deitel & Deitel, 2001, Java How to write a program. USA, Prentice Hall.

[5] Z Onderka, The efficiency analysis of the object oriented realization of the client server systems based on the CORBA standard publication published online January 23, 2012. DOI 10.4467/20838476SI.11.010.0296.

[6] Hill, J.H.; "Towards Heterogeneous Composition of Distributed Real-Time and Embedded (DRE) Systems Using the CORBA Component Model," Software Engineering and Advanced Applications (SEAA), 2011 37th EUROMICRO Conference on , vol., no., pp. 73-80, Aug. 30 2011-Sept. 2 2011.

[7] A Yahiaoui, J Hensen, L Soethout; Developing CORBA-Based Distributed control and building performance environments by run-time coupling , International Conference on Computing in Civil and Building Engineering , ICCCBE , 10 , 2004.06.02-04 , Weimar.

[8] Object Management Group; Object Management Architecture Guide, OMG Document Number 92.11.1, Revision 2.0, 1992.

[9] Object Management Group; The Common Object Request Broker: Architecture and Speciation, OMG Document, Version 2.0., 1995.

[10] IIOP. NET-Documentation URL at http://iiop-net.sourceforge.net/documentation.html

[11] The omniORB version 4.1, User's Guide Duncan Grisby, Apasphere Ltd., Sai-Lai Lo, David Riddoch, AT&T Laboratories Cambridge, July 2009.

[12] Object Management Group (OMG), Object management architecture guide: revision 2.0

[13] Corba 3 fundamentals and programming, John Wiley & Sons, 2000 - Computers.

[14] M. K. Pawar, Dr. Ravindra Patel, Dr. N. S. Chaudhari; "Way to Component-based Vending Machine," CIIT International Journal of software engineering, 2012 , Vol.4, no.10. , pp. 447-451, Nov. 2012.

M. K. Pawar, Assistant Professor, Department of Information Technology at Rajiv Gandhi Proudyogiki Vishwavidyalaya (State Technological University of Madhya Pradesh), Bhopal, India. He has M. Tech. Degree in Information Technology. He posses more than 12 years of experience in the industry as well as teaching of graduate and postgraduate classes. He has published 02 papers in international journals and conference proceedings. He is a member IEEE.

Dr. Ravindra Patel, Associate Professor and Head, Department of Computer Applications at Rajiv Gandhi Proudyogiki Vishwavidyalaya (State Technological University of Madhya Pradesh), Bhopal, India. He has been awarded Ph.D. degree in Computer Science. He posses more than 12 years of experience in teaching postgraduate classes. He has published more than 15 papers in international journals and conference proceedings. He is a member of the International Association of Computer Science and Information Technology (IACSIT) & IEEE.

Dr. Narendra S. Chaudhari Professor, Department of Computer Science, Indian Institute of Technology (IIT) Indore, MP and Member - Advisory Board, ITM University, Gwalior (M.P.). He has been referee and reviewer for a number of premier conferences and journals including IEEE Transactions, Neurocomputing, etc. Dr. Chaudhari is Fellow of the Institution of Engineers, India (IE- India), as well as Fellow of the Institution of Electronics and Telecommunication Engineers (IETE) (India), senior member of Computer Society of India, Senior Member of IEEE, USA, Member of Indian Mathematical Society (IMS), Member of Cryptology Research Society of India (CRSI), and many other professional societies.

Intelligent Car Parking Management System On FPGA

Rehanullah Khan[a], Yasir Ali Shah[b], Zeeshan Khan[c], Kashif Ahmed[ad], Muhammad Asif Manzoor[c], Amjad Ali[a]

[a] Sarhad University of Science and IT, Peshawar, Pakistan

[b] Institute of Business & Management Sciences, AU, Peshawar, Pakistan

[c] UET, Peshawar, Pakistan

[d] *i*Fahja Limited, Peshawar

Abstract— Car parking has become an immense issue, especially in big cities. There are two main reasons: Firstly, the growth in population, secondly, the security. Moreover, the car theft has become an evil art haunting drivers. In this paper, we provide an interface and a software/ hardware module for Intelligent Car Park Management System (ICPMS). The ICPMS will provide an extensive management for vehicles including parking facilities and security. The ICPMS is validated using a test case scenario and extensive experimentation proves the feasibility of the approach.

Index Terms—Car park management system, Verilog HDL, wireless sensor network, FPGA.

I. INTRODUCTION

Due to the technological innovations man is leading a comfortable life. But at the same moment these advancements have at times become troublesome. The number of people using their own cars has increased exponentially in the past ten or fifteen years. The car parking has become an immense issue especially in big cities. Two main reasons can be cited for this. One reason is the growth in population and the other is the security. Car theft has become an evil art nowadays. Now the question arises, is it possible to introduce such a system that would solve all these issues and will be intelligent too. We have provided an interface and software/ hardware module which is validated using a test case scenario. The extensive experimentation proves the feasibility of the approach. ICPM solves all the issues related to car parking such as finding free parking slots, improved invoice system and certainly the security issues. The work is aimed at providing such a system that would be feasible in the third world countries like Pakistan. Our approach is cost effective and it covers all the features of a complete intelligent car parking management system.

The central idea of the project came from the troubles we face in parking our cars in our daily routine. The inspiration was always there but it required a rock-solid approach. The nuisance of parking cars is escalating day by day. Indeed a good design was required. For that a literature survey was done so as to confirm that that this effort should not be a repetition of anything accomplished before. The reference paper [1] is about the Car-Park Occupancy Information System. This is implemented in Matlab and used cameras for finding the free parking slots. With this system, images captured by a surveillance camera were processed in real-time to identify the occupancies of the parking lots. This occupancy information is further processed by a central control unit and distributed to display panels located at strategic locations at the parking area. The drivers can easily find a vacant parking lot based on the information displayed on the panels. An approach using WSN (Wireless Sensor Network) based intelligent car parking system [2], in which wireless sensors are deployed into a car park field, with each parking lot equipped with one sensor node, which detects and monitors the occupation of the parking lot. The status of the parking field detected by sensor nodes is reported

periodically to a database via the deployed wireless sensor network and its gateway. A camera based surveillance system [3], uses sensor nodes equipped with low-cost microphones to localize acoustic events such as car alarms or car crash sounds. [4] Presents another parking scheme for the car parking management systems, it uses the vehicular communication for finding the free slot in a congested car park and theft prevention. It provides real-time parking navigation service and also helpful in theft prevention and provides the drivers parking information. Dusan et al. [5] proposed a technique for car park management based on combination of fuzzy logic and integer programming techniques provided an online mechanism for the acceptance and rejection of car driver's request for parking. Firstly it developed a number of best parking strategies for different situations and then used learning algorithm to choose best solution a specific situation based on the training data. A parking system for guiding the drivers to an appropriate parking using the PGI (Parking Guidance Information) signs and the arrival time estimation at the park was based on driver characteristics, trip patterns, car park attributes and the car park availability [6]. The underlying assumption of this model is that the choice of the car park does not change after entering the city even if the statistics are changed then the initially perceived. This model does not provide any security and theft prevention. Shuo-Yan et al. [7] proposed an intelligent agent system that helped in selecting the optimal price car park. Bong et al. [8] used an image based parking system for finding out the vacant parking lot in a congested car park. Security surveillance cameras were used for acquiring the images. This background study has provided us an in-depth knowledge of the current existing car parking systems around the world.

II. ALGORITHM

We have divided the proposed ICPMS into following different modules.

A. Car Entering Module

In Car Entering Module, as the car enters the lot, it is detected by the IR Sensors. The IR Sensors provide the pulse to the FPGA which assumes that an input is detected and thus the car is entered into the parking lot. Now as the car enters the lot, the car is directed to park in the first empty slot available. This is an important feature because the user doesn't need to search for the empty slot rather it is directed to park in the empty slot number. Thus our parking system's approach is to provide ease to the users.

B. Car Exiting Module

In Car Exiting Module, as the car leaves the lot, it is detected by the IR Sensors. The IR Sensors provide the pulse to the FPGA which assumes that an input is detected and thus the car is exited out of the parking lot. A significant task here is to keep track of the slot number from which the car leaves. This slot number should be tracked so that at exit we can display the right invoice and the security code, which the user will provide, is correctly matched. As the car exits, it is shown an invoice depending upon its stay in the lot. Similarly, the user is asked to provide the security code which was assigned to it initially at the time of entering. As he enters the code, it is matched via the Security Code Module and if found correct, the car is allowed to exit.

C. Security Module

In Security Code Module, as the car enters the lot, it is assigned a security code. Now the user needs to keep this code with him at a safe place because when he will go out of the lot with his car, the Car Parking

Management System will ask for that code. He will be only allowed to leave the Parking Lot, if the given code is correct. Looking at the implementation point of view, the Security Code needs to be saved in a memory for further usage or requirements. Each slot number should have its own Security Code to distinguish between the slots and this code generation should be random and tough to crack because this code is the basis of all the security our system is providing. The type of code is an issue. We can have numbers or alphanumerical values or bits. We have chosen the bits as our code type, as they are difficult to crack.

D. Invoice Module

In Invoice Module, as the car exits the lot, it is shown the invoice or the bill. In other words the payment details are displayed to the user, who leaves the car park. If a day was passed, then the invoice changes accordingly. We have developed a procedure to calculate invoice. This formula keeps track of the time spent by each car in the Parking Lot.

III. DESIGN AND IMPLEMENTATION

In the design section the most significant was to outline the algorithm. Simplest possible algorithm was adopted. Fig. 1 is the flow chart of Entering module When the car enters the Car park the sensors at the main entrance detects the arrival of the car, After detecting the car's arrival by the sensors the capacity of the car park is checked, if there is a free slot in the parking lot, the car is allowed to enter the car park and a security token is assigned to it, otherwise it is shown that the park is full. After assigning the security token the car is allowed to park in specific location which is shown on the display. As our intelligent Car Park System keeps track of all the parked cars, free locations and the location where the next car should be parked so when the car is parked the corresponding values are updated, the free locations or free slots are decremented by one and the allotted slots values are incremented by one. And if this was the last free location in the lot then it is displayed that the park is full. At the end a counter is started to create invoice for it.

Fig. 2 shows the flow chart of Exiting Module. Initially the car is in the parked state, when it exits, the sensor in the slot detects it. After detecting the car the security token assigned is checked. If it is found correct the car is allowed to go to the next state which is the invoice payment, otherwise it is not allowed to exit. As our Intelligent Car Park Management System keeps track of all the parked cars and free locations, the corresponding values are updated. The free locations or free slots are incremented by one and the allotted slots values are decremented by one. The invoice is displayed to the car leaving the lot according to its stay in the Parking System.

IV. PRACTICAL REALIZATION

We have implemented the ICPMS using FPGA: Though we can realize the CMS on Microcontroller. However, the ICPMS is intended to be a modular system. In the next phase, we plan to integrate image based solution, therefore, the amount of data to be handled will be huge. A simple microcontroller is therefore not an option. Moreover, the intercommunication between the vehicles and the ICPMS will require a feasible solution and FPGA provides such an interface. The uniqueness of the ICPMS is mainly due to the target hardware i-e FPGA and splashing of the empty slot's number for a new user entering in the car park system. In this way the troubles of finding empty slot in a gigantic parking lot are resolved. Displaying the free slots available in the parking lot in different parts of a city is also an innovation in itself. The ICPMS was tested for 16 spaces.

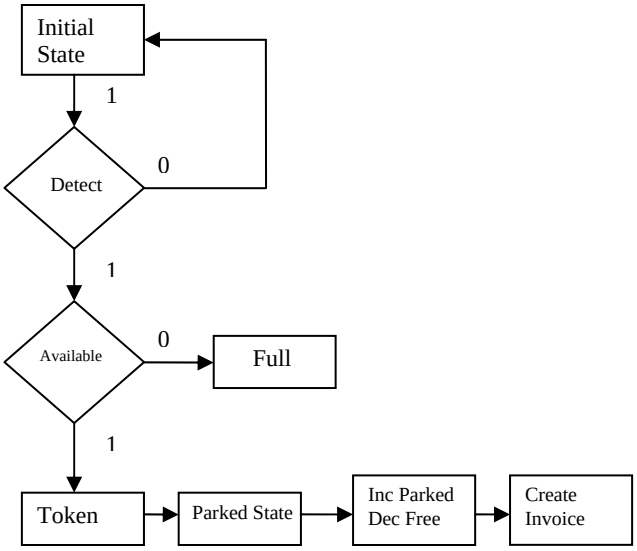


Fig.1 Flow Chart of the Entering Module

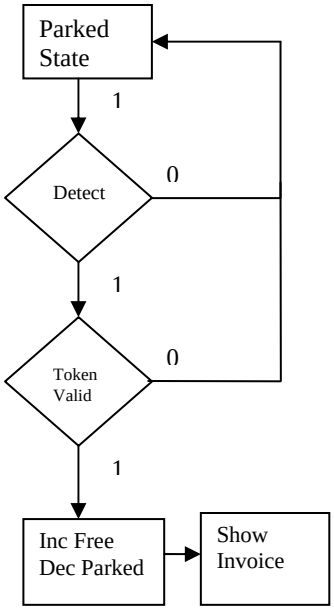


Fig.2 Flow chart of the Exit Module

Number of spaces can vary according to the capacity of a parking lot. The parking lot will sense or detect car's arrival or departure. As soon as a car enters the lot, a space is reserved for it and the space number flashed on the Display. This would guide the user to the allotted space in the lot. The system monitors the in and out traffic and updates the available spaces. When the parking lot is full, this information is flashed on the electronic boards. The system also monitors the spaces that are parked and are free. It has the knowledge about the number of cars parked, number of free slots in the Parking Lot at any instant. It creates Invoice for each entering car and splashing on the screen the net invoice whenever a car leaves. It also keeps track of the time a car stayed in the parking lot. For the security purpose, a security code is assigned to the arrived car and it is checked at the exit time, ruling out any thefts or security lapses.

V. RESULTS

After thoroughly analyzing the algorithms and design features, the programming code was written in Verilog HDL and implemented on FPGA. A hardware model of the project was accomplished. It was a complete Intelligent Car Parking Management System Model, which was tested and deployed in a parking lot. It was tested for 16 spaces. Number of spaces can vary according to the capacity of a parking lot. External hardware was interfaced with the FPGA. All the above mentioned features were checked and the project was thus acknowledged as a success. The security code feature, accurate calculation of invoice, space reservation for a newly arrived user in the parking system are the worth mentioning triumphs of our work.

VI. CONCLUSIONS

Our approach for finding the free parking slots is simpler as we don't need any camera and there is no involvement of image processing. Feeding data to a database can complicate the design even though will provide more services but we can achieve those features even without using it. Similarly WSN is an expensive and complicated technique when compared with

FPGA. We have used IR sensors for the detection of car's arrival and for the car's departure avoiding the expensive wireless sensors. Our design is simple yet it yields the desired result. It is cost effective and can be practically implementable even in countries like Pakistan. The idea was to keep it simple and innovative so that the parking system is cheap and at the same time provides the functionalities as per the constraints set at the time of designing. The automation of the project avoids any theft or mishaps which are very much probable in a common car park. Security has become a concern in our age and at the time of designing such ideas it should be scrutinized to the extent of avoiding it to the fullest. Such kind of Car Park Systems together with the security feature is really the need of the hour.

VII. FUTURE WORK

There is always room for improvement and our work can be modified and improved further. The pressure sensors can be used which are expensive but will yield accurate results. Every slot will have a separate sensor to detect the car's presence and will send the signals at the time of car's departure. Another feature that is number plate recognition can also be added to enhance the security of the Parking Lot. As mentioned in [1], cameras can be used to take the pictures of entering cars' number plates and then applying different algorithms for matching it with the originals at the time of exit. A surveillance system can be added to the system to make it more powerful. Feeding data to a database and keeping track of the frequent users visiting the parking lot and can give them discount. Many other ideas can be merged with the original project to improve its functionality and attain the perfection which is always desirable in works like these.

REFERENCES

[1] Bong, D.B.L. K. C. Ting, N. Rajaee, 2006. "Car-Park Occupancy Information System." Third Real-Time Technology and applications symposium, RENTAS 2006, Serdang, Selangor, December 2006

[2] V. Tang, Y. Zheng, and J. Cao, "An intelligent car park management system based on wireless sensor networks, "in Proc. of the First International Symposium on Pervasive Computing and Applications, Urumchi, Xinjiang, P. R. China, pp.65-70, August 2006.

[3] K. Na, Y. Kim, and H. Cha, "Acoustic sensor network-based parking lot surveillance system". In Proceedings of the 6th European Conference on Wireless Sensor Networks, EWSN (2009), ACM.

[4] Lu, R., Lin, X., Zhu, H., Shen, X.: "SPARK: A New VANET-Based Smart Parking Scheme for Large Parking Lots", In INFOCOM (2009)1413-1421

[5] Dusan Teodorovic and Panta Lucic, "Intelligent parking systems" in European Journal of Operational Research, 2006, vol. 175, issue 3, pages 1666-1681

[6] Thompson, Russell G., Takada, Kunimichi and Kobayakawa, Saturo, "Optimization of parking guidance and information systems display configurations" 2001, Transportation Research Part C 9, pp. 69-85.

[7] Shuo-Yan Chou, Shih-Wei Lin, Chien-Chang Li: Dynamic parking negotiation and guidance using an agent-based platform. Expert Syst. Appl. 35(3): 805-817 (2008)

[8] Bong, D.B.L. K. C. Ting, K. C. Lai, 2008. Integrated Approach in the Design of Car Park Occupancy Information System. IAENG Int. J. Comput. Sci., 35: 1-8.

Improvement in Accuracy for Three-Dimensional Sensor (Faro Photon 120 Scanner)

Mohd Azwan Abbas[1], Halim Setan[2], Zulkepli Majid[2], Albert K. Chong[3], Lau Chong Luh[2], Mohd Farid Mohd Ariff[2], Khairulnizam M. Idris[2]

[1] Department of Geomatics Science, Universiti Teknologi MARA Arau, Perlis, Malaysia

[2] Department of Geomatic Engineering, Universiti Teknologi Malaysia Skudai, Johor, Malaysia

[3] Department of Geomatic Engineering, University of Southern Queensland, Australia

Abstract

The ability to provide actual information and attractive presentation, three-dimensional (3D) information has been widely used for many purposes especially for documentation, management and analysis. As a non-contact 3D sensor, terrestrial laser scanners (TLSs) have the capability to provide dense of 3D data (point clouds) with speed and accuracy. However, similar to other optical and electronic sensors, data obtained from TLSs can be impaired by errors coming from different sources. In order to ensure the high quality of the data, a calibration routine is crucial for TLSs to make it suitable for accurate 3D applications (e.g. industrial measurement, reverse engineering and monitoring). There are two calibration approaches available: 1) component, and 2) system calibration. Due to the requirement of special laboratories and tools to perform component calibration, the task cannot be carried out by most TLSs users. In contrast, system calibration only requires a room with appropriate targets. Through self-calibration, this study involved a system calibration for Faro Photon 120 scanner in a laboratory with dimensions of 15.5m x 9m x 3m and 138 well-distributed planar targets. Four calibration parameters were derived from well-known error sources of geodetic instruments. Data obtained using seven scan stations were processed, and statistical analysis (e.g. t-test) shows that all error models, the constant error (8.9mm), the collimation axis error (-4.3”), the trunnion axis error (-11.6”) and the vertical circle index error (8.0”) were significant for the calibrated 3D sensor.

Keywords: *3D sensor, terrestrial laser scanner, accuracy, systematic errors, self-calibration.*

1. Introduction

Recently, three-dimensional (3D) model has been widely used for many purposes such as reverse engineering, medical, accident mapping, facility management, industrial measurement, monitoring and city modeling. In order to provide 3D information, there are several methods which can be used to acquire 3D data either using contact or noncontact scanners. Coordinates Measurement Machines (CMMs) is an example of contact scanner, which is very popular among mechanical engineers. However, there is restriction on the size of the object part scanned and also it can be slow in data acquisition rate because each point is generated sequentially at the tip of the probe has become the main drawback of CMMs method [1]. The tacheometer is a noncontact based scanner which can give better accuracy but it’s not only slow and cumbersome (during data collection phase) but most of the time this method also fail to provide the amount of detailed required [2]. Photogrammetry also noncontact scanner can be used to obtained 3D data but it required extensive manual editing and refinement for modeling purposes. With the rapid increase in speed and accuracy, and capability to provide 3D data (point clouds) directly, terrestrial laser scanners (TLSs) make it much easier to produce 3D models. Furthermore, their costs and sizes also have been continuously decreased. For that reason, TLSs have been chosen by many researchers for 3D modeling applications.

However, similar to other surveying instruments, TLSs have to be examined and calibrated regarding the instrumental and non-instrumental errors. Furthermore, the precision and the accuracy of the measurements should be determined regularly. As discussed earlier, the performance of TLSs is impressive regarding the data acquisition rate and accuracy is at centimetre level or better. However, the user needs to understand which scanner is the best-suited for a specific application. Schulz [3] in his study has listed some typical applications for TLSs with respect to the scanner precision (Figure 1).

According to Abdul and Halim [4], precision is defined as the closeness of the agreement between independent test results obtained compared and the mean value. Accuracy is defined as the closeness of the agreement between the result of a measurement and its true value. That means, even if a scanner is able to give better precision, it is not necessarily able to provide accurate measurement. This

argument arises because all electronic and optical instruments contain systematic errors. The precision can be determined by referring to manufacturer specification or by independent testing. Accuracy is different, it has to be evaluated through the deviation between the nominal and real value. In order to ensure the high quality of information provided by TLSs, calibration routine is very essential. Furthermore, the calibration process is very crucial to guarantee the data provided by the scanner meet the requirements of the job specifications.

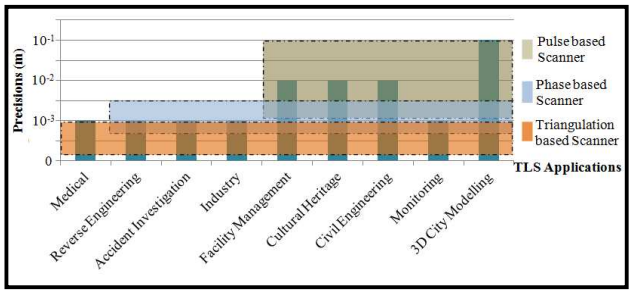


Fig. 1 : Applications of scanner with respect to the measurement precision [3].

2. Terrestrial Laser Scanners

TLSs is a non-contact sensor, optics-based instrument technology that collects three-dimensional (3D) data of a defined region of an object surface automatically and in a systematic pattern with a high data collecting rate. This capability has made TLSs widely applied for robust 3D reconstruction. In order to capture 3D point clouds that covering its entire field of view, laser source direction should be changed during scanning process. This can be performed either by rotating the laser source itself, or by using a system of rotating mirrors. The latter method is commonly used because mirrors are much lighter, faster and gives higher accuracy. This method may consist of either two scanning mirrors or one scanning mirror and a servomechanism. There are three different types of beam deflection units used in TLSs (Figure 1) as follows:

i. Oscillating mirrors;
ii. Rotating polygonal mirrors; and
iii. Monogon (flat) rotating mirrors.

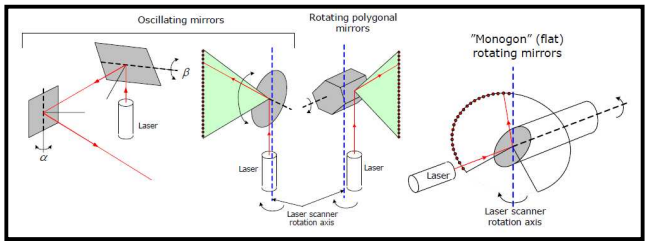


Fig. 2: Beam deflection units used in TLSs [6].

Figure 2 shows that the type of laser beam deflection unit which represents the field of view (FOV) of the TLSs. According to Staiger [5] and Reshetyuk [6], there are three classifications of TLSs based on FOV as follows (Figure 3):

i. Camera scanner;
ii. Hybrid scanner; and
iii. Panoramic scanner.

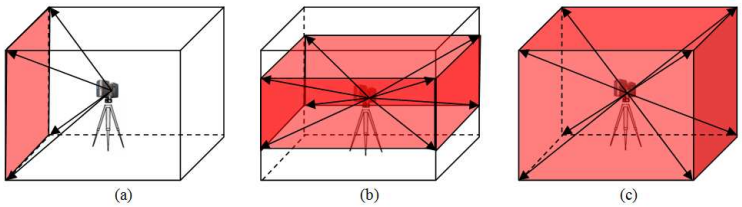


Fig. 3: Classification of TLSs based on field of view, (a) Camera scanner, (b) Hybrid scanner and (c) Panoramic scanner.

Camera scanner uses oscillating mirrors to deflect the laser beam about the horizontal and vertical axes of the scanner. The scanning head remains stationary during the scanning process. The system carry out their distance and angle measurement over a much more limited angular range and must be within a specific FOV (Figure 3a) of e.g. 40x40°, comparable to a photogrammetric camera [5].

Hybrid scanner has a horizontal FOV of 360° but a limited vertical FOV (Figure 3b). This scanner employs the oscillating or rotating polygonal mirrors (Figure 2) to deflect the laser beam in vertical and horizontal axes. With aid of servomotor, hybrid scanner is capable to be rotated by a small step around the vertical axis (horizontally). It works by scanning the vertical profile using a mirror system, and this process is repeated around the vertical axis until the scanner rotates a full 360°.

Monogon mirror used in panoramic scanner has improved the vertical FOV compared to hybrid scanner (Figure 3c). Using the same mechanism as hybrid scanner which is based on servomotor, this scanner is also capable of providing 360° horizontal FOV. These advantages of having a 360° horizontal FOV and nearly the same amount for vertical FOV has made panoramic scanner very useful for indoors scanning.

3. Calibration of Terrestrial Laser Scanners

There are many error sources to be modeled in TLSs measurements as discussed by Schulz [3], Böhler et al. [7], Gordon et al. [8] and Lichti [9]. Two approaches are available to investigate those errors, either separately (component calibration) or simultaneously (system calibration) which are based on statistical analysis (Figure 4).

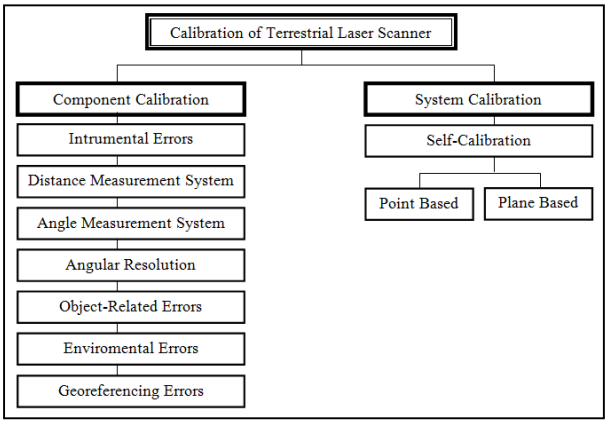


Fig. 4: Calibration procedures for terrestrial laser scanners.

2.1 Component Calibration

According to Schulz [3], component calibration requires precise knowledge of the scanner error model, and individual error is investigated separately in a specific experimental setup. All of these errors are identified separately in component calibration. In order to carry out this type of calibration, special facilities and device are required (Figure 5). Other than being used for calibration purposes, component calibration also performed to compare the performance of scanners from different models and manufacturers. Many studies regarding component calibration were made by Schulz [3], Gordon et al. [8], Brian et al. [10] and Kersten and Mechelke [11].

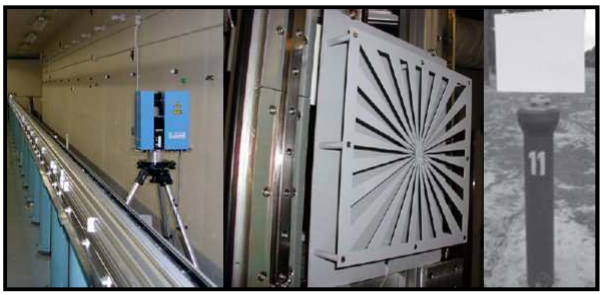


Fig. 5: Facilities and devices required for component calibration [3,8,10].

2.2 System Calibration

System calibration is generally used for the determination of all geometric parameters of a complete measurement system, which includes the interior (calibration parameters) and exterior orientation parameters of all the system components [12]. This calibration can be performed through self-calibration techniques. According to Reshetyuk [6], self-calibration for TLSs is the determination of all systematic errors of a terrestrial laser scanner simultaneously with all other system parameters.

In contrast to the component calibration, performing self-calibration doesn't require special facilities or devices, only a room with appropriate targeting is required [13]. In order to de-correlate model variables and also to maximise the accuracy of the estimated systematic error parameters, the network used for the calibration should be designed carefully as discussed in Lichti [9].

4. Geometric Model for Self-Calibration

Due to the very limited knowledge regarding the inner functioning of modern terrestrial laser scanners, most researchers have made assumptions about a suitable error model for TLSs based on errors involve in reflectorless total stations [9]. Since the data measured by TLSs are range, horizontal and vertical angle, the equations for each measurement are augmented with systematic error correction model as follows [6]:

$$\text{Range}, r = \sqrt{x^2 + y^2 + z^2} + \Delta r \quad (1)$$

$$\text{Horizontal\_direction}, \varphi = \tan^{-1}\left(\frac{x}{y}\right) + \Delta\varphi \quad (2)$$

$$\text{Vertical\_angle}, \theta = \tan^{-1}\left(\frac{z}{\sqrt{x^2 + y^2}}\right) + \Delta\theta \quad (3)$$

Where,

x, y, z = Cartesian coordinates of point in scanner space.

Δr, $\Delta\varphi$, $\Delta\theta$ = Systematic error model for range, horizontal angle and vertical angle, respectively.

Since this study was conducted on panoramic scanners (Faro Photon 120), the angular observations computed using Eq. (2) and (3) must be modified. This is due to the scanning procedure applied by panoramic scanner, which rotates only through 180° to provide 360° information for horizontal and vertical angles as depicted in Figure 6.

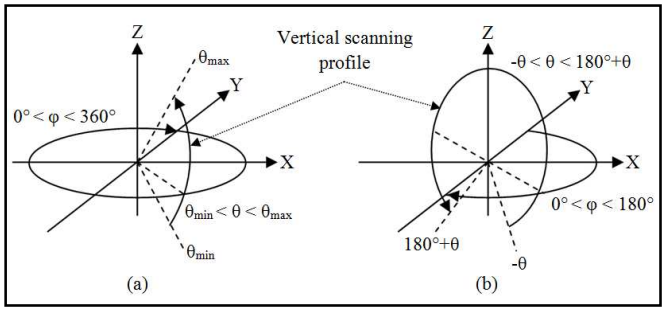


Fig. 6: Angular observation ranges for (a) Hybrid scanner and (b) Panoramic scanner.

Based on Lichti (2010), the modified mathematical model for a panoramic scanner can be presented as follows:

$$\varphi = \tan^{-1}\left(\frac{x}{y}\right) - 180° \quad (4)$$

$$\theta = 180° - \tan^{-1}\left(\frac{z}{\sqrt{x^2 + y^2}}\right) \quad (5)$$

The modified models above (Eq. 4 and Eq. 5) are only applicable when horizontal angle is more than 180° as shown in Figure 4. Otherwise, Eq. (2) and (3) will be used, which means that panoramic scanner has two equations for both angular observations.

According to Lichti [13], the systematic error models can be classified into two groups, physical and empirical parameters. The first group can be considered as basic calibration parameters which have been derived from the total station systematic error models. This group includes the constant, cyclic, collimation axis and, vertical circle index errors and others as described in Lichti and Licht [14]. The other group of error models is not necessarily apparent and may be due to geometric defects in construction and/or electrical cross-talk and may be system dependent. These are inferred from systematic trends visible in the residuals of a highly-redundant and geometrically strong, minimally-constrained least-square adjustment. Lichti [9] has identified 21 systematic errors model from phase-based scanner (Faro 880).

However, this study will focuses on the most significance systematic errors model as applied by Reshetyuk [6] in his study as follows:

i. Systematic error model for range.
$$\Delta r = a_0 \quad (6)$$

ii. Systematic error model for horizontal angle.
$$\Delta\varphi = b_0 \sec\theta + b_1 \tan\theta \quad (7)$$

Where,
b_0 = Collimation axis error
b_1 = Trunnion axis error

iii. Systematic error model for vertical angle.
$$\Delta\theta = c_0 \quad (8)$$

Lichti et al. [15] mentioned that systematic error models for panoramic scanner can be recognised based on the trends in the residuals from a least squares adjustment that excludes the relevant calibration parameters. In most cases, the trend of un-modelled systematic error closely resembles the analytical form of the corresponding correction model. Figure 7 shows the trend of the adjustment residuals for systematic error model.

Based on Figure 7, all systematic error models are identified by plotting a graph of adjusted observations against residuals. The graph of adjusted range against its residuals (Figure 7a) will indicate a constant error (a_0) if the trends appear like an sloping line. When residuals of the horizontal observations are plotted against the adjusted vertical angles a trend like the secant function, mean that the scanner has significant collimation axis error (Figure 7b). Trunnion axis error can be identified by having a trend like tangent function as shown in Figure 7c. For vertical index error, by plotting a graph of adjusted horizontal angles against vertical angles residual, this systematic error model is considered exist when the trend looks like the big curve as depicted in Figure 7d.

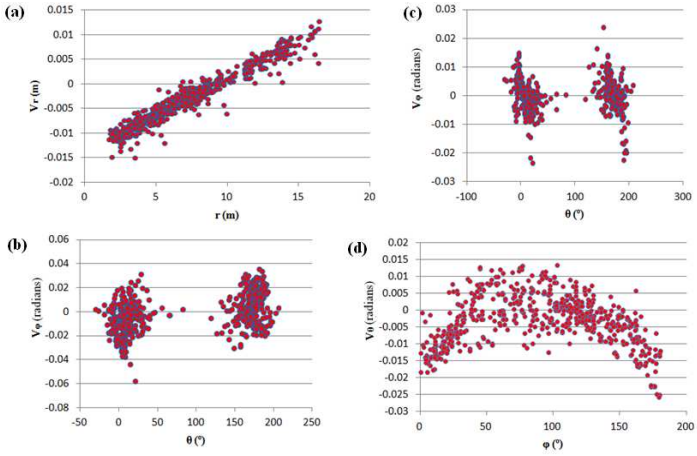


Fig. 7: Systematic errors for terrestrial laser scanner, (a) Un-modelled constant error, a_0, (b) Collimation axis error, b_0, (c) Trunnion axis error, b_1 and (d) Vertical circle index error, c_0.

In order to perform self-calibration bundle adjustment, the captured x, y, z of the laser scanner observations need to be expressed as functions of the position and orientation of the laser scanner in a global coordinate system [16]. Based on rigid-body transformation, for the j^{th} target scanned from the i^{th} scanner station, the equation is as follows:

$$\begin{aligned} x &= R_{11}(X_j - X_{Si}) + R_{21}(Y_j - Y_{Si}) + R_{31}(Z_j - Z_{Si}) \\ y &= R_{12}(X_j - X_{Si}) + R_{22}(Y_j - Y_{Si}) + R_{32}(Z_j - Z_{Si}) \\ z &= R_{13}(X_j - X_{Si}) + R_{23}(Y_j - Y_{Si}) + R_{33}(Z_j - Z_{Si}) \end{aligned} \quad (9)$$

Where,
$[x \quad y \quad z]$ = Coordinates of the target in the scanner coordinate system
${}_3R_3$ = Components of rotation matrix between the two coordinate systems for the i^{th} scanner station

$[X_j \quad Y_j \quad Z_j]$ = Coordinates of the j^{th} target in the global coordinate system

$[X_{Si} \quad Y_{Si} \quad Z_{Si}]$ = Coordinates of the i^{th} scanner station in the global coordinate system

5. Experiment Description

As shown in Figure 8, a self-calibration target field has been established in a laboratory with dimensions 15.5m x 9m x 3m. The 138 black and white targets were distributed on the four walls and ceiling based on conditions stated by Lichti [9].

Seven scan stations were used to observe the targets. As shown in Figure 9, five scan stations were located at the each corner and centre of the room. The other two were positioned close to the two corners with the scanner orientation manually rotated 90° from scanner orientation at the same corner. In all cases the height of the scanner was placed midway between the floor and the ceiling.

Fig. 8: Self-calibration for the Faro Photo 120 scanner.

In this experiment, the scan resolution was set to the 1/4 setting which is equivalent to the medium resolution. Higher resolution scans were not captured due to the longer time required to complete the scanning. Furthermore, medium resolution also was sufficient for Faroscene software to extract all targets except for those which have high incidence angle.

After the scanning and target measurement processes were completed, a bundle adjustment was performed with precision settings based on the manufacturer's specification, which were 2mm for distance and 0.009º for both angle measurements. After two iterations, the bundle adjustment process converged.

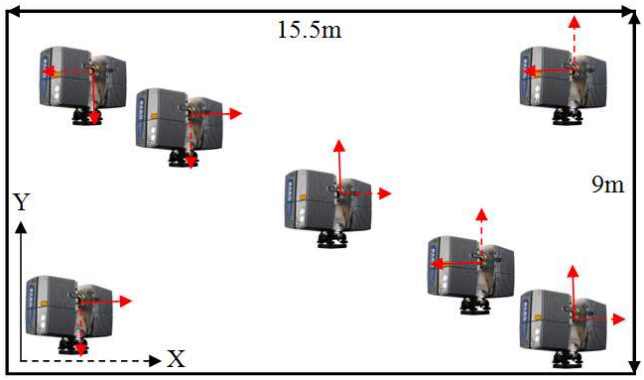


Fig. 9: Scanner locations during self-calibration.

5. Self-Calibration Results

In contrast with the hybrid scanner, the residual patterns of a panoramic scanner bundle adjustment can be used to detect the systematic error trends. As a result, other than statistical analysis, observation residual patterns are also used in this analysis. After performing the bundle adjustment process without any calibration parameters, residual patterns were plotted as a function of the adjusted observations as shown in Figures 10, 11 and 12.

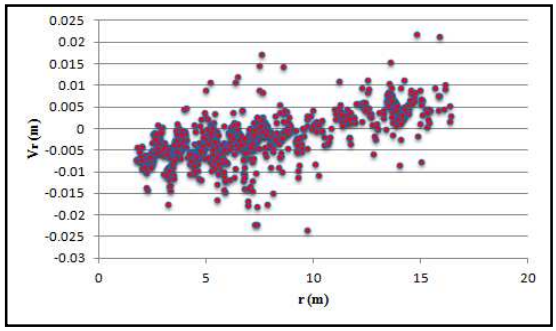


Fig. 10: Range residuals as a function of adjusted range for the adjustment without calibration parameters.

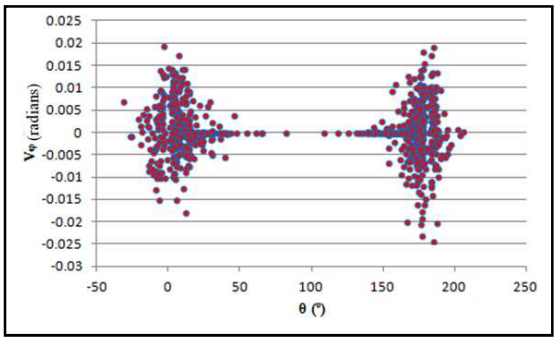


Fig. 11: Horizontal angle residuals as a function of adjusted vertical angles for the adjustment without calibration parameters.

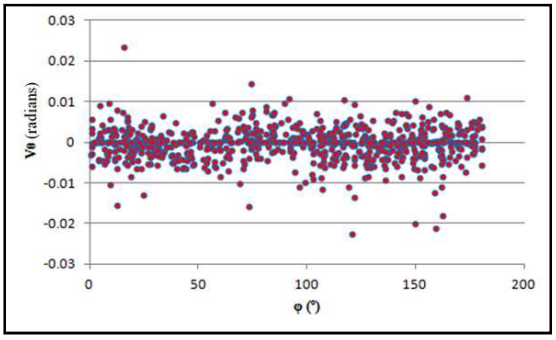


Fig. 12: Vertical angle residuals as a function of adjusted horizontal angles for the adjustment without calibration parameters.

Based on the sample of residual patterns shown in Figure 7, all significant systematic errors were investigated using the graphs from Figures 10 to 12. There are no systematic errors exhibited in both horizontal and vertical angles observations except for the range. The residual pattern graph has obviously demonstrated the trend of inclining line. Further analysis has been performed by running the bundle adjustment again using the calibration parameters. Results of the calibration parameters are shown in Table 1.

Table 1: Calibration parameters and their standard deviation

Calibration Parameters	$a_0 \pm \sigma_{a_0}$	$b_0 \pm \sigma_{b_0}$	$b_1 \pm \sigma_{b_1}$	$c_0 \pm \sigma_{c_0}$
Values (mm/")	8.9 ± 0.3	-4.3 ± 0.9	-11.6 ± 3.4	8.0 ± 1.1

Table 2 presents the RMS of residuals for each observable group for the cases without and with the self-calibration. The results of RMS have shows the improvement in accuracy for up to 27% by implementing self-calibration procedure.

Table 2: RMS of residuals from the adjustments without and with self-calibration.

Observable	RMS (without self-calibration)	RMS (with self-calibration)
Range (mm)	5.9	4.3
Horizontal angle (")	33.5	33.2
Vertical angle (")	21.4	21.3

In order to have a high accuracy solution regarding the significant of the estimated systematic error models, statistical tests were performed. All calibration parameters were tested to investigate their significant. The hypotheses were set as follows:

H_0 : The parameter is not significant.
H_A : The parameter is significant.

Using 95% confidence level, the results of the test are shown in Table 3.

Table 3: Significant test for calibration parameters

Number of scanner stations	7	
Degree of freedom	1928	
Critical value for 't' (95%)	1.645	
Calibration Parameters	**Calculated 't'**	**95%**
- Constant error (a_0)	2.967	Yes
- Collimation axis error (b_0)	4.778	Yes
- Trunnion axis error (b_1)	3.412	Yes
- Vertical circle index error (c_0)	7.273	Yes

Note: Yes – Significant, No – Not Significant

Results from Table 3 above show that null hypothesis was rejected for all parameters. This indicates that those parameters are significant to the scanner observations. Even though the graphs of residual pattern above (Figure 10, 11 and 12) illustrated that only constant error was present in the observation, but mathematically all of error models are significant. As a conclusion, to ensure that the calibrated scanner (Faro Photon 120) gives accurate measurement, all point clouds are needed to be refined by applying all four systematic error models of a_0, b_0, b_1 and c_0.

6. Conclusion

A self-calibration of the Faro Photon has been conducted over a dense 3D target field. The adjustment results were evaluated using graphs, which were based on residual pattern graph and mathematically utilising statistical analysis procedures. The differences between the RMS of residuals for adjustment with and without calibration parameters show an improvement up to 27%. Using the (t-test), the significant test was performed and the results show that all calibration parameters are statistically significant.

Acknowledgments

Authors would like to acknowledge the UiTM for the financial support for my PhD study. Special thanks goes to the Photogrammetry & Laser Scanning Research Group, INFOCOMM Research Alliance, UTM for the facility support in this project.

References

[1] Raja, V. and Fernandes, K.J. (2008). Reverse Engineering: An Industrial Perspective. Springer-Verlag London Limited 2008.

[2] Rabbani, T. (2006). Automatic Reconstruction of Industrial Installations Using Point Clouds and Images. A thesis for the degree of Doctor of Philosophy at TU Delft.

[3] Schulz, T. (2007). Calibration of Terrestrial Laser Scanner for Engineering Geodesy. A Dissertation submitted for the degree of Doctor of Scinces, Technical University of Berlin.

[4] Abdul, W. I. and Halim, S. (2001). Pelarasan Ukur. Kuala Lumpur: Dewan Bahasa dan Pustaka.

[5] Staiger, R. (2003). Terrestrial Laser Scanning: Technology, Systems and Applications. Second FIG Regional Conference, Marrakech, Morocco.

[6] Reshetyuk, Y. (2009). Self-Calibration and Direct Georeferencing in Terrestrial Laser Scanning. Doctoral Thesis in Infrastructure, Royal Institute of Technology (KTH), Stockholm, Sweden.

[7] Böhler, W., Bordas, V. M. and Marbs, A. (2003). Investigating Laser Scanner Accuracy. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXIV (Part 5/C 15), pp. 696-701.

[8] Gordon, S., Davies, N., Keighley, D., Lichti, D. and Franke, J. (2005). A Rigorous Rangefinder Calibration Method for Terrestrial Laser Scanners. Journal of Spatial Science, Vol. 50:2, 91-96.

[9] Lichti, D. D. (2007). Error Modelling, Calibration and Analysis of an AM-CW Terrestrial Laser Scanner System. ISPRS Journal of Photogrammetry & Remote Sensing 61 (2007) 307-324.

[10] Brian, F., Catherine, L. C. and Robert, R. (2004). Investigation on Laser Scanners. IWAA2004, CERN, Geneva.

[11] Kersten, T. and Mechelke, K. (2008). Geometric Accuracy Investigation of the Latest Terrestrial Laser Scanning System. FIG Working Week 2008, Stockholm, Sweden.

[12] Luhmann, T., Robson, S., Kyle, S. and Harley, I. (2006). Close Photogrammetry: Principles, Methods and Applications. Whittles Publishing, Dunbeath Mains Cottages, Dunbeath, Scoland, UK.

[13] Lichti, D. D. (2010). A Review of Geometric Models and Self-Calibration Methods for Terrestrial Laser Scanner. Bol. Ciênc. Geod., sec. Artigos, Curitiba (2010) 3-19.

[14] Lichti, D.D. and Licht, M. G. (2006). Experiences with Terrestrial Laser Scanner Modelling and Accuracy Assessment. IAPRS Volume XXXVI, Part 5, Dresden.

[15] Lichti, D. D., Chow, J. and Lahamy, H. (2011). Parameter De-Correlation and Model-Identification in Hybrid-Style Terrestrial Laser Scanner Self-Calibration. ISPRS Journal of Photogrammetry and Remote Sensing 66 (2011) 317-326.

[16] Schneider, D. (2009). Calibration of Riegl LMS-Z420i based on a Multi-Station Adjustment and a Geometric Model with Additional Parameters. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 38 (Part 3/W8)(2009) 177-182.

Mohd Azwan Abbas is currently PhD student at Department of Geomatic Engineering, Universiti Teknologi Malaysia. He received the B.Sc. (2004) And M.Sc. (2006) in Geomatic Engineering from Universiti Teknologi Malaysia. His current interests include the calibration and 3D modeling using terrestrial laser scanner.

Halim Setan is a Professor at Department of Geomatic Engineering, Universiti Teknologi Malaysia. He received M.Sc. (1988) in Geodetic Science from The Ohio State University, Columbus, USA and Ph.D. (1995) in Engineering Surveying from The City University, London, England. His current interests include the use of optical and range sensors for 3D reconstruction.

Zulkepli Majid is a Senior Lecturer at Department of Geomatic Engineering, Universiti Teknologi Malaysia. He received the M.Sc (1998), B.Sc. (2004) And Ph.D. (2006) in geomatic engineering from Universiti Teknologi Malaysia. His current interests include the use of optical and range sensors for 3D reconstruction.

Albert K. Chong is a Senior Lecturer at Department of Geomatic Engineering, University of Southern Queensland, Australia. He received the Ph.D. (1986) from University of Washington. His primary research focus is on the use of optical and range imagery for automated 3D object reconstruction.

Lau Chong Luh is currently PhD student at Department of Geomatic Engineering, Universiti Teknologi Malaysia. He received the B.Sc. (2012) in Geomatic Engineering from Universiti Teknologi Malaysia. His current interests include the use of terrestrial laser scanner for 3D topography.

Mohd Farid Mohd Ariff is a Senior Lecturer at Department of Geomatic Engineering, Universiti Teknologi Malaysia. He received the M.Sc (2005), B.Sc. (2002) And Ph.D. (2012) in geomatic engineering from Universiti Teknologi Malaysia. His current interests include the calibration and 3D reconstruction using photogrammetry technique.

Khairulnizam M. Idris is a Senior Lecturer at Department of Geomatic Engineering, Universiti Teknologi Malaysia. He received the M.Sc (2003), B.Sc. (2001) And Ph.D. (2011) in geomatic engineering from Universiti Teknologi Malaysia. His current interests include the 3D mapping via unmanned aerial vehicle and spatial adjustment.

A Method of Neural Network Internal Model Control in Unstable Time-lag Process

Liu Qi[1], Zhang Honghui[1], Shao Yonggang[2], Liu Kuili[3], Wang Jie[4], Chen Zhanwei[5] and Huang Zhenzhen[6]

**[1] Department of Physics and Electronic Engineering, Zhoukou Normal University
Zhoukou, Henan, PR China**

**[2] Henan Electric Power Industry School
Zhengzhou, Henan, PR China**

**[3] Deparement of Laboratory and Equipment Management, Zhoukou Normal University,
Zhoukou, Henan, PR China**

**[4] School of Electric Engineering, Zhengzhou Unisersity,
Zhengzhou, Henan, PR China**

**[5] Department of Computer Science, Zhoukou Normal University,
Zhoukou 466001, China**

**[6] Department of politics and law, Zhoukou Normal University,
Zhoukou 466001, China**

Abstract

The phenomenon of unstable time-lag process is usually familiar in the process industry, but it is hard to be controlled by the conventional method. In this paper a control method called double-loop feedback is put forward, first internal feedback stabilization is adopted, then neural network is used to form the internal model control system, finally it solves the problem of bias and instability between the model and the real process. Through the simulation, it is seen that the method has short adjusting time and high control accuracy, which shows the validity and superiority of neural network internal model control.

Keywords: *Unstable time-lag process; Internal feedback; RBF; Neural network; Internal model control; Double-loop control.*

1. Introduction

Superheated stem temperature of power plant has the highest temperature in steam-water channel of boiler, and the temperature of superheater is close to the limiting temperature of metal materials. If the temperature of superheated steam is too high, strength of the metal materials and service life of steam pipeline will decrease, also excessive thermal expansion in steam turbine will be caused, as a result metal of the high-pressure part will be damaged, but if the temperature of it is too low, the thermal efficiency of the equipments will be reduced, and when the steam temperature changes greatly, fatigue in piping material and related components will be caused, in consequence the steam turbine rotor and differential expansion will change, when serious, turbine vibration will occur, which is dangerous to production safety. The over-heat steam temperature system has some features, such as large delay, large inertia, integration, time-varying and so on, and the control quality directly affect safety and economy of the electric power production [1].

Time-lag process has the traits of great inertia, nonlinearity and uncertainty of model structure, for stable time-lag process, Smith predictor control system is utilized [2]. But when the controlled plant is unstable time-lag process, such as some chemical process, the conventional method can not reach satisfying control effect. According to this problem, large of research work has been done [3], [4], [5], [6], [7].

Phenomenon and things with uncertainty are generally existing in the nature and society. But how to express and deal with the uncertainty is a hot-spot and key point in the research on nature science, which is also a blockage at the same time. In all kinds of uncertainty, fuzziness and randomness are most important, which are paid more attention to.

Neural network has strong ability of nonlinear mapping, which can be used to approach the nonlinear model. In this paper, according to the unstable time-lag process, first internal feedback [8] is adopted to stabilize the generalized controlled plant, then neural network is used to form the internal model control system, which solves the problem

of bias between the model and the real process, as well as the problem that robustness and stability of closed-loop system are hard to be determined.

2. Internal feedback stabilization of unstable time-lag process

2.1 Unstable time-lag process

Superheated steam temperature system is a multiple input and single output object. There are several influence factors for temperature changes, such as steam flow rate, the heat of flue gas, water flow rate.

When the boiler load is disturbed, change of the steam flow will make the steam flow velocity change almost at the same time along different points of the entire superheater pipeline, thus if the convective heat transfer coefficient of the superheater changes, the steam temperature of each points of the superheater changes almost the same time.

Therefore the steam temperature responses fast, which has properties of time-delay, inertial and integration. Suppose both τ and T are small.

the dynamic characteristics of steam flow which is influenced by the change of steam flow is showed as equation (1).

$$G(s) = \frac{K}{Ts-1} e^{-\tau s} \tag{1}$$

As the integration process in equation (1) is difficult to control, some self-turning method [9], [10], [11], [12], [13] of integration process has large overshoot and long adjusting time. Because there are structural defects in the integration process, it is difficult to be controlled by traditional PID controller [8].

2.2 Control structure of internal feedback

According to unstable time-lag process, the double feedback circuit is put forward by Sung and Lee[8], for this algorithm, first a control structure of internal feedback is introduced, shown as fig.1.

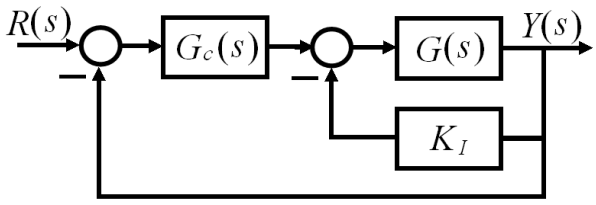


Fig. 1. The diagram for unstable time-lag process

The proportional controller K_I in the internal feedback circuit can change the unstable process into generalized stable processes, and the external feedback circuit can be designed by the expectation performance index.

By adding the proportional controller K_I in the internal feedback circuit, the close-loop transfer function is shown as equation (2).

$$\begin{aligned} G_p(s) &= \frac{G(s)}{1+K_I G(s)} \\ &= \frac{Ke^{-\tau s}}{Ts-1+KK_I e^{-\tau s}} \end{aligned} \tag{2}$$

Use Taylor series to expanse, it can obtain:

$$e^{-\tau s} \cong 1-\tau s+0.5\tau^2 s^2 \tag{3}$$

Combine equation (3) with $e^{-\tau s}$ in the denominator of (2), the model of second order delay is shown as equation (4).

$$G_p(s) \cong \frac{Ke^{-\tau s}}{0.5KK_I\tau^2 s^2+(T-KK_I\tau)s+KK_I-1} \tag{4}$$

According to louts criterion, to achieve stability of the system, the following equation should be meet.

$$\frac{1}{K} < K_I < \frac{T}{K\tau} \tag{5}$$

The gain of proportion controller which can greatly suppress disturbance is brought forward by Sung and Lee.

$$K_I = \frac{1}{K}\sqrt{\frac{T}{\tau}} \tag{6}$$

3 Method of neural network internal model control

3.1 Internal model control

Equation (2) is the mathematical model of the first-order unstable process after stabilization, and the second-order Taylor approximation is shown as equation (7).

$$G_p(s) \cong \frac{K(1-\tau s+0.5\tau^2 s^2)}{0.5KK_I\tau^2 s^2+(T-KK_I\tau)s+KK_I-1} \tag{7}$$

The general structural diagram of internal model control system [14], [15] is shown as Fig.2, in which $G_c(s)$ is the internal model controller, $G_p(s)$ is the controlled plant after internal feedback stabilization, and $\hat{G}_p(s)$ is the internal model.

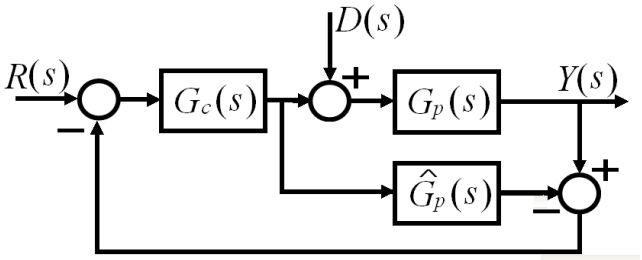


Fig. 2. Sstructure of internal model control

The closed-loop response of the system shown in figure 2 is:

$$\begin{aligned} Y(s) &= \frac{G_c(s)G_p(s)}{1+G_c(s)[G_p(s)-\hat{G}_p(s)]}R(s) \\ &+ \frac{1-G_c(s)\hat{G}_p(s)}{1+G_c(s)[G_p(s)-\hat{G}_p(s)]}D(s) \end{aligned} \tag{8}$$

If there is no bias in the model, that is $G_p(s) = \hat{G}_p(s)$, then equation (8) can be simplified as:

$$\begin{aligned} Y(s) &= G_c(s)G_p(s)R(s) \\ &+ [1-G_c(s)\hat{G}_p(s)]D(s) \end{aligned} \tag{9}$$

Overcoming the disturbance is a main task in the industrial process control, if the change of balance point is totally removed, the following equation has to be meet:

$$G_c(s) = \frac{1}{\hat{G}_p(s)} \tag{10}$$

3.2 Method of neural network internal model control

The structure of internal model control based on neural network is shown as Fig.3, separately two RBF is used to replace $G_c(s)$ and in Fig.2.

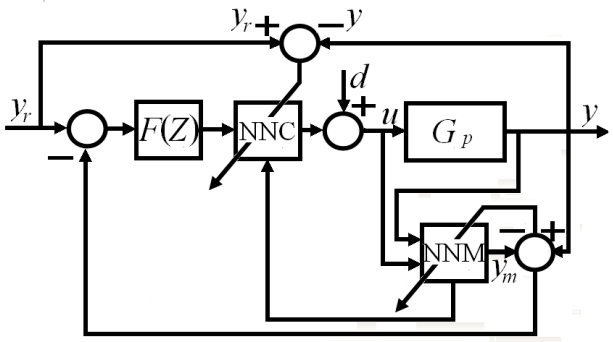


Fig. 3. Structure of neural network internal model control

NNM is the state estimator of RBF, which is parallel set according to the real system.

NNC is an inverse system model of RBF, neural network can correct the weighting coefficient according to the inputs and outputs, and finally control system the parameters.

The return signal is obtained from the difference of output between system and model, and is handled by NNC.

F(z) a linear filter, which is used to satisfy necessary robustness.

3.3 Design of the neural network internal model (NNM)

Generally, NNM is expressed by the following discrete-time nonlinear system:

$$y_m(k) = f[y(k-1),...,y(k-n), u(k-1),...,u(k-m)] + d(k) \quad (11)$$

The internal model is formed by RBF, and input layer is described as:

$$x_i(k) = \begin{cases} y(k-i), & 1 \le i \le n \\ u(k+n-i), & n+1 \le i \le n+m \end{cases} \quad (12)$$

The hidden layer is:

$$s_j(k) = \exp\left[\frac{\|x(k) - c_j(k)\|}{\sigma_j^2(k)}\right] \quad (13)$$

The output layer is:

$$y_m(k) = \sum_{j=1}^{a} s_j(k-1) v_j(k-1) \quad (14)$$

The performance index function is:

$$J = \frac{1}{2}[y(k) - y_m(k)]^2 \quad (15)$$

3.4 Design of neural network internal model controller (NNC)

The internal model controller is an inversion of the object model, and the inverse dynamic model is:

$$u(k) = f^{-1}[y_r(k+1),...,y_r(k-n+1), u(k-1),...,u(k-m+1), e_m(k)] \quad (16)$$

The nonlinear function of equation (16) is obtained by RBF, and the input layer is described as:

$$x_i(k) = \begin{cases} y(k-i), & 1 \le i \le n \\ u(k+n-i), & n+1 \le i \le n+m \\ e_m(k) \end{cases} \quad (17)$$

The hidden layer is:

$$h_j(k) = \exp\left[-\frac{\|x(k) - a_j(k)\|}{\varphi_j^2(k)}\right] \quad (18)$$

The output layer is:

$$u(k) = \sum_{j=1}^{b} h_j(k-1) v_j(k-1) \quad (19)$$

The performance index function is:

$$J_r = \frac{1}{2}[y_r(k+1) - y(k+1)]^2 \quad (20)$$

The output equation of close-loop system is determined by (21).

$$y(k) = \frac{u(k) G_p [y_r(k) - d(k)]}{1 + u(k)[G_p - y_m]} + d(k) \quad (21)$$

In the above equation G_p is the controlled plant.

The error equation of close-loop system output is:

$$E(k) = \frac{u(k) y_m - 1}{1 + u(k)[G_p - y_m]} [y_r(k) - d(k)] \quad (22)$$

Seen from equation (22) it can be known that NNM can totally describe the dynamic response, and NNC can totally describe the inverse dynamic response, at this time the error between step input and disturbance is $E(\infty)=0$, and the system can realize unbiased tracking to the input signal.

3.5 Procedure of RBF neural network internal model control algorithm

The concrete steps of algorithm are shown as follows:

Step1: Set k=1, select value domain and initialize the network function;

Step2: Calculate $u(k)$ by NNC;

Step3: According to equations (12)-(14), use $y_r(k)$, $y(k)$ to calculate $y_m(k)$;

Step4: Train forward model NNM by RBF neural network;

Step5: Train inverse model NNC by RBF neural network;

Step6: Set k= k+1, and return to Step2.

4 Simulation

The method of neural network internal model control is applied in unstable time-lag process.

The change of steam flow makes a fast reaction of the steam temperature, generally the gain is K=1~3, delay is τ=10~20s, and time constant is T=30~60s, in this paper, it is deemed that K=2, τ=10s and T=40s.

As a result the dynamic characteristics of steam flow is shown as equation (23).

$$G(s) = \frac{2}{40s-1} e^{-10s} \tag{23}$$

After adding the proportional controller K_I into the internal feedback circuit, according to the gain of K_I , the generalized stable processes can be obtained, shown as equation (24).

$$G_0(s) \cong \frac{e^{-10s}}{50s^2+10s+1} \tag{24}$$

The unit step response is shown as Fig.4, and it can be seen that the control method put forward in this paper has great control performance.

The unit ramp response is shown as Fig.5, and it can be seen that the control method put forward in this paper has great control performance.

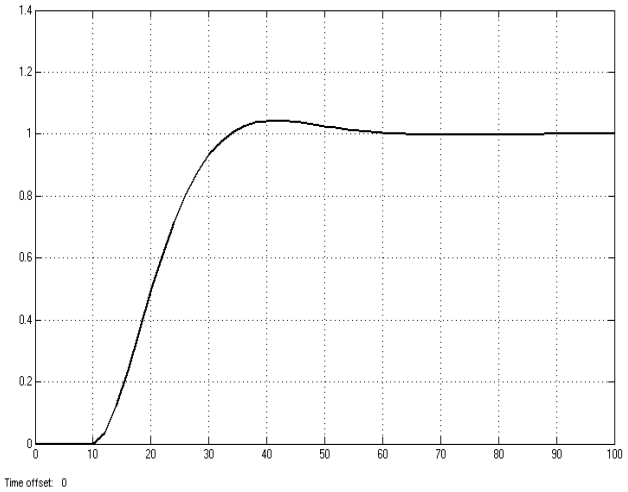


Fig. 4. Unit step response of neural network internal model control

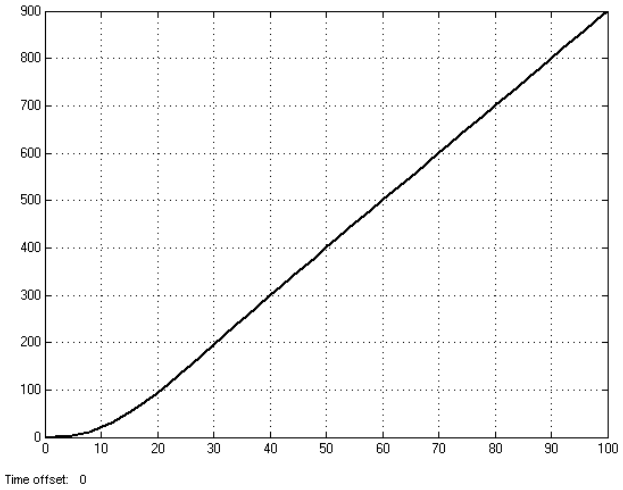


Fig. 5. Unit ramp response of neural network internal model control

5 Conclusion

The neural network has great learning function. In this paper, according to the problem that unstable time-lag process can not be well controlled by conventional method, a new algorithm is put forward, first internal feedback stabilization is adopted, then neural network is used to form the internal model control system, which solves the bias and instability between the model and real process. Through simulation of the first-order time-lag process,

method of neural network internal model control has got satisfying results, which shows the validity and superiority of the method.

5 Acknowledge

The research of this paper has been sponsored by Henan Provincial Research Foundation for Basic Research, China (Grant No.122300410168), Henan Provincial Research Foundation for Science and Technological Breakthroughs, China (Grant No.112102210485), Natural Science Foundation of He'nan Educational Committee, China (Grant No.2011B510021), Scientific Research Innovation Foundation for youth teachers of Zhoukou Normal University, China (Grant No. 2012QNA02).

References

[1] Han P, Wang G.Y., Wang D.F.. "On the application of predictive function control in steam temperature system of thermal power plant". IEEE Proc-Control Theory, 2004, 148(6): 135-138

[2] Smith O J M. "A controller to overcome dead time".ISA, 1959,6(2):28-33.

[3] Venkatashankar V, Chidambaram M. "Design of P and PI controllers for unstable first-order plus time delay systems". INT. J. Control, 1994, 60(1): 137-144.

[4] ZHANG Jian-hai, ZHANG Sen-lin, LIU Mei-qin. "Robust stability analysis of delayed discrete-time standard neural network".Journal of Zhejiang University: Engineering Science, 2009,43(8): 1383-1388 (In Chinese).

[5] Park J H, Sung S W, Lee I B. "An enhanced PID control strategy for unstable processes". Automatica,1998, 34(6):751-756.

[6] HAN An-tai, WANG Shu-qing. "Decentralized fuzzy control for a class of nonlinear interconnected large-scale systems with time-delay based on LMI approach".Control and Decision, 2004,19(4): 416-428 (In Chinese).

[7] YUAN Yu-hao, ZHANG Qing-ling, CHEN Bing. "Delay-dependent fuzzy control for nonlinear descriptor systems". Acta Automatica Sinica, 2006,32(5): 824-828 (In Chinese).

[8] SUNG S W, LEE I. "Limitations and countermeasures of PID controllers". Industrial & Engineering Chemistry Research, 1996, 35(8):2596-2610.

[9] KWAK H J, SUNG S W，LEE I B. "On-line Process identification and autotuning for integrating Processes". Industrial Engineering Chemistry Research, 1997, 36(12): 5329-5338.

[10] LUYBEN W L. "Tuning Proportional-Integral-Derivative Controllers for Integrator/Deadtime Processes". Industrial Engineering Chemistry Research, 1996, 35(10):3480-3483.

[11] WANG L, CLUETT W R. "Tuning PID controllers for integrating processes". IEE Proceedings-Control Theory & Applications, 1997, 144(5):385-392.

[12] SONG S H, CAI W J, WANG Y G."Auto-tuning of cascade control systems". ISA Transaction, 2003, 42(l):63-72.

[13]M. Mahlouji and A. Noruzi, "Human Iris Segmentation for Iris Recognition in Unconstrained Environments", IJCSI International Journal of Computer Science Issues, Vol. 9, No 3, 2012.

[14]Garcia C E, Morari M. "Internal model control 1. A, unifying review and some new results". Ind Eng Chem Process Des Dev, 1982, 21(2): 308-323.

[15]S. Nithyanandam, K. S. Gayathri, P. L. K. Priyadarsini, "A New IRIS Normalization Process For Recognition System With Cryptographic Techniques", IJCSI International Journal of Computer Science Issues, Vol. 8, No 4, 2011.

First Author: Liu Qi graduated from the School of Electric Engineering, Zhengzhou Unisersity, Zhengzhou, Henan, PR China, with a Bachelor degree in engineering science in 2004. He then, obtained PGD and MSc in Control Theory and Engineering from Zhengzhou University, Zhengzhou, Henan, PR China, in 2010. He joined the services of the Department of Physics and Electronic Engineering, Zhoukou Normal University, Zhoukou, Henan, PR China, from 2004. He has more than 15 published papers. His current research interests are in control theory and its application, pattern recognition.

Second Author: Zhang Honghui graduated from the North China Institute of Water Conservancy and Hydroelectric power, Zhengzhou, Henan, PR China, with a Bachelor degree in engineering science in 2004. He then, obtained PGD and MSc in Control Engineering from Zhengzhou University, Zhengzhou, Henan, PR China, in 2012. He joined the services of the Department of Physics and Electronic Engineering, Zhoukou Normal University, Zhoukou, Henan, PR China, from 2004. He has more than 10 published papers. His current research interests are in control theory and its application.

Third Author: Shao Yonggang graduated from the School of Electric Engineering, Zhengzhou Unisersity, Zhengzhou, Henan, PR China, with a Bachelor degree in engineering science in 2007. He then, obtained PGD and MSc in Systems Engineering from Zhengzhou University, Zhengzhou, Henan, PR China, in 2010. He joined the services of Henan Electric Power Industry School, Zhengzhou, Henan, PR China, from 2010. He has more than 10 published papers. His current research interests are in systems engineering.

Integration of Public Transportation through National e-Governance Service Delivery Framework

Ajay Kumar Bharti [1], Sanjay K. Dwivedi[2]

[1] Department of Computer Science, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, Uttar Pradesh - 226025, India

[2] Department of Computer Science, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, Uttar Pradesh - 226025, India

Abstract

Government of India has taken major initiatives and policy plans to accelerate the development and implementation of e-Governance to provide an appropriate environment by introducing G2G, G2B, G2C and G2E services within the country. Impact of e-Governance is gradually changing our life, from day to day access of information to access various services at our door steps. Public Transportation is also improving their mechanism for service delivery using ICT in their service delivery process. This paper discusses the current scenario of public transportation in India and various issues involved therein. It gives a brief idea of government initiative regarding structure and service delivery framework for e-Governance and their basic components in India. Further it discusses the integration and nationalization of public transportation through effective implementation of e-Governance in the sector. Finally we focused on integration of various State Road Transport Corporation's through a common service delivery gateway using existing National e-Governance Service Delivery Framework.

Keywords: *Integration, e-Governance, public transportation, service delivery gateway, domain gateway.*

1. Introduction

e-Governance can be defined [1] as "E-governance is the application of information & communication technologies to transform the efficiency, effectiveness, transparency and accountability of informational & transactional exchanges with in government, between govt. & govt. Agencies of National, State, Municipal & Local levels, citizen & businesses, and to empower citizens through access & use of information". Government of India trying to utilize ICT to improve its efficiency in service delivery, through implementation of e-Governance. The transport sector is one of them but it is limited to the Vahan and Sarathi e-Governance projects [2] for vehicle registration, driving license and for various certifications for drivers and conductors. The central government also promotes the use of ICT in public transportation too, by providing financial assistance to the sector for infrastructure developments to provide nationalized transportation up to year 2032 [3]. This aims to achieve improvement in service delivery mechanism to empower citizens or commuters through greater access to information and services through transparent and accountable governance in public transportation. The objective of the paper is to discuss various components of National e-Governance service delivery framework, issues and integration of public transportation with it. Further it discussed some core benefits of the integration of public transportation and finally concluded.

2. Public Transportation

The public transport sector (Bus based) provides an alternate mode of transportation. It makes the most optimum use of the available road space and fossil fuel by transporting the maximum number of people per unit of road space. Public transportation sector in the developing countries like India carries more than 90 percent of passengers by buses and about 65 percent of freight [4], even though the sector faces severe problems such as lack of infrastructure, comfortable buses and financial resources which restrict investment and up gradation of the existing transportation system. Moreover the lack of proper and effective planning in public transportation sector India has led to rapid growth in cars and two wheeled motor vehicle which causes congestion on roads that slows down the bus services or public transportation, ultimately increases the operating cost and discourage the use of public transportations [5]. Economical pressure and deficit

IJCSI
www.IJCSI.org

budgets pressurizes the public transport sector to improve operational effectiveness and efficient services. The public transport sector should use information and communication technology as a powerful tool to achieve operational effectiveness. Web technologies enabled the government and administration to reduce efforts and costs for their services. Effective implementation of e-Governance in the public transport sector will be able to minimize the economic pressure.

3. Issues in Public Transportation

Issues of public transportation have been raised in the extensive survey performed over commuters as well as on officials of public transportation on state of Uttar Pradesh in India to identify the needs of customers as well as officials to provide effective services to their commuters [6].

• Interconnectivity: Officials as well as commuter is enthusiastically requiring the need of interconnectivity between states to access interstate services. This will be helpful to improve business prospects for citizens as well as State Road Transport Corporation (SRTC) of the operating state.

• ICT enabled depots: Capacity building for development of ICT enabled depots for easy and rapid and service delivery to the commuters.

• Online services: Effective and efficient online applications will help to access G2B and G2C services to their commuters. This will help to provide services at the citizen's doorstep.

• Time bound grievance system: Time bound grievance system will force the SRTC to deliver quality services.

• GIS based system: Required a GIS based tracking services to foresight the approximate arrival of services at any source station.

The above issues can be resolved by using an integrated approach of public transportation because it can improve the interconnectivity between states that helps to improve business prospects for citizens as well as SRTC's of the states. Integration in public transportation will also help to improve the service delivery mechanism by using ICT enabled application. Using NeGP's National e-Governance service delivery framework an effective and efficient integrated model has been proposed in Fig. 3.

4. National e-Governance Service Delivery Framework

The National e-Governance Plan (NeGP) of the Government of India aims to make a framework for all government services accessible to the common man in his locality, through common service delivery outlets which ensure efficiency, transparency & reliability of such services at affordable cost.

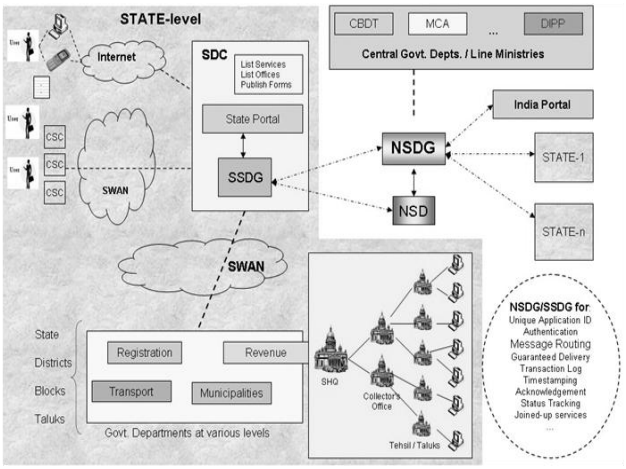


Fig. 1 NeGP's National e-Governance service delivery framework [7]

To meet this vision government's needed to cooperate, collaborate and integrate information across different departments in the Centre, States and Local levels. The given framework in figure1 is the NeGP's framework to deliver e-Governance services to the citizens as well as numerous department of the India, transport sector is also among them.

National e-Governance Service Delivery Gateway (NSDG): It is one of the Mission mode project (MMP) under the NeGP, NSDG can act as a standard based messaging switch to provide flawless interoperability and exchange of data across the departments. It acts as a nerve centre or middleware, would handle large number of transactions and helps in tracking, time stamping transaction log, joined up of services for all transactions of the governments.

State e-Governance Service Delivery Gateway (SSDG): SSDG is an attempt to reduce point to point connections between departments and provide a standardized interfacing, messaging and routing switch through which various players such as departments, front-end service access providers and back-end service providers can make their applications and data interoperable. The State e-Governance Service Delivery Gateway (SSDG) aims to achieve a high order of interoperability among autonomous and heterogeneous entities of the states based on a framework of e-Governance Standards as in figure 2.

National Service Directory (NSD): The NSD has utilized by all gateways across the country for address resolution of services. The primary function of the National Services Directory (NSD) is to provide a registry, which acts as a

IJCSI
www.IJCSI.org

service resolution point for all the services in the Gateway constellation. NSD is a collection of service hosting information outside the Gateway. All the Gateways that need to resolve services, which are not in their domain, need to resolve it at the NSD. The Gateways need to register with the NSD before they can attempt to resolve a service from the directory.

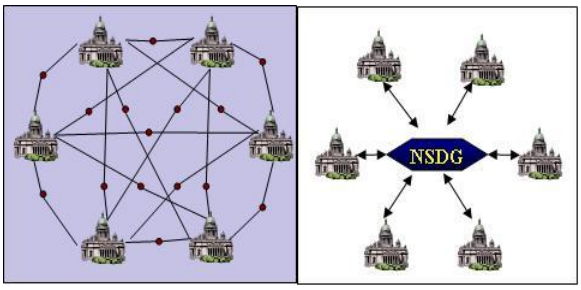


Fig. 2 SSDG interoperability mechanism by reducing point of connections

Domain Gateway: Domain gateways are purpose specific or department specific gateways to provide G2G, G2C and G2B service to the citizens. These are implemented for specific business needs requirements for the perspective projects to route request between front end and backend applications are known as domain gateways. Many government departments of centre and state have required domain gateways to satisfy their specific needs.

5. Integration of Public Transportation

SRTC's are geographically dispersed in all states of India. Technically it is not feasible to transform the all SRTC's together to integrate the services of public transportation. Therefore it needs a middleware like NSDG, so that all communications or information's are routed through the NSDG which controls communication, authentication, authorizations and security of application.

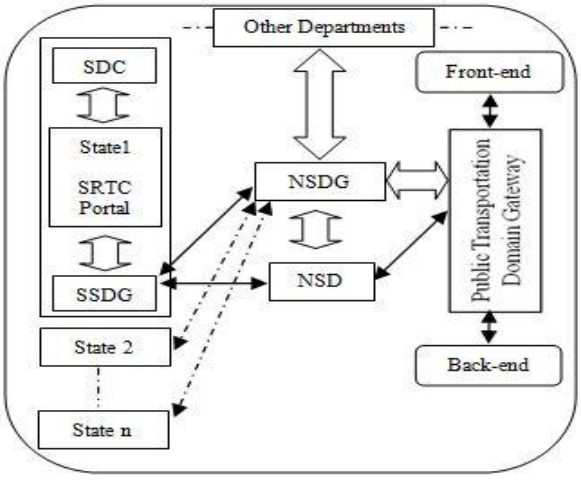


Fig. 3 Design for Integration of public transportation

Integration of public transportation is a step towards nationalization of public transportation using National e-Governance Service Delivery Framework as in figure 1. We propose to establish and setup a domain gateway to provide integrated services for specific needs of public transportation as in figure 3. Public Domain gateway will provides integration and interoperability between SRTC's of the states in India. The model represents centralized access of information with distributed environment.

Figure 4 and 5 represents the communication between commuter and domain gateway. Figure 4 represents the information flow within state or State Road Transport Corporation (SRTC) for communication where as figure 5 represent the inter-state information flow between SRTC's of the states. Domain gateway (D.G.) for public transportation will enables the public transportation to integrate all the SRTC's of states of India which are geographically connected. It acts as central web application server to integrate all SRTC's of the states in India. Domain gateway of public transportation has its own Front end application and backend to manage user or commuter's request. In figure 4, the commuter's request is analyses at domain gateway by web application server and forwards the request to the NSDG middleware after keeping necessary records in their database. NSDG is used to connect the intended State Road Transport Corporation (SRTC) portal through SSDG using middleware standard based interoperability to full fill commuter's request.

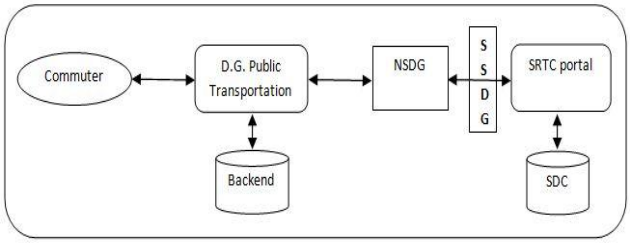


Fig 4. Information flow for within state Public Transportation

In case of inter-state communication, domain gateway web application will analyses the users or commuters request. Forward the request to the source state through NSDG after keeping necessary information in their database. The desired request has to be served at source state's portal through SSDG and update their SDC (State Data Centre). Further state portal retrieves the in between states from source state to destination station from user request except source state. The list of in between states would be passed from the commuter's request to the NSDG middleware server for interoperable services. NSDG will forward message along with user request to all the SSDG's of the

IJCSI
www.IJCSI.org

intended states in the incoming list from source state and portals have to update respective information's.

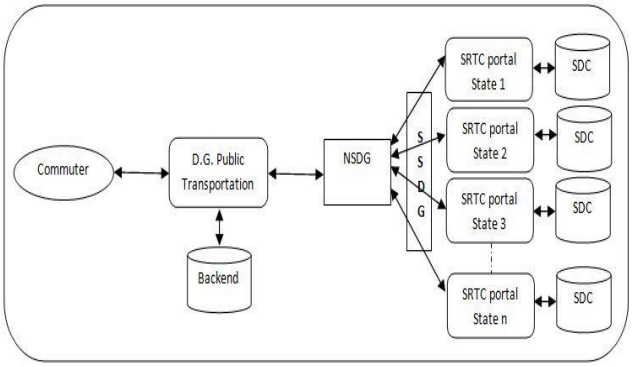


Fig. 5 Information flow for inter-state Public Transportation

6. Core Advantages

The Integration of the public transportation as suggested in this paper may enhance the functionality of the sector and would provide the following benefits:-

1. The implementation of the model through the NeGP's architecture for e-governance may provide flawless inter-connectivity between states which may enhance the business prospects of SRTU's.
2. Commuters are benefited by the one-stop, integrated services of public transportation to access, interact and performed online transaction with any source to the desired destination of India.
3. SWAN acts as the backbone for the National Service Delivery Gateway (NSDG) to support a national network infrastructure for e-Governance service delivery. Use of integrated public transportation saves cost incurred by third parties for application and network management.
4. SRTU's effectiveness can be improved by providing integrated services to the commuters of public transportation because of the competitive environment.
5. Effectiveness and efficiency can be achieved by one-stop integrated could improve government response time to citizens and reduce paperwork burden of public transportation.

4. Conclusions

The implementation of e-Governance in any sector can enhance the quality of service delivery. It could be beneficial for public transportation. NeGP's National e-Governance service delivery framework will acts as the backbone for integrating the SRTC of the state in India. The integration of public transportation permits the commuters and officials to access services of any states from one stop. This will boost up the business prospects for citizens by better connectivity between SRTU's of all the states inoder to fulfill the dream of nationalized public transportation.

References

[1] Business Intelligence And E-Governance, Analytics & Modeling Division National Informatic Centre Department of Information Technology, Ministry of Communication & IT, New Delhi, India
[2] Vahan and Sarathi E-Governance at Regional Transport Offices in Taminnadu www.tn.nic.in/tnhome/projectfiles/brochure-transport.pdf
[3] Report of the sub-group on SRTU under group on Road Transport Constituted by planning commission: 12th Five year plan 2012-2017, http://morth.nic.in/writereaddata/.../Report%20SRTUs-3081731927.pdf
[4] India Transport Sector - Roads http://web.worldbank.org/WBSITE/EXTERNAL/COUNTRIES/SOUTHASIAEXT/EXTSARREGTOPTRANSPORT/0,,contentMDK:20703625~menuPK:868822~pagePK:34004173~piPK:34003707~theSitePK:579598,00.html
[5] Ajay Kumar Bharti, Sanjay K. Dwivedi, Design Of An Analytical And Foresight Based Strategic Model For E-Governance In Public Transportation, Springer link, Communications in Computer and Information Science, Volume 250, Part 2, 2011 pp 615-620
[6] Ajay Kumar Bharti, Sanjay K. Dwivedi, E-Governance in Public Transportation: U.P.S.R.T.C. - A Case Study, Proceedings of ICSCA-2011, IPCSIT vol.9, PP 7-12 2011, ISSN 2010-460X
[7] National e-Governance Service Delivery Framework http://www.nsdg.gov.in/administration/images/Slide4.PNG

Ajay Kumar Bharti has obtained his MCA degree in 2000 from Kamla Nehru Institute of Technology, Sultanpur (UP). His research interest is Information technology and e-Governance. Worked in M.I.E.T. Meerut as Sr. Lecturer and Assistant Professor. He has published some of the research papers in refereed Journals and international conferences. Recently he is doing research from Babasaheb Bhimrao Ambedkar University (A CENTRAL UNIVERSITY) Lucknow - 226025 , UP, India

Dr. S.K. Dwivedi has obtained his Ph.D. Degree from Banasthali Vidyapeeth in the year 2006. He has completed his Ph.D. in the area of Web Mining. His research interests are Web content Mining, Semantic Web, Search Engine performance evaluation, e-Governance etc. He has published many of the valuable research papers in various national and international Journals. He is presently working as a Associate Professor of Computer Science dept, of BBAU, Lucknow, India.

IJCSI
www.IJCSI.org

Numerical Simulation of Two Phase Flow in Reconstructed Pore Network Based on Lattice Boltzmann Method

Song Rui[1], Liu Jianjun[1,2], Qin Dahui[1]

1, School of Civil Engineering and Architecture, Southwest Petroleum University, Chengdu, China;

2, State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation (Southwest Petroleum University), Chengdu, China;

Abstract

Accurate prediction and understanding of the disorder microstructures in the porous media contribute to acquiring the macroscopic physical properties such as conductivity, permeability, formation factor, elastic moduli etc. Based on the rock serial sectioning images of Berea sandstone acquired by the core scanning system developed by our research group, the reconstructed rock model is established in the Mimics software and the extracted pore network of the porous rock is accomplished by the self-programming software in C++ programming language based on the revised Medial axis based algorithm and the Maximal ball algorithm. Using a lattice Boltzmann method, the single and two – phase flow are accomplished. Both of the pore-scale networks and the seepage mechanism of the single- and two –phase flow are identical with the benchmark experimental data.

Keywords: *Berea sandstone; serial sectioning; reconstructed porous media; extracted network; single - phase flow; two – phase flow.*

1. Introduction

Accurate prediction and understanding of the disorder microstructures in the porous media, such as rocks [1], soils [2], biomedical field [3], ceramics [4], and composites [5], contribute to acquiring the macroscopic physical properties such as conductivity, permeability, formation factor, elastic moduli etc. [6 - 8]. Though those transport properties can be obtained by experiments, it is hard to gain the detailed information of the fluid flow in the pore and to conduct the three – phase flow experiment in the current conditions.

The pore-scale network model describing the disorder system in the porous media is considered as a starting point emphasized by many scholars [9-12]. Fortunately, with the developing of electronic computer and the technology of porous media imaging in recent years, such as Scanning Electron Microscopy (SEM)[13,14], serial sectioning[15-17], confocal laser scanning microscopy[18], micro X-ray computerized tomography (micro-CT)[19]and reconstructed porous media by mathematical methods[1,20], it is applicable to acquire the pore space images mapping the real interior structure of its original sample, on basis of which the pore-scale numerical simulation can be carried on to make the above study complete. However, it is necessary to develop an effective algorithm to extract the pore network from these three dimensional porous media. In this paper, based on the rock serial sectioning images of Berea sandstone acquired by the core scanning system developed by our research group, the reconstructed rock model is established in the Mimics software and the extracted pore network of the porous rock is accomplished by the self-programming software in C++ programming language based on the revised Medial axis based algorithms [21-22] and the Maximal ball algorithm [23-24]. Using a lattice Boltzmann method, the single and two – phase flow are accomplished. Both of the pore-scale networks and the seepage mechanism of the single- and two –phase flow are identical with the benchmark experimental data.

2. Reconstructed Berea Sandstone and the Extracted Pore Network Model

2.1 Serial Sectioning Imaging of the Berea Sandstone and Reconstructed Model

Serial sectioning provides a direct way to visualize 3D microstructures when successive layers of materials are removed and exposed surfaces are imaged at high resolution. The 3D image of pore media can be obtained by stacking serial sections [25]. The workflow of the serial sectioning is illustrated in Fig.1.

In this paper, the images of the Berea sandstone section are obtained by the self – developed core scanning system. The rock matrix and the pore space are identified using the optical properties of different minerals in the scanning system, the cell size is 3.45 μm ×3.45 μm for the maximum resolution of 2454 Pixel × 2056 Pixel. For the sake of the storage space increasing sharply along with the resolution and the feasibility of the numerical simulation, the image resolution for this paper is 9 μm, in which case the storage space of the reconstructed

sandstone is 1.2GB and the extracted pore network is 526 MB. Fig.2 shows two of the sandstone slice images using the scanning system.

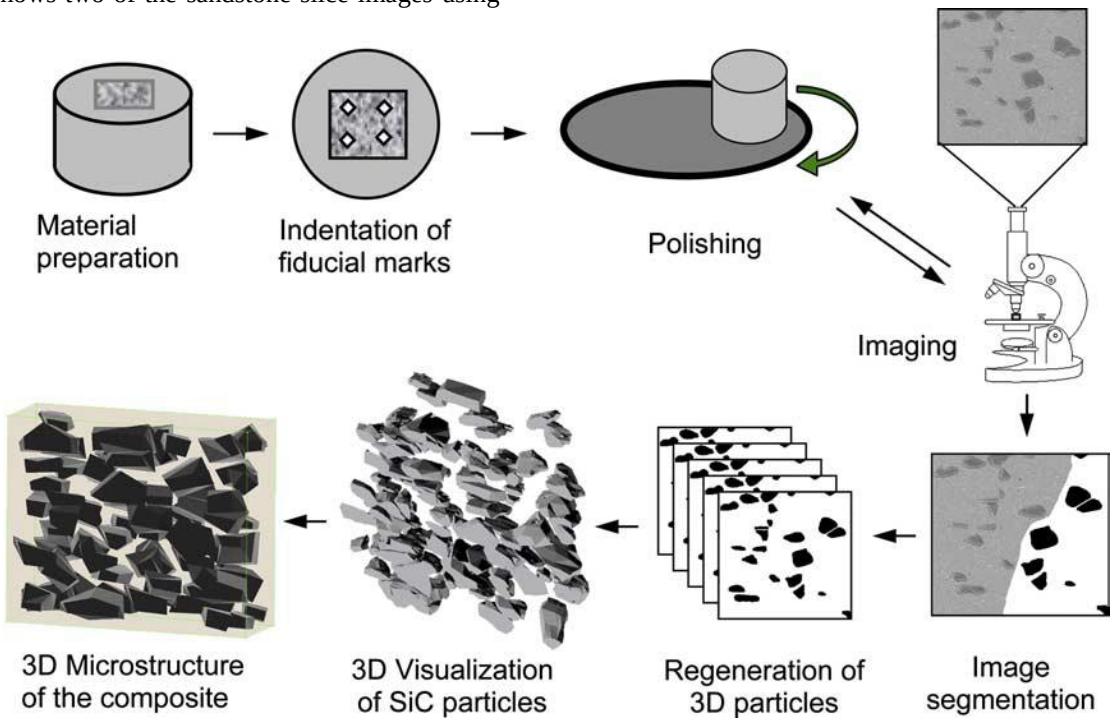
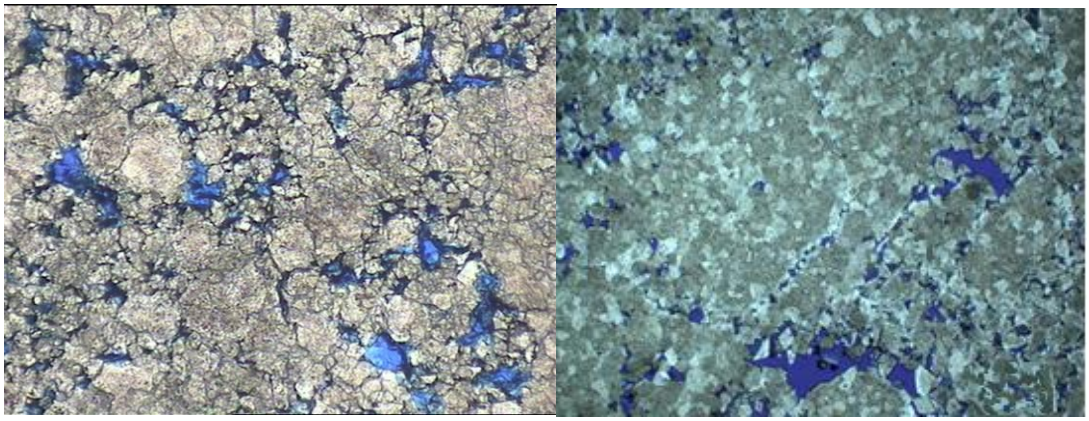


Figure 1. Flow chart of serial sectioning and 3D reconstruction process [25].



(a) (b)

Figure 2. The sandstone slice images

Berea sandstone is selected as the original sample, and a total of 97 sections are polished in this study. And the central part of the images for 300 Pixels × 300 Pixels is selected as the basic data to reconstruct the porous media. Due to being storage individually, as a result of which the image data is discrete, these images are imported into the ImageJ software by the National Institutes of Health in sequence. Then Binary images are obtained by utilize the information of shadow of stone and the gray image histogram of the origin SEM image, by which the rock matrix and the pore space can be resolved from the white part and the black, respectively. Followed that, the images are de-noised and smoothed and converted into the standard CT format(.raw) in ImageJ. Four of the processed images are shown in Fig.3. Finally, the images data with a *.raw* suffix is imported into Mimics software to accomplish the reconstructed Berea sandstone model, the size of which is 2700 μ m ×2700 μ m × 2700 μ m in Fig.4.

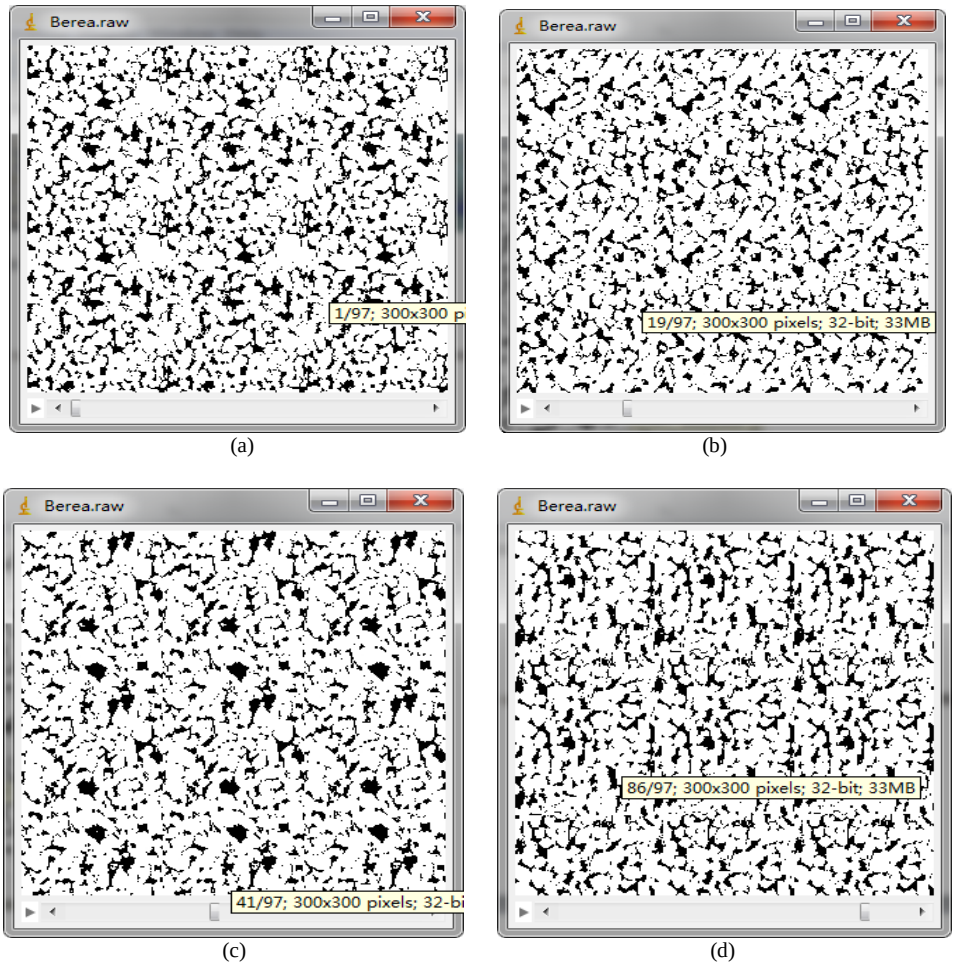


Figure 3. The processed sandstone images (300 Pixels × 300 Pixels). The white part is the rock matrix and the black part is the pore space. (a) is the top image of the 97; (b) is the 19th; (c) is the forty-first; (d) is the 86th.

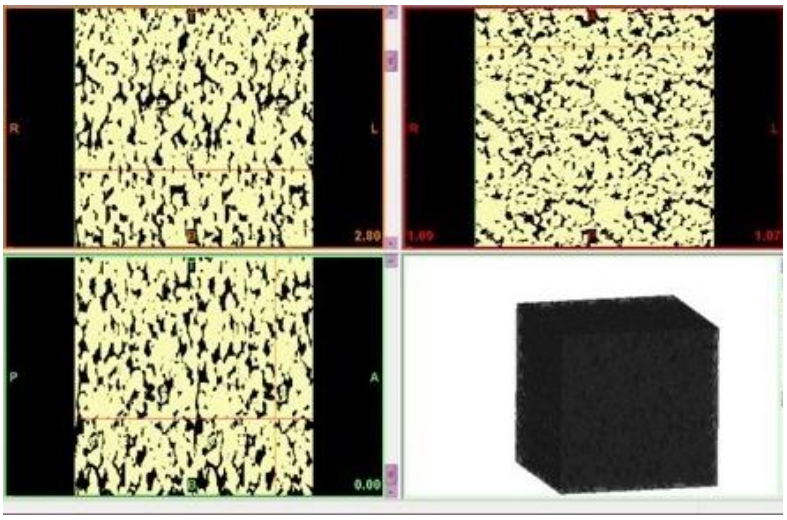


Figure 4. The three view of the reconstructed Berea sandstone and the three – dimensional model in Mimics software.

2.2 Extracted Pore Network of the Reconstructed Berea Sandstone Model

As established in the previous section, the three - dimensional porous media can be acquired in many ways. However, for further study on the transport properties in the porous media, it is essential to develop an effective algorithm to extract the pore network of the three – dimensional porous media images, which is considered as a breakthrough point by many scholars. The synthetical and revised pore network algorithm evaluated from the Medial axis based algorithm and the Maximal ball algorithm, the workflow of which is as follows:

1. The three – dimensional porous media image is segmented and exported as the txt image, the maximum RGB value of which is limited as 1, that is 0 represents the pore space and 1represents the rock matrix. Each number represents a voxel in the image. Part of the txt image is shown in the Fig.5.

2. Along the status "0" in the txt image the medial axis will be obtained and located accurately, that is, the center of the pore is known. Taking the central point as the center of sphere, the sphere is expanded until touching the rock matrix voxel. Then every sphere is numbered according with the radius of the sphere.

3. The largest ball is defined as pore, then the second. The connected balls form pore chains, and the ball where the two different chains meets are defined as throat and replace by the cylinder with the same radius.

4. The same sorting processes are applied for all the ball chains until reaching the minimum size set for a pore in our self-programming system. Below this threshold, the ball may provide connections (be throats) but cannot be pores.

Based on the theory above, we develop the pore network extraction system and provide visualization model in Rhinoceros software, and complete the pore network and one of the cross sections of the model shown in the Fig.6. The geometrical parameters are listed in the table 1, by which we can find the algorithm is feasible.

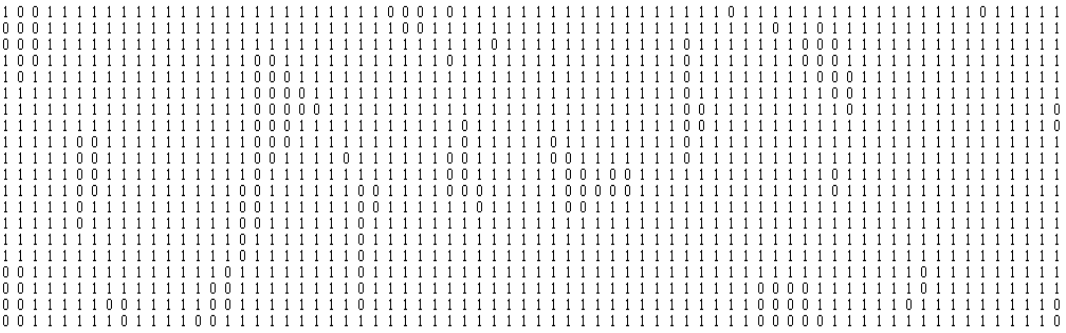
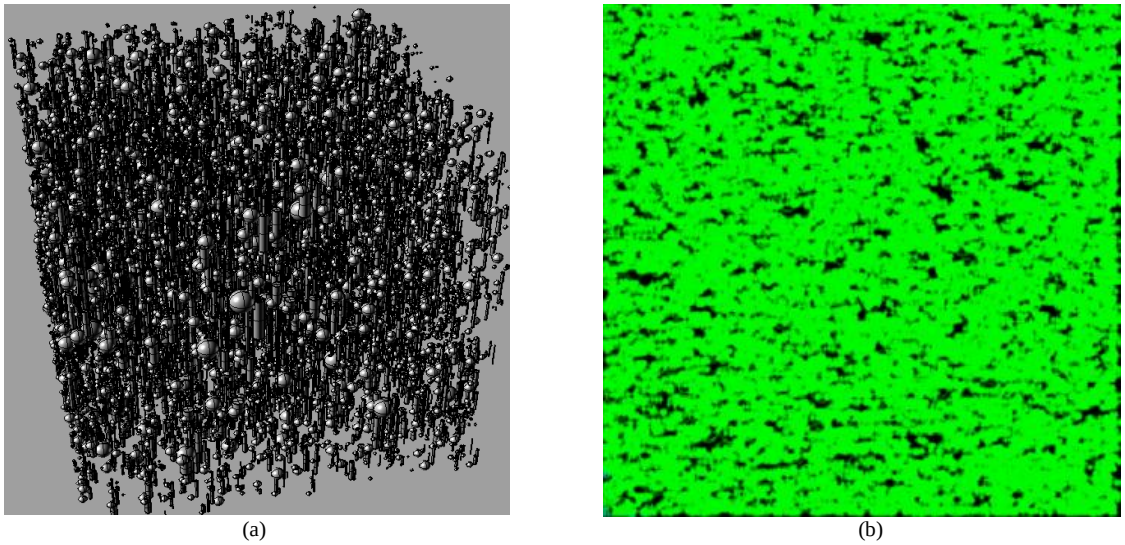


Figure 5. The txt image of the porous media



(a) (b)

Figure 6. The extracted pore network. (a) is model in Rhinoceros software; (b) is the one of the cross sections of the model.

Table 1. The geometrical parameters of the extracted pore network

Geometrical Parameters	Count or Size
Porosity	14.06%
Number of the Pores	1868
Number of the throat	3053
Average connection number	3.15685
Minimum connection number	0
Maximum connection number	12
Number of connections to inlet	124
Number of connections to outlet	138
Average pore radius	19.04 μ m
Average throat radius	7.29 μ m

3. Numerical study on the extracted pore network based on lattice Boltzmann method

To verify the feasibility of our pore network extraction algorithm, the single – and two – phase flow experiment and simulation study are conducted in this study. For the sake of the disconnection of different pore chains, it is impossible to be meshed in the commercial FEM software. Unlike conventional numerical schemes based on discretizations of macroscopic continuum equations, the lattice Boltzmann method is based on microscopic models and mesoscopic kinetic equations. The fundamental idea of the LBM is to construct simplified kinetic models that incorporate the essential physics of microscopic or mesoscopic processes so that the macroscopic averaged properties obey the desired macroscopic equations [26-29].

Single-phase flow is simulated across the extracted network and the absolute permeability is calculated on the network by the by the self-programming software in C++ programming language using Lattice Boltzmann and compared to the experimental absolute permeability results on the same rock sample.

3.1 Mathematical model for flow in the pore network using lattice Boltzmann method

The absolute permeability K of the network is derived from Darcy's law [10]:

$$K = \frac{\mu_p q_{tsp} L}{A(\Phi_{inlet} - \Phi_{outlet})} \quad (1)$$

where the network is fully saturated with a single phase p of viscosity μ_p ; q_{tsp} is the total single phase flow rate through the pore network of length L with the potential drop $(\Phi_{inlet} - \Phi_{outlet})$. A is the cross-sectional area of the model.

Then relative permeability is

$$k_{rp} = \frac{q_{tmp}}{q_{tsp}} \quad (2)$$

where q_{tmp} is the total flow rate of phase p in multiphase conditions with the same imposed pressure drop.

The conductance of the single phase g_p is given by the Hagen-Poiseuille formula:

$$g_p = k\frac{A^2 G}{\mu_p} \quad (3)$$

For a circular, an equilateral and a square tube, the constant k is 0.5, 0.6 and 0.5623 respectively.

The conductance between two pore bodies (i, j) via throat (t) is given as:

$$\frac{l_{ij}}{g_{p,ij}} = \frac{l_i}{g_{p,i}} + \frac{l_t}{g_{p,t}} + \frac{l_j}{g_{p,j}} \quad (4)$$

where l_{ij} is the distance from pore i center to pore j center (throat total length); l_i and l_j are the pore body lengths which are the lengths from the pore-throat interface to the pore centers, as illustrated in Fig. 7

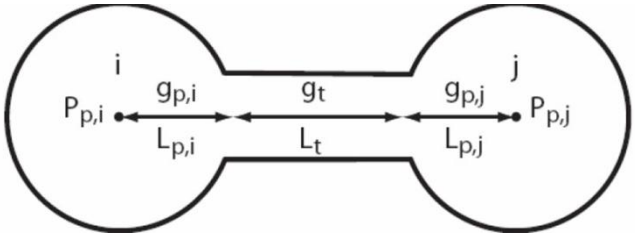


Figure 7. Conductance between two pores [30].

The lattice Boltzmann equation is:

$$f_i(\vec{x}+\vec{e}_i, t+1) - f_i(\vec{x}, t) = \frac{1}{t}(f_i^0 - f_i) \quad (5)$$

where $f_i(\vec{x}, t)$ is the particle distribution function at location x and time t along the ith direction *(i=0,1,2…18)*

using three-dimensional nineteen velocity model and where τ is the single time relaxation parameter and f_i^o is the local equilibrium state depending on the local density and velocity [26].

3.2 Comparison between the experimental results and the simulation

The comparison between the experimental results and the simulation can be seen in the table 2, by which we can find the porosity and the absolute permeability of the extracted pore network is quite close to the benchmark experimental data for the same Berea sandstone sample. In the same way, the relative permeability curve for the two – phase flow in the extracted pore network and the experiment is shown in the Fig.8. These indicate that, even though the micro structure, which may be in different geological shapes, is substituted by the balls and cylinder, the extracted network approaches to the microstructure in the origin sample for the size of 2.7 mm × 2.7 mm ×2.7 mm, which verifies that the extraction algorithm and the simulation software developed by our group is feasible and applicable in the porous media study.

Table 2. Single – phase flow result of the experiment and the simulation

Method	Porosity	Absolute permeability
Experiment	14.53%	1633.52md
Simulation	14.06%	1752.18md

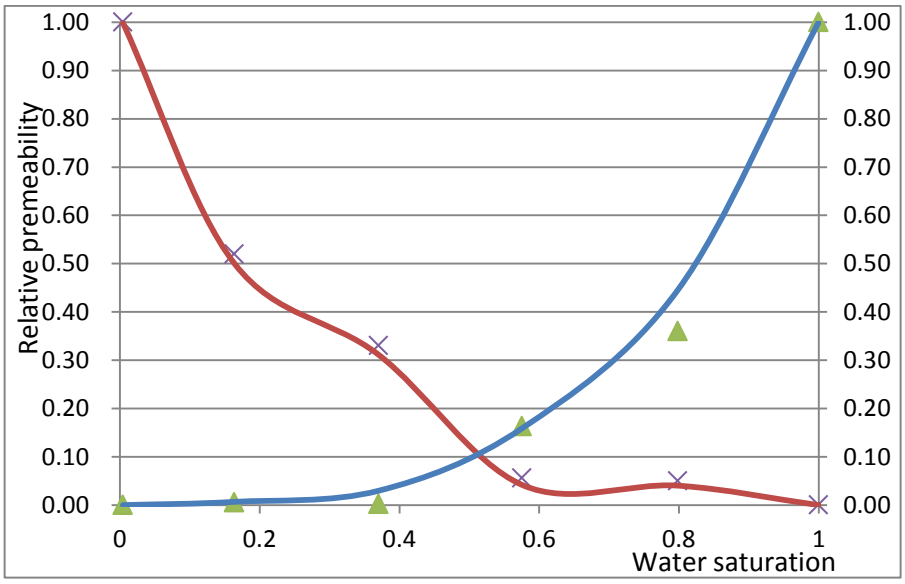


Figure 8. Relative permeability vs. water saturation; the splashes are experimental results while the curves are the simulation.

4. Conclusion

In this paper, an effective method of reconstructing the three - dimensional model from the serial sectioning image of Berea sandstone and an effective pore network extraction algorithm is presented. Using the lattice Boltzmann method, the single and two – phase flow are accomplished. Both of the pore-scale networks and the seepage mechanism of the single- and two –phase flow are identical with the benchmark experimental data. The study aims to provide an ideal pore network model describing the same microstructure of the origin sample, and apply the network to numerical simulation. In future, we will develop the visualized distribution cloud chart of the different fluid for the multi – phase flow to make the study more reliable and applicable.

Acknowledgement

This paper is financially supported by Natural Science Foundation of China (Grant No.51174170) and National Science and Technology Major Project of China under Grant No. 2011ZX05013-006.

Reference

[1] D. Bauer, S. Youssef, M. Fleury, S. Bekri, E. Rosenberg, O. Vizika, Improving the Estimations of

Petrophysical Transport, Transp Porous Med 94, 2012,pp.505–524.
[2] M. Kataja, K. Hiltunen, and J. Timonen, J. Phys. D 25, 1992,1053.
[3] Zollikofer Christoph P.E. et al. Tools for rapid prototyping in the biosciences. IEEE Computer Graphics and Applications, 12, 1995, pp.48-57.
[4] H. Kamiya, K. Isomura, G. Jimbo, and T. Jun-ichiro, J. Am.Chem. Soc. 78, 49, 1995.
[5] Michele Panico. Modeling of Shape Memory Alloys and Application to Porous Materials. PHD thesis, NORTHWESTERN UNIVERSITY, EVANSTON, ILLINOIS , 2008.
[6] R. Hilfer and Th. Zauner. High-precision synthetic computed tomography of reconstructed porous media, PHYSICAL REVIEW E 84, 062301, 2011.
[7] Blunt, M.J.. Flow in porous media pore-network models and multiphase flow.Current Opinion in Colloid & Interface Science, 6, 2001,pp. 197-207
[8] D.Bauer, S.Youssef, M. Fleury, S. Bekri, E.Rosenberg, O. Vizika. Improving the Estimations of Petrophysical Transport Behavior of Carbonate Rocks Using a Dual Pore Network Approach Combined with Computed Microtomography. Transp Porous Med, 94, 2012, pp.505–524.
[9] R. Elliot, J. Krumhansl, and P. Leath, The theory and properties of randomly physical systems, Rev. Mod. Phys., vol. 46, no. 3, pp. 465-543, 1974
[10] Hu Dong. Micro-CT imaging and pore network extraction. PhD thesis, Imperial College London, 2007.
[11] B. Biswal, P.-E. Øren, R. J. Held, S. Bakke, and R. Hilfer. Stochastic multiscale model for carbonate rocks. PHYSICAL REVIEW E 75, 061303, 2007.
[12] Wenddabo Olivier Sawadogo, Noureddine Alaa, Blaise Somé Numerical simulation of groundwater level in a fractured porous medium and sensitivity analysis of thehydrodynamic parameters using grid computing: application of the plain of Gondo (Burkina Faso). International Journal of Computer Science Issues, v 9, n 1 1-2, 2012, p 227-236.
[13] McMullan, D. Von Ardenne and the scanning electron microscope. Proc Roy Microsc Soc 23, 1988, pp.283-288.
[14] 14. McMullan, D. Scanning electron microscopy 1928–1965. Scanning 17, 1995, pp.175–185.
[15] Lymberopoulos, D.P. and Payatakes, A.C. Derivation of topological, geometrical, and correlational properties of porous media from pore-chart analysis of serial section data. Journal of Colloid and Interface Science, 150(1):61-80, doi: 10.1016/0021-9797(92)90268Q, 1992.
[16] Vogel, H.-J. Roth, K. Quantitative morphology and network representation of soil pore structure. Advances in Water Resources 24(3-4), 2001, pp. 233-242.
[17] Tomutsa, L. and Silin, D.. Nanoscale Pore Imaging and Pore Scale Fluid Flow Modeling in Chalk. Lawrence Berkeley National Laboratory: Paper LBNL56266 (2004). http://repositories.cdlib.org/lbnl/LBNL-56266.
[18] Fredrich, J.T., Menendez, B. and Wong, T.-F. Imaging the pore structure of geomaterials. Science, 268, 1995, pp. 276-279.
[19] Dunsmuir, J.H., Ferguson, S.R., D'Amico, K.L. and Stokes, J.P. X-ray microtomography: a new tool for the characterization of porous media, Paper SPE 22860, Proceedings of 66th Annual Technical Conference and Exhibition of the Society of Petroleum Engineers, Dallas, TX (1991).
[20] D.Bauer, S.Youssef, M. Fleury, S. Bekri, E.Rosenberg, O. Vizika. Improving the Estimations of Petrophysical Transport Behavior of Carbonate Rocks Using a Dual Pore Network Approach Combined with Computed Microtomography. Transp Porous Med, 94, 2012, pp.505–524.
[21] Lindquist, W.B., Lee, S.M., Coker, D., Jones, K. and Spanne, P. Medial axis analysis of void structure in three-dimensional tomographic images of porous media. Journal of Geophysical Research, 101B: 8297, 1996.
[22] Lindquist, W.B. and Venkatarangan, A. Investigating 3D Geometry of Porous Media from High Resolution Images. Phys. Chem. Earth (A), 25(7), 1999, pp. 593-599.
[23] Silin, D. and Patzek, T. Pore space morphology analysis using maximal inscribed spheres. Physica A, 371, 2006, pp. 336-360.
[24] Silin, D.B., Jin, G. and Patzek, T.W. Robust Determination of Pore SpaceMorphology in Sedimentary Rocks, Paper SPE 84296, Proceedings of SPE Annual Technical Conference and Exhibition, Denver, Colorado, U.S.A, 2003.
[25] Chawla, N., Sidhu, R.S. and Ganesh, V.V. Three-dimensional visualization and microstructure-based modeling of deformation in particle-reinforced composites. Acta Materialia, 54(6), 2006, pp.1541-1548.
[26] Shiyi Chen, Gary D. Doolen. LATTICE BOLTZMANN METHOD FOR FLUID FLOWS. Annu. Rev. Fluid Mech. 30, 1998, pp.329–64.
[27] Dyntar, Jakub, Soucek, Ivan; Gros, Ivan. Application of discrete event simulation in LPG storage operation and optimization. International Journal of Computer Science Issues, v 9, n 3, 2012, pp. 33-42.
[28] Rastaghi, Roohallah,); Oskouei, Hamid R. Dalili. Cryptanalysis of a public-key cryptosystem using lattice basis reduction algorithm. International

Journal of Computer Science Issues, v 9, n 5 5-1, 2012, pp. 110-117.

[29] Rakhshani, Mohammad Reza, Mansouri-Birjandi, Mohammad Ali. Design and optimization of photonic crystal triplexer for optical networks. International Journal of Computer Science Issues, v 9, n 4 4-1, 2012, pp. 24-28.

[30] Valvatne, P.H. Predictive pore-scale modelling of multiphase flow. PhD thesis, Department of Earth Science and Engineering, Imperial College London, 2004.

New Delay-dependent Stability Criteria for Linear Systems with Time-varying Delay

Weiwei Zhang[1], Chao Ge[2], Hong Wang[3]

[1]College of Information Engineering, Hebei United University, Tangshan, Hebei 063009, PR China

[2]College of Information Engineering, Hebei United University, Tangshan, Hebei 063009, PR China

[3]College of Qing Gong, Hebei United University, Tangshan, Hebei 063009, PR China

Abstract

This paper is concerned with the problem of asymptotic stability for linear systems with time-varying delays. With the introduction of delay-partition approach, some new delay-dependent stability criteria are established and formulated in the form of linear matrix inequalities. Both constant time delays and time-varying delays have been taken into account. Numerical examples are given to demonstrate the effectiveness and less conservativeness of the proposed methods.

Keywords: *Linear systems, Time-varying delay, Delay-partition, Asymptotic stability, Linear matrix inequalities(LMIs).*

1. Introduction

Time delay is commonly encountered in various physical and engineering systems such as aircraft, biological systems, networked control systems, and so on. Since the existence of time-delays causes poor performance, oscillation, or even instability, it is very important to investigate stability analysis for systems with time-delays before designing control systems, see for example [1] and references therein.

On the other hand, neutral time-delay systems contain delays both in its state, and in its derivatives of the state. Such a system can be found in population ecology [2], distributed networks containing lossless transmission lines [3], heat exchangers, robots in contact with rigid environments [4], etc. Stability of these systems was proved to be a more complex issue because the system involves the derivative of the delayed state. Because of its wider application, the problem of the stability for neutral time-delay systems has attracted considerable attention during the last two decades.

By using the Lyapunov--Razumikhin functional approach or the Lyapunov--Krasovskii functional approach, several stability criteria have been proposed for delay-independent [5,6] and delay-dependent stability criteria [7,8] cases. Since delay independent conditions are usually more conservative than the delay-dependent conditions, more attention has been paid to the study of delay-dependent conditions. For example, a delay-dependent stability criterion for uncertain neutral systems with time-varying discrete delay was obtained in [9] based on a model transformation and Park's inequality [10].By taking an augmented model which included the original system and the model obtained by taking the time-derivative of original system, Ariba et al. [11] proposed a new delay-dependent stability criteria for time-varying delay systems. In [12], the triple integral Lyapunov-Krasovskii functional approaches which utilize more information about states and delayed states have been proposed. Suplin et al.[13] proposed delay-dependent stability conditions for time-delay systems based on the augmented Lyapunov-Krasovskii's functional and Finsler's lemma. Therefore, it is strongly needed that some new methods should be studied to improve the upper bounds of stability criteria.

Motivated by the above, in this paper, a new delay-decomposition method for neutral systems with time-varying delays will be proposed. By constructing a suitable Lyapunov-Krasovskii's functional, some novel delay-dependent stability criteria are derived in terms of LMIs which can be solved efficiently. In order to derive less conservatively results, by using the delay decomposition approach, the delay interval $[-\tau,0]$ is decomposed into $[-\tau,-\alpha\tau]$ and $[-\alpha\tau,0]$. Since a tuning parameter is introduced, the information about $x(t-\alpha\tau)$ can be taken into full consideration. Then we chosen different weighting matrices in each subinterval, which yields less

conservative delay-dependent stability criteria. Finally, numerical examples are included to show the effectiveness of the proposed method.

Notation. Throughout this paper, R^n is the n-dimensional Euclidean space, $R^{m\times n}$ denotes the set of $m\times n$ real matrix. X_{ij} denotes the element in row i and column j of matrix X. I is the identity matrix. The notation $*$ always denotes the symmetric block in one symmetric matrix. Matrices, if not explicitly stated, are assumed to have compatible dimensions.

2. Problem statement and preliminary

Consider the following neutral system with time-varying delay:

$$\begin{aligned}\dot{x}(t) &= Ax(t) + A_1x(t-\tau(t))\\ x(s) &= \phi(s), s\in[-\tau,0]\end{aligned} \tag{1}$$

where $x(t)\in R^n$ is the vector, A, A_1 are known constant matrices with appropriate dimensions, $\phi(s)\in C_{n,\tau}$ is a given continuous vector-valued initial function, and $\tau(t)$ is a time-varying continuous function that satisfies the conditions

$$0\le\tau(t)\le\tau \quad \dot{\tau}(t)\le\mu<1 \tag{2}$$

The purpose of this paper is to establish delay-dependent stability conditions for neutral system (1). To obtain the main results, the following lemmas are needed.

Lemma 2.1[14]: For any constant matrix $X\in R^{n\times n}$, $X=X^T>0$, a scalar function $h:=h(t)>0$, and a vector valued function $\dot{x}:[-h,0]\to R^n$ such that the following integrations are well defined, then

$$-h\int_{-h}^{0}\dot{x}^T(t+s)X\dot{x}(t+s)ds\le\xi_1^T(t)\begin{bmatrix}-X & X\\ X & -X\end{bmatrix}\xi_1(t) \tag{3}$$

$$-\frac{h^2}{2}\int_{-h}^{0}\int_{t+\theta}^{t}\dot{x}^T(s)X\dot{x}(s)dsd\theta\le\xi_2^T(t)\begin{bmatrix}-X & X\\ X & -X\end{bmatrix}\xi_2(t) \tag{4}$$

where $\xi_1^T(t)=[x^T(t) \quad x^T(t-h)]$

and $\xi_2^T(t)=[h\,x^T(t) \quad \int_{t-h}^{t}x^T(s)\,ds]$

Lemma 2.2[19]: Let $f_1, f_2,..., f_N: R^m\mapsto R$ have positive values in an open subset D of R^m. Then, the reciprocally convex combination of f_i over D satisfies

$$\min_{\{\alpha_i|\alpha_i>0,\sum_i\alpha_i=1\}}\sum_i\frac{1}{\alpha_i}f_i(t)=\sum_i f_i(t)+\max_{g_{i,j}(t)}\sum_{i\ne j}g_{i,j}(t)$$

subject to

$$\left\{g_{i,j}:R^m\longmapsto R, g_{j,i}(t)\triangleq g_{i,j}(t),\begin{bmatrix}f_i(t) & g_{i,j}(t)\\ g_{i,j}(t) & f_j(t)\end{bmatrix}\ge 0\right\} \tag{5}$$

3. Main results

In this section, we propose a new delay-dependent stability criteria for neutral system (1). Both constant time delays and time-varying delays are treated. In order to obtain some less conservative sufficient conditions, we decompose the delay interval $[-\tau,0]$ into $[-\tau,-\alpha\tau]$ and $[-\alpha\tau,0]$, and we consider the both condition $\tau(t)\in[-\tau,-\alpha\tau]$ and $\tau(t)\in[-\alpha\tau,0]$. For convenience, we define $e_i(i=1,2,...,9)$ as block entry matrices. For example, $e_3^T=[0 \quad 0 \quad I \quad 0 \quad ... \quad 0]$. The other notations for some vectors and matrices are defined as:

$\zeta_1^T(t)=[x^T(t)\,x^T(t-\tau(t))\,x^T(t-\alpha\tau)\,x^T(t-\tau)\,\dot{x}^T(t-\alpha\tau)$
$\dot{x}^T(t-\tau)\int_{t-\tau(t)}^{t}x^T(s)ds\int_{t-\alpha\tau}^{t-\tau(t)}x^T(s)ds\int_{t-\tau}^{t-\alpha\tau}x^T(s)ds]$,
$\zeta_2^T(t)=[x^T(t)\,x^T(t-\tau(t))\,x^T(t-\alpha\tau)\,x^T(t-\tau)\,\dot{x}^T(t-\alpha\tau)$
$\dot{x}^T(t-\tau)\int_{t-\alpha\tau}^{t}x^T(s)ds\int_{t-\tau(t)}^{t-\alpha\tau}x^T(s)ds\int_{t-\tau}^{t-\tau(t)}x^T(s)ds]$,
$\zeta_0^T(t)=[x^T(t)\;x^T(t-\alpha\tau)\,x^T(t-\tau)\,\dot{x}^T(t-\alpha\tau)\,\dot{x}^T(t-\tau)$
$\int_{t-\alpha\tau}^{t}x^T(s)ds\int_{t-\tau}^{t-\alpha\tau}x^T(s)ds]$,
$\varepsilon^T(t)=[x^T(t)\,x^T(t-\alpha\tau)\,x^T(t-\tau)\int_{t-\alpha\tau}^{t}x^T(s)ds$
$\int_{t-\tau}^{t-\alpha\tau}x^T(s)ds]$
$\eta^T(t)=[x^T(t)\,\dot{x}^T(t)]$
$\Pi_0^1=[e_1\,e_2\,e_3\,e_6\,e_7]$, $\Pi_0^2=[A_{c0}\,e_4\,e_5\,e_1-e_3\,e_3-e_4]$,
$\Pi_1^2=[e_1\,e_3\,e_4\,e_7\,e_8+e_9]$, $\Pi_1^1=[e_1\,e_3\,e_4\,e_7+e_8\,e_9]$
$\Pi_3^2=[e_8\;e_3-e_2\;e_9\;e_2-e_4]$,
$\Pi_2=[A_c\,e_6\,e_7\,e_1-e_3\,e_3-e_4]$, $\Pi_3^1=[e_7\;e_1-e_2\;e_8\;e_2-e_3]$.
$A_c=[A\,A_1\,0\,0\,0\,0\,0\,0]$, $A_{c0}=[A\,A_1\,0\,0\,0\,0\,0]$

Now, we have the following theorem.

Theorem 3.1: For given scalars τ, μ and $0<\alpha<1$, the system (1) with (2) is asymptotically stable if there exist positive definite matrices $P=[P_{ij}]_{5\times5}$, $\Omega_1=[Q_{1,ij}]_{2\times2}$, $\Omega_2=[Q_{2,ij}]_{2\times2}$, $\Psi_1=[W_{1,ij}]_{2\times2}$, $\Psi_2=[W_{2,ij}]_{2\times2}$, T_j, Q_0, $Q_i(i=3,4,5,6)$, any matrices $\Theta_1=[S_{1,ij}]_{2\times2}$, $\Theta_2=[S_{2,ij}]_{2\times2}$, $N_i(i=1,2,3,4,5,6)$, with appropriate dimensions such that the following LMIs hold:

$$\Phi_1^{(k)}<0,\begin{bmatrix}\Psi_k & \Theta_k\\ * & \Psi_k\end{bmatrix}\ge0, k=1,2$$

$$\begin{bmatrix} Q_3 & N_1 \\ * & Q_4 \end{bmatrix}>0, \begin{bmatrix} Q_3 & N_2 \\ * & Q_4 \end{bmatrix}>0, \begin{bmatrix} Q_5 & N_3 \\ * & Q_6 \end{bmatrix}>0,$$
$$\begin{bmatrix} Q_3 & N_4 \\ * & Q_4 \end{bmatrix}>0, \begin{bmatrix} Q_5 & N_5 \\ * & Q_6 \end{bmatrix}>0, \begin{bmatrix} Q_5 & N_6 \\ * & Q_6 \end{bmatrix}>0,$$

where

$$\begin{aligned}\Phi_1^{(1)}=&\Pi_1^1 P\Pi_2^T+\Pi_2 P(\Pi_1^1)^T+(e_1Q_{1,11}e_1^T+2e_1Q_{1,12}A_c^T+A_cQ_{1,22}A_c^T)\\&+(e_3Q_{2,11}e_3^T+2e_3Q_{2,12}e_5^T+e_5Q_{2,22}e_5^T)\\&-(e_4Q_{2,11}e_4^T+2e_4Q_{2,12}e_6^T+e_6Q_{2,22}e_6^T)\\&+(\alpha\tau)^2(e_1W_{1,11}e_1^T+2e_1W_{1,12}A_c^T+A_cW_{1,22}A_c^T)\\&+(1-\alpha)^2\tau^2(e_1W_{2,11}e_1^T+2e_1W_{2,12}A_c^T+A_cW_{2,22}A_c^T)\\&-e_9W_{2,11}e_9^T-2e_9W_{2,12}e_3^T+2e_9W_{2,12}e_4^T-e_3W_{2,22}e_3^T\\&+2e_3W_{2,22}e_4^T-e_4W_{2,22}e_4^T-\Pi_3^1\begin{bmatrix}\Psi_1 & \Theta_1\\ * & \Psi_1\end{bmatrix}(\Pi_3^1)^T\\&+A_c(\gamma_1^2T_1+\gamma_2^2T_2)A_c^T-(\alpha\tau)^2e_1T_1e_1^T+2\alpha\tau e_1T_1e_7^T\\&-e_7T_1e_7^T+e_1Q_0e_1^T-(1-u)e_2Q_0e_2^T-e_7T_1e_7^T\\&-2e_7T_1e_8^T-e_8T_1e_8^T-(1-\alpha)^2\tau^2e_1T_2e_1^T-e_9T_2e_9^T\\&+2(1-\alpha)\tau e_1T_2e_9^T+\alpha\tau(e_1Q_3e_1^T+A_cQ_4A_c^T)\\&-(e_3Q_{1,11}e_3^T+2e_3Q_{1,12}e_5^T+e_5Q_{1,22}e_5^T)\\&+(1-\alpha)\tau(e_1Q_5e_1^T+A_cQ_6A_c^T)+e_1N_1e_1^T\\&-e_2N_1e_2^T+e_2N_2e_2^T-e_3N_2e_3^T+e_3N_3e_3^T-e_4N_3e_4^T\end{aligned}$$

$$\begin{aligned}\Phi_1^{(2)}=&\Pi_1^2 P\Pi_2^T+\Pi_2 P(\Pi_1^2)^T+(e_1Q_{1,11}e_1^T+2e_1Q_{1,12}A_c^T+A_cQ_{1,22}A_c^T)\\&+(e_3Q_{2,11}e_3^T+2e_3Q_{2,12}e_5^T+e_5Q_{2,22}e_5^T)\\&-(e_4Q_{2,11}e_4^T+2e_4Q_{2,12}e_6^T+e_6Q_{2,22}e_6^T)\\&+(\alpha\tau)^2(e_1W_{1,11}e_1^T+2e_1W_{1,12}A_c^T+A_cW_{1,22}A_c^T)\\&+(1-\alpha)^2\tau^2(e_1W_{2,11}e_1^T+2e_1W_{2,12}A_c^T+A_cW_{2,22}A_c^T)\\&-e_7W_{1,11}e_7^T-2e_7W_{1,12}e_1^T+2e_7W_{1,12}e_3^T-e_1W_{1,22}e_1^T\\&+2e_1W_{1,22}e_3^T-e_3W_{1,22}e_3^T-\Pi_3^2\begin{bmatrix}\Psi_2 & \Theta_2\\ * & \Psi_2\end{bmatrix}(\Pi_3^2)^T\\&+A_c(\gamma_1^2T_1+\gamma_2^2T_2)A_c^T-(\alpha\tau)^2e_1T_1e_1^T+2\alpha\tau e_1T_1e_7^T\\&-e_7T_1e_7^T+e_1Q_0e_1^T-(1-u)e_2Q_0e_2^T\\&-(1-\alpha)^2\tau^2e_1T_2e_1^T+2(1-\alpha)\tau(e_1T_2e_8^T+e_1T_2e_9^T)\\&-e_8T_2e_8^T-2e_8T_2e_9^T-e_9T_2e_9^T+e_1N_4e_1^T-e_2N_5e_2^T\\&+\alpha\tau(e_1Q_3e_1^T+A_cQ_4A_c^T)+(1-\alpha)\tau(e_1Q_5e_1^T+A_cQ_6A_c^T)\\&-(e_3Q_{1,11}e_3^T+2e_3Q_{1,12}e_5^T+e_5Q_{1,22}e_5^T)\\&+e_2N_6e_2^T-e_3N_4e_3^T+e_3N_5e_3^T-e_4N_6e_4^T\end{aligned}$$

Proof: Let us consider the following candidate for the appropriate Lyapunov-Krasovskii functional:

$$V=\sum_{i=1}^{6}V_i, \tag{6}$$

where

$$V_1=\varepsilon^T(t)P\varepsilon(t),$$

$$V_2=\int_{t-\alpha\tau}^{t}\eta^T(t)\Omega_1\eta(t)ds+\int_{t-\tau}^{t-\alpha\tau}\eta^T(t)\Omega_2\eta(t)ds,$$

$$\begin{aligned}V_3=&\alpha\tau\int_{-\alpha\tau}^{0}\int_{t+\theta}^{t}\eta^T(t)\Psi_1\eta(t)dsd\theta\\&+(1-\alpha)\tau\int_{-\tau}^{-\alpha\tau}\int_{t+\theta}^{t}\eta^T(t)\Psi_2\eta(t)dsd\theta,\end{aligned}$$

$$\begin{aligned}V_4=&\gamma_1\int_{-\alpha\tau}^{0}\int_{\theta}^{0}\int_{t+\lambda}^{t}\dot{x}^T(s)T_1\dot{x}(s)dsd\lambda d\theta\\&+\gamma_2\int_{-\tau}^{-\alpha\tau}\int_{\theta}^{0}\int_{t+\lambda}^{t}\dot{x}^T(s)T_2\dot{x}(s)dsd\lambda d\theta,\end{aligned}$$

$$V_5=\int_{t-\tau(t)}^{t}x^T(s)Q_0x(s)ds,$$

$$\begin{aligned}V_6=&\int_{-\alpha\tau}^{0}\int_{t+\theta}^{t}[x^T(s)Q_3x(s)+\dot{x}^T(s)Q_4\dot{x}(s)]dsd\theta\\&+\int_{-\tau}^{-\alpha\tau}\int_{t+\theta}^{t}[x^T(s)Q_5x(s)+\dot{x}^T(s)Q_6\dot{x}(s)]dsd\theta,\end{aligned}$$

$$\gamma_1=\frac{(\alpha\tau)^2}{2},\ \gamma_2=\frac{(1-\alpha)^2\tau^2}{2}$$

From V_1, V_2, V_5, and V_6, we have their time-derivatives as:

$$\dot{V}_1=2\varepsilon^T(t)P\dot{\varepsilon}(t) \tag{7}$$

$$\begin{aligned}\dot{V}_2=&\eta^T(t)\Omega_1\eta(t)-\eta^T(t-\alpha\tau)\Omega_1\eta(t-\alpha\tau)\\&+\eta^T(t-\alpha\tau)\Omega_2\eta(t-\alpha\tau)-\eta^T(t-\tau)\Omega_2\eta(t-\tau)\\=&\zeta_1^T(t)[(e_1Q_{1,11}e_1^T+2e_1Q_{1,12}A_c^T+A_cQ_{1,22}A_c^T)\\&-(e_3Q_{1,11}e_3^T+2e_3Q_{1,12}e_5^T+e_5Q_{1,22}e_5^T)\\&+(e_3Q_{2,11}e_3^T+2e_3Q_{2,12}e_5^T+e_5Q_{2,22}e_5^T)\\&-(e_4Q_{2,11}e_4^T+2e_4Q_{2,12}e_6^T+e_6Q_{2,22}e_6^T)]\zeta_1(t)\end{aligned} \tag{8}$$

$$\dot{V}_5\le\zeta_1^T(t)[e_1Q_0e_1^T-(1-u)e_2Q_0e_2^T]\zeta_1(t) \tag{9}$$

$$\begin{aligned}\dot{V}_6\le&\zeta_1^T(t)[\alpha\tau(e_1Q_3e_1^T+A_cQ_4A_c^T)\\&+(1-\alpha)\tau(e_1Q_5e_1^T+A_cQ_6A_c^T)]\zeta_1(t)\\&-\int_{t-\alpha\tau}^{t}[x^T(s)Q_3x(s)+\dot{x}^T(s)Q_4\dot{x}(s)]ds\\&-\int_{t-\tau}^{t-\alpha\tau}[x^T(s)Q_5x(s)+\dot{x}^T(s)Q_6\dot{x}(s)]ds\end{aligned} \tag{10}$$

Also, by Eq.(4) in Lemma 2.1, we can obtain $\dot{V}_3$, and $\dot{V}_4$ as follows:

$$\begin{aligned}\dot{V}_3=&(\alpha\tau)^2\eta^T(t)\Psi_1\eta(t)+(1-\alpha)^2\tau^2\eta^T(t)\Psi_2\eta(t)\\&-\alpha\tau\int_{t-\alpha\tau}^{t}\eta^T(s)\Psi_1\eta(s)ds-(1-\alpha)\tau\int_{t-\tau}^{t-\alpha\tau}\eta^T(s)\Psi_2\eta(s)ds\end{aligned} \tag{11}$$

$$\begin{aligned}\dot{V}_4=&\gamma_1^2\dot{x}^T(t)T_1\dot{x}(t)+\gamma_2^2\dot{x}^T(t)T_2\dot{x}(t)\\&-\gamma_1\int_{-\alpha\tau}^{0}\int_{t+\theta}^{t}\dot{x}^T(s)T_1\dot{x}(s)dsd\theta\\&-\gamma_2\int_{-\tau}^{-\alpha\tau}\int_{t+\theta}^{t}\dot{x}^T(s)T_2\dot{x}(s)dsd\theta\end{aligned} \tag{12}$$

and

$$-\gamma_1\int_{-\alpha\tau}^{0}\int_{t+\theta}^{t}\dot{x}^T(s)T_1\dot{x}(s)dsd\theta\le \begin{bmatrix}\alpha\tau x(t)\\ \int_{t-\alpha\tau}^{t}x(s)ds\end{bmatrix}^T\begin{bmatrix}-T_1 & T_1\\ * & -T_1\end{bmatrix}\begin{bmatrix}\alpha\tau x(t)\\ \int_{t-\alpha\tau}^{t}x(s)ds\end{bmatrix} \quad (13)$$

$$-\gamma_2\int_{-\tau}^{-\alpha\tau}\int_{t+\theta}^{t}\dot{x}^T(s)T_2\dot{x}(s)dsd\theta\le \begin{bmatrix}(1-\alpha)\tau x(t)\\ \int_{t-\tau}^{t-\alpha\tau}x(s)ds\end{bmatrix}^T\begin{bmatrix}-T_2 & T_2\\ * & -T_2\end{bmatrix}\begin{bmatrix}(1-\alpha)\tau x(t)\\ \int_{t-\tau}^{t-\alpha\tau}x(s)ds\end{bmatrix} \quad (14)$$

Here, we will consider the time-derivative of V for two cases, $0\le\tau(t)\le\alpha\tau$ and $\alpha\tau\le\tau(t)\le\tau$.

Case I: $0\le\tau(t)\le\alpha\tau$ We can get V_1 as follow

$$\dot{V}_1=\zeta_1^T(t)\left[\Pi_1^1P\Pi_2^T+\Pi_2P(\Pi_1^1)^T\right]\zeta_1(t) \quad (15)$$

From Eq.(11), by use of Eq.(5) in Lemma 2.2, we can get

$$-\alpha\tau\int_{t-\alpha\tau}^{t}\eta^T(s)\Psi_1\eta(s)ds=$$
$$-\alpha\tau\int_{t-\tau(t)}^{t}\eta^T(s)\Psi_1\eta(s)ds-\alpha\tau\int_{t-\alpha\tau}^{t-\tau(t)}\eta^T(s)\Psi_1\eta(s)ds$$
$$\le-\frac{\alpha\tau}{\tau(t)}\int_{t-\tau(t)}^{t}\eta^T(s)ds\Psi_1\int_{t-\tau(t)}^{t}\eta(s)ds$$
$$-\frac{\alpha\tau}{\alpha\tau-\tau(t)}\int_{t-\alpha\tau}^{t-\tau(t)}\eta^T(s)ds\Psi_1\int_{t-\alpha\tau}^{t-\tau(t)}\eta^T(s)ds$$
$$\le-\begin{bmatrix}\int_{t-\tau(t)}^{t}\eta^T(s)ds\\ \int_{t-\alpha\tau}^{t-\tau(t)}\eta^T(s)ds\end{bmatrix}^T\begin{bmatrix}\Psi_1 & \Theta_1\\ * & \Psi_1\end{bmatrix}\begin{bmatrix}\int_{t-\tau(t)}^{t}\eta^T(s)ds\\ \int_{t-\alpha\tau}^{t-\tau(t)}\eta^T(s)ds\end{bmatrix} \quad (16)$$

where Θ_1 is the matrix satisfying $\begin{vmatrix}\Psi_1 & \Theta_1\\ * & \Psi_1\end{vmatrix}\ge 0$. It should be noted that when $\tau(t)=0$ or $\tau(t)=\alpha\tau$, we have $\int_{t-\tau(t)}^{t}x(s)ds=0$ or $\int_{t-\alpha\tau}^{t-\tau(t)}x(s)ds=0$, respectively. Thus, Eq.(15) still holds. From (11) and (16), $\dot{V}_3$ satisfies:

$$\dot{V}_3\le\zeta_1^T(t)[(\alpha\tau)^2(e_1W_{1,11}e_1^T+2e_1W_{1,12}A_c^T+A_cW_{1,22}A_c^T)$$
$$+(1-\alpha)^2\tau^2(e_1W_{2,11}e_1^T+2e_1W_{2,12}A_c^T+A_cW_{2,22}A_c^T)$$
$$-e_9W_{2,11}e_9^T-2e_9W_{2,12}e_3^T+2e_9W_{2,12}e_4^T-e_3W_{2,22}e_3^T$$
$$+2e_3W_{2,22}e_4^T-e_4W_{2,22}e_4^T-\Pi_3^1\begin{bmatrix}\Psi_1 & \Theta_1\\ * & \Psi_1\end{bmatrix}(\Pi_3^1)^T]\zeta_1(t) \quad (17)$$

From (12), (13) and (14), $\dot{V}_4$ satisfies:

$$\dot{V}_4\le\zeta_1^T(t)[A_c(\gamma_1^2T_1+\gamma_2^2T_2)A_c^T-(\alpha\tau)^2e_1T_1e_1^T+2\alpha\tau e_1T_1e_7^T$$
$$+2\alpha\tau e_1T_1e_8^T-e_7T_1e_7^T-2e_7T_1e_8^T-e_8T_1e_8^T-$$
$$(1-\alpha)^2\tau^2e_1T_2e_1^T+2(1-\alpha)\tau e_1T_2e_9^T-e_9T_2e_9^T]\zeta_1(t) \quad (18)$$

Inspired by the work of [17], the following four zero equalities with any symmetric matrices $N_i(i=1,2,3)$, are considered:

$$0=x^T(t)N_1x(t)-x^T(t-\tau(t))N_1x(t-\tau(t))$$
$$-2\int_{t-\tau(t)}^{t}x^T(s)N_1\dot{x}(s)ds$$
$$0=x^T(t-\tau(t))N_2x(t-\tau(t))$$
$$-x^T(t-\alpha\tau)N_2x(t-\alpha\tau)-2\int_{t-\alpha\tau}^{t-\tau(t)}x^T(s)N_2\dot{x}(s)ds$$
$$0=x^T(t-\alpha\tau)N_3x(t-\alpha\tau)-x^T(t-\tau)N_3x(t-\tau)$$
$$-2\int_{t-\tau}^{t-\alpha\tau}x^T(s)N_3\dot{x}(s)ds \quad (19)$$

By use of Eq.(10) and Eq.(19), we have

$$\dot{V}_6=\zeta_1^T(t)[\alpha\tau(e_1Q_3e_1^T+A_cQ_4A_c^T)+(1-\alpha)\tau(e_1Q_5e_1^T+A_cQ_6A_c^T)$$
$$+e_1N_1e_1^T-e_2N_1e_2^T+e_2N_2e_2^T-e_3N_2e_3^T+e_3N_3e_3^T-e_4N_3e_4^T]\zeta_1(t)$$
$$-\int_{t-\tau(t)}^{t}[\eta^T(s)\begin{bmatrix}Q_3 & N_1\\ * & Q_4\end{bmatrix}\eta(s)]ds$$
$$-\int_{t-\alpha\tau}^{t-\tau(t)}[\eta^T(s)\begin{bmatrix}Q_3 & N_2\\ * & Q_4\end{bmatrix}\eta(s)]ds$$
$$-\int_{t-\tau}^{t-\alpha\tau}[\eta^T(s)\begin{bmatrix}Q_5 & N_3\\ * & Q_6\end{bmatrix}\eta(s)]ds \quad (20)$$

Then combining Eqs.(8)-(9), (15), (17)-(18), (20) yields $\dot{V}\le\zeta^T(t)\Phi_1^{(1)}\zeta(t)$. If $\Phi_1^{(1)}<0$ and $0\le\tau(t)\le\alpha\tau$, then $\dot{V}<0$, the system(1) is asymptotically stable.

Case II: $\alpha\tau\le\tau(t)\le\tau$ We can get $\dot{V}_1$ as follow

$$\dot{V}_1=\zeta_1^T(t)\left[\Pi_1^2P\Pi_2^T+\Pi_2P(\Pi_1^2)^T\right]\zeta_1(t) \quad (21)$$

From Eq.(11), by use of Eq.(5) in Lemma 2.2, we can get

$$-(1-\alpha)\tau\int_{t-\tau}^{t-\alpha\tau}\eta^T(s)\Psi_2\eta(s)ds$$
$$=-(1-\alpha)\tau[\int_{t-\tau(t)}^{t-\alpha\tau}\eta^T(s)\Psi_2\eta(s)ds+\int_{t-\tau}^{t-\tau(t)}\eta^T(s)\Psi_2\eta(s)ds]$$
$$\le-\frac{(1-\alpha)\tau}{\tau(t)-\alpha\tau}\int_{t-\tau(t)}^{t-\alpha\tau}\eta^T(s)ds\Psi_2\int_{t-\tau(t)}^{t-\alpha\tau}\eta(s)ds$$
$$-\frac{(1-\alpha)\tau}{\tau-\tau(t)}\int_{t-\tau}^{t-\tau(t)}\eta^T(s)ds\Psi_2\int_{t-\tau}^{t-\tau(t)}\eta^T(s)ds$$
$$\le-\begin{bmatrix}\int_{t-\tau(t)}^{t-\alpha\tau}\eta^T(s)ds\\ \int_{t-\tau}^{t-\tau(t)}\eta^T(s)ds\end{bmatrix}^T\begin{bmatrix}\Psi_2 & \Theta_2\\ * & \Psi_2\end{bmatrix}\begin{bmatrix}\int_{t-\tau(t)}^{t-\alpha\tau}\eta^T(s)ds\\ \int_{t-\tau}^{t-\tau(t)}\eta^T(s)ds\end{bmatrix} \quad (22)$$

where Θ_2 is the matrix satisfying $\begin{bmatrix} \Psi_2 & \Theta_2 \\ * & \Psi_2 \end{bmatrix} \geq 0$. It should be noted that when $\tau(t) = \alpha\tau$ or $\tau(t) = \tau$, we have $\int_{t-\tau(t)}^{t-\alpha\tau} x(s)ds = 0$ or $\int_{t-\tau}^{t-\tau(t)} x(s)ds = 0$, respectively. Thus, Eq.(22) still holds. From (11) and (22), $\dot{V}_3$ satisfies:

$$\begin{aligned}\dot{V}_3 \leq & \zeta_1^T(t)[(\alpha\tau)^2(e_1W_{1,11}e_1^T+2e_1W_{1,12}A_c^T+A_cW_{1,22}A_c^T) \\ & +(1-\alpha)^2\tau^2(e_1W_{2,11}e_1^T+2e_1W_{2,12}A_c^T+A_cW_{2,22}A_c^T) \\ & -e_7W_{1,11}e_7^T-2e_7W_{1,12}e_1^T+2e_7W_{1,12}e_3^T-e_1W_{1,22}e_1^T \\ & +2e_1W_{1,22}e_3^T-e_3W_{1,22}e_3^T-\Pi_3^2\begin{bmatrix}\Psi_2 & \Theta_2 \\ * & \Psi_2\end{bmatrix}(\Pi_3^2)^T]\zeta_1(t)\end{aligned} \quad (23)$$

From (12), (13) and (14), $\dot{V}_4$ satisfies:

$$\begin{aligned}\dot{V}_4 \leq & \zeta_1^T(t)[A_c(\gamma_1^2T_1+\gamma_2^2T_2)A_c^T-(\alpha\tau)^2e_1T_1e_1^T+2\alpha\tau e_1T_1e_7^T \\ & -(1-\alpha)^2\tau^2e_1T_2e_1^T+2(1-\alpha)\tau(e_1T_2e_8^T+e_1T_2e_9^T) \\ & -e_7T_1e_7^T-e_8T_2e_8^T-2e_8T_2e_9^T-e_9T_2e_9^T]\zeta_1(t)\end{aligned} \quad (24)$$

Inspired by the work of [17], the following four zero equalities with any symmetric matrices $N_l(l = 4,5,6)$, are considered:

$$\begin{aligned}0 = & x^T(t)N_4x(t)-x^T(t-\alpha\tau)N_4x(t-\alpha\tau) \\ & -2\int_{t-\alpha\tau}^{t} x^T(s)N_4\dot{x}(s)ds \\ 0 = & x^T(t-\alpha\tau)N_5x(t-\alpha\tau)-x^T(t-\tau(t))N_5x(t-\tau(t)) \\ & -2\int_{t-\tau(t)}^{t-\alpha\tau} x^T(s)N_5\dot{x}(s)ds \\ 0 = & x^T(t-\tau(t))N_6x(t-\tau(t))-x^T(t-\tau)N_6x(t-\tau) \\ & -2\int_{t-\tau}^{t-\tau(t)} x^T(s)N_6\dot{x}(s)ds\end{aligned} \quad (25)$$

By use of Eq.(10) and Eq.(25), we have

$$\begin{aligned}\dot{V}_6 = & \zeta_1^T(t)[\alpha\tau(e_1Q_3e_1^T+A_cQ_4A_c^T)+(1-\alpha)\tau(e_1Q_5e_1^T+A_cQ_6A_c^T) \\ & +e_1N_4e_1^T-e_2N_5e_2^T+e_2N_6e_2^T-e_3N_4e_3^T+e_3N_5e_3^T-e_4N_6e_4^T]\zeta_1(t) \\ & -\int_{t-\tau(t)}^{t}[\eta^T(s)\begin{bmatrix}Q_3 & N_4 \\ * & Q_4\end{bmatrix}\eta(s)]ds \\ & -\int_{t-\alpha\tau}^{t-\tau(t)}[\eta^T(s)\begin{bmatrix}Q_5 & N_5 \\ * & Q_6\end{bmatrix}\eta(s)]ds \\ & -\int_{t-\tau}^{t-\alpha\tau}[\eta^T(s)\begin{bmatrix}Q_5 & N_6 \\ * & Q_6\end{bmatrix}\eta(s)]ds\end{aligned} \quad (26)$$

Then combining Eqs.(8)-(9), (21), (23)-(24), (26) yields $\dot{V} \leq \zeta^T(t)\Phi_1^{(2)}\zeta(t)$. If $\Phi_1^{(2)} < 0$ and $\alpha\tau \leq \tau(t) \leq \tau$, then $\dot{V} < 0$, the system(1) is asymptotically stable. Thus, the proof is completed.

Remark 3.2: In order to reduce the conservatism, a new delay-dependent stability criterion is obtained in Theorem 3.1 by constructing a new Lyapunov-Krasovskii functional. In Eq.(6), V_2, V_3 and V_4 are constructed by using such an idea that the whole delay interval $[-\tau, 0]$ is decomposed into two partitions. We consider the time-varying delay $\tau(t)$ in each partition, then on each partition we choose different weighting matrices, which yields less conservative delay-dependent stability criteria, and that will be illustrated through the examples in the next section.

Remark 3.3: Recently, the reciprocally convex optimization technique was proposed in [15] and [17] to reduce the conservatism of stability criteria for systems with time-varying delays. Motivated by this work, the proposed methods of [15] and [17] were applied to the delay-decomposition method as shown in Eq. (15) and (20).

In many cases, the information on the delay derivative may not be available. Considering this case, the following result can be obtained from Theorem 3.1 by omitting V_5.

Corollary 3.4: For given scalars τ and $0 < \alpha < 1$, the system (1) with (2) is asymptotically stable if there exist positive definite matrices $P = [P_{ij}]_{5\times5}$, $\Omega_1 = [Q_{1,ij}]_{2\times2}$, $\Omega_2 = [Q_{2,ij}]_{2\times2}$, $\Psi_1 = [W_{1,ij}]_{2\times2}$, $\Psi_2 = [W_{2,ij}]_{2\times2}$, T_j, $Q_i(i = 3,4,5,6)$, any matrices $\Theta_1 = [S_{1,ij}]_{2\times2}$, $\Theta_2 = [S_{2,ij}]_{2\times2}$, $N_i(i = 1,2,3,4,5,6)$, with appropriate dimensions such that the following LMIs hold:

$$\Phi_2^{(k)} < 0, \begin{bmatrix}\Psi_k & \Theta_k \\ * & \Psi_k\end{bmatrix} \geq 0, k = 1,2$$

$$\begin{bmatrix}Q_3 & N_1 \\ * & Q_4\end{bmatrix} > 0, \begin{bmatrix}Q_3 & N_2 \\ * & Q_4\end{bmatrix} > 0, \begin{bmatrix}Q_5 & N_3 \\ * & Q_6\end{bmatrix} > 0,$$

$$\begin{bmatrix}Q_3 & N_4 \\ * & Q_4\end{bmatrix} > 0, \begin{bmatrix}Q_5 & N_5 \\ * & Q_6\end{bmatrix} > 0, \begin{bmatrix}Q_5 & N_6 \\ * & Q_6\end{bmatrix} > 0,$$

where

$$\begin{aligned}\Phi_2^{(1)} = & \Pi_1^1P\Pi_2^T+\Pi_2P(\Pi_1^1)^T+(e_1Q_{1,11}e_1^T+2e_1Q_{1,12}A_c^T \\ & +A_cQ_{1,22}A_c^T)+(e_3Q_{2,11}e_3^T+2e_3Q_{2,12}e_5^T+e_5Q_{2,22}e_5^T) \\ & -(e_4Q_{2,11}e_4^T+2e_4Q_{2,12}e_6^T+e_6Q_{2,22}e_6^T) \\ & +(\alpha\tau)^2(e_1W_{1,11}e_1^T+2e_1W_{1,12}A_c^T+A_cW_{1,22}A_c^T) \\ & +(1-\alpha)^2\tau^2(e_1W_{2,11}e_1^T+2e_1W_{2,12}A_c^T+A_cW_{2,22}A_c^T) \\ & -e_9W_{2,11}e_9^T-2e_9W_{2,12}e_3^T+2e_9W_{2,12}e_4^T-e_3W_{2,22}e_3^T \\ & +2e_3W_{2,22}e_4^T-e_4W_{2,22}e_4^T-\Pi_3^1\begin{bmatrix}\Psi_1 & \Theta_1 \\ * & \Psi_1\end{bmatrix}(\Pi_3^1)^T \\ & +A_c(\gamma_1^2T_1+\gamma_2^2T_2)A_c^T-(\alpha\tau)^2e_1T_1e_1^T+2\alpha\tau e_1T_1e_7^T \\ & -e_7T_1e_7^T+e_1Q_0e_1^T-(1-u)e_2Q_0e_2^T \\ & -e_7T_1e_7^T-2e_7T_1e_8^T-e_8T_1e_8^T-(1-\alpha)^2\tau^2e_1T_2e_1^T\end{aligned}$$

$$+2(1-\alpha)\tau e_1T_2e_9^T-e_9T_2e_9^T+\alpha\tau(e_1Q_3e_1^T+A_cQ_4A_c^T)$$
$$+(1-\alpha)\tau(e_1Q_5e_1^T+A_cQ_6A_c^T)+e_1N_1e_1^T$$
$$-(e_3Q_{1,11}e_3^T+2e_3Q_{1,12}e_5^T+e_5Q_{1,22}e_5^T)$$
$$-e_2N_1e_2^T+e_2N_2e_2^T-e_3N_2e_3^T+e_3N_3e_3^T-e_4N_3e_4^T$$
$$\Phi_2^{(2)}=\Pi_1^2P\Pi_2^T+\Pi_2P(\Pi_1^2)^T+(e_1Q_{1,11}e_1^T+2e_1Q_{1,12}A_c^T$$
$$+A_cQ_{1,22}A_c^T)-e_7W_{1,11}e_7^T-2e_7W_{1,12}e_1^T$$
$$+(e_3Q_{2,11}e_3^T+2e_3Q_{2,12}e_5^T+e_5Q_{2,22}e_5^T)$$
$$-(e_4Q_{2,11}e_4^T+2e_4Q_{2,12}e_6^T+e_6Q_{2,22}e_6^T)$$
$$+(\alpha\tau)^2(e_1W_{1,11}e_1^T+2e_1W_{1,12}A_c^T+A_cW_{1,22}A_c^T)$$
$$+(1-\alpha)^2\tau^2(e_1W_{2,11}e_1^T+2e_1W_{2,12}A_c^T+A_cW_{2,22}A_c^T)$$
$$+2e_7W_{1,12}e_3^T-e_1W_{1,22}e_1^T+2e_1W_{1,22}e_3^T-e_3W_{1,22}e_3^T$$
$$-e_7T_1e_7^T+e_1Q_0e_1^T-(1-u)e_2Q_0e_2^T-\Pi_3^2\begin{bmatrix}\Psi_2 & \Theta_2\\ * & \Psi_2\end{bmatrix}(\Pi_3^2)^T$$
$$+A_c(\gamma_1^2T_1+\gamma_2^2T_2)A_c^T-(\alpha\tau)^2e_1T_1e_1^T+2\alpha\tau e_1T_1e_7^T$$
$$-(1-\alpha)^2\tau^2e_1T_2e_1^T+2(1-\alpha)\tau(e_1T_2e_8^T+e_1T_2e_9^T)$$
$$+\alpha\tau(e_1Q_3e_1^T+A_cQ_4A_c^T)+(1-\alpha)\tau(e_1Q_5e_1^T+A_cQ_6A_c^T)$$
$$-(e_3Q_{1,11}e_3^T+2e_3Q_{1,12}e_5^T+e_5Q_{1,22}e_5^T)$$
$$+e_1N_4e_1^T-e_2N_5e_2^T+e_2N_6e_2^T-e_3N_4e_3^T+e_3N_5e_3^T$$
$$-e_4N_6e_4^T-e_8T_2e_8^T-2e_8T_2e_9^T-e_9T_2e_9^T$$

Proof. The proof of this corollary immediately follows form Theorem 3.1.

When $\tau(t)$ is constant: $\tau(t)\equiv\tau$, we have the following theorem.

Theorem3.5 : For given scalars τ and $0<\alpha<1$, the system (1) with (2) is asymptotically stable if there exist positive definite matrices $P=[P_{ij}]_{5\times5}$, $\Omega_1=[Q_{1,ij}]_{2\times2}$, $\Omega_2=[Q_{2,ij}]_{2\times2}$, $\Psi_1=[W_{1,ij}]_{2\times2}$, $\Psi_2=[W_{2,ij}]_{2\times2}$, T_j, $Q_i(i=3,4,5,6)$, any matrices $N_i(i=1,2)$, with appropriate dimensions such that the following LMIs hold:

$$\Phi_3<0,\ \begin{bmatrix}Q_3 & N_1\\ * & Q_4\end{bmatrix}>0,\ \begin{bmatrix}Q_5 & N_2\\ * & Q_6\end{bmatrix}>0,\forall j,k=1,2$$

where

$$\Phi_3=\Pi_0^1P(\Pi_0^2)^T+\Pi_0^2P(\Pi_0^1)^T$$
$$+(e_1Q_{1,11}e_1^T+2e_1Q_{1,12}A_{c0}^T+A_{c0}Q_{1,22}A_{c0}^T)$$
$$+(e_2Q_{2,11}e_2^T+2e_2Q_{2,12}e_4^T+e_4Q_{2,22}e_4^T)$$
$$-(e_3Q_{2,11}e_3^T+2e_3Q_{2,12}e_5^T+e_5Q_{2,22}e_5^T)$$
$$+(\alpha\tau)^2(e_1W_{1,11}e_1^T+2e_1W_{1,12}A_{c0}^T+A_{c0}W_{1,22}A_{c0}^T)$$
$$-(e_2Q_{1,11}e_2^T+2e_2Q_{1,12}e_4^T+e_4Q_{1,22}e_4^T)$$
$$-e_6W_{1,11}e_6^T-2e_6W_{1,12}e_1^T+2e_6W_{1,12}e_2^T$$
$$-e_1W_{1,22}e_1^T+2e_1W_{1,22}e_2^T-e_2W_{1,22}e_2^T$$
$$-e_7W_{2,11}e_7^T-2e_7W_{2,12}e_2^T+2e_7W_{2,12}e_3^T$$
$$-e_2W_{2,22}e_2^T+2e_2W_{2,22}e_3^T-e_3W_{2,22}e_3^T$$
$$+(1-\alpha)^2\tau^2(e_1W_{2,11}e_1^T+2e_1W_{2,12}A_{c0}^T+A_{c0}W_{2,22}A_{c0}^T)$$
$$+\alpha\tau(e_1Q_3e_1^T+A_{c0}Q_4A_{c0}^T)+e_1N_1e_1^T-e_2N_1e_2^T$$
$$-e_3N_2e_3^T+(1-\alpha)\tau(e_1Q_5e_1^T+A_{c0}Q_6A_{c0}^T)$$
$$-e_6T_1e_6^T+A_{c0}(\gamma_1^2T_1+\gamma_2^2T_2)A_{c0}^T+e_2N_2e_2^T$$
$$-e_7T_2e_7^T-(\alpha\tau)^2e_1T_1e_1^T+2\alpha\tau e_1T_1e_6^T$$
$$-(1-\alpha)^2\tau^2e_1T_2e_1^T+2(1-\alpha)\tau e_1T_2e_7^T$$

Proof: Let us consider the following candidate for the appropriate Lyapunov-Krasovskii functional:

$$V=\sum_{i=1}^{6}V_i,$$

where

$$V_1=\varepsilon^T(t)P\varepsilon(t),$$
$$V_2=\int_{t-\alpha\tau}^{t}\eta^T(t)\Omega_1\eta(t)ds+\int_{t-\tau}^{t-\alpha\tau}\eta^T(t)\Omega_2\eta(t)ds,$$
$$V_3=\alpha\tau\int_{-\alpha\tau}^{0}\int_{t+\theta}^{t}\eta^T(t)\Psi_1\eta(t)dsd\theta$$
$$+(1-\alpha)\tau\int_{-\tau}^{-\alpha\tau}\int_{t+\theta}^{t}\eta^T(t)\Psi_2\eta(t)dsd\theta,$$
$$V_4=\gamma_1\int_{-\alpha\tau}^{0}\int_{\theta}^{0}\int_{t+\lambda}^{t}\dot{x}^T(s)T_1\dot{x}(s)dsd\lambda d\theta$$
$$+\gamma_2\int_{-\tau}^{-\alpha\tau}\int_{\theta}^{0}\int_{t+\lambda}^{t}\dot{x}^T(s)T_2\dot{x}(s)dsd\lambda d\theta,$$
$$V_5=\int_{-\alpha\tau}^{0}\int_{t+\theta}^{t}[x^T(s)Q_3x(s)+\dot{x}^T(s)Q_4\dot{x}(s)]dsd\theta$$
$$+\int_{-\tau}^{-\alpha\tau}\int_{t+\theta}^{t}[x^T(s)Q_5x(s)+\dot{x}^T(s)Q_6\dot{x}(s)]dsd\theta,$$
$$\gamma_1=\frac{(\alpha\tau)^2}{2},\gamma_2=\frac{(1-\alpha)^2\tau^2}{2}$$

From V_1,V_2, we have their time-derivatives as:

$$\dot{V}_1=2\varepsilon^T(t)P\dot{\varepsilon}(t)=\zeta_0^T(t)\left[\Pi_0^1P(\Pi_0^2)^T+\Pi_0^2P(\Pi_0^1)^T\right]\zeta_0(t) \quad (27)$$
$$\dot{V}_2=\eta^T(t)\Omega_1\eta(t)-\eta^T(t-\alpha\tau)\Omega_1\eta(t-\alpha\tau)$$
$$+\eta^T(t-\alpha\tau)\Omega_2\eta(t-\alpha\tau)-\eta^T(t-\tau)\Omega_2\eta(t-\tau)$$
$$=\zeta_0^T(t)[(e_1Q_{1,11}e_1^T+2e_1Q_{1,12}A_{c0}^T+A_{c0}Q_{1,22}A_{c0}^T)$$
$$-(e_2Q_{1,11}e_2^T+2e_2Q_{1,12}e_4^T+e_4Q_{1,22}e_4^T)$$
$$+(e_2Q_{2,11}e_2^T+2e_2Q_{2,12}e_4^T+e_4Q_{2,22}e_4^T)$$
$$-(e_3Q_{2,11}e_3^T+2e_3Q_{2,12}e_5^T+e_5Q_{2,22}e_5^T)]\zeta_0(t) \quad (28)$$

By Eq.(4) in Lemma 2.1, we can obtain $\dot{V}_3$ as follows

$$\dot{V}_3=(\alpha\tau)^2\eta^T(t)\Psi_1\eta(t)+(1-\alpha)^2\tau^2\eta^T(t)\Psi_2\eta(t) \quad (29)$$
$$-\alpha\tau\int_{t-\alpha\tau}^{t}\eta^T(s)\Psi_1\eta(s)ds-(1-\alpha)\tau\int_{t-\tau}^{t-\alpha\tau}\eta^T(s)\Psi_2\eta(s)ds$$
$$\le(\alpha\tau)^2\eta^T(t)\Psi_1\eta(t)+(1-\alpha)^2\tau^2\eta^T(t)\Psi_2\eta(t)$$
$$-\int_{t-\alpha\tau}^{t}\eta^T(s)ds\Psi_1\int_{t-\alpha\tau}^{t}\eta(s)ds-\int_{t-\tau}^{t-\alpha\tau}\eta^T(s)ds\Psi_2\int_{t-\tau}^{t-\alpha\tau}\eta^T(s)ds$$
$$=\zeta_0^T(t)[(\alpha\tau)^2(e_1W_{1,11}e_1^T+2e_1W_{1,12}A_{c0}^T+A_{c0}W_{1,22}A_{c0}^T)$$

$$+(1-\alpha)^2\tau^2(e_1W_{2,11}e_1^T+2e_1W_{2,12}A_{c0}^T+A_{c0}W_{2,22}A_{c0}^T)$$
$$-e_6W_{1,11}e_6^T-2e_6W_{1,12}e_1^T+2e_6W_{1,12}e_2^T$$
$$-e_1W_{1,22}e_1^T+2e_1W_{1,22}e_2^T-e_2W_{1,22}e_2^T$$
$$-e_7W_{2,11}e_7^T-2e_7W_{2,12}e_2^T+2e_7W_{2,12}e_3^T$$
$$-e_2W_{2,22}e_2^T+2e_2W_{2,22}e_3^T-e_3W_{2,22}e_3^T]\zeta_0(t)$$

Also, we can get $\dot{V}_4$ as follows:

$$\dot{V}_4=\gamma_1^2\dot{x}^T(t)T_1\dot{x}(t)+\gamma_2^2\dot{x}^T(t)T_2\dot{x}(t)$$
$$-\gamma_1\int_{-\alpha\tau}^{0}\int_{t+\theta}^{t}\dot{x}^T(s)T_1\dot{x}(s)dsd\theta$$
$$-\gamma_2\int_{-\tau}^{-\alpha\tau}\int_{t+\theta}^{t}\dot{x}^T(s)T_2\dot{x}(s)dsd\theta \quad (30)$$

and

$$-\gamma_1\int_{-\alpha\tau}^{0}\int_{t+\theta}^{t}\dot{x}^T(s)T_1\dot{x}(s)dsd\theta\le$$
$$\begin{bmatrix}\alpha\tau x(t)\\ \int_{t-\alpha\tau}^{t}x(s)ds\end{bmatrix}^T\begin{bmatrix}-T_1 & T_1\\ * & -T_1\end{bmatrix}\begin{bmatrix}\alpha\tau x(t)\\ \int_{t-\alpha\tau}^{t}x(s)ds\end{bmatrix}$$
$$-\gamma_2\int_{-\tau}^{-\alpha\tau}\int_{t+\theta}^{t}\dot{x}^T(s)T_2\dot{x}(s)dsd\theta\le$$
$$\begin{bmatrix}(1-\alpha)\tau x(t)\\ \int_{t-\tau}^{t-\alpha\tau}x(s)ds\end{bmatrix}^T\begin{bmatrix}-T_2 & T_2\\ * & -T_2\end{bmatrix}\begin{bmatrix}(1-\alpha)\tau x(t)\\ \int_{t-\tau}^{t-\alpha\tau}x(s)ds\end{bmatrix}$$

then

$$\dot{V}_4\le\zeta_0^T(t)[A_{c0}(\gamma_1^2T_1+\gamma_2^2T_2)A_{c0}^T-(\alpha\tau)^2e_1T_1e_1^T+2\alpha\tau e_1T_1e_6^T-e_6T_1e_6^T$$
$$-(1-\alpha)^2\tau^2e_1T_2e_1^T+2(1-\alpha)\tau e_1T_2e_7^T-e_7T_2e_7^T]\zeta_0(t) \quad (31)$$

From V_5 we can obtain

$$\dot{V}_5\le\zeta_0^T(t)[\alpha\tau(e_1Q_3e_1^T+A_{c0}Q_4A_{c0}^T)$$
$$+(1-\alpha)\tau(e_1Q_5e_1^T+A_{c0}Q_6A_{c0}^T)]\zeta_0(t)$$
$$-\int_{t-\alpha\tau}^{t}[x^T(s)Q_3x(s)+\dot{x}^T(s)Q_4\dot{x}(s)]ds$$
$$-\int_{t-\tau}^{t-\alpha\tau}[x^T(s)Q_5x(s)+\dot{x}^T(s)Q_6\dot{x}(s)]ds \quad (32)$$

Inspired by the work of [17], the following four zero equalities with any symmetric matrices $N_i(i=1,2)$, are considered:

$$0=x^T(t)N_1x(t)-x^T(t-\alpha\tau)N_1x(t-\alpha\tau)$$
$$-2\int_{t-\alpha\tau}^{t}x^T(s)N_1\dot{x}(s)ds$$
$$0=x^T(t-\alpha\tau)N_2x(t-\alpha\tau)-x^T(t-\tau)N_2x(t-\tau)$$
$$-2\int_{t-\tau}^{t-\alpha\tau}x^T(s)N_2\dot{x}(s)ds \quad (33)$$

By use of Eq.(32) and Eq.(33), we have

$$\dot{V}_5\le\zeta_0^T(t)[\alpha\tau(e_1Q_3e_1^T+A_{c0}Q_4A_{c0}^T)$$
$$+(1-\alpha)\tau(e_1Q_5e_1^T+A_{c0}Q_6A_{c0}^T)+e_1N_1e_1^T$$
$$-e_2N_1e_2^T+e_2N_2e_2^T-e_3N_2e_3^T]\zeta_0(t)$$
$$-\int_{t-\alpha\tau}^{t}[\eta^T(s)\begin{bmatrix}Q_3 & N_1\\ * & Q_4\end{bmatrix}\eta(s)]ds$$
$$-\int_{t-\tau}^{t-\alpha\tau}[\eta^T(s)\begin{bmatrix}Q_5 & N_2\\ * & Q_6\end{bmatrix}\eta(s)]ds \quad (34)$$

Then combining Eqs.(27)-(29), (31) and (34) yields $\dot{V}\le\zeta_0^T(t)\Phi_3\zeta_0(t)$. If $\Phi_3<0$ then $\dot{V}<0$, the system(1) is asymptotically stable, which completes the proof.

4. Numerical examples

In this section, we provide two examples to show the less conservativeness of the proposed new stability criteria in this paper.

Example 1. Consider the following neutral time-delay system $\dot{x}(t)=Ax(t)+A_1x(t-\tau(t))$ with

$$A=\begin{bmatrix}-2 & 0\\ 0 & -0.9\end{bmatrix},\ A_1=\begin{bmatrix}-1 & 0\\ -1 & -1\end{bmatrix}$$

When $\dot{\tau}\le\mu<1$, applying Theorem 3.1, the corresponding maximum admissible upper bounds are given in Table 1 which clearly shows that the effectiveness of the delay-decomposition approach.

Example 2. Consider the following nominal neutral system with constant time-delay

$$\dot{x}(t)-\begin{bmatrix}-0.2 & 0\\ 0.2 & -0.1\end{bmatrix}\dot{x}(t-\tau)$$
$$=\begin{bmatrix}-0.9 & 0.2\\ 0.1 & -0.9\end{bmatrix}x(t)+\begin{bmatrix}-1.1 & -0.2\\ -0.1 & -1.1\end{bmatrix}x(t-\tau)$$

For above system, the maximum delay bounds for asymptotic stability were investigated in [25], [26], [27] and [16]. From Table 2, it can be seen that the obtained delay bounds by Theorem 3.5 are larger than those of [25], [26], [27] and [16].

5. Conclusion

It this paper, new delay-dependent stability criteria for neutral time-delay systems are proposed. In order to obtain less conservative results, a new delay-decomposition method is used to improve the maximum admissible upper bounds of stability criterion. Numerical examples have been

given to show that our stability are less conservative than some existing ones in the literatures.

References

[1] H.R. Karimi, M. Zapateiro, N. Luo, Stability analysis and control synthesis of neutral systems with time-varying delays and nonlinear uncertainties, Chaos, Solitons and Fractals 42 (2009) 595--603.

[2] Y. Kuang, Delay Differential Equations with Applications in Population Dynamics, Academic Press, Boston, 1993.

[3] R.K. Brayton, Bifurcation of periodic solutions in a nonlinear difference--differential equation of neutral type, Quarterly of Applied Mathematics 24 (1996) 215--224.

[4] S.I. Niculescu, Delay Effects on Stability: A Robust Control Approach, Springer, Berlin, 2001.

[5] G.D. Hu, Some simple stability criteria of neutral delay-differential systems, Applied Mathematics and Computation 80 (1996) 257--271.

[6] M.S. Mahmoud, Robust H∞ control of linear neutral systems, Automatica 36 (2000) 757--764.

[7] J.D. Chen, C.H. Lien, K.K. Fan, J.S. Cheng, Delay-dependent stability criterion for neutral time-delay systems, Electronics Letters 22 (2000) 1897--1898.

[8] S. Xu, J. Lam, Y. Zou, Further results on delay-dependent robust stability conditions of uncertain neutral systems, International Journal of Robust and Nonlinear Control 15 (2005) 233--246.

[9] Q.-L. Han, On robust stability of neutral systems with time-varying discrete delay and norm-bounded uncertainty, Automatica 40 (2004) 1087--1092.

[10] P. Park, A delay-dependent stability criterion for systems with uncertain time-invariant delays, IEEE Trans. Autom. Control 44 (1999) 876--877.

[11] Y. Ariba, F. Gouaisbaut, An augmented model for robust stability analysis of time-varying delay systems, International Journal of Control 82 (2009) 1616-1626.

[12] J. Sun, G.P. Liu, and J. Chen, Delay-dependent stability and stabilization of neutral time-delay systems, International Journal of Robust and Nonlinear Control 15 (2009) 1364-1375.

[13] V. Suplin, E. Fridman, U. Shaked, H∞ control of linear uncertain time-delay systems-a projection approach, IEEE Trans. Autom. Control 51 (2006) 680--685.

[14] X.M. Zhang, Q.L. Han, New Lyapunov--Krasovskii functionals for global asymptotic stability of delayed neural networks, IEEE Trans. Neural Netw. 20 (2009) 533--539.

[15] P.G. Park, J.W. Ko, C. Jeong, Reciprocally convex approach to stability of systems with time-varying delays, Automatica 47 (2011) 235-238.

[16] P. Balasubramaniam, R. Krishnasamy, R. Rakkiyappan, Delay-dependent stability of neutral systems with time-varying delays using delay-decomposition approach, Applied Mathematical Modelling 36 (2012) 2253-2261.

[17] S.H. Kim, P. Park, and C. Jeong, Robust H∞ stabilisation of networked control systems with packet analyser, IET Control Theory and Applications 4 (2010) 1828-1837.

[18] C. Lin, Q.-G. Wang, T.H. Lee, A less conservative robust stability test for linear uncertain time-delay systems, IEEE Trans. Autom. Control l51 (1) (2006) 87--91.

[19] H. Yan, X. Huang, M. Wang, H. Zhang, New delay-dependent stability criteria of uncertain linear systems with multiple time-varying state delays, Chaos, Solitons and Fractals 37 (1) (2008) 157--165.

[20] Y. He, Q.-G. Wang, C. Lin, M.Wu, Delay-range-dependent stability for systems with time-varying delay, Automatica 43 (2) (2007) 371--376.

[21] P. Park, J.W. Ko, Stability and robust stability for systems with a time-varying delay, Automatica 43 (10) (2007) 1855--1858.

[22] J.W. Ko, P.G. Park, Delay-dependent stability criteria for systems with asymmetric bounds on delay derivative, Journal of the Franklin Institute 348 (2011) 2674-2688.

[23] M.Wu, Y. He, and J.H. She, New delay-dependent stability criteria and stabilizing method for neutral systems, IEEE Trans. Autom. Control 49 (2004) 2266-2271.

[24] M.N.A. Parlakci, Robust stability of uncertain neutral systems: a novel augmented Lyapunov functional approach, IET Control Theory and Applications 1 (2007) 802-809.

[25] W. Qian, S. Cong, Y. Sun, and S. Fei, Novel robust stability criteria for uncertain systems with time-varying delays, Appl. Math. Comput. 215 (2009) 866-872.

[26] X. Nian, H. Pang, W. Gui, and H. Wang, New stability analysis for linear neutral system via state matrix decomposition, Appl. Math. Comput. 215 (2009) 1830-1837.

[27] M.J. Park, O.M. Kwon, J.H. Park, and S.M. Lee, A new augmented Lyapunov-Krasovskii functional approach for stability of linear systems with time-varying delays, Appl. Math. Comput. 217 (2011) 7197-7209.

Weiwei Zhang received the B.S. and M.S. degrees in college of information from Yan Shan University, Qinhuangdao, China, in 2003 and 2006, respectively. Currently she is an Assistant Professor at the Hebei United University, Tangshan, China. Her current research interests include stability analysis of linear systems and synchronization of complex dynamic networks.

Chao Ge received the B.S. and M.S. degrees in college of information from Hebei Polytechnic University, Tangshan, China, in 2003 and 2006, respectively. Currently he is an Assistant Professor at the Hebei United University, Tangshan, China. His current research interests include nonlinear control systems, control systems design over network and teleoperation systems.

Hong Wang received the B.S. and M.S. degrees in college of chemical engineering from Hebei Polytechnic University, Tangshan, China, in 2003 and 2006, respectively. Currently she is an Assistant Professor at the Hebei United University, Tangshan, China. Her current research interests include environmental ecology control systems.

Table 1: The maximum admissible upper bounds of time-varying delays with different values of μ (Example1)

μ	**0**	**0.1**	**0.5**	**0.9**	**unknown**
[18]	4.47	3.60	2.00	1.18	-
[13]	4.47	3.60	2.00	1.18	-
[19]	4.47	3.60	2.00	1.18	-
[20]	4.47	3.60	2.04	1.37	-
[21]	4.47	3.66	2.33	1.87	-
[22]	5.55	4.41	2.40	2.12	2.12
Theorem 3.1	5.87 ($\alpha = 0.59$)	4.43 ($\alpha = 0.46$)	2.46 ($\alpha = 0.22$)	2.22 ($\alpha = 0.51$)	2.22 ($\alpha = 0.51$)

Table 2: The maximum admissble upper bounds of constant time delays (Example2)

Method	τ
[25]	1.8037
[26]	1.9132
[27]	2.0054
[16]	2.1046
Theorem 3.5	2.1445 $\alpha = 0.57$

Routing Protocol in Urban Environment for V2V communication Vanet

My Driss LAANAOUI, Pr. Said RAGHAY

University Cadi Ayyad Marrakech, Department of Mathematics LAMAI
B.P 549, Av.Abdelkarim Elkhattabi, Marrakech, Morocco

Abstract

The vehicle-to-vehicle communication is a very actual and challenging topic. Vehicles equipped with devices capable of short-range wireless connectivity can form a particular mobile ad-hoc network, VANET − Vehicular Ad-hoc Network. The existence of such networks opens the way for a wide range of applications. Two of the most important classes of such applications are those related to route planning and traffic safety. Route planning aims to provide drivers with real-time traffic information, which, in the absence of a VANET, would require an expensive infrastructure.

In this work we evaluate our VANET routing protocol that is especially designed for city environments. This protocol is based on the localization of the node, the cost assigned to the section and score for each vehicle.

Keywords: *VANET, routing, simulation, Dijkstra, urban environment, Delivery Ratio, end to end delay, IDM.*

1. Introduction

A critical aspect in a simulation study of VANET is the need for a mobility model that would reflect the real behavior of vehicular traffic, as vehicular mobility significantly impacts the networking shape of VANET.

The majority of the VANET convenience applications are more or less directly related to a navigation system. Prime examples are again a distributed traffic information system for finding routes with short travel times based on the current traffic situation and a system for finding free parking places. From the perspective of information generation in VANETs, the fact that more and more vehicles are equipped with a navigation system means that more and more vehicles have a particularly powerful and sophisticated kind of ‘sensor’ at their disposal: a navigation system not only has quite accurate position and speed information available, but also detailed map data and information about the intended driving direction.

2. Communication requirements

We are interested to the Vehicle to vehicle communication. One way to propagate information between vehicles very fast is to use flooding. In a naive implementation every node that receives this information will simply rebroadcast it. To avoid infinite packet duplication, each node will broadcast a given packet at most once. In addition a time to live (TTL) counter may be used to limit the area where the packet is distributed. This naive approach will transmit a large amount of redundant packets, potentially leading to severe congestion. This is known as the ‘broadcast storm problem’ [1]. Many approaches have been proposed to deal with this problem.

We begin at first by description of the V2V communication with the diagram figure 1:

- A vehicular system is composed of one or more area.
- An area consists of several junctions and cells.
- A junction may be source or candidate.
- A vehicle can be elected and belongs to one or more cells

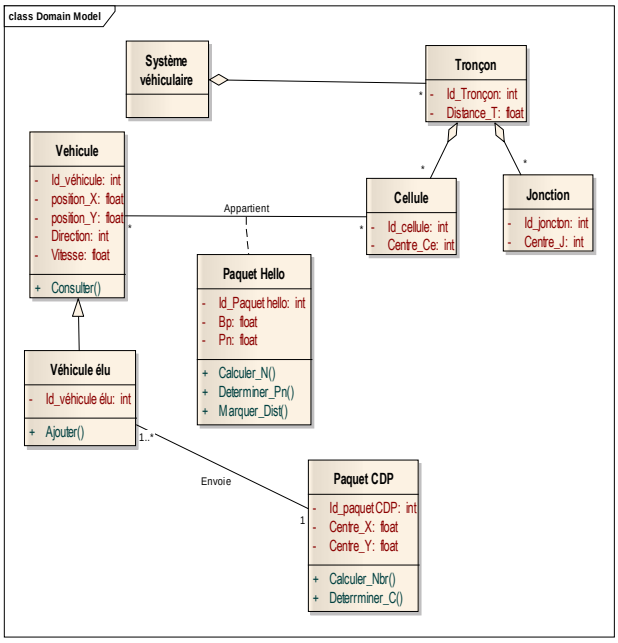


Figure 1: UML class diagram for the vehicular system

3. Routing performance requirements

In Ad-hoc vehicle-to-vehicle communication, where no supporting infrastructure is required, vehicles communicate when they are within the radio range of each other, or when multiple hop relay via other vehicles are available. Messages need to be routed from the source to one or several destinations. Desirable characteristics of routing protocols include [3] :

- Minimal control overhead
- Loop-free routing paths
- Low complexity
- Multicast capabilities

Beside the above requirements, the vehicular environment poses new challenging requirements to vehicle-to-vehicle routing protocol design, including [4]:

- Adapting routing information in highly mobile topologies
- Short convergence time of the routing algorithms
- Short delay for neighbor discovery
- Scalability

In this work, we present a new geographic routing protocol VANET called “Intelligent Routing protocol in Urban environment for Vanet “(IRUV). To evaluate the performances of this protocol, we compare it with GyTAR and LAR in terms of: End to End Delay, Delivery Ratio and efficacy. GyTAR and LAR are efficient in comparison with GSR [2].

4. Description of the IRUV protocol

The IRUV protocol uses Multipoint Relays (elected vehicle) in each zone. The elected vehicle sends information to the neighbor’s vehicles and updates the CDP packet where only the links that lead to the elected vehicle are authorized to enriching CDP [4]. The frequency of control packets increased with mobility.
IRUV adopts Dijkstra's algorithm to choice the optimal way for destination. So, we calculate the cost of junction instead of the score in GyTAR
The approach adopted by IRUV protocol is given in three parts:

- Collecting information on traffic segment "between source and candidate junctions".
- Calculating the score for the candidate junction which represents the cost of the section of road.
- Apply Dijkstra's algorithm to choose the best path to the destination

The UML sequence diagram below describes the purpose of electing vehicle to vehicle V2V communication that can enrich the CDP package.

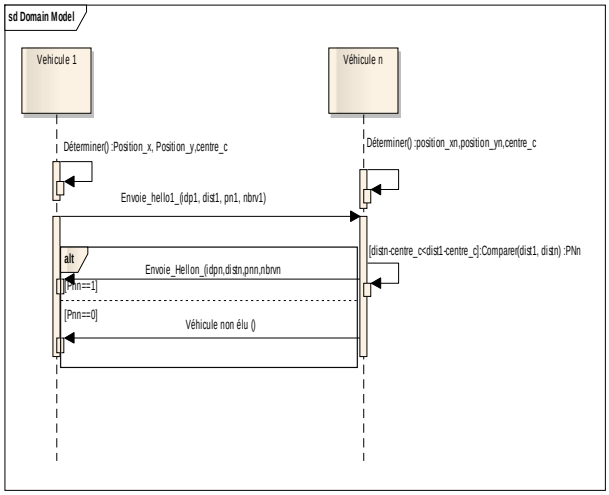


Figure 2: UML sequence diagram for V2V communications

The process of communication is composed of the following steps:

- Election of the vehicle which is near to the candidate junction to send the message
- Electing the vehicle closest to the center of the cell

Enrichment of the CDP package elected by the vehicle, by adding information about: traffic density, the position of the center, the identifier of the cell i, and send the packet to the CDP vehicle closest to the cell i+1. So, it can quickly reach the range of the vehicle so as to arrive faster to elected vehicle in the cell i-1 (repeat the procedure until getting the vehicle to a cell 1 which is the source junction)

The procedure is composed of following step:

- Take into consideration the distance between the vehicle, the next cell, the speed of candidate nodes and the geographical position of these nodes
- The notion of score will also be assigned to the node according to time:

tp=(xp-xi)/vi+ti (1)

The node will be selected is the node with the minimal score (figure 2)

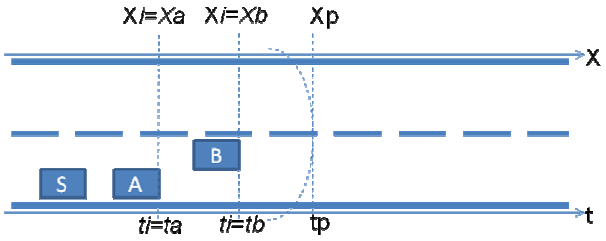


Figure 3 : sending the packet to the node closest to the next zone

With:

- tp: time needed to reach the limited scope of the vehicle S
- ti: Vhi moment the vehicle and the position xi
- Xi: Vhi position of the car at time t
- Xp: scope of the vehicle S
- Vi: Car speed Vhi

To send data between two junctions, we consider:

✓ The direction of the next vehicle
✓ The speed of this vehicle

Each vehicle maintains a neighbor table where all information mentioned above is registered.

Computing the score of the section of the road between the source junction and candidate ones is based on the collected information in the first step.

$$score(Ni) = \alpha(1 - D_p) + \beta(1 - D) + \gamma N_d \quad (2)$$

$$\alpha + \beta + \gamma = 1 \quad (3)$$

With:

- DJ (Di respectively): the curvilinear distance between J (respectively I) and the destination
- Dp = Dj / Di: Dp determines the proximity of the intersection relative to the candidate destination
- D = Dij / Sun: The distance between the source and the branch candidate from the source junction
- α, β, γ are constants

After this step, we calculate the cost of the candidate junction, to identify the section between source and candidate junctions using the following equation:

$$\cos t(Ni) = \frac{1}{score(Ni)} \quad (4)$$

This parameter will be used later to identify the best path using Dijkstra's algorithm implemented in the IRUV protocol. IRUV protocol selects the junctions based on the cost.

5. Simulation

5.1 Present simulation plan

In present study, we work on a VANET simulation using VanetMobiSim / NS-2 application. In VanetMobiSim, we use a micro-mobility model belonging to Intelligent Driver Model (IDM), using lane Changing scenario (IDM-IM) in mobility model building.

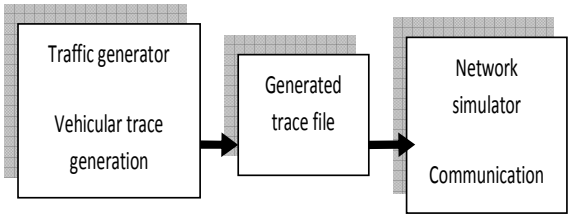


Figure 4: Simulation basic architecture

This diagram describes the simulation steps:

- Traffic simulation tool that generates a vehicular mobility traces (using VanetMobiSim).
- Network simulator that generates the communication environment (using NS2).

5.1.1. The network simulator NS2

NS2 is a discrete event simulator targeted at networking research. NS2 provides substantial support for simulation of TCP routing and multicast protocols over wired and wireless (local and satellite) networks.

5.1.2 The Traffic simulator

To generate realistic motion, we use a mobility emulator well known: VanetMobiSim.

VanetMobiSim is an extension of CanuMobiSim, which focuses on road mobility, offering more realistic mobility models at microscopic and macroscopic
VanetMobiSim Processes:

Input:

- Defining xml file with VANET parameters such as nodes, area, speed, Time etc.
- Output:
- Generate mobility model (Traffic generator) trace file in formats of Ns-2 file
- Provides visualization of the mobility scenario.
- A screen shot of the model in an xfig figure (figure5).

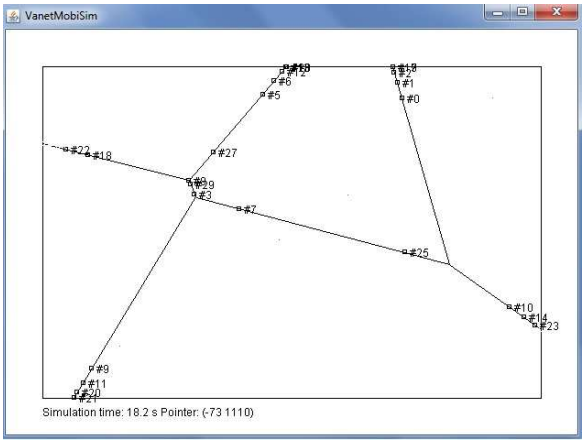


Figure 5: IDM_IM VanetMobiSim model

5.2 Simulation results

To evaluate our proposition, we implement IRUV, GyTAR and LAR protocols in NS2 and we compared those protocols by applying them on a real traffic generated by VanetMobisim.
For measuring performance of routing protocols we chose metrics that we considered most significant to measure the performance of a routing protocol: Packet Delivery Fraction (PDF), Average End to End Delay (AVG), and efficiency.
All the key parameters of our simulation are summarized in the following table:

Table 1: Simulation parameters

Parameter	**Setting**
Traffic model	IDM_IM.xml
Vehicle velocity	30 to 60 Km/H
Transmission range	250m
Map size	1000x1000 m^2
Number of vehicle	50 to 400
Packet sending rate	0.2 s
MAC protocol	IEEE 802.11
Simulation time	200s

For Displaying Results, We use awk to extract information to display from trace file generated by NS2.

5.2.1 Delivery Ratio

In this part, we calculate the delivery ratio for IRUV and LAR. We obtained the following graph (figure 6). It is the amount (total size) of packet received to the amount of packets sent by all nodes. The estimation is made by using UNIX 'awk'.
In figure 6, we have demonstrate that the IRUV protocol provides a delivery ratio better than LAR and GyTAR, especially in the case of a mobility less than 200 vehicles, and significantly higher than LAR and GyTAR protocols in the case of a denser mobility. This is due to the implementation of Dijkstra's algorithm, which has good convergence, and improved delivery because of using to the Dijkstra algorithm used to choose the path with the lowest cost.

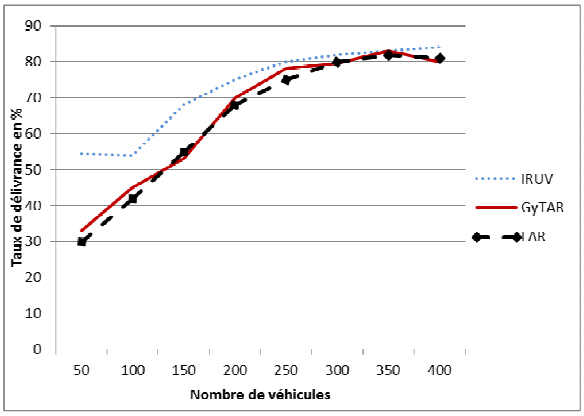


Figure 6: Delivery ratio in terms of vehicles number

5.2.2 End to End Delay

The delay of end to end latency includes the discovery of roads, transit time in the queues for intermediate nodes and the transmission time of a jump to another. We measure the average time from start to finish over all packets received during the simulation and then compute the average. This metric represents the efficiency of the protocol in terms of response time and in terms of choice for optimal paths.
We extract information to calculate the end to end delay by using UNIX 'awk'.

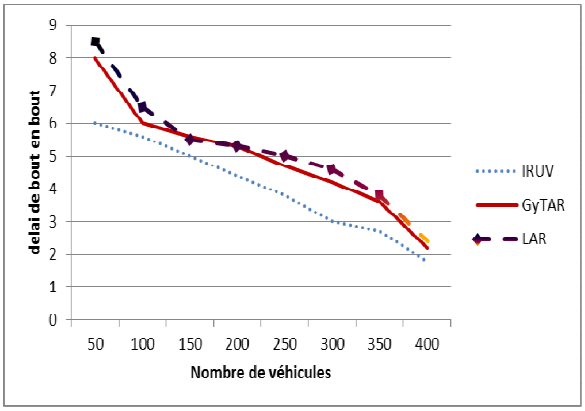


Figure 7: End to end delay in term of vehicle number

Note that IRUV protocol gives end to end delay much lower than the Protocol LAR and GyTAR where traffic is less than 250 vehicles, and significantly lower in the case of a denser mobility. This can be explained by the addition of the concept of score for vehicles to choose the vehicle closest to the next zone for forwarding the packet as soon as possible to the elected vehicle in the next zone, which improves the greedy forwarding algorithm [Lakshimi12] adopted by protocol GyTAR and LAR, hence the fast exchanging CDP packet for IRUV protocol to collect information specific to the zone and choose the best section to transfer data.

6. Conclusion:

In this work, we proposed a new geographic protocol (IRUV), and we compare this protocol with LAR and GyTAR protocols, all of them uses real time traffic density information and movement prediction to route data in VANET.
Our protocol selects the junctions by comparing the cost, more traffic is high, and more the cost given to the candidate junction of the link is low.

It was demonstrated that our proposition has a best delivery ratio and end to end delay. So our protocol IRUV selects the fastest and shortest route in the road network as the best way

References

[1] Ni SY, Tseng YC, Chen YS and Sheu JP 1999 The broadcast storm problem in a mobile ad hoc network. MobiCom '99: Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking, pp. 151–162.

[2] Protocoles pour les communications dans les réseaux de véhicules en environnement urbain :Routage et GeoCast bases sur les intersections ; Jerbi , these 2008

[3] A New Mobile Infrastructure Based VANET Architecture for Urban Environment ; Jie Luo; Xinxing Gu; Tong Zhao; Wei Yan; 2010

[4] Impact of directional density on GyTAR routing protocol for VANETs in city environments ; Bilal, Sardar M.; Mustafa, Saad; Saeed, Uzma; 2011

[5] An Intelligent Routing Protocol For VANET; MD Laanaoui, S.Raghay; International Journal of Computer Information Systems, Vol. 3, No. 1, 2011

[6] Comparison of Three Greedy Routing Algorithms for Efficient Packet Forwarding in VANET ; K. Lakshmi1, K.Thilagam2, K. Rama3, A.Jeevarathinam4, S.Manju Priya; Int.J.Computer Techology & Applications,Vol 3 (1),146-151, 2012

My Driss Laanaoui received his bachelor on IEEA (Informatics, Electronics, Electrical, and Automation) in 2003 from the University Cadi Ayyad, Faculty of Science and Technology of Marrakech, and Master on Telecommunications and networks in 2007 from the University Cadi Ayyad, Marrakech. He has published five papers in International conferences. He is currently a Ph.D student in the university CADI AYYAD Marrakech, Morroco.

Said Raghay is a Professor in Mathematical modeling and informatics . University Cadi Ayyad, Faculty of Science and Technology of Marrakech.

Mining User Similarity Using Spatial-temporal Intersection

Yimin Wang[1], Ruimin Hu[1], Wenhua Huang[1] and Jun Chen[1]

**[1] National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University
Wuhan, Hubei, 430072, China**

Abstract

The booming industry of location-based services has accumulated a huge collection of users' location trajectories and also brings us opportunities and challenges to automatically discover valuable knowledge from these trajectories. In this paper, we investigate the problem of measuring the similarity between users. Such user similarity is significant to individuals, communities and businesses by helping them effectively retrieve the information. To achieve this goal, we firstly propose a storage structure to represent the user's trajectories, which not only stores the sequence of user's trajectory, but also stores regions with indexing of trajectories which pass the regions. After that, we give the similarity function between users using the spatial-temporal intersection in regions which are passed by the two users. Finally, we develop a spatial-temporal intersection algorithm to measure user similarity based on the definition and storage structure, and we illustrate the results and performance of the algorithm by extensive experiments.

Keywords: *Trajectory Analysis, User Similarity, Spatial-temporal Data Mining*

1. Introduction

Recently, the increasing pervasiveness of location-acquisition technologies, like GPS and GSM networks, are leading to the collection of large spatial-temporal trajectory data, which bring the opportunity of discovering valuable knowledge about users' movements [1]. A number of interesting applications are being developed based on the analysis of trajectories. For example, it is possible to determine migration patterns of animals by analyzing similar trajectories of them; in a city traffic system, it is helpful to locate popular routes for programming a new route by mining the trajectories of the citizens. The basis of these applications is determining the similarity among trajectories of moving object.

The trajectory of a moving object could be defined as a sequence of positions of the moving object over a period of time, see Fig. 1. We will say that A and B are more similar, because they pass more common locations. As a trajectory is a time series essentially, many sequence similarity search techniques were used to address this problem, such as Longest Common Subsequences (LCSS)[5], Dynamic Time Warping (DTW)[6], and Edit Distance on Real sequence (EDR)[7] etc. These methods address the problem of different length, noise and local shift, which often appear on comparing the similarity between trajectories. However, the complexity of the algorithm is O(n2), where n is the length of trajectory. Many existing methods were developed based on these classical methods, most of which optimize the efficiency of k-nearest neighbors search algorithm by using index structure and pruning techniques, nevertheless these optimization are restricted by the complexity of computing similarity between two trajectories.

Fig. 1 Trajectories of Persons

In this paper, we proposed a new approach inspired by text retrieval methods, in which a document is seen as a set of many words and all the documents are organized as an inverted file to facilitate efficient retrieval. An inverted file is a structure that has an entry for each word in the corpus followed by a list of all the documents in which occurs. In our methods, the persons are analogy to documents and the locations are the words. Different to words which are discrete, locations are continuous. We compensate this difference by region dividing. Specially, we divide the whole activity area on which the persons move on into many small regions, and then we use the sequence of regions to approximately represent the trajectory. While computing the similarity between two persons, we use the number of regions passed by both the two persons, with the region-based representation.

The major contributions of this paper are as follows:

1. We propose Compression representation of user's trajectories and a storage structure which not only stores

the grid sequence of a trajectory, but also stores regions with indexing of trajectories which pass the regions. With the storage structure, we can easily determine whether and when a trajectory passes through a certain grid.

2. We give three definitions of similarity function, named Maximum Co-occurrence Time (MCT): Absolute Similarity, Relative Similarity and Partial Similarity based on the compression representation. The detail description of these definitions is given in Section 2.

3. Based on the storage structure, we propose a fast k-nearest neighbor search algorithm. The detailed description of these algorithms is given in Section 3.

The rest of the paper is organized as follows. Section 2 describes the grid representation of trajectory, a storage structure and a novel similarity function. In Section 3, we give the k-nearest neighbor search algorithm. In Section 4, we evaluate the accuracy and efficiency of MCT by comparing it with other algorithms. Section 5 is a brief conclusion.

2. Related Work

Similarity search has been well studied in the context of time series and trajectory data. The simplest approach to define the similarity between two sequences is to convert sequence into a vector and then use a p-norm distance to define the similarity measure. The p-norm distance between two n-dimensional vectors $\overline{x}$ and $\overline{y}$ is defined as $L_p(\overline{x},\overline{y}) = (\sum_{i=1}^{n}(x_i - y_i)^p)^{\frac{1}{p}}$. For p=2 it is the well known Euclidean distance.

Agrawal et al. firstly proposed an approach for sequence similarity measure. Their method transforms the sequence into vector with Discrete Fourier Transformation and uses Euclidean distance to determine the similarity [2]. Chen et al. [3] took Discrete Wavelet Transformation to convert the sequence, while Cai et al. [4] transformed the sequence with Chebyshev Polynomials. However, these methods based on p-norm distance require the trajectories with the same length.Several typical similarity functions for different length sequence include Longest Common Subsequence (LCSS) [5], Dynamic Time Warping (DTW) [6], Edit Distance on Real Sequences (EDR) [7]and Edit Distance with Real Penalty (ERP) [8]. Time warping technique first has been used to match signals in speech recognition.

Berndt and Clifford[6] proposed to exploit this technique to measure the similarity of time-series data, and its basic idea is to allow 'repeating' some points as many times as needed to achieve the best alignment. Sakurai et al. [9] improved DTW by using an index structure with segmentation and lower bounded distance measure. Vlachos et al. [5] used LCSS to compare two trajectories with the consideration of spatial space shifting. Longest Common Subsequence is a classical distance function for two strings; another distance function is Edit Distance. The edit distance between two strings of characters is the number of operations required to transform one of them into the other. The allowed operations include 'replace', 'delete', 'insert' and so on. Each operation's cost is 1. EDR and ERP are both based on Edit Distance and proposed by Chen et al. EDR directly utilize Edit Distance to measure similarity of two trajectories. Compare to EDR, ERP use the Euclidean distance as operation's cost instead of 1[8]. These method could be used in both time series and trajectory data and the time complexity is basically $O(n^2)$.

Lin et al. [10] use OWD to compute two trajectories similarity. They only consider the spatial shape of trajectory without the time ordering, and utilize grid to approximate represent trajectories. Grid representation is also used in this paper. Frentzos et al. [11] define a dissimilarity metric (DISSIM) for the measurement of the spatiotemporal dissimilarity between two trajectories. Pelekis et al. [12] introduce the Locality in-between Polylines (LIP) function. The idea is to calculate the area of the shape formed by two 2D polylines. This method requires that the area is finite.

Recently, Tiakas et al. [13] consider similarity search for moving object trajectories in spatial networks. They use the distance in networks instead of the Euclidean distance between two points. Chen et al. [14] study a problem of searching trajectories by locations, in which the query is a set of locations.

3. Representation and Storage of Trajectory

This section first introduces a region-based representation for user's trajectories, and then proposes a storage structure for the region-based trajectories.

3.1 Dividing Region

For transfer continuous location information to discrete regions, the cluster method could be used for creating the regions from location information by merging the locations with a small distance in spatial [1]. In real world, the region usually dividing by manual or auto divided though the semantic function of an area, such as super market, restaurant, park and Stadium etc. For easy to explain, we use a grid-based region dividing methods in this paper, with which the whole area is divided to many grids with the same size, and a grid is seen as a region, see Fig. 2. This method was also exploited by [10]

3.2 Region-based Representation

We divide the whole activity area of moving object into disjoined small regions, and then we merge the continuous points in a region, formulated as (R_i, t_i, t_i), where R_i means the region, t_i means the time of the first point of the continuous points, and t_i means the interval between the last point and the first point.

Therefore, the user's trajectory can be compressed represented as:

$$C(o) = \{(R_0, t_0, t_0), (R_1, t_1, t_1), \mathbf{L} \\ , (R_m, t_m, t_m)\}, t_i + t_i < t_{i+1}, 0 \le i \le m \quad (1)$$

Further, (t_i, t_i) is denoted by T_i for short, that is $T_i = (t_i, t_i) = [t_i, t_i + t_i]$, and then (R_i, t_i, t_i) is denoted as (R_i, T_i), and we have $t_i = |T_i|$.

The diagram of compressed representation is shown as Fig 2. We don't restrict the way of dividing the activity area, but use the grid with the same size in this paper merely.

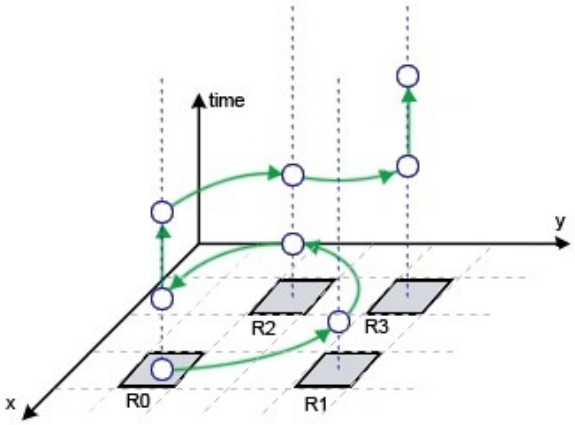


Fig. 2 compressed representation

3.3 Storage Structure

Storage structure is divided into two parts, shown in Fig 3. The first part is a users' trajectories table for storing the users' trajectories which store index of regions which are passed by the users. The second part stores a set of regions, in which indexing of trajectories and the intervals of users passing the region are stored. For example, as shown in the Fig 3, User3 passes three regions, R3, R5 and R7; and R7 is passed by three users, User1, User3 and User5, while User5 passes R7 at three interval [t1,t2], [t3,t4], [t5,t6].

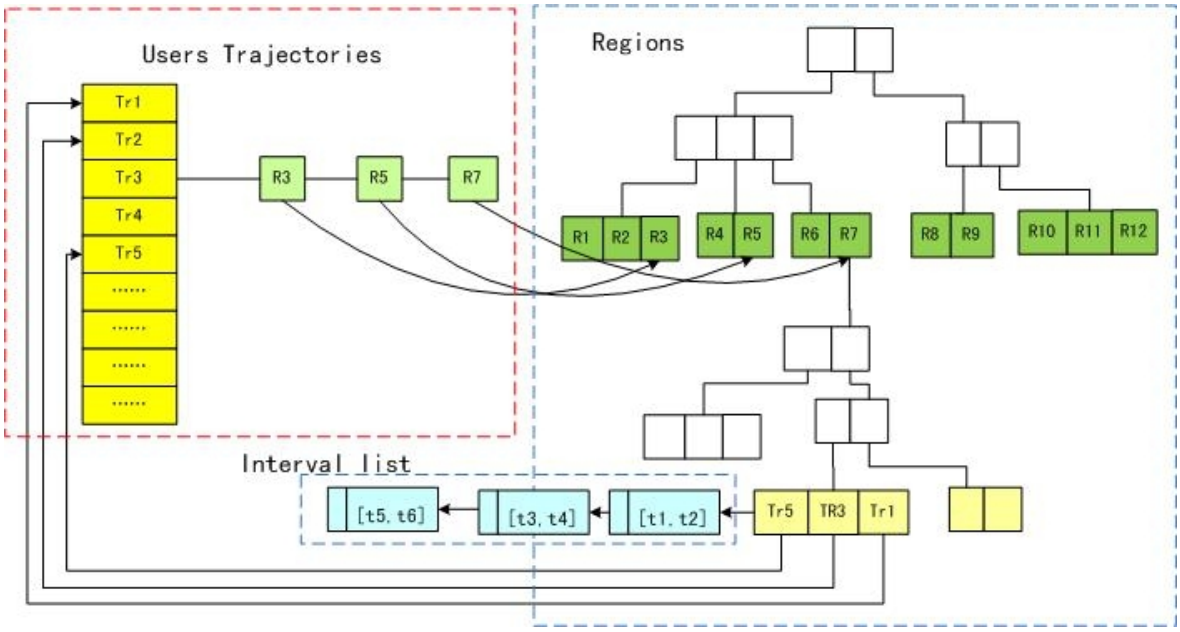


Fig. 3 The diagram for storage structure

4. Similarity Function and KNN Search

This section first gives three similarity functions between trajectories and then provides an algorithm for KNN search with the storage structure proposed in Section 3.

4.1 Similarity Function

Before defining the similarity function, we first introduce a variable to measure the user's activity scope in a region with several operations.

Definition 1 : The user's activity scope in the region R_i is defined as $V(R_i,t_i,t_i)$, and $V(R_i,t_i,t_i)=S(R_i)\times t_i$, where $S(R_i)$ represents the area of the region R_i .

Intersection operation:

$$V((R_i,t_i,t_i)\cap(R_j,t_j,t_j)) = \begin{cases} 0 & R_i \neq R_j \\ V(R_i,T_i\cap T_j)=S(R_i)\times\left|T_i\cap T_j\right| & R_i=R_j \end{cases} \quad (2)$$

Union operation:

$$V((R_i,t_i,t_i)\cup(R_j,t_j,t_j)) = \begin{cases} V(R_i,t_i,t_i)+V(R_j,t_j,t_j) & R_i \neq R_j \\ V(R_i,T_i\cup T_j)=S(R_i)\times\left|T_i\cup T_j\right| & R_i=R_j \end{cases} \quad (3)$$

Based on the above operations, we define three similarity functions: Absolute Similarity, Relative Similarity and Partial Similarity.

Definition 2: we define the Absolute Similarity as the absolute value of intersection of the user's activity scope. The formula is as follows.

$$aSim(o_i,o_j)=V(C(o_i)\cap C(o_j)) \quad (4)$$

Definition 3: we define the Relative Similarity as the ratio of the intersection and union of the user's activity scope. The formula is as follows:

$$rSim(o_i,o_j)=\frac{V(C(o_i)\cap C(o_j))}{V(C(o_i)\cup C(o_j))} \quad (5)$$

From the definition of Relative Similarity, we can easily obtain that the similarity between two trajectories is between 0 and 1.

Definition 4: we define the Partial Similarity as the ratio of the intersection and one user's activity scope. The formula is as follows:

$$pSim_{o_i}(o_j)=\frac{V(C(o_i)\cap C(o_j))}{V(C(o_i))} \quad (6)$$

Obviously, *Absolute Similarity* and *Relative Similarity* function is symmetrical, where *Partial Similarity* is not, and *Relative Similarity* can be seen as Normalized form of *Absolute Similarity*. In some applications, *Partial Similarity* is valuable. This paper focuses on *Absolute Similarity*.

4.2 Similarity Search Problems

K-nearest neighbors search problem can be described as follows:

Given a set of Users, S={O_1,O_2,……, O_n}, a query User Q, and a positive integer k, find the k Users in S to form a new set S', meeting for any User O_i in S' and any User O_j in S-S' , Sim(O_i,Q)>=Sim(O_j,Q).

4.3 KNN-search Algorithm

This subsection gives the k-nearest neighbor search algorithm, see Algorithm 1. Inputs include target user, the number of the nearest neighbor and user set with compressed trajectories. And the output is the k nearest users. Lines 1-2 are used to initialize memory space used for storing similarity values, Lines 4-10 are used to sum up the co-occurrence time of user at each region to calculate the similarity of users. The "Com_Time" in Line 6 is a function to calculate intersection of two interval lists.

```
Algorithm 1  MCT_KNN(Q,K,Os)
input:  Q—Target User
        K—The number of the nearest neighbor
        Os—user set with compressed trajectory
Output: the user set of K nearest neighbor
1.  Initialize V to store the trajectories similarity
2.  initialize Vk  to store the k most similar trajectories
3.  convert Q into compressed representation
4.  For each Ri in Q do
5.    For each Oi in Ri do
6.      V(Oi) + = Com_Time(Tqi,Toi)
7.  update Vk using V(Oi)
8.    End for
9.  End for
10. Return Vk
```

5. Experimental Evaluation

In this Section, we present experimental results to evaluate our techniques, Maximum Co-occurrence Time (MCT) by comparing it with well known method: LCSS, DTW and EDR. For experimental purposes, we used the synthetic datasets which are generated by Network generator—a moving object dataset generator developed by Brinkhoff [15]. This trajectory generator is also used by [10]. Our experiments were run on a PC AMD Athlon at 2.09GHz with 1.87G RAM and 160G hard disk.

5.1 Accuracy Evaluation

Accuracy is one of the most important aspects of similarity function. In this work, we use an objective evaluation method recently exploited by [5-7,10] to evaluate the accuracy of our techniques. The idea is to use a k-nearest neighbor classifier on labeled data to evaluate the efficacy of the distance measure used. We used the trajectories generator create 1000 trajectories with 500 time steps and 20 Categories. We continue to adjust the size of e and select the optimal value for LCSS and EDR, and set the parameter d for LCSS to be ∞ to get the optimal accuracy. For our method, we set the size of grid to be 2 times of e for LCSS. The ratio of train set is 20%. In order to avoid possible random, we repeated the experiments 20 times and took the average.

Fig. 4 shows the results. Our method-MCG has nearly classification accuracy with LCSS, while has higher accuracy than DTW and EDR.

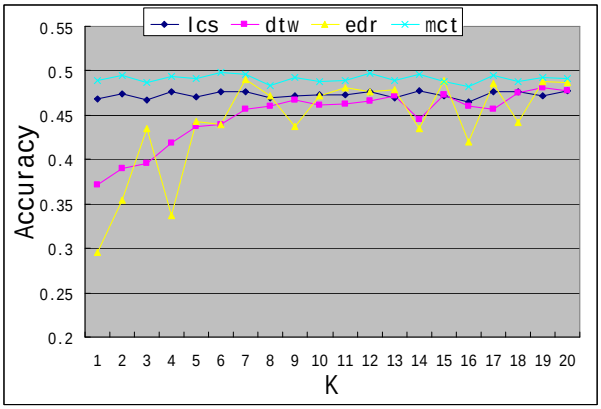


Fig. 4 Classify accuracy with different train ratio

5.2 The number of Trajectories in grid

By algorithm 3, the key of KNN search efficiency is the number of trajectories in grid. It is easy to know, the number is relative to the size of grid cells and dataset. In this section, we analyze the impact of the two factors.

5.2.1 Impact of grid size

In this subsection, we do experiments on Trucks dataset, changing the size of grid cell from 1 to 20, and recording the maximum and average of the number of trajectories in grid cells. Fig. 5(a) shows changes of maximum value. When the size is smaller than 9, maximum value grows as a linear almost, while when the size is greater than 9, the value is steady. Fig. 5(b) shows average value grows as a linear as size of grid is increase.

This is consistent to our conscious. By increasing the size of grid, two trajectories which were in different grids originally, may be in the same grid now, so the maximum and average value may increase. To reduce the computation, it should be reduce grid size, but too small grid cells may lead to greater errors and more Space consumption. Therefore, we should set the grid size dependent on the application and the dataset.

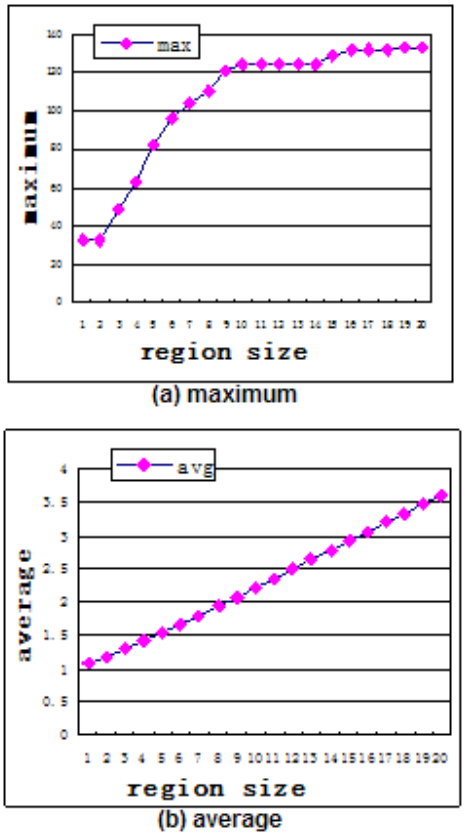


Fig. 5 Impact of grid size to the number of trajectories in one grid

5.2.2 Impact of dataset size

In this section, we use Network generator to create datasets with different size, concluding 1000, 2000, 3000, 4000, 6000, and 8000.

Fig. 6 shows that the maximum and average value is almost steady, when the size of datasets increases. The reason is that new trajectories increase the number of grid cells, thus maximum and average value change little. Further, the average number of trajectories in one region is much smaller than the total number of trajectories set. This property makes our methods significantly faster than other methods with matching trajectory directly. In a sense, we use a hash technology when we use an inverted file.

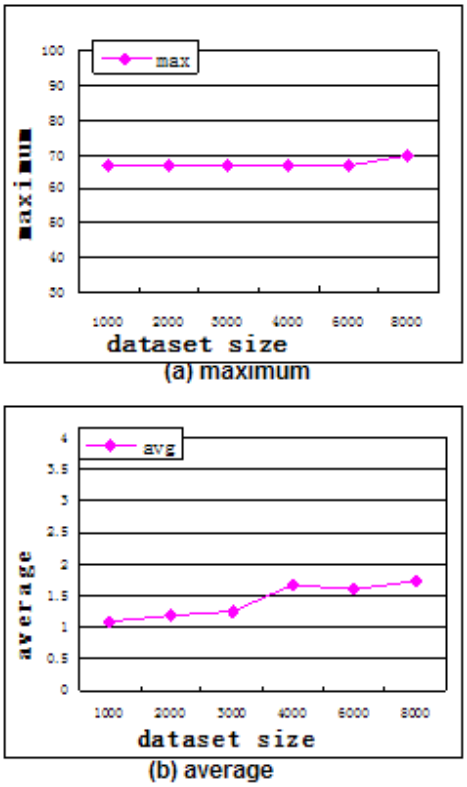


Fig. 6 Impact of dataset size to the number of trajectories in one grid

5.3 Efficiency Evaluation

Efficiency is the other important aspect. We use Network generator to create datasets with different size, concluding 1000, 2000, 3000, 4000, 6000, and 8000.

Fig. 7 presents the results. The time of DTW is similar to EDR, and is longer than LCS. Our method is much faster than other methods: LCS, DTW and EDR, even not in an order of magnitude. There are two factors to get the fast speed for our method. First, the number of trajectories in grid cells is much smaller than the number of trajectories; and second, we only compute the grid cells passed by the target users, which is also much smaller than the number of grids, instead of comparing them with all users.

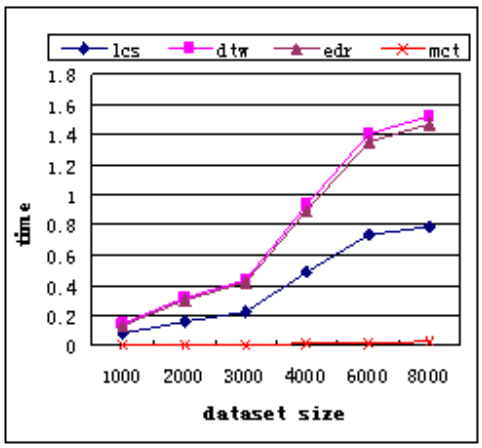


Fig. 7: Efficiency of KNN-search

6. Conclusion

In this paper, we present efficient techniques to compute the similarity between users by mining the historical trajectories. We first divide the space into small regions and compressed represent user's trajectories. Then, we presented a storage structure which not only stores the sequence of user's trajectory, but also stores regions with indexing of trajectories which pass the region. Based on compressed representation, we defined three similarity functions between users. At last, we give the k-nearest neighbor search algorithm and evaluation the accuracy and efficiency. Our experiments indicate that our similarity function has good accuracy comparing with LCSS, DTW and EDR and our algorithm is significantly faster than other algorithms.

Acknowledgments

Acknowledgments The research was supported by the major national science and technology special projects (2010ZX03004-003-03, 2010ZX03004-001-03), the National Basic Research Program of China (973 Program) (2009CB320906), the National Natural Science Foundation of China (60970160, 61070080, 61003184, 61172173, 61170023, 61231015).

References

[1] Zheng, Y., et al. , Recommending friends and locations based on individual location history. ACM Trans. Web, 2011,5(1), pp5-42.

[2] Agrawal, R., C. Faloutsos and A.N. Swami. Efficient Similarity Search In Sequence Databases. in FODO '93. 1993. London, UK: Springer-Verlag.

[3] Efficient Time Series Matching by Wavelets. in ICDE '99. 1999. Washington, DC, USA: IEEE Computer Society.
[4] Cai, Y. and R. Ng. Indexing spatio-temporal trajectories with Chebyshev polynomials. in SIGMOD '04. 2004. New York, NY, USA: ACM.
[5] Vlachos, M., D. Gunopoulos and G. Kollios, Discovering Similar Multidimensional Trajectories, in 18th International Conference on Data Engineering (ICDE'02). 2002: Los Alamitos, CA, USA. p. 0673.
[6] Berndt, D.J. and J. Clifford. Using Dynamic Time Warping to Find Patterns in Time Series. 1994.
[7] Chen, L., et al. Robust and fast similarity search for moving object trajectories. in Proceedings of the 2005 ACM SIGMOD international conference on Management of data. 2005. Baltimore, Maryland: ACM.
[8] Chen, L. and R. Ng. On the marriage of Lp-norms and edit distance. in VLDB '04. 2004: VLDB Endowment.
[9] Sakurai, Y., M. Yoshikawa and C. Faloutsos. FTW: fast similarity search under the time warping distance. in Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. 2005. Baltimore, Maryland: ACM.
[10] Lin, B. and J. Su, One Way Distance: For Shape Based Similarity Search of Moving Object Trajectories. GeoInformatica, 2008. 12(2): p. 117-142.
[11] Frentzos, E., K. Gratsias and Y. Theodoridis. Index-based Most Similar Trajectory Search. in Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. 2007.
[12] Pelekis, N., et al. Similarity Search in Trajectory Databases. in Proceedings of the 14th International Symposium on Temporal Representation and Reasoning. 2007: IEEE Computer Society.
[13] Tiakas, E., et al., Searching for similar trajectories in spatial networks. 2009. 82(5): p. 772-788.
[14] Chen, Z., et al. Searching trajectories by locations: an efficiency study. in SIGMOD '10. 2010. New York, NY, USA: ACM.
[15] Brinkhoff, T. Generating Traffic Data, IEEE Data Eng. Bull., 2003, 26(2), pp. 19-25.

Yimin Wang received the B.S. degree in computer school of Wuhan University, Wuhan, China, in 2008. He is currently working as a doctor at National Engineering Research Center for Multimedia Software, Wuhan, China. His research interests include public safety, multimedia content analysis, machine learning, and Data Mining.

Ruimin Hu received the B.S and M.S degrees from Nanjing University of Posts and Telecommunications, Nanjing, China, in 1984 and in 1990, and Ph.D degree in Communication and Electronic System from Huazhong University of Science and Technology, Wuhan, China, in 1994. Dr. Hu is the director of National Engineering Research Center for Multimedia Software, Wuhan University and Key Laboratory of Multimedia Network Communication Engineering in Hubei province. He is Executive Chairman of the Audio Video coding Standard (AVS) workgroup of China in Audio Section. He has published two books and over 100 scientific papers. His research interests include audio and video coding and decoding, video surveillance and multimedia data processing.

Wenhua Huang received the B.S. degree in computer school of Wuhan University, Wuhan, China, in 2010. He is currently working as a Graduate at National Engineering Research Center for Multimedia Software, Wuhan, China. His research interests include public safety and multimedia content analysis.

Jun Chen received the Ph.D degree in computer school of Wuhan University, Wuhan, China. He is currently a professor at National Engineering Research Center for Multimedia Software, Wuhan, China. His research interests include public safety and multimedia content analysis

Network Security Using Job Oriented Architecture (SUJOA)

Tariq Ahamad[1], Abdullah Aljumah[2]

College Of Computer Engineering & Sciences
Salman Bin Abdulaziz University, KSA

ABSTRACT

In the modern world operating system, various security systems (Collection of various security components) are already installed or can be directly installed in it. They are responsible for checking the complete node for suspicious behaviour. There are some intrusions having the ability to hide themselves from being checked called armouring intrusions. In this research article we present alternative organisation of security systems. To distinguish operating system applications and security systems, the node is completely virtualized with current virtualized systems. The node is then checked by security systems from outside and the right security components are provided through job oriented architecture. Since they run on a virtual machine the infected nodes can be halted, duplicated and moved to other nodes for further analysis and legal aspects. The coordinated architecture analysed in this research article and the results of a preliminary implementation with positive results are discussed.

Key Words: ***Network Security, Network Security Architecture.***

INTRODUCTION

A lot of network security systems protect a computer network against various types of network or electronical attacks. The attacks are e.g. worms and viruses but also hacker or internal attacks performed by the normal users of the network. The security systems are a collection of various security components as antivirus software, firewalls, and intrusion detection systems [5] and in the modern world operating systems these security components are directly installed in and monitor the nodes for suspicious behaviour. The components lack from cooperative workflows in order to correlates events for abnormal behaviour detection [2, even redundant checks lead to an increase in the required resources. This organisation of security components is a major weakness especially in coping with upcoming intrusions [6]. E.g. the Bradley virus is computational hard to detect when such an organisation is used [1]. We concern with the organisation of security components and introduce a more sophisticated way. The whole node is virtualized using an virtualization system as VMware. The operating and security systems run in a node in different virtual machines so that the security components check the node from outside. The features of current virtualization systems are used with the ability to halt, duplicate, and move virtual machines. This leads to a different handling of infections because the infected virtual machine is duplicated and saved for further

analysis and legal aspects. Job oriented architecture (JOA) provides on demand the right security components, which is implemented through the exchange or adding of virtual machines. We concern the advantages and disadvantages of such an environment and conclude with the current project status.

ARCHITECTURE

Architecture for each node is organised in four layers: the first layer is the hardware and the second layer is a core operating system providing the kernel to access the hardware and the virtualization system to run several virtual machines simultaneously[3][5]. The third layer contains the virtual machines with different operating systems and one virtual machine for security. This implements the security environment, which ensures the access to all other virtual machines for scanning purposes. The fourth layer is the application layer with user's applications and installed security components, which are connected to the security environment[4]. In the network, one or more security servers exist, which provide the right virtual machine with installed security environment and security components. This server provides the virtual machine to new nodes and ensures that each node is properly secured. Furthermore, it contains more analysis systems to scan virtual machines deeply. With this implementation, the security components are seen as different services provided by the security server and distributed to the node where they are required. This is expanded to a service oriented architecture where the right security components are provided on demand according to the current situation in the network [6]. Especially with the novel more and more different node types as mobile handhelds connected to the network and thin clients, the required security components in a node has to be adapted due to different available resources and security issues. With the introduced architecture, the security components can be easily adapted to the security requirements of the node. The maintenance of security systems changes according to the architecture. If a new node connects [4][6] to the network, the security server checks the node if the right architecture is installed and provides an virtual machine with the security environment and installed security components. Through integrating this workflow in the DHCP, a network is properly secured because only when the node has a running security system it receives access to the network. New security components or changes in the required components in a node are quickly resolved: the security server provides a new virtual machine that exchanges the current machines on the node [14].

IMPLEMENTATION ISSUES

The implementation of the architecture is feasible. Current virtualization systems as VMware or KVM provide most of the required features. VMware also provides main boards where the layer one and two is directly installed. The only missing feature is the ability that the security components of the security environment are able to access the other virtual machines for scanning purposes. However, this can be implemented through an extension of the virtualization system [7][10]. With the right implementation, well known security components - e.g. antivirus software, firewall, and intrusion detection system - are facilitated. These are installed in the security environment and a guard measures the required data from the operating

system and presents it to the security component. This analyses the data and defines the response, which is executed by the guard accordingly. Consequently, all existing security components are reusable [17].This architecture provides an improvement in the implementation of security components. These are platform independent running in the security environment and must consequently not adapted when the used operating system or used hardware platform changes. In addition, the security environment gathers the data to analyse and perform the response accordingly, which must not be implemented in the security component. This leads to a faster deployment of novel approaches.

The project status is that the architecture of the nodes and of the network is designed and theoretically analysed. Different proof-of-concept implementations of the virtualization system have been realised to analyse the features of these. Various parts of the implementation are still missing due to the early stage of the project. The preliminary results are discussed in the next section [8].

PRELIMINARY RESULTS

The first results are promising. The implementation of the node is feasible where the only hard task is to ensure that the security components are able to access the operating system. The installation of the security components in the security environment installed in a virtual machine is also challenging because this influences the performance and the security of the security system[9]. Security issues of this organisation are analysed and they are solvable when up-to-date approaches from cryptography are used. Especially the more and more emerging integrity checks and the complete installation of a distributed public/private key infrastructure are important. With this, an adversarial is not able to use the distributed security system for attacking the network [11].

A. Infection Handling

One main advantage of the proposed architecture is the infection handling. If some security component identifies a virtual machine of a node as infected, the following workflow is processed: the security environment of the node halts the infected virtual machine to prevent propagation [12]. It duplicates the virtual machine and sends this to the security server to analyse it more deeply and to save the evidences for legal aspects - this is a weakness in current systems: either the node is disinfected and the evidences are destroyed or the evidences are saved but the node is still infected. Afterwards, the security server provides a clean virtual machine with the desired applications installed in order to limit the downtime of the node [17].

B. Security Issues

The proposed architecture is used to implement a distributed security system with integrated components. The security components are able to roam through the network and cooperative workflows enable the detection of unforeseen intrusions. This has a major drawback that adversaries may use the system to attack a network, i.e. to propagate intrusions through it. This must be prevented through the design of the architecture and is discussed now [13]. The layer one containing the hardware is not an aim of adversaries. Layer two with the core operating and virtualization system is furthermore highly dependent on the facilitated hardware and changes therefore only

when the hardware changes. This is ensured through public/private key signatures used in cryptography (this is also integrity checking called). When an adversarial installs an intrusion in this layer, it is immediately recognised and prevented [14]. Layer three and four contains the operating system with application of the user, which are protected as the operating system in nowadays implementations. The virtual machine containing the security environment and the security components is additionally secured: the security environment does not change and is therefore protected using a cryptographic signature with integrity checking [15].The security components access resources of the node. These are secured using public/private keys organised in a distributed public key infrastructure. Only when the security components have the right keys, they receive access to the resources where security components initiated by the adversarial does not have these. To summarise the security issues, the adversarial is still able to install intrusions in the operating system and in the security system when the implementation provides bugs or when the adversarial knows internal knowledge.

CONCLUSION

In this research article we discussed how to increase the performance of a network security system by features of virtualized system with job oriented architecture. The advantages especially help to identify novel more and more intelligent intrusions and provides a more sophisticated infection handling. Furthermore, the article faces several unsolved problems, which are of interested for novel network security systems. The next step in the project is to implement a prototype of the architecture and to build up a testbed with some nodes to simulate the whole workflow. For this, a VMware implementation is first desired due to the reduced time for setting up all of these nodes. The architecture is also usable in current approaches of network security systems facilitating artificial immune systems , multi-agent systems, and distributed systems to distinguish the normal operating system and the security system on each node .

References

[1] A. Aho, R. Sethi, J. Ullman, “Compilers, Principles, Techniques, and Tools”. Addison-wesley Publishing Company.

[2] G. Ammons, J. Larus. “Improving Data-flow Analysis with Path Profiles”. In the Proceeding of the 1998 ACM SIGPLAN Conference on Programming Language Design and Implementation, Montreal Canada, June 17-19, 1998.

[3] T. Ball, J. Larus. “Efficient Path Profiling”, In the proceeding of MICRO-29, December 2-4, 1996, Paris, France.

[4] T. Ball, J. R. Larus. “Optimally Profiling and Tracing Programs”. ACM Transactions on Programming Languages and Systems, Vol 16, No. 4, July 1994, pp1319-1360.

[5] S. Berkovits, J. Guttman, V. Swarup. “Authentication for Mobile Agents”, in: Giovanni Vigna (Ed.): Mobile Agents and Security. pp 114-136. Springer-Verlag, 1998.

[6] C. Cifuentes. “Structuring Decompiled Graphs. Personal Communication. Proceedings of the

International Conference on Compiler Construction (CC'96), Lecture Notes in Computer Science 1060. Linkoping, Sweden. 22-26 April 1996, pp 91-105.

[7] C Cifuentes, M Van Emmerik, and N. Ramsey, *The Design of a Resourceable and Retargetable Binary Translator*. Proceedings of the Sixth Working Conference on Reverse Engineering, Atlanta, USA, October 1999, IEEE-CS Press, pp 280-291.

[8] Cloakware Systems. "Building Cloakware". Five minute presentation at the 2000 IEEE Symposium of Security and Privacy. May, 2000. Berkeley, California.

[9] D. Chess. "Security issues in mobile code systems". in: Giovanni Vigna (Ed.): Mobile Agents and Security. pp 1-14. Springer-Verlag, 1998.

[10] S. Cheung, R. Crawford, M. Dilger, J. Frank, J. Hoagland, K. Levitt, S. Staniford-Chen, R. Yip, D. Zerkle. GrIDS: "*A Graph-Based Intrusion Detection System"*. National Information System and Security Conference, Baltimore, 1997.

[11] T. Cormen, C. E. Leiserson, R. Rivest, "Introduction to Algorithms". The MIT Press, 1993. Tenth edition.

[12] M. Elder, J. Knight, "Security Attacks on Critical Infrastructure Systems". Computer Science Technical Report. CS-98-23.

[13] M. Hennessey, J. Riely, "Type Safe Execution of Mobile Agents in Anonymous Networks", in: Jan Vitek; Christian Jensen (Eds.): Secure Internet Programming, LNCS 1603, Springer-Verlag, pp. 95-116, 1999.

[14] F. Hohl. "Time Limited Blackbox Security: Protecting Mobile Agents from Malicious Hosts". In Lecture Notes in Computer Science, vol. 1419, Mobile Agents and Security. Edited by G. Vigna. Springer-Verlag, 1998.

[15] J. Knight, K. Sullivan, M. Elder, C. Wang. "Survivability Architectures: Issues and Approaches" In Proceedings: DARPA Information Survivability Conference and Exposition. IEEE Computer Society Press. Los Alamitos, CA, January 2000, pp. 157-171.

[16] B. Yee. "A Sanctuary for Mobile Agents". Technical Report CS97-537. Computer Science Department, University of California in San Diego, USA.

[17] http://gita.state.az.us/enterprise_architecture/NEW/Security_Arch/

On the local controllability of a discrete-time inhomogeneous multi-input bilinear systems

Omar Balatif , Mohamed El hia, Jamal Bouyaghroumni, Mostafa Rachik

**Laboratory of Analysis Modeling and Simulation,
Department of Mathematics and Computer Science,
Faculty of Sciences Ben M.Sik, University Hassan II
Mohammedia, BP 7955, Sidi Othman, Casablanca, Morocco**

Abstract

This paper studies the local controllability of a class of discrete-time inhomogeneous bilinear systems. A sufficient condition for the local controllability is proposed and the form of the optimal control is also presented. Furthermore, the established results are illustrated by an example and numerical simulation.

Keywords: Bilinear systems, discrete time, local controllability, optimal control.

1. Introduction

Bilinear systems are a special class of nonlinear systems; they form a transitional class between the linear and the general nonlinear systems. Through nearly half a century, they have received great attention by researchers. The importance of such systems lies in the fact that many important processes, not only in engineering [1], but also in biology [2], socio-economics [3], and chemistry [4-5], can be modeled by bilinear systems[6].

In the literature, several papers address the problem of controllability for bilinear systems. In [7], we raise two conditions for controllability: one for necessity and the other for sufficiency. Such approach is a local one and consists in decomposing the bilinear system model into a linear system and a multiplicative feedback. However, it requires that rank(Q) = 1 where Q must be factorized in two vectors; in other terms this technic needs orthogonality property. The same problem as considered in [8] gives rise to a global necessary and sufficient condition. In addition to decomposing the system as in [7], the approach involves forward and backward composition of the transition function. It still ensues a condition of orthogonality on the matrix Q, plus an inversibility condition on the matrix A. etc

The present paper deals with the question of local controllability for discrete time inhomogeneous multi-input bilinear systems. We adopt a method based on the linearization of the system and the definition of an appropriate operator that leads to the control transferring the system to a desired given state with a minimum energy.

The paper is organized as follows. In section 2, we present an approximation of the final state. Section 3 is aimed to the presentation of a sufficient condition for local controllability of an inhomogeneous multi-input bilinear discrete-time system. Section 4 provides an expression of an optimal control that can transfer the system from the initial state to a final desired state. Finally, an example of controllable bilinear systems is provided in section 5.

2. An approximation of the final state

In this article we consider the following inhomogeneous discrete-time bilinear system:

$$x(k+1) = Ax(k) + \sum_{i=1}^{p} u_i(k) B_i x(k) + Bu(k) \qquad (1)$$

where $x(k)$ is the n-dimensioned state vector at time k, $u(k) = (u_i(k))$ is the p-dimensioned control vector at time k, B is a matrix of dimension $n \times p$, A and $B_1; \ldots; B_p$ are square matrices of order n.

Let x_N denotes the final state and

$$x(k+1) = Ax(k) + \sum_{i=1}^{p} u_i(k) B_i x(k) + Bu(k) = F(x(k), u(k))$$

where F is a continuous vector function.

Let also $B = (b_{ij})$ with $b_{ij} \in \mathbb{R}$ for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$, then the system (1) becomes

$$x(k+1) = Ax(k) + \sum_{i=1}^{p} u_i(k) B_i x(k) + V_{u_i(k)} = F(x(k), u(k))$$

where $V_{u_i(k)} = \left(\sum_{i=1}^{p} b_{1i} u_i(k) \ \cdots \ \sum_{i=1}^{p} b_{ni} u_i(k) \right)^T$

Consider the following function composition

$$x(N) = F_{u(N-1)} \circ \ldots \circ F_{u(1)} \circ F_{u(0)}(x(0)) \qquad (2)$$

with $F_{u(k)}(x(k)) = F(x(k), u(k))$

Using Taylor's development to expand the right-hand side of the previous equation yields:

$$x(N) = F_{u(N-1)} \circ \ldots \circ F_{u(1)} \circ F_{u(0)}(x(0))\big|_{\underline{u}=0} + \left[\frac{\partial F_{u(N-1)}(x(N-1))}{\partial u(N-1)} \; \frac{\partial F_{u(N-1)}(x(N-1))}{\partial u(N-2)} \; \ldots \frac{\partial F_{u(N-1)}(x(N-1))}{\partial u(0)}\right]_{\underline{u}=0} = 0 + O\left(u^2\right)$$

with $\underline{u} = \left(u(N-1) \ldots u(1) \;\; u(0)\right)^T$, which can be rewritten as:

$$\bar{x}(N) = F_{u(N-1)} \circ \ldots \circ F_{u(1)} \circ F_{u(0)}(x(0))\big|_{\underline{u}=0} + P\big|_{\underline{u}=0} \; \underline{u} + O\left(u^2\right)$$

where

$$P = \left[P_{N-1} \; P_{N-2} \; \ldots P_0\right] = \left[\frac{\partial F_{u(N-1)}(x(N-1))}{\partial u(N-1)} \; \frac{\partial F_{u(N-1)}(x(N-1))}{\partial u(N-2)} \; \ldots \frac{\partial F_{u(N-1)}(x(N-1))}{\partial u(0)}\right]$$

From the equation (1) we have

$$\frac{\partial F_{u(k)}(x(k))}{\partial x(k)} = A + \sum_{i=1}^{p} u_i(k) B_i \text{ and}$$

$$\frac{\partial F_{u(k)}(x(k))}{\partial u_i(k)} = B_i x(k) + V_i \text{ with } V_i = \frac{\partial V_{u_i(k)}}{\partial u_i(k)} = \begin{pmatrix} b_{1i} \\ \vdots \\ b_{ni} \end{pmatrix}$$

for $i = 1, \ldots, p$

So by computing P after function composition, when controls are assumed to be equal to zero, we obtain

$$P = \begin{bmatrix} \left(B_1 A^{N-1} x(0) + V_1 \ldots B_p A^{N-1} x(0) + V_p\right) \\ \left(A\left(B_1 A^{N-1} x(0) + V_1\right) \ldots A\left(B_p A^{N-2} x(0) + V_p\right)\right) \\ \vdots \\ \left(A^{N-1}\left(B_1 A^{N-1} x(0) + V_1\right) \ldots A^{N-1}\left(B_p x(0) + V_p\right)\right) \end{bmatrix}^T$$

In other words, an approximation of the final state x_N when neglecting higher order control terms can be expressed as:

$$\bar{x}(N) = A^N x(0) + \sum_{k=0}^{N-1} \sum_{i=1}^{p} A^{N-1-k} (B_i A^k x(0) + V_i) u_i(k) \quad (3)$$

Note that, in the rest of this work, we neglect higher order control terms. This assumption gives a local criterion of controllability.

3. A sufficient condition of local controllability

In this section we propose a sufficient condition of local controllability for the system (1).

First recall the definition of the local controllability for the systems (1).

Definition 1

The system (1) is said to be locally controllable on $I = \{0, 1, \ldots, N-1\}$ *for any* x_0 *and* x_d *from* $\mathbb{R}^n$*; there exists a control* $u = (u_0, u_1, \ldots, u_{N-1})$ *as* $\bar{x}_N = x_d$*; where* $\bar{x}_N$*; given by (3), is the approximate solution of (1) at instant N corresponding to the initial state* x_0 *and the control u.*

Then, let consider the operator defined by

$$H : \left(\mathbb{R}^p\right)^N \to \mathbb{R}^n$$

$$u = (u(0), \ldots, u(N-1))^T \to Hu = \sum_{k=0}^{N-1} \sum_{i=1}^{p} A^{N-1-k} (B_i A^k x(0) + V_i) u_i(k) \quad (4)$$

with $u(k) = \left(u_1(k) \ldots u_p(k)\right)^T \quad \forall u_i(k) \in \mathbb{R}$

Proposition 2

The operator H is linear, continuous and its adjoint operator H^* *is given by:*

$$H : \mathbb{R}^n \to \left(\mathbb{R}^p\right)^N$$
$$x \to H^* x = P^T x$$

Proof.

- The linearity of H is obvious
- For the continuity of H we show the existence of a constant $\alpha > 0$ such as

$$\|Hu\| \le \alpha \|u\| \quad \forall u \in L^2\left(0, N-1, \mathbb{R}^p\right)$$

$$\|Hu\| = \left\| \sum_{k=0}^{N-1} \sum_{i=1}^{p} A^{N-1-k} (B_i A^k x(0) + V_i) u_i(k) \right\|$$

$$\le \sum_{k=0}^{N-1} \left(\sum_{i=1}^{p} \left\| A^{N-1-k} (B_i A^k x(0) + V_i) \right\|^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^{p} \|u_i(k)\|^2 \right)^{\frac{1}{2}}$$

$$\le \sum_{k=0}^{N-1} \|T(k)\| \|u(k)\|$$

$$\le \left(\sum_{k=0}^{N-1} \|T(k)\|^2 \right)^{\frac{1}{2}} \left(\sum_{k=0}^{N-1} \|u(k)\|^2 \right)^{\frac{1}{2}}$$

$$\le \alpha \|u\|$$

So H is linear

- For the adjoint operator we have

$$\langle Hu, x \rangle = \langle Pu, x \rangle = \langle u, P^T x \rangle = \langle u, x \rangle$$

Hence the expression of H^*.

Proposition 3
If H is surjective then (1) is locally controllable.

Proof.
Let $x_0 ; x_d \in \mathbb{R}^n$. Thus $x_d - A^N x_0 \in \mathbb{R}^n$
As H is surjective, there exists a control $u \in L^2\left(0, N-1, \left(\mathbb{R}^p\right)^N\right)$ such as $Hu = x_d - A^N x_0$, so $x_d = A^N x_0 + Hu$, then $x_u^0(N) = x_d$
Hence the result according to the definition 1.

Proposition 4
If rank [P] = n; then the system (1) is locally controllable.

Proof.
Let H be the operator defined by (4) and H^* is the adjoint operator.
We know that $\operatorname{Im} H = \mathbb{R}^n \Leftrightarrow \ker H^* = \{0\}$
If $rank\left[P\right] = rank\left[P_{N-1}\ P_{N-2}\ \cdots P_0\right] = n$
Then $\ker \begin{bmatrix} P_{N-1}^T \\ P_{N-2}^T \\ \vdots \\ P_0^T \end{bmatrix} = \{0\}$
Hence $\ker H^* = \{0\}$
So if *rank [P] = n* then H is surjective and according to the proposition (3), the system (1) is locally controllable.

4. Optimal control

In this section we focus on the characterization of optimal control for the case of system (1).
Let introduce the matrix W defined by

$$W = \sum_{k=0}^{N-1} P_k P_k^T \qquad (5)$$

We have the following result

Theorem 5
The system (1) is locally controllable if the matrix P has full rank. Furthermore, the control $u^*(\cdot)$ which can transfer the system from the initial state x_0 to the final state x_d with a minimum energy, is given by

$$\begin{cases} u^*(k) = -P_k^T W^{-1}\left(A^N x(0) - x_d\right) \\ k \in \{0,1,\ldots,N-1\} \end{cases} \qquad (6)$$

Before we prove this theorem, we first prove the following lemma:

Lemma 6
The matrix P has full rank if and only if the matrix W is positive definite.

Proof.
$\Rightarrow$| We have

$$\langle Wx, x\rangle = \left\langle \sum_{k=0}^{N-1} P_k P_k^T x, x \right\rangle = \sum_{k=0}^{N-1} \left\langle P_k^T x, P_k^T x \right\rangle = \left\| P_k^T x \right\|^2 \geq 0$$

If $\langle Wx, x\rangle = 0$ then $P_k^T x = 0$, $\forall k \in \{0,1,\ldots,N\text{-}1\}$

Then $x \in \ker \begin{bmatrix} P_{N-1}^T \\ P_{N-2}^T \\ \vdots \\ P_0^T \end{bmatrix} = \{0\}$ because *rank [P] = n*

Hence *x=0* and therefore *W* is positive definite.

$\Leftarrow$| Let $x \in \ker \begin{bmatrix} P_{N-1}^T \\ P_{N-2}^T \\ \vdots \\ P_0^T \end{bmatrix}$

so $P_{N-1}^T x = P_{N-2}^T x = \cdots = P_0^T x = 0$
then $P_k P_k^T x = 0$, $\forall k \in \{0,1,\ldots,N\text{-}1\}$ and $Wx = 0$
hence *x=0* (because *W* is positive definite)
thus $\ker \begin{bmatrix} P_{N-1}^T \\ P_{N-2}^T \\ \vdots \\ P_0^T \end{bmatrix} = \{0\}$
and finally we get *rank [P] = n*

Proof (of theorem 5).
• First, suppose the matrix P has full rank then, according to the proposition 4, the system (1) is locally controllable. Furthermore, using the previous lemma, the matrix W is positive definite which implies it inversibility. Consequently u^* defined by (6) is well defined.

• Then by replacing u^* in (3) by the expression (6) one can easily check that $\bar{x}(N) = x_d$.

• Finally we show that $\|u^*\| = \inf U$, with
U={ $\|v\|/v$ is a control that allows the transfer of the system from x_0 to x_d }.

Let $u \in U$, then

$$A^N x(0) + \sum_{k=0}^{N-1} \sum_{i=1}^{p} A^{N-1-k} \left(B_i A^k x(0) + V_i \right) u_i(k) = x_d$$

then

$$\sum_{k=0}^{N-1} \sum_{i=1}^{p} A^{N-1-k} \left(B_i A^k x(0) + V_i \right) \left(u_i(k) - u_i^*(k) \right) = 0$$

$$\Rightarrow \left\langle \sum_{k=0}^{N-1} \sum_{i=1}^{p} A^{N-1-k} \left(B_i A^k x(0) + V_i \right) \left(u_i(k) - u_i^*(k) \right); -W^{-1} \left(A^N x(0) - x_d \right) \right\rangle = 0$$

$$\Rightarrow \sum_{k=0}^{N-1} \left\langle P_k \left(u(k) - u^*(k) \right); -W^{-1} \left(A^N x(0) - x_d \right) \right\rangle = 0$$

$$\Rightarrow \sum_{k=0}^{N-1} \left\langle \left(u(k) - u^*(k) \right); -P_k^T W^{-1} \left(A^N x(0) - x_d \right) \right\rangle = 0$$

$$\Rightarrow \sum_{k=0}^{N-1} \left\langle \left(u(k) - u^*(k) \right); u^*(k) \right\rangle = 0$$

$$\Rightarrow \left\langle u - u^*; u^* \right\rangle = 0$$

$$\Rightarrow \left\langle u; u^* \right\rangle - \left\| u^* \right\|^2 = 0$$

$$\Rightarrow \left\| u^* \right\|^2 = \left\langle u; u^* \right\rangle$$

$$\Rightarrow \left\| u^* \right\|^2 \le \|u\| \left\| u^* \right\|$$

$$\Rightarrow \left\| u^* \right\| \le \|u\|$$

Hence the result.

5. Example

Consider the dynamical system (Mohler, 1973)

$$\dot{x} = Ax + u_1 B_1 + Bu \qquad (7)$$

Where

$$A = \begin{bmatrix} \frac{-R_a}{L_a} & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -\frac{D}{J} \end{bmatrix} ; \quad B_1 = \begin{bmatrix} 0 & 0 & \frac{-K_a}{L_a} \\ 0 & 0 & 0 \\ \frac{K_y}{J} & 0 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} 0 & \frac{1}{L_a} \\ 0 & 0 \\ 0 & 0 \end{bmatrix}; \; x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} i_a \\ \theta \\ \omega \end{bmatrix} \text{ and } u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} i_e \\ \upsilon_a \end{bmatrix}$$

J is the moment of inertia, D is the viscous damping ratio, R_a is the armature resistance, L_a is the applied armature inductance, K_y , K_a are motor characteristics, K_a is the motor const, i_a is the armature current, i_e is the field current, υ_a is the armature voltage, ω is the angular velocity, and θ is the angular position.

Equation (7) can be discretized by use of a first-order Euler expansion to give

$$x(k+1) = x(k) + TAx(k) + u_1(k) TB_1 x(k) + TBu(k) \qquad (8)$$

where T is the sampling interval. Equation (8) can be rewritten as

$$x(k+1) = A^* x(k) + u_1(k) B_1^* x(k) + B^* u(k) \qquad (9)$$

With $A^* = I + TA$, $B_1^* = TB_1$ and $B^* = TB$

The parameter values chosen for the model are taken from [9] and are T = 0.1, K_a =0.156, K_y =37.7, L_a =0.05, $J = 2.4 \times 10^{-4}$, D = 0.0032 and.
Then the system (9) becomes

$$x(k+1) = \begin{bmatrix} 0.880 & 0 & 0 \\ 0 & 1 & 0.1 \\ 0 & 0 & -0.334 \end{bmatrix} x(k) + u_1(k) \begin{bmatrix} 0 & 0 & -75.4 \\ 0 & 0 & 0 \\ 15708.334 & 0 & 0 \end{bmatrix} x(k) + \begin{bmatrix} 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} u(k) \qquad (10)$$

Consider $x_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ the initial state and $x_d = \begin{bmatrix} 10 \\ 10 \\ 10 \end{bmatrix}$ the desired state.

We present numerical results obtained using Matlab.
For N = 20; we have *rank[P]* = 3, so the system (10) is locally controllable and the optimal control is given by the following table.

Table 1: optimal control

k	$u_1(k)$	$u_2(k)$
0	-0.864	0.028
1	0.554	0.032
2	-0.011	0.036
3	0.179	0.041
4	0.086	0.047
5	0.103	0.053
6	0.080	0.060
7	0.074	0.068
8	0.064	0.078
9	0.057	0.088
10	0.050	0.100

11	0.044	0.114
12	0.039	0.130
13	0.034	0.147
14	0.030	0.168
15	0.026	0.190
16	0.023	0.216
17	0.020	0.246
18	0.019	0.280
19	0.012	0.318

6. Conclusion

In this paper, we have studied the local controllability of a bilinear discrete-time system. The method that we present in this paper is based on a linearization of the system and then the definition of a suitable operator which can lead to control transferring the system to a desired given state with a minimum energy.

Acknowledgments

This work was supported by: "Le Réseau de la Théorie des Systèmes".

References

[1] R.R. Mohler. Bilinear Control Processes, volume 106 of Mathematics in Science and Engineering. Academic Press, New York, 1973.
[2] D. Williamson. Observation of bilinear systems with application to biological control. Automatica, 13 :243 254, 1977.
[3] R.R. Mohler. Nonlinear systems : Applications to Bilinear Control, volume 2. Prentice Hall, Englewood Clis, New Jersey, 1991.
[4] M. España and I.D. Landau. Reduced order bilinear models for distillations columns. Automatica, 14 :345 355, 1977.
[5] M. V. Basin and A. Alcorta-Garcia, Optimal filtering for bilinear system states and its application to polymerization process identification, in Proceedings of the American Control Conference, pp. 1982.1987, Denver, Colo, USA, June 2003.
[6] M. Ekman : Modeling and Control of Bilinear Systems: Applications to the Activated Sludge Process. Written in English. ACTA UNIVERSITATIS UPSALIENSIS. Uppsala Dissertations from the Faculty of Science and Technology 65. 231 pp. Uppsala, Sweden, 2005. ISBN 91-554-6342-8.
[7] T. Goka, T.J. Tarn and J. Zaborszky. On the controllability of a class of discrete bilinear systems. Automatica, vol 9, 1973.
[8] M.E. Evans and D.N.P. Murthy. Controllability of a class of discrete time bilinear systems. IEEE Trans on Automatic Control, AC-22, 78-83, February 1977.
[9] B. Gerard. Observers and control based on an observer for bilinear systems. Doctoral thesis of Henri Poincare university -Nancy1, November 2008.

Two-terminal Fault Location Method Based on the Lines Converted Midpoint and HHT

Yutian Wang, Huixin Wang, Shuqing Zhang, Hanlu Shangguan

**Institute of Electrical Engineering, Yanshan University,
Qinhuangdao, Hebei Province, 066004, China**

Abstract

Aiming at the problems of travelling wave's speed velocity discontinuity problems in the hybrid transmission line composed by cables and overhead lines, a new method of two-terminal fault location based on the converted midpoint of the transmission line and HHT is presented in this paper. First, the hybrid transmission line was reduced to a single parameter line to get the midpoint of the line. Then, the HHT (Hilbert- Huang Transform) was used to detect the travelling waves' heads. The search direction of the fault was calculated according to the time difference Δt between two measurement endpoints from travelling wave of the fault point. When travelling waves moved $\Delta t/2$ from the converted midpoint along the search direction, the point was the fault point. The simulation results by ATP and Matlab show that this method is correct and accurate.

Keywords: *Error fault location, Travelling wave, two-terminal location, converted midpoint, HHT*

1. Introduction

Electricity distribution network has many branches and complex structure, and its fault location has been the difficult problem studied. With the rapid development of power systems, cable-overhead line hybrid transmission lines have applied to power cables widely, increasing the difficulty of fault location.

The theoretical study of fault location method based on the principle includes intelligent ranging methods, fault analysis methods and travelling wave methods [1]. Fault analysis methods are mainly represented by the impedance method; intelligent Ranging methods include Kalman filtering, pattern recognition, probability and statistics decision-making, fuzzy theory and optical ranging, intelligent simulated annealing algorithm ranging method. All these are in the research stage currently, having not applied to practice. Travelling wave method is divided into single-terminal and two-terminal location methods. Single terminal location has larger error ranging as it is difficult to overcome the system impedance and resistance of the transition in principle; The fault location based on two-terminal electrical quantities has good prospects as the ability of eliminating transition resistances and impedances of the system in principle[2]-[3].

The velocity of travelling waves in the cable and overhead lines is inconsistent, difficult to range directly. Considering the impact of a hybrid circuit to traveling wave fault location, a new two-terminal fault location method combined line converted middle points and the HHT was proposed, and the problem of wave velocity discontinuous was solved.

2. An effective fault location method by converted midpoint

The principle of two-terminal fault location is based on the time difference between the first travelling waves arrive at both ends of the distance generated by fault voltage and current. When the fault of one-phase ground occurs, the voltage and current travelling waves will transfer along the way to both ends. Distance measurement devices installed at both ends of the bus record the initial time when travelling waves reach the terminals and calculate the distances. The Nomenclatures used in this paper are given bellow:

v is the travelling wave speed of overhead lines.

u is the travelling wave speed of cable lines.

D is the length of the fault line.

D_{MF} is the theoretical distance from the fault point F to the bus M.

D_{NF} is the theoretical distance from the fault point F to the bus N.

L_{MF} is the actual distance from the fault point F to the bus M.

L_{NF} is the actual distance from the fault point F to the bus N.

Δt is the time difference of travelling waves arrive at two measurement endpoints.

t_m is the time when the first wave of travelling waves reach M-side of the bus.

t_n is the time when the first wave of travelling waves reach N-side of the bus.

P_0 is the midpoint of the line

P_1 is the converted midpoint of line *MN*

The formulae following express the two-terminal fault location Method

$$\begin{cases} \dfrac{D_{MF}}{v} - \dfrac{D_{NF}}{v} = t_m - t_n \\ D_{MF} + D_{NF} = D \end{cases} \quad (1)$$

$$\begin{cases} D_{MF} = \dfrac{1}{2}[v(t_m - t_n) + D] \\ D_{NF} = \dfrac{1}{2}[v(t_n - t_m) + D] \end{cases} \quad (2)$$

Wave velocity has a relationship with the line medium and is basically constant in a line with the single electrical parameters [4]. The formulae of (1) and (2) have good application effects to the line with single electrical parameters such as high voltage power lines. However, the formulae above cannot be applied to measure distances directly in the small current grounding system, because the speed of wave propagation is discontinuous due to the existence of the alternating lines of overhead and cable.

The cable length is converted as the base v, that is, the length D of the cables will be vD/u after conversion. The crossing connected lines of overhead lines and cable would then be regarded as the unity of the overhead line. The two-terminal fault location formulae (1) and (2) could be applied to the equivalent line and the fault point's position in the converted line could be obtained. The accurate location could be realized after the conversion to the original real line, eliminating the influence of the discrete wave velocity.

The converted line's midpoint of the overhead lines and cable lines is the point from which the time the wave reaches the both line ends is the same. The line structure is symmetry for a transmission line with single electrical parameters, and the midpoint of the line is the converted line's midpoint as travelling wave signals take the same time from the point to both ends. As for the mixed distribution lines, the distribution of overhead lines and cable is complex and asymmetry, the time of the travelling wave move to the both ends from the line's midpoint is different, that is, the converted midpoint is not coincident to the midpoint of the line. Fig.1 is a hybrid structure diagram of overhead line and cable line.

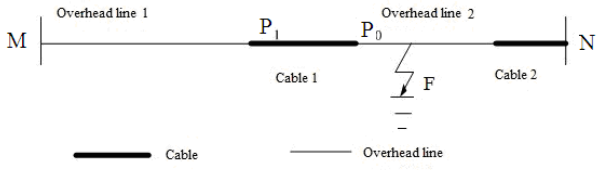


Fig .1 Hybrid structure diagram of overhead line and cable line

The steps of the new two-terminal fault location method are as follows:

(1)Determine the lengths of overhead and cable lines.

(2)Determine the wave velocity of the travelling wave propagating in various segments.

Because the unit length distribution parameters of cables are much different to the overhead lines, especially the distributed capacitance are about two orders of magnitude higher than overhead lines, resulting the apparent discontinuity of the travelling wave speed in the cable lines and the overhead lines.

According to wave propagation velocity formula (3), the wave velocity of the travelling wave propagating in various segments could be calculated.

$$v = \frac{1}{\sqrt{LC}} \quad (3)$$

Where, C and L are the unit length distributed capacitance and inductance in the various segment.

(3) Find the line converted midpoint P_1 of the line section MN. It could be determined according to the specific structure of lines and the lengths of overhead lines and cable in the various sections. If the time when the waves arrive at the line ends M and N are Tm and t_n respectively, and the corresponding time deference Δt =0, the starting point of the travelling wave is the line converted midpoint P_1.

(4) Determine the fault searching direction. When fault occurs, if the calculated parameters Δt <0, the fault is in the line section between P_1M, and the fault point could be searched from P_1 to M side. If Δt >0, the fault point could be searched from P_1 to N side. If Δt =0, P_1 is the fault point.

(5)Determine the fault point. When travelling waves moved Δt /2 from the converted midpoint along the search direction, the point was the fault point.

3. Travelling waves' heads detection by HHT

The parameters of electric transmission Lines would vary with the change of the frequency. The cable's frequency-dependence is serious as the result of its own characters, and the wave head would have serious attenuation and distortion. HHT method has good purpose on the detection of the mutation, nonlinearity and non-stationary signals. The frequency cluster of HHT is adaptive produced, needless of the selection to primary functions. So it is suitable for the detection of electric transmission Lines signal [5]-[6].

The signal was first decomposed into a limited number of Intrinsic Mode Function (IMF) by the Empirical Mode Decomposition (EMD) method. Then, every IMF components was transformed by HHT and the instantaneous amplitude and frequency were obtained.

3.1 Signals classifying by EMD

The IMF is defined as the component satisfies the following definition.

In the whole dataset, the number of the extreme points and zero-crossing points are equal or differ by one.

The maxima and the minima envelopes are obtained by cubic spline-interpolate, and the local mean value at any point of the maxima and the minima envelopes is zero.

Obviously, most of the signal does not meet the conditions above and is not the IMF component, and it is necessary to decompose signals to the IMF by algorithm EMD, Which steps are:

(1) Calculate all the local maxima and the minima points of a time series $X(t)$, and the maxima and the minima envelopes are fitted by cubic spline-interpolate. The local mean value m1(t) of the maxima and the minima envelopes are got, and it is eliminated from the original signal $X(t)$. The residual component h1(t) is obtained:

$$h_1(t) = x(t) - m_1(t) \tag{4}$$

(2) On ideal occasion, $h_1(t)$ is the first IMF component. But, as not all the local maxima and the minima points are included by the spline-interpolate accurately, it will induct error as the result that the missing point would become a new maxima or minima point at the next spline-interpolate step. So, the second classifying will take $h_1(t)$ as a new time series, and $m_{11}(t)$ is the maxima and the minima envelopes value of $h_1(t)$. The residual component $h_{10}(t)$ is obtained。

$$h_{10}(t) = h_1(t) - m_{11}(t) \tag{5}$$

Repeated the steps for k until $h_{1k}(t)$ meet the two conditions of IMF above, and let

$$x(t)=h_1(t) \tag{6}$$

3.2 The HHT Transform to the classified signals

The Hilbert-Huang Transform Function $Y(t)$ of a time series $X(t)$ is [5]:

$$Y(t) = \frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{X(t)}{t-\tau}d\tau \tag{7}$$

Whereas $X(t)$ could expressed by $Y(t)$, which is:

$$X(t) = \frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{Y(t)}{\tau-t}d\tau \tag{8}$$

It can be seen from the formulae (7) and (8), $X(t)$ and $Y(t)$ constitute a pair of complex conjugate. Meanwhile, $X(t)$ and $Y(t)$ consist of the information related to time series, and the relations followed could be available:

$$Z(t) = X(t) + jY(t) = A(t)e^{j\theta(t)}$$

$$A(t) = \sqrt{X^2(t)+Y^2(t)} \tag{10}$$

$$\theta(t) = \arctan(\frac{Y(t)}{X(t)}) \tag{11}$$

From the equations above, the two important instantaneous parameters are obtained, where $A(t)$ is the instantaneous amplitude and $\theta(t)$ is the phase and another important parameter, the instantaneous frequency could be obtained by the relation of the amplitude and frequency:

$$f(t) = \frac{1}{2\pi}\frac{d\theta(t)}{dt} \tag{12}$$

4. Example simulation

The small current grounding system simulation model was established using the electromagnetic transient simulation software ATP [6], shown in Fig.2. The system was coil grounding via arc suppression system, and is set over compensation. The inductance of the arc suppression is L=8.02H, and the resistance is $R_L = 80\Omega$.

Assuming the line 1 A-phase ground fault occurred, the cable length is *D*1=20km, and the over head lines length is *D*2=40km. The distance from the fault point to the M side is 11km. The cable lines' electrical parameters are:

$$R_1 = 2.415\times10^{-5}\Omega/\text{m} \quad L_1 = 5.163\times10^{-4}\text{mH/m}$$
$$R_0 = 1.965\times10^{-4}\Omega/\text{m} \quad L_0 = 3.976\times10^{-4}\text{mH/m}$$
$$C = 3.175\times10^{-4}\mu\text{F/m}$$

The overhead lines' electrical parameters are:

$$R_1 = 2.084\times10^{-5}\Omega/\text{m} \quad L_1 = 8.981\times10^{-4}\text{mH/m}$$
$$R_0 = 1.168\times10^{-4}\Omega/\text{m} \quad L_0 = 2.285\times10^{-3}\text{mH/m}$$
$$C = 1.29\times10^{-8}\mu\text{F/m}$$

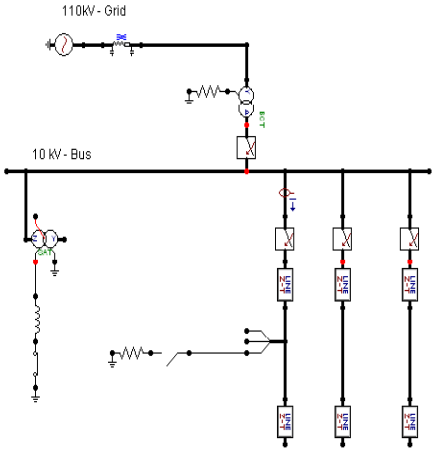


Fig. 2. Simulation model of the small current grounding system

The detected signals were converted into the mat format and then input into MATLAB and were decomposed into six IMF adaptively by EMD. The first IMF component was proceeded Hilbert Transform and time-frequency diagrams were got. The first IMF component is a frequency cluster of the travelling wave signals. The fault travelling waves head is high-frequency mutations in performance in time-frequency diagram, and the time corresponding was the time when the initial failure of the travelling wave arrived at the bus of M-side and N-side.

Fig.3 is the fault transient current travelling wave diagram of M-side and Fig.4 is the decomposed IMF components of M-side by EMD. The frequency plot of the travelling wave detected by HHT is shown in Fig.5, and the instantaneous frequency plot of the travelling wave detected at N-side is shown in Fig. 6.

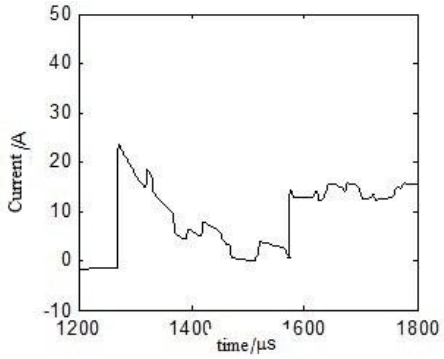


Fig .3. The fault transient current travelling wave of M-side

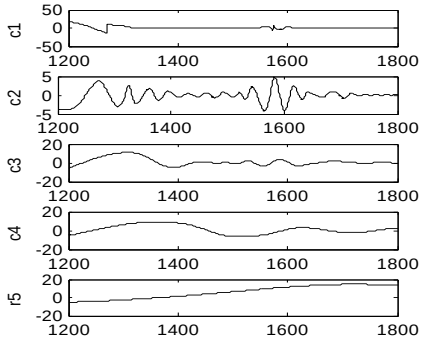


Fig.4 The decomposed IMF components of M-side by EMD

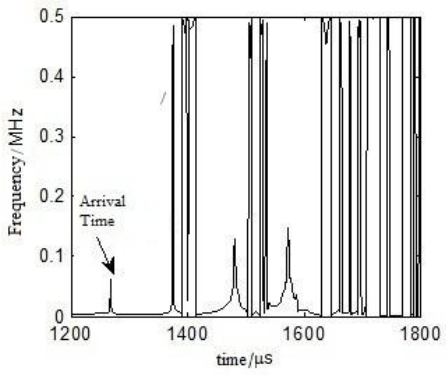


Fig. 5 The frequency of the travelling wave detected at M-side

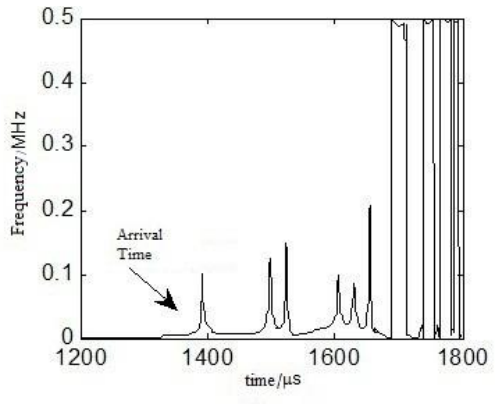


Fig. 6 The frequency of the travelling wave detected at N-side

From the Fig.5 and Fig.6, it can be seen that:

$t_m = 1.270 \times 10^{-3} s$, $t_n = 1.391 \times 10^{-3} s$

By the formulae (2), D_{MF} and D_{NF} could be calculated as:

$$\begin{cases} D_{MF}=21.85\text{km} \\ D_{NF}=58.15\text{km} \end{cases} \quad (10)$$

After converted to the actual lines, the actual distances from the fault point F to the M-side and N-side were:

$$\begin{cases} L_{MF} = 10.925km \\ L_{NF} = 49.075km \end{cases} \quad (11)$$

Experiments show that the measurement error is 0.075km, with a high accuracy.

5. Conclusion

This paper proposed a new method of two-terminal fault location based on the lines converted midpoint and HHT. The two-terminal fault location algorithm based on the lines converted midpoint is not affected by the dielectric media, solving the discontinuous problem in mixed lines with a high ranging accuracy. Considering the factors of the signal itself, the travelling wave was decomposed by EDM and the wave's head was detected by HHT, improving the measuring accuracy. It provides a correct and accurate method for the fault location in the small current grounding system.

Acknowledgements

This work is supported by the Key Program of National Natural Science Foundation of China (61077071, 61071202,51075349) and Program of National Natural Science Foundation of Hebei Province (F2011203207, F2010001312).

References

[1]Ge Yaozhong. New relay protection and fault ranging theory and technology, Xi'an Jiaotong University Press, 2007. 2

[2] E. G. Silveira, C. Pereira. Transmission Line Fault Location Using Two-terminal Data without Time Synchronization. IEEE Trans. on Power Delivery. 2007,22(1):498-499

[3] LUO Jing-nian, YAN Ting-chun. Error analysis and solution of line fault location based on traveling wave. East China Electric Power 2006,34 (10): 31-33

[4] YANG Jun, WU Yong-hong, JIANG Wen-bo etc. A Fault Location Algorithm for Hybrid Transmission Line Composed by High Voltage Cable and Overhead Line Based on Two-Terminal Information. Power System Technology, 2010, 34(1):208-212

[5] LI Tian-yun, ZHAO Yan, LI Nan. Apply Empirical Mode Decomposition Based Hilbert Transform to Power System Transient Signal Analysis, Automation of Electric Power Systems 2005,29(4):49-52

[6] Shangguan Hanlu, Study ON the Technology of Fault Location in high voltage measurement to power Network. Dissertation for the Master Degree in Engineering, 2011

Yutian Wang received the Master degree at Harbin Industrial University in 1981, the Doctor degree at Harbin Industrial University in 1995; He has been employed at Yanshan University since, and currently, he is a professor at Yanshan University; He has been supported by the National Natural Science Foundation of China, Natural Science Foundation of Hebei Province, and has received the awards of science and technology progress of Hebei Province and the Mechanical Industry Community. His articles published are over 50. His interests are in Intelligent Information Processing and photoelectric measurement.

Performance Analysis of web page recommendation algorithm based on weighted sequential patterns and markov model

[1]K. Suneetha and [2]M. Usha Rani
[1]Assistant Professor [SL], Department of Master of Computer Applications
Sree Vidyanikethan Engineering College, A.Rangampet, Tirupati, Andhra Pradesh, INDIA-517102

[2]Associate Professor, Department of Computer Science
Sri Padmavati Mahila Viswavidyalayam, (SPMVV Woman's' University), Tirupati
Andhra Pradesh, INDIA-517501

Abstract

Web usage mining techniques helps the users to predict the required Web page recommendations. In recent times, there has been a considerable significance given to sequential mining approaches to construct web page recommendation systems. This paper focuses on developing a web page recommendation approach for accessing related web pages more efficiently and effectively using weighted sequential pattern mining and markov model. Here we have developed an algorithm called, W-PrefixSpan, that is the modification of traditional Prefixspan algorithm including the constraints of spending time and recent visiting to extract weighted sequential patterns. Then by utilizing weighted sequential patterns recommendation model is constructed based on
Patricia-trie data structure. Later the web page recommendation of the current users is done with the help of markov model.

Experimentation is done with the help of synthetic dataset and we present the performance report of web page recommendation algorithm in terms of precision, applicability and hit ratio. The results have shown that, the precision of our algorithm is improved by 5% than the previous algorithm. Also we have achieved high applicability in the support of 50 % and in terms of hit ratio, the proposed algorithm ensured that the performance is considerably improved for various support values.

Keywords:- *Prefixspan, Web page recommendation, Weighted sequential pattern, Patricia-trie, Markov model.*

1. INTRODUCTION

With the explosive growth of information available on the World Wide Web, it has become much more difficult to access relevant

information from the Web. Web Mining is one of the most propitious fields of Data Mining, which deals with the extraction of meaningful or relevant knowledge from the World Wide Web [3]. More specifically, Web Content Mining is the branch of Web Mining that focuses on extracting the raw information available in web pages. Web Usage Mining is also the branch of Web mining, which deals with the extraction of relevant information from server log files. Here, the source data is mainly composed of the (textual) logs that are gathered when the users access the web servers and might be depicted in standard formats; and classic applications are those based on user modeling approaches, namely web personalization, adaptive web sites, and user modeling [6].

Web personalization [15] refers to any action that adapts the information or services provided by a Web site to the needs of a particular user or a set of users by using the knowledge procured from the navigational activities and individual interests of users recorded in the web usage logs, in conjunction with the content and the structure of the Web site [5]. The role of Web personalization system is to provide the users with an information they desire or need, without expecting from them to inquire for it explicitly [6]. Web Recommender system is one type of personalized web application, which provides substantial user value by

personalizing numerous sites on the Web [2]. Recently, Web-based Recommender Systems (RS) are widely applied to provide diverse type of customized information to the users. Generally, there are many data mining techniques such as association rule mining, sequential pattern discovery, clustering, and classification. Among them, sequential pattern-mining method is an extensively used data analysis technique in web usage mining [1]. Sequential pattern mining [14], an advance of association rule mining, is an imperative subject of data mining, often applied for extracting the useful information [7]. In recent times, there has been a considerable significance given to sequential mining approaches to construct web page recommendation systems. This paper focuses on developing a web page recommendation approach for accessing related web pages more efficiently and effectively. The main goal of this approach is to determine which web pages are more likely to be accessed next by the current user in the near future.

The paper is organized as follows: Section 2 presents the motivating algorithms of the proposed algorithm. Section 3 presents the contribution made in the paper. Section 4 provides the proposed technique of web page recommendation and section 5 presents experimentation and the results obtained. Section 6 concludes the paper.

2. MOTIVATING ALGORITHMS

This section describes the motivating algorithms of the proposed web recommendation approach. Here, we have mentioned three different algorithms that are based on weighted association rules, markov model and closed sequential patterns.

1. Weighted association rule-based web page recommendation algorithm

Web page recommendation based on weighted association rules was proposed by R. Forsati, M. R. Meybodi [11]. Here, they have proposed three algorithms to clear up the web page recommendation problems. In the first algorithm, a distributed learning machine has been employed to study the behavior of previous users' and to recommend pages to the current user based on the learned patterns. In the second algorithm, Weighted Association Rule mining algorithm has been applied for recommendation purposes. Finally, in the third algorithm, the above two algorithms have been combined to enhance the competence of web page recommendation. The general block diagram of the hybrid algorithm based on distributed learning automata and weighted association rule mining algorithm is given in figure 1.

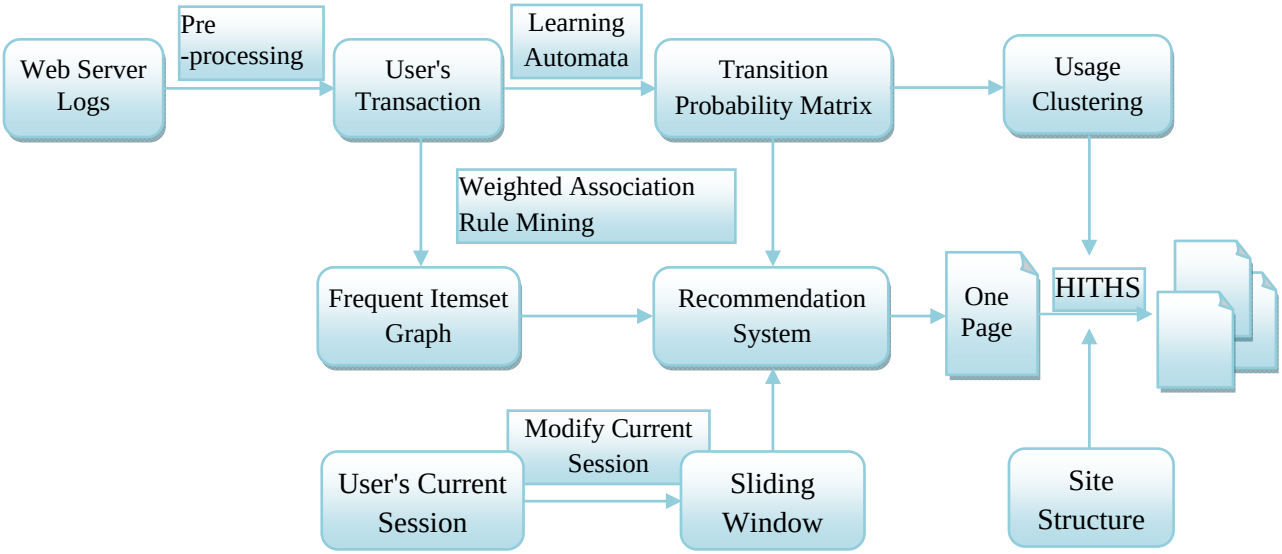


Fig.1. Weighted association rule-based web page recommendation algorithm

2. Markov model-based web page recommendation algorithm

The probability theory-based Markov model is effectively utilized by Faten Khalil *et al.* [9] for web page recommendation. Here, the Web page access prediction accuracy has been enhanced by including three prediction models such as Markov model, Clustering, and association rules according to certain constraints. They have integrated these three models using 2-Markov model computed on clusters achieved by means of k-means clustering algorithm and Cosine distance measures for states that belong to the majority class and performing association rule mining on the rest. The algorithmic procedure is described as follows.

Training:

(1) Combine functionally related web pages according to services requested
(2) Group user sessions into l-clusters
(3) Construct a k-Markov model for each cluster
(4) For Markov model states where the majority is not clear
(5) Mine association rules for each state
(6) End For

Prediction:

(1) For each coming session
(2) Find its closest cluster
(3) Use relevant Markov model to make prediction
(4) If the predictions are made by states that do not belong to a majority class
(5) Use association rules to make a revised prediction
(6) End If
(7) End For

3. Closed Sequential Pattern-based web page recommendation algorithm

Closed sequential pattern, one of the variants of sequential pattern is used by the
U. Niranjan *et al.* [12, 8] for web page recommendation. The proposed system was mainly based on discovering the closed sequential web access patterns. Firstly, the PrefixSpan algorithm has been applied on the preprocessed web server log data for extracting the sequential web access patterns. Then, the closed sequential web access patterns have been mined from the complete set of sequential web access patterns via post-pruning approach. Subsequently, a pattern tree, a compact representation of closed sequential patterns, has been build from the mined closed sequential web access patterns. Moreover, the Patricia trie based data structure has been employed in the construction of the pattern tree. Based on the constructed pattern tree, the proposed system has provided recommendations for a given user's web access sequence. The general block diagram of the recommendation algorithm based on closed sequential pattern mining algorithm is given in figure 2.

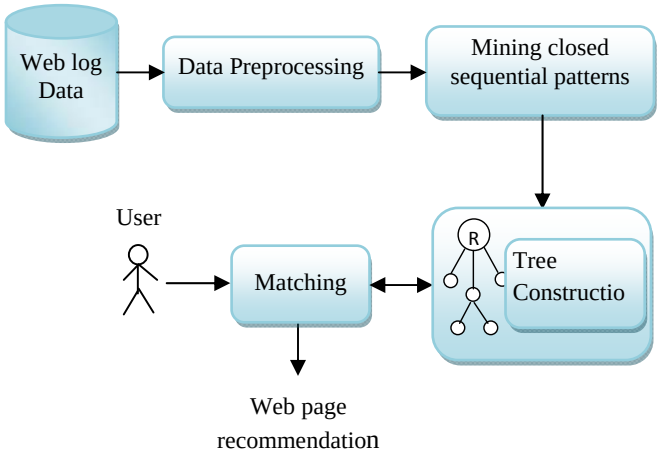


Fig.2. Closed sequential pattern-based web Page recommendation algorithm

3. CONTRIBUTIONS OF THE PAPER

The main contributions of the paper are given as follows,

- We have presented an algorithm for web page recommendation by combining the weighted sequential pattern and markov model.
- We have developed an algorithm called, W-prefixSpan, that is the modification of traditional Prefixspan algorithm including the constraints of spending time and recent visiting.
- We analyze the performance of W-prefixspan algorithm with the Prefixspan algorithm in terms of computation time and memory usage.
- We present the performance report of web page recommendation algorithm in terms of precision, applicability and hit ratio.

4. WEB PAGE RECOMMENDATION ALGORITHM BASED ON WEIGHTED SEQUENTIAL ACCESS PATTERNS

In recent years there have been increasing interests in applying web usage mining

techniques to build web page recommendations. With the intention of real world applicability, we have developed an approach for web page recommendation using weighted sequential pattern and markov model. Here, the traditional sequential pattern mining algorithm called Prefixspan is modified significantly by incorporating the significant measures such as spending time and recent view to mine more useful patterns. Then, the markov model [9] is used to recommend the web pages. The steps in the algorithm for generating recommendations to the user could be briefly summarized as follows [13]:

Step 1: Data Preprocessing: This step is used to extract the useful and relevant information from raw web logs. This raw web logs need to be processed analyzed and converted into proper format of sequential database to mine the weighted sequential patterns.

Step 2: W-PrefixSpan for mining of weighted sequential web access patterns: To identify the interesting sequential patterns from a weighted sequential database, the proposed recommendation system utilizes a traditional Prefixspan [10], which is a well known pattern-growth algorithm by incorporating two measures ***spending time*** and ***recent view*** into the mining procedure.

Step 3: Building Pattern tree model: Once weighted sequential patterns are mined, a pattern tree is constructed using the procedure defined in [12, 8] is applied to the proposed approach for constructing tree structure. Patricia-based data structure is used for web page recommendation due to the advantages of particia structure over the trie structure.

Step 4: Generation of recommendations using markov model: Here, we make use of the markov model described in [9] that is used in the identification of the next page to be accessed by the Web site user based on the sequence of previously accessed pages. The accurate recommendations can be found using the definitions of the probability described in [13].

4.1 Data Preprocessing: Due to large amount of irrelevant information available in the web log, the original log data cannot be directly used in the web mining procedure. A web log file consists of, IP address, access time, HTTP request method used, URL of the referring page and browser name. To mine the required sequential patterns, it is very difficult to directly use the web log data. Hence the following preprocessing techniques can be used to convert the data into proper format.

User Identification: In this step a sequential database is constructed by identifying each user accessing web pages. Users may be tracked based on IP address and user session. A new IP address is used to identify the new user but at the same time, the user session should be fixed for a particular time period.

*Weighted Sequential database generation: T*he weighted sequential database is generated including the sequence of web pages visited by the user, time spent by the user on corresponding web page and its recent information.

4.2 W-PrefixSpan for Mining of Weighted sequential web access pattern

To identify the interesting sequential patterns from a weighted sequential database, the proposed recommendation system utilizes a traditional Prefixspan [10], which is a well known pattern-growth algorithm by incorporating two measures ***spending time*** and ***recent view*** into the mining procedure.

Spending time is an important measure for the researchers who are attempting to identify the interest of the users. Time spent by the user within a particular page is necessary to identify the importance of web pages.

Recent view is another important measure to find whether the page is accessed recently or not. More importance should be given for the web pages which are accessed recently because the behavior of the user surely varies depend on the time so the recent behavior of the user is significant for finding the sequence analysis.

W-PrefixSpan algorithm: An efficient sequential pattern mining algorithm called W-Prefixspan [13] is developed by modifying traditional sequential pattern mining algorithm Prefixspan for finding frequent sequential patterns.

Initially, the weighted sequential database W_{ij} is given as an input to the proposed W-PrefixSpan algorithm that discovers the 1-length weighted sequential patterns from the weighted sequential

database by scanning the database once. The 1-SR patterns (spending time with recent view) which satisfy the predefined support threshold are mined from the sequential database by simply scanning the database. The W-support for the 1-length pattern is computed as follows,

$$W\_\sup(p) = \frac{1}{N} \frac{\sum_{i=1}^{N_T \in p} I_s(i) * R(i)}{\sum_{i=1}^{N_T \in p} R(i)}$$

Where, N → Number of user transaction in the weighted sequential database

N_T → Number of transaction that contains the web page p

$R(i)$ → Recent information

$$I_s(i) = \sum_{i=1}^{N_T} \left(\frac{s_i}{\sum_{i=1}^{M_T} s_i} \right)$$

constructed pattern tree is based on Patricia-trie data structure. The procedure for constructing a pattern tree defined in [13] is applied to the proposed approach.

4.4 Generation of Recommendations using markov model

Markov models are the most effective techniques for Web page access prediction and to improve the Web server access efficiency. The markov model described in [9] is used for the identification of next page to be accessed by the user based on the sequence of previously accessed pages. Whenever a new user comes to get the recommendation, the sequence path of the new user is matched with the Patricia-trie structure. Then, the subsequent web page whether it may be from same node or from its child node is retrieved. Now, the sequence path of the new user is used to find the accurate recommendation using the probability definition used in the previous work [9]. The probability of computation is carried out to find the most important sequence for the user [13]. The probability, $pro(s_{n+1} \mid s)$, is estimated by using all sequences of all users in tree structure constructed from the weighted sequential database W_{ij} .

$$P(s_{n+1} = (s \in s_1, s_2, s_n, s_{n+1}, \ldots, s_m) \mid s_1, s_2, \ldots, s_n) = \frac{W\_\sup(s_1, s_2, \ldots, s_n)}{W\_\sup(s_1, s_2, s_n, s_{n+1}, \ldots, s_m)}$$

Then, the final recommendation is based on the

$$s_{n+1} = \arg\ sort\ \{s_{n+1}^{(1)}, s_{n+1}^{(2)}, s_{n+1}^{(3)}\}$$

5. RESULTS AND DISCUSSION

This section presents the detailed discussion about the results which are obtained from the experimentation. The experimentation is done on the proposed approach using synthetic dataset and the results are evaluated with the precision, applicability and hit ratio.

1. Experimental set up and dataset description

The proposed web page recommendation approach is implemented in Java (jdk 1.6) with I3 processor of 2GB RAM. Here, the synthetic dataset is generated as like the same format of real datasets and the performance of the proposed approach is evaluated with the evaluation metrics. The generated synthetic dataset is divided into two parts such as, Training dataset (It is used for building the pattern tree model and test dataset (It is used for testing the web recommendation approach).

2. Evaluation metrics

For evaluating the proposed approach, we have used three measures such as precision, applicability and hit ratio [12, 8]. The formal definition of these three measures are given as,

$$\text{Precision} = \frac{C^+}{C^+ + I^-}$$

Where, C^+ → Number of correct recommendations.

I^- → Number of incorrect recommendations.

Definition:

Let $S = s_1\ s_2 \cdots s_j\ s_{j+1} \cdots s_n$ be a web access sequence of test dataset. The recommendation $R = \{r_1, r_2, \cdots, r_k\}$ is generated by using the constructed pattern tree for the subsequence $S_{sub} = s_1\ s_2 \cdots s_j$ $(\text{minlen} \le j \le \text{maxlen})$. The recommendation R is said to be correct, if it contains s_{j+1} ($s_{j+1} \in R$). Otherwise, R is said to be incorrect recommendation.

$$Applicability = \frac{C^+ + I^-}{|N|}$$

Where, $|N|$ → Total number of given requests.

$$Hit\ ratio = Precision \times Applicability = \frac{C^+}{|N|}$$

3. Performance of the web page recommendation algorithm

The proposed web page recommendation approach is analyzed with the help of precision, applicability and hit ratio. Here, the testing dataset is given to the tree model constructed with the help of training data. Subsequently, the precision is computed based on the result obtained for the test dataset. Here, the results are taken for the proposed approach, PrefixSpan algorithm- based approach and the previous algorithm [12, 8] and the graphs are plotted for the taken results, which have shown in figure 3, 4 and figure 5. In figure 3, the W-prefixSpan algorithm has achieved the precision of about 70 % where, the Niranjan et al's algorithm has achieved only 65%. In the figure 4, we have achieved high applicability in the support of 50 % and in terms of hit ratio, the proposed algorithm ensured that the performance is considerably improved for various support values.

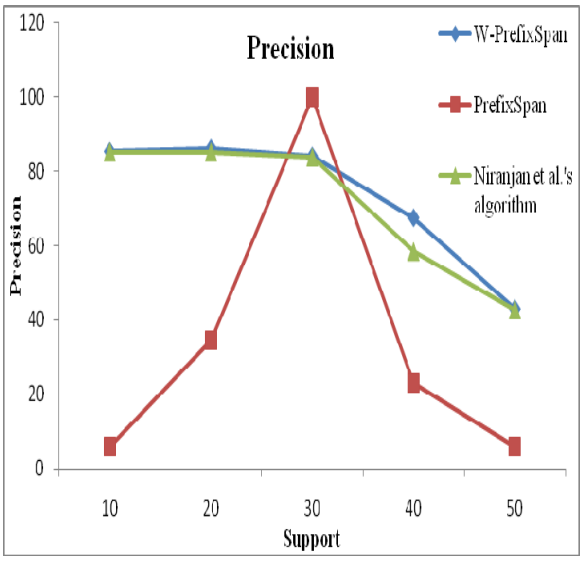


Fig.3. Precision

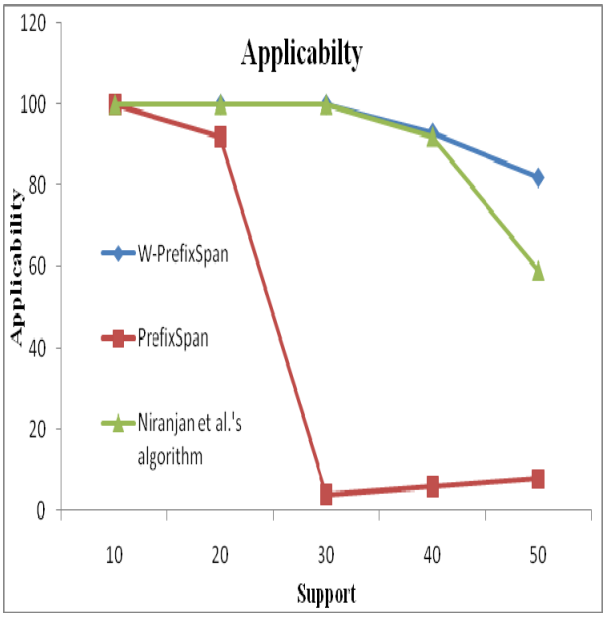


Fig.4. Applicability

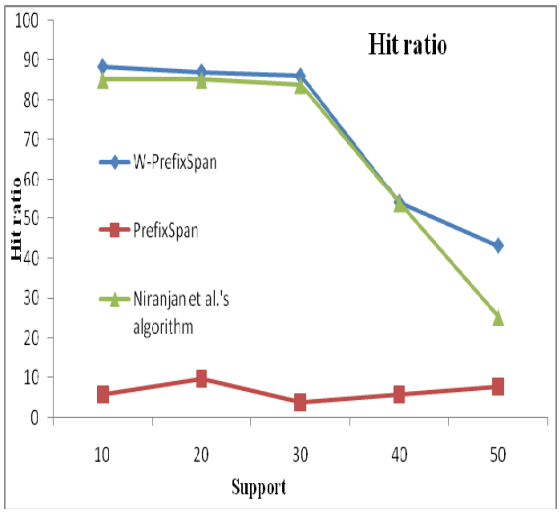


Fig.5. Hit ratio

6. CONCLUSION

We have proposed a web page recommendation algorithm using weighted sequential patterns and markov model.

Here, we have presented W-PrefixSpan algorithm that is the developed by incorporating the weightage constraints such as, spending time and recent visiting with the prefixspan algorithm. The mined weighted sequential patterns are then utilized to construct the recommendation model using the Patricia-trie based tree structure. At last, markov model-based recommendation is carried out for the current users by matching the visiting path with the tree and markov model. The experimentation is done with the help of synthetic dataset and the performance of W-Prefixspan algorithm as well as web page recommendation algorithm is analyzed. From the results, the precision of our algorithm is improved by 5% than the previous algorithm. Also achieved high applicability in the support of 50 % and in terms of hit ratio, the proposed algorithm ensured that the performance is considerably improved for various support values.

REFERENCES

[1] Feng-Hsu Wang and Hsiu-Mei Shao, “Effective personalized recommendation based on time-framed navigation clustering and association mining”, Expert Systems with Applications, vol. 27, no.3, pp. 365–377, 2004.

[2] J. Ben Schafer, Joseph A. Konstan and John T. Riedl, "Recommender Systems for the Web", In Visualizing the Semantic Web, Springer, pp.102-123, 2006.

[3] Oren Etzioni, “The world-wide web: Quagmire or gold mine?”, Communications of the ACM, vol. 39, no.11, pp.65–68, 1996.

[4] Federico Michele Facca and Pier Luca Lanzi, "Recent Developments in Web Usage Mining Research", Lecture Notes in Computer Science, vol. 2737, pp. 140-150, 2003.

[5] M. Eirinaki, M. Vazirgiannis, “Web Mining for Web Personalization”, ACM Transactions on Internet Technology vol.3, no.1, pp.1-27, February 2003.

[6] Mulvenna. M. D, Anand. S. S, and Buchner. A. G, “Personalization on the Net using Web Mining”, Communications of the ACM, vol. 43, no. 8, pp. 123–125, August 2000.

[7] Sizu Hou, Xianfei Zhang, "Alarms Association Rules Based on Sequential Pattern Mining Algorithm," In proceedings of the Fifth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 2, pp.556-560, Shandong, 2008.

[8] Utpala Niranjan, R. B. V. Subramanyam and V. Khanaa, "Developing a Web Recommendation System Based on Closed Sequential Patterns", Communications in Computer and Information Science, Vol. 101, no. 1, pp. 171-179, 2010.

[9] Faten Khalil, Jiuyong Li and Hua Wang, "Integrating Recommendation Models for Improved Web Page Prediction Accuracy", in Proceedings of the thirty-first Australasian conference on Computer science, Vol. 74, 2008.

[10] Jian Pei; Jiawei Han; Mortazavi-Asl, B.; Pinto, H.; Qiming Chen; Dayal, U.; Mei-Chun Hsu, "PrefixSpan,: mining sequential patterns efficiently by prefix-projected pattern growth", in Proceedings of 17th International Conference on Data Engineering, 2001.

[11] R. Forsati, M. R. Meybodi, "Effective Page Recommendation Algorithms Based on Distributed Learning Automata and Weighted Association Rules",

[12] Utpala Niranjan, R.B.V. Subramanyam, V.Khana, "An Efficient System Based On Closed Sequential Patterns for Web Recommendations", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 4, May 2010.

[13] K. Suneetha and M. Usha Rani, “Web Page Recommendation Approach Using Weighted Sequential Patterns and Markov Model”, in Global Journal of Computer Science and Technology, Vol. 12, Issue 9, Version 1.0, April 2012.

[14] Qiankun Zhao and Sourav S. Bhowmick, "Sequential Pattern Mining: A Survey" , Technical Report, CAIS, Nanyang Technological University, Singapore, No.118, 2003.
[15]Sarabjot Singh Anand and Bamshad Mobasher,"Intelligent Techniques for Web Personalization,"ITWP 2003, LNAI 3169, pp. 1–36, Springer-Verlag Berlin Heidelberg, 2005.

Ms. K.Suneetha obtained her Bachelor's Degree in Sciences from S.V. University Tirupathi. Then she obtained her Master's degree in Computer Applications from S.V.University. She is working as Assistant Professor [SL] in the Department of Master of Computer Applications at Sree Vidyanikethan Engineering College, A.Rangampet, Tirupati. She is pursuing her Ph.D. in Computer Science in the area of Data Warehousing and Data Mining. She is in teaching since 2000. She presented many papers at National and Internal Conferences and published articles in National & International journals.

Dr. M. Usha Rani is an Associate Professor in the Department of Computer Science and HOD for MCA, Sri Padmavati Mahila Viswavidyalayam (SPMVV Woman's' University), Tirupati. She did her Ph.D. in Computer Science in the area of Artificial Intelligence and Expert Systems. She is in teaching since 1992. She presented many papers at National and Internal Conferences and published articles in national & international journals. She also has written 4 books like Data Mining - Applications: Opportunities and Challenges, Superficial Overview of Data Mining Tools, Data Warehousing & Data Mining and Intelligent Systems & Communications. She is guiding M.Phil. and Ph.D. in the areas like Artificial Intelligence, Data Warehousing and Data Mining, Computer Networks and Network Security etc.

Predicting the Effects of Medical Waste in the Environment Using Artificial Neural Networks: A Case Study

Qeethara Al-Shayea[1] and Ghaleb El-Refea[2]

[1] MIS Department, Al-Zaytoonah University of Jordan
Amman, Jordan

[2] Al Ain University of Science and Technology
Abu Dhabi, United Arab Emirates

Abstract

Protection of the environment from medical waste hazards is becoming a serious problem. There is a big relation between medical waste and disease injury. The main idea of this study is predict the relation between medical wastes and diseases in Hashemite Kingdom of Jordan using Artificial Neural Networks (ANNs) model. There are six predictor parameters associated with solid and liquid wastes in the medical services sector which are affecting the diseases injury. This study deals with two types of diseases the first one is acute hepatitis and the other is typhoid. Generalized Regression Neural Network (GRNN) is used to predict the diseases injury. It is noticed a significant improvement in the prediction made by GRNN due to its generalization property. Results showed that all six parameters associated with solid and liquid medical wastes which have the largest regression value affect the acute hepatitis injuries and the typhoid injuries. It is also showed that the medical waste affected the typhoid injuries in large percentage so the regression is very large.

Keywords: *Regression, Artificial neural networks, General Regression Network, Prediction, Medical Wastes.*

1. Introduction

As in many other developing countries, the generation of regulated medical waste (RMW) in Jordan has increased significantly over the last few decades. Despite the serious impacts of RMW on humans and the environment, only minor attention has been directed to its proper handling and disposal [1].

The waste produced in the in the course of health care activities carries a higher potential for infection and injury than any other type of waste [2].

A. Puss, E. Giroult and P. Rushbrook [3] presented an overview of these environmental concerns from landfilling practices and their adverse environmental effects. In their paper, a number of remedial measures needed to minimize these environmental and socio-economic effects are suggested, with in total ten long term and eight short term measures for improving of the solid waste management system of Jordan.

Awad et. al. [4] presented research under the assumption that wastes generated from hospitals in Jordan and Irbid were hazardous.

Jahandideh et. al. [5] presented two predictor models including artificial neural networks and multiple linear regression were applied to predict the rate of medical waste generation totally and in different types of sharp, infectious and general.

Al-Habash and Al-Zu'bi [6] proposed an Idea about the medical waste management in the health sector and its impact on the environment in Jordan, the right and the safe management which include, segregate, classify, collect, processing of these waste may contribute to achieve the main goal which is to reduce the hazardous effect on the local community.

2. Artificial Neural Networks

An artificial neural network (ANN) is a computational model that attempts to account for the parallel nature of the human brain. An (ANN) is a network of highly interconnecting processing elements (neurons) operating in parallel. These elements are inspired by biological nervous systems. As in nature, the connections between elements largely determine the network function. A subgroup of processing element is called a layer in the network. The first layer is the input layer and the last layer is the output layer. Between the input and output layer, there may be additional layer(s) of units, called hidden layer(s). Fig. 1 represents the typical neural network. You can train a neural network to perform a particular function by adjusting the values of the connections (weights) between elements.

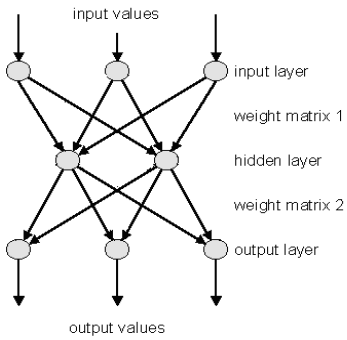


Fig. 1 A typical neural network.

For the researcher and the financial analyst, the main advantage of ANNs is that there is no need to specify the functional relation between variables. Since they are connectionist-learning machines, the knowledge is directly imbedded in a set of weights through the linking arcs among the processing nodes. In order to train a neural network properly one needs a large set of representative 'good quality' examples. In the case of bankruptcy problems, the researcher should be cautious when drawing conclusions from neural networks trained with only one or two hundred cases, as observed in most previous studies [7].

2.1 Generalized Regression Neural Network

The GRNN was applied to solve a variety of problems like prediction, control, plant process modeling or general mapping problems [8]. General regression neural network Specht [9], Nadaraya [10] and Watson [11], does not require an iterative training procedure as in back-propagation method.
The GRNN is used for estimation of continuous variables, as in standard regression techniques. It is related to the radial basis function network and is based on a standard statistical technique called kernel regression. By definition, the regression of a dependent variable y on an independent x estimates the most probable value for y, given x and a training set. The regression method will produce the estimated value of y, which minimizes the mean-squared error. GRNN is a method for estimating the joint probability density function (pdf) of x and y, given only a training set. Because the pdf is derived from the data with no preconceptions about its form, the system is perfectly general. Furthermore, it is consistent; that is, as the training set size becomes large, the estimation error approaches zero, with only mild restrictions on the function. In GRNN, instead of training the weights, one simply assigns to wij the target value directly from the training set associated with input training vector i and component j of its corresponding output vector [12]. GRNN architecture is given in Fig. 2.
GRNN is based on the following formula [13]:

$$E[y \mid x] = \frac{\int_{-\infty}^{\infty} y.f(x,y).dy}{\int_{-\infty}^{\infty} f(x,y).dy} \quad (1)$$

where y is the output of the estimator, x is the estimator input vector, $E[y|x]$ is the expected output value, given the input vector x and $f(x,y)$ is the joint probability density function (pdf) of x and y.
The function value is estimated optimally as follows:

$$y_j = \frac{\sum_{i=1}^{n} h_i . w_{ij}}{\sum_{i-1}^{n} h_i} \quad (2)$$

Where w_{ij}= the target output corresponding to input training vector x_i,

$h_i = e^{\frac{-D_i^2}{2.spread^2}}$, the output of the hidden layer neuron,

$D_i^2 = (x-u_i)^T (x-u_i)$, the squared distance between the input vector x and the training vector u, x= the input vector,

u_i=training vector i, the center of neuron i, spread=a constant controlling the size of the receptive region.

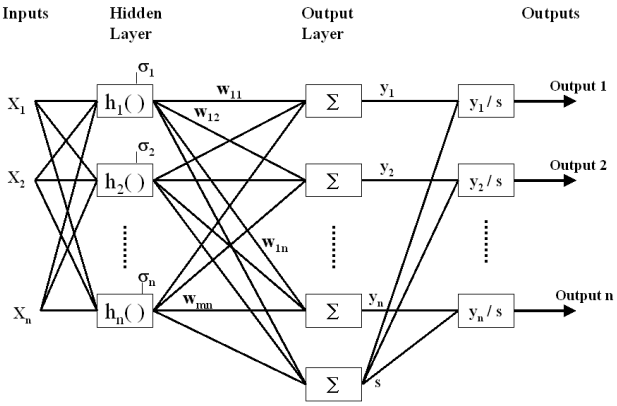


Fig. 2 Generalized Regression Neural Network (GRNN) Architecture.

3. Experimental Results

3.1 Data

This study was conducted at Hashemite Kingdom of Jordan which is split the database to north, central and south regions. The environment data is obtained from the department of statistics-GIS during 2002-2009. This dataset contains six parameters which represent the quantity of solid and liquid wastes in medical services sector are the outputs of the network. The input is number of cases to two diseases. The first is acute hepatitis and the other is typhoid and para typhoid. Table 1 presents the wastes in medical service sector which are considered as predictor variables used in the study.

Table 1: Predictor variable of datasets used in the study

Environmental data		
S. No.	***Predictor Variable Name***	***Measurement***
1	Body human waste	number
2	Waste water	liter
3	Waste of chemical and medicine	liter
4	Fluids resulting from dialysis unit	liter
5	Lab tests residues	liter
6	Medical waste	number

3.2 Results Analysis

A generalized regression neural network (GRNN) with a radial basis layer and a special linear layer and linear output neurons was created using the neural network toolbox from Matlab 7.10 as shown in Fig. 3.

Generalized regression neural networks are a kind of radial basis network that is often used for function approximation. The use of a probabilistic neural network is especially advantageous due to its ability to converge to the underlying function of the data with only few training samples available. GRNN is adopted to discover the association between medical wastes and diseases.
The first layer has as many neurons as there are input/ target vectors. Each neuron's weighted input is the distance between the input vector and its weight vector. Each neuron's net input is the product of its weighted input with its bias. Each neuron's output is its net input passed through radial basis transfer function. Radial basis transfer function is a neural transfer function which calculates a layer's output from its net input. If a neuron's weight vector is equal to the input vector (transposed), its weighted input will be 0, its net input will be 0, and its output will be 1. The second layer also has as many neurons as input/target vectors.
A spread slightly lower than the distance between input values, in order, to get a function that fits individual data points fairly closely is used. A smaller spread would fit data better but be less smooth.

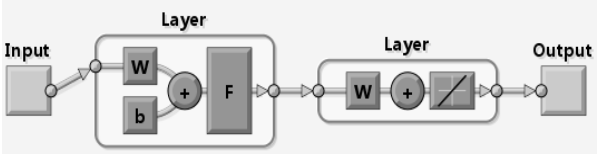


Fig. 3 A generalized regression neural network (GRNN).

Six GRNN with medical wastes as input and acute hepatitis as a target was been created. Then simulate the network with 8 samples. The spread value was chosen 0.1. The percent correctly predicted in the simulation sample for all medical wastes are shown in table 2.

Table 2

Predictor Variable Name (Medical wastes)	***Regression value***
Body human waste	R=0.65727
Waste water	R=0.41544
Waste of chemical and medicine	R=0.34625
Fluids resulting from dialysis unit	R=0.4763
Lab tests residues	R=0.62707
Medical waste	R=0.40936

Six GRNN with medical wastes as input and typhoid as a target was been created. Then simulate the network with 8 samples. The spread value was chosen 0.1. The percent correctly predicted in the simulation sample for all medical wastes are shown in table 3.

Table 3

Predictor Variable Name (Medical wastes)	***Regression value***
Body human waste	R=0.88687
Waste water	R=0.62446
Waste of chemical and medicine	R=0.69068
Fluids resulting from dialysis unit	R=0.73998
Lab tests residues	R=0.86629
Medical waste	R=0.93993

It is clear from the results, that all medical wastes are very influential factors in the pathogenesis of disease, viral hepatitis acute typhoid.
Fig. 4 shows the multiple regressions for the six affecting factors as input and acute hepatitis as target. The spread value was chosen 0.1. The percent correctly predicted in the simulation sample is approximately 53 percent.

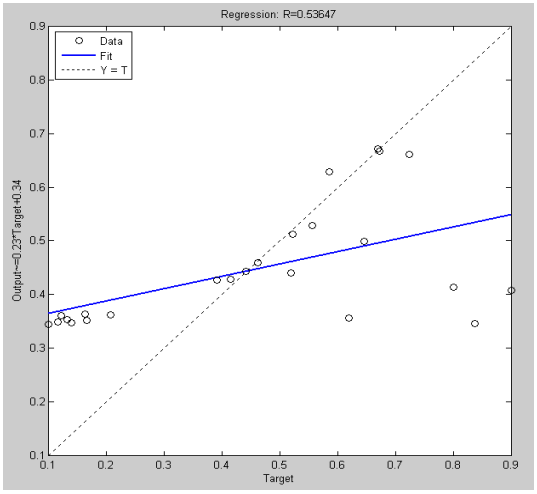


Fig. 4

Fig. 5 shows the multiple regressions for the six affecting factors as input and typhoid as target. The spread value was chosen 0.1. The percent correctly predicted in the simulation sample is approximately 97 percent.

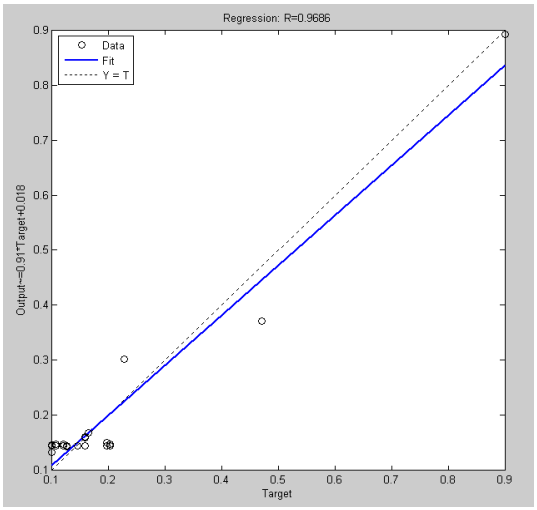


Fig. 5

4. Conclusions

In this paper the general regression neural network is used for the prediction of the diseases injury. The advantage of using the GRNN in the prediction is its generalization property. The results of this study provide evidence that the human body waste and lab tests residues are the most affecting in acute hepatitis. While the medical waste is the most affecting in typhoid injure.

The results also indicate that the variables examined in the study provided a significant contribution in predicting disease injure human.

References

[1] R. Oweis, M. Al-Widyan and O. Al-Limoon, "Medical waste management in Jordan: A study at the King Hussein Medical Center", Waste Management Journal, Vol. 25, Issue 6, 2005, pp. 622-625.
[2] M. Aljaradin and K. M. Persson, "Environmental Impact of Municipal Solid Waste Landfills in Semi-Arid Climates - Case Study – Jordan", The Open Waste Management Journal, Vol. 5, 2012, pp. 28-39
[3] A. Puss, E. Giroult and P. Rushbrook, Safe Management of Wastes From Health-Care Activities, Would Health Organization, 1999.
[4] A. R. Awad, M. Obeidat and M. Al-Shareef, "Mathematical-Statistical Models of Generated Hazardous Hospital Solid Waste", Journal of Environmental Science & Health, Vol. 39, Issue 2, 2004, pp. 315-327.
[5] S. Jahandideh, S. Jahandideh, E. Asadabadi, M. Askarian, M. Movahedi, S. Hosseini and M. Jahandideh, "The use of artificial neural networks and multiple linear regression to predict rate of medical waste generation", Waste Management Journal, Vol. 29, Issue 11, 2009, pp. 2874-2879.
[6] M. Al-Habash and A. Al-Zu'bi, "Efficiency and Effectiveness of Medical Waste Management Performance, Health Sector and its Impact on Environment in Jordan Applied Study", World Applied Sciences Journal, Vol. 19, No. 6, 2012, pp. 880-893.
[7] J. C. Neves and A. Vieira, "Improving Bankruptcy Prediction with Hidden Layer Learning Vector Quantization", European Accounting Review, Vol. 15, No. 2, 2006, pp. 253–271.
[8] D. W. Patterson, Artificial Neural Networks, Theory, and Applications, Englewood Cliffs, NJ: Prentice-Hall, 1995.
[9] D. F. Specht, "A general regression neural network", IEEE Trans. Neural Networks, vol. 2, 1991, pp. 568–576.
[10] E. A. Nadaraya, "On estimating regression", Theory of Probability Applicant, vol. 9, 1964, pp. 141–142.
[11] G. S. Watson, "Smooth regression analysis", Sankhya Series A, vol. 26, 1964, pp. 359–372.
[12] M. T. Hagan, H. B. Demuth, M. Beale, Neural network design, PWS Publishing Company, Boston, 1996.
[13] K. Kayaer and T. Yildirim, "Medical Diagnosis on Pima Indian Diabetes Using General Regression Neural Networks", web page available at: www.yildiz.edu.tr/~tulay/publications/Icann-Iconip2003-2.pdf.

Qeethara Kadhim Abdulrahman Al-Shayea is Associate Professor in Department of Management Information Systems Faculty of Economics & Administrative Sciences Al-Zaytoonah University of Jordan. She has received Ph. D. in Computer Science, Computer Science Department, University of Technology, Iraq, 2005. She received her M.Sc. degree in Computer Science, Computer Science Department from University of Technology, Iraq, 2000. She has received her High Diploma degree in information Security from Computer Science Department, University of Technology, Iraq, 1997. She has received B. Sc. Degree in Computer Science Department from University of Technology, Iraq, 1992. She is interested in Artificial Intelligent, Business Intelligence, Image Processing, Computer Vision, Coding Theory and Information Security. She has already published many papers in international journals and conferences

Ghaleb A. El-Refae is a Professor. He is president of Al Ain University of Science and Technology in United Arab of Emirates. Ghaleb A. El-Refae, has a Ph. D. and M.A in Financial Economics form USA, M. Sc and B. Sc in Accounting. His research interest is in the application of IT and IS in Business and Economics. He has already published over 30 papers in international journals and conferences

Some Models for Multiple Attribute Decision Making with Intuitionistic Fuzzy Information and Uncertain Weights

Yujun Luo[1], Xianfu Li[2], Ying Yang[1] and Zhenglong Liu[1]

[1] Department of Computer Science and Mathematics, North Sichuan Medical College, Nanchong, Sichuan 637007, China

[2] Department of Imaging Medicine, North Sichuan Medical College, Nanchong, Sichuan 637000, China

Abstract

Multiple attribute decision making problems with uncertain weights in intuitionistic fuzzy setting are investigated. Some concepts related to the theory of intuitionistic fuzzy set (IFS), including intuitionistic fuzzy weighted averaging (IFWA) operator, score function, and accuracy function, are reviewed. Based on the technology for order preference by similarity to idea solution (TOPSIS) method and the score matrix converted from decision matrix given in the form of intuitionistic fuzzy number (IFN), some quadratic programming models, by which the attribute weights can be derived, are established. Then, the alternatives are ranked, and the most desirable one is selected according to the score and accuracy degree of the collective IFN aggregated by IFWA operator. Finally, an example about evaluation of teaching quality is discussed to verify the effectiveness of the proposed approach.

Keywords: *Multiple Attribute Decision Making, Intuitionistic Fuzzy Set, Uncertain Attribute Weights, TOPSIS, Quadratic Programming Models, Evaluation of Teaching Quality.*

1. Introduction

As a generalization of fuzzy set proposed by Zadeh, Atanassov [1, 2] introduced intuitionistic fuzzy set (IFS), which models the various uncertainty or vagueness by membership function and nonmembership function. As a more suitable way to deal with vagueness than classical fuzzy set, IFS plays an important role in solving the complicated multiple attribute decision making (MADM) problems, especially in the circumstances where the assessment information given by decision makers is imprecise or uncertain due to time pressure, lack of data, or the decision maker's limited attention and information processing capabilities [3]. Recently, some researchers addressed the intuitionistic fuzzy MADM problems in the situations where attribute weights are completely known, developed intuitionistic fuzzy aggregation operators [4-9] and proposed score function and accuracy function to solve the intuitionistic fuzzy MADM [10]. The other researchers showed great interest in intuitionistic fuzzy MADM problems with uncertain attribute weights and established some optimization models to derive the vector of attribute weights [11-15].

In this paper, based on the score matrix converted from decision matrix given in the form of IFN by the decision makers, based on the technology for order preference by similarity to idea solution (TOPSIS) method, we establish some quadratic programming models, by which the optimal attribute weights can be derived from the given incomplete information about attribute weights. Then, the procedures for ranking the alternatives in intuitionistic setting are developed.

The remainder of this paper is organized as follows. In section 2, we briefly review some concepts about IFS. In section 3, the MADM problems with attribute assessment given in the form of IFN and attribute weights incompletely known are discussed. Based on TOPSIS and score matrix, some programming models for determining the optimal attribute weights are established. Then, we rank the alternatives and select the most desirable one according to score function and accuracy function of the overall IFN. In section 4, an illustrative example is presented to verify the proposed method. In section 5, we conclude the paper.

2. Preliminaries

Atanassov [1, 2] introduced the concept of IFS, which is defined as follows:

Definition 1 Let X be a finite set. An IFS in X is an object having the form:

$$A = \{\langle x, m_A(x), n_A(x)\rangle | x \in X\}, \quad (1)$$

where $\mu_A(x): X \to [0,1]$ and $v_A(x) \to [0,1]$ are the degree of membership and the degree of nonmembership of the element x to the set X, respectively, such that $0 \le \mu_A(x) + v_A(x) \le 1, \forall x \in X$.

For convenience, we call $a=(m_a,n_a)$ an intuitionistic fuzzy number (IFN), where $\mu_a \in [0,1]$, $v_a \in [0,1]$, and $0 \le \mu_a(x)+v_a(x) \le 1$.

Definition 2 [4] Let $a_i=(m_i,n_i)\,(i=1,2,...,n)$ be a collection of the IFNs. The intuitionistic fuzzy weighted averaging (IFWA) operator is a mapping, which is defined as follows:

$$IFWA_w(a_1,a_2,\mathbf{L},a_n)=\sum_{i=1}^{n} w_i a_i$$
$$=(1-\prod_{i=1}^{n}(1-u_i)^{w_i},\prod_{i=1}^{n} v_i^{w_i}), \quad (2)$$

where $w_i \in [0,1]$ and $\sum_{i=1}^{n} w_i = 1$ is the weight of IFN $a_i\,(i=1,2,\mathbf{L}\,n)$.

Definition 3 [10] Let $a=(m_a,n_a)$ be an IFN, then we call $S(a)=m_a-n_a$ and $H(a)=m_a+n_a$ a score function of a and an accuracy degree function of a, respectively.

Let $S(a)$ and $S(b)$ be the scores of IFN a and b. Let $H(a)$ and $H(b)$ be the accuracy degree of a and b. We can compare two IFNs according to the following principles:

1) If $S(a)<S(b)$, then a is smaller than b, denoted by $a<b$;

2) If $S(a)=S(b)$ and $H(a)<H(b)$, then $a<b$;

3) If $S(a)=S(b)$ and $H(a)=H(b)$, then $a=b$.

3. MADM Problems with Uncertain Attribute Weights in Intuitionistic Fuzzy Setting

Let $A=\{a_1,a_2,\mathbf{L},a_m\}$ be a finite set of alternatives, $C=\{c_1,c_2,\mathbf{L},c_n\}$ be a finite set of attributes, and $W=\{w_1,w_2,\mathbf{L},w_n,\}$ be a finite set of attribute weights. Let $R=(r_{ij})_{m\times n}$ be a decision matrix with attribute assessment values given in the form of IFN, where IFN $r_{ij}=(m_{ij},n_{ij})$ denotes the degree to which the alternative a_i satisfies and does not satisfy the attribute c_j.

Obviously, if the information about attribute weights given by the decision makers is crisp real value, we can weight each attribute value and aggregate all the weighted attribute values corresponding to each alternative into the collective attribute values by formula (2). According to the score and accuracy degree of the collective values, we can rank the alternatives and choose the most desirable one. In the real world, however, it is very often that the information about attribute weights is incompletely known because of the inherent complexities of the MADM problems, time pressure or lack of knowledge. Let W be the set of the known information about attribute weights, which can be constructed by the following forms [16]:

1) A weak ranking: $w_i \ge w_j$;

2) A strict ranking: $w_i - w_j \ge d_i,\ d_i > 0$;

3) A ranking of differences: $w_i - w_j \ge w_k - w_l$, for $j \ne k \ne l$;

4) A ranking with multiples: $w_i \ge k_i w_j,\ 0 \le k_i \le 1$;

5) An interval form: $0 \le d_i \le w_i \le d_i + e_i \le 1$.

In the MADM problems with incomplete attribute weights, before choosing the most desirable one among the candidate alternatives, we must determine the weight of each attribute. In the following, let us construct some optimization models based on TOPSIS method to derive the optimal attribute weights from uncertain information about attribute weights.

Step 1 Transform the intuitionisitc fuzzy decision matrix into the score matrix.

Base on the score function of IFN given in definition 3, we can transform the IFN decision matrix into the score matrix $S=(s_{ij})_{m\times n}$, where $s_{ij}=m_{ij}-n_{ij}$.

Step 2 Decide the positive idea solution and the negative idea solution.

Based on the score matrix, the positive idea solution S^+ and the negative idea solution S^- can be written as follows:

$$S^+=(s_1^+,s_2^+,\mathbf{L},s_n^+), \quad (3)$$
$$S^-=(s_1^-,s_2^-,\mathbf{L},s_n^-), \quad (4)$$

where $s_j^+=\max_i(s_{ij})$ and $s_j^-=\min_i(s_{ij})$.

Step 3 Calculate the distance between each alternative and the positive/negative idea solution.

The distance between alternative a_i and the positive idea solution S^+ can be calculated as follows:

$$d_i^+ = \sqrt{\sum_{j=1}^{n}\left[w_j\left(s_{ij} - s_j^+\right)\right]^2}. \quad (5)$$

The distance between alternative a_i and the negative idea solution S^- can be calculated as follows:

$$d_i^- = \sqrt{\sum_{j=1}^{n}\left[w_j\left(s_{ij} - s_j^-\right)\right]^2}. \quad (6)$$

Step 4 Decide attribute weights from the given information about attribute weights.

According to the traditional TOPSIS method, the shorter the distance between each alternative and the positive ideal solution is, the better the alternative is; the longer the distance between each alternative and the negative ideal solution is, the better the alternative is. Therefore, for each alternative, we can construct the multi-objective programming models (7) and (8) to maximize d_i^+ and minimize d_i^- ($i = 1,2,...m$).

$$\begin{cases} \min d_i^+(w) = \sqrt{\sum_{j=1}^{n}\left[w_j\left(s_{ij} - s_j^+\right)\right]^2}, i = 1,2,...m \\ \text{s.t. } w = (w_1, w_2, ..., w_n) \in W \end{cases}, \quad (7)$$

$$\begin{cases} \max d_i^-(w) = \sqrt{\sum_{j=1}^{n}\left[w_j\left(s_{ij} - s_j^-\right)\right]^2}, i = 1,2,...m \\ \text{s.t. } w = (w_1, w_2, ..., w_n) \in W \end{cases}. \quad (8)$$

The models (7) and (8) can be reduced to the models (9) and (10).

$$\begin{cases} \min D_i^+(w) = \sum_{j=1}^{n}\left[w_j\left(s_{ij} - s_j^+\right)\right]^2, i = 1,2,...m \\ \text{s.t. } w = (w_1, w_2, ..., w_n) \in W \end{cases}, \quad (9)$$

$$\begin{cases} \max D_i^-(w) = \sum_{j=1}^{n}\left[w_j\left(s_{ij} - s_j^-\right)\right]^2, i = 1,2,...m \\ \text{s.t. } w = (w_1, w_2, ..., w_n) \in W \end{cases}. \quad (10)$$

Because we have none of the preference for any alternative, the models (9) and (10) can be transformed into the two single objective programming models as follows:

$$\begin{cases} \min D^+(w) = \sum_{i=1}^{m}\sum_{j=1}^{n} w_j^2\left(s_{ij} - s_j^+\right)^2 \\ \text{s.t. } w = (w_1, w_2, ..., w_n) \in W \end{cases}, \quad (11)$$

$$\begin{cases} \max D^-(w) = \sum_{i=1}^{m}\sum_{j=1}^{n} w_j^2\left(s_{ij} - s_j^-\right)^2 \\ \text{s.t. } w = (w_1, w_2, ..., w_n) \in W \end{cases}. \quad (12)$$

Combining model (11) and (12), we can get the model (13), which is a quadratic programming model.

$$\begin{cases} \min D(w) = \sum_{i=1}^{m}\sum_{j=1}^{n} w_j^2[(s_{ij} - s_j^+)^2 \\ \qquad -(s_{ij} - s_j^-)^2] \\ \text{s.t. } w = (w_1, w_2, ..., w_n) \in W \end{cases}. \quad (13)$$

Solving the quadratic programming model (13), we can get the vector of the attribute weights.

Step 5 Aggregate the attribute assessment values of each alternative into collective attribute value. Then calculate the score of collective attribute value corresponding to each alternative and order the alternatives.

4. Evaluation of Teaching Quality by the Proposed Approach

Suppose that we want to solve an evaluation of teaching quality problems in which the alternatives are four young teachers to be evaluated ($a_1 \sim a_4$) according to their teaching performances by the expert committee. The evaluation system includes the four indexes: teaching altitude (c_1), teaching content (c_2), teaching method (c_3) and teaching result (c_4). The assessment information provided in the form of IFN in the following denotes the membership degree and nonmenbership degree to which the alternatives corresponding to each attribute belong to the fuzzy concept “excellence”.

$$\begin{pmatrix} (0.80,0.10) & (0.75,0.20) & (0.80,0.05) & (0.85,0.10) \\ (0.65,0.25) & (0.85,0.05) & (0.80,0.05) & (0.75,0.20) \\ (0.75,0.20) & (0.75,0.15) & (0.85,0.10) & (0.70,0.25) \\ (0.80,0.15) & (0.70,0.20) & (0.75,0.10) & (0.75,0.15) \end{pmatrix}.$$

Step 1 Get the score decision matrix according to score function.

According to definition 3, we can transform the decision matrix with intuitionistic fuzzy information into the score matrix shown as follows:

$$\begin{pmatrix} 0.70 & 0.55 & 0.75 & 0.75 \\ 0.40 & 0.80 & 0.75 & 0.55 \\ 0.55 & 0.60 & 0.75 & 0.45 \\ 0.65 & 0.50 & 0.65 & 0.60 \end{pmatrix}.$$

Step 2 Get the positive idea solution and the negative idea solution.

The positive idea solution S^+ and the negative idea solution S^- can be calculated as follows:

$$S^+ = (0.70,0.80,0.75,0.75),$$
$$S^- = (0.40,0.50,0.65,0.45).$$

Step 3 Construct the optimization model to decide the vector of attribute weights.

Suppose that the known information about attribute weights is given in the form of the following:

$$W = \{\ 0.2 \le w_1 \le 0.3,\ 0.15 \le w_2 \le 0.3,$$
$$w_2 - w_1 \ge 0.05, w_4 - w_3 \ge 0.1, w_3 \ge 0.5w_2.$$
$$\sum_{j=1}^{4} w_j = 1, w_j \ge 0,\ (\mathrm{j}=1,2,3,4)\ \}$$

According to model (13), we can construct the quadratic programming model (14)

$$\begin{cases} \min\ D(w) = -0.06w_1^2 + 0.09w_2^2 \\ \qquad\qquad -0.02w_3^2 + 0.03w_4^2 \\ \text{s.t. } 0.2 \le w_1 \le 0.3, 0.15 \le w_2 \le 0.3, \\ \quad w_3 \ge 0.5w_2, w_2 - w_1 \ge 0.05, \\ \quad w_4 - w_3 \ge 0.1, \\ \quad \sum_{j=1}^{4} w_j = 1, w_j \ge 0, (j = 1,2,3,4) \end{cases} . \qquad (14)$$

Solving model (14), we can get the vector of attribute weights w = (0.200, 0.250, 0.225, 0.325).

Step 4 Calculate the collective values of each alternative and rank the alternatives according to their scores.

The collective values of the alternatives in the form of IFN are a_1: (0.8074, 0.1017), a_2: (0.7762, 0.1083), a_3: (0.7635, 0.1712), a_4: (0.7498, 0.1471).

The scores of collective attribute value of each alternative are:

$$s_1 = 0.7057, s_2 = 0.6679, s_3 = 0.5923, s_4 = 0.6027.$$

Since $s_1 > s_2 > s_4 > s_3$, hence, $a_1 \succ a_2 \succ a_4 \succ a_3$ and the most desirable alternatives is a_1.

5. Conclusions

In this paper, we investigate the MADM problems with attribute values given in the form of IFN and uncertain attribute weights. In order to derive attribute weights, some quadric programming models based on TOPSIS method are constructed. We aggregate the IFNs by IFWA operator into the collective values, based on which, the alternatives are ranked, and the most desirable one is chosen. The proposed models can be extended to solve the MADM problems with interval-valued intuitionistic fuzzy information and uncertain attribute weights.

Acknowledgments

This work was supported by Sichuan Center for Education Development Research (CJF10018) and Sichuan Provincial Education Commission (11ZA214).

References

[1] K. T. Atanassov, "Intuitionistic fuzzy sets", Fuzzy sets and systems, Vol. 20, No. 1, 1986, pp. 87-96.
[2] K. T. Atanassov, "More on intuitionistic fuzzy sets", Fuzzy sets and systems, Vol. 33, 1989, pp. 37-45.
[3] Z. S. Xu, and R. R. Yager, "Dynamic intuitionistic fuzzy multi-attribute decision making", International Journal of Approximate Reasoning, Vol. 48, 2008, pp. 246-262.
[4] Z. S. Xu, "Intuitionistic fuzzy aggregation operators", IEEE Transactions on Fuzzy Systems, Vol. 15, 2007, pp. 1179-1187.
[5] Z. S. Xu, and M. M. Xia, "Induced generalized intuitionistic fuzzy operators", Knowledge-Based Systems, Vol. 24, 2011, pp. 197-209.

[6] W. Yang, and Z.P. Chen, "The quasi-arithmetic intuitionistic fuzzy OWA operators", Knowledge-Based Systems, Vol. 27, 2012, pp. 219-233.
[7] M. M. Xia, Z. S. Xu, and B. Zhu, "Generalized intuitionistic fuzzy Bonferroni means", International Journal of Intelligent Systems, Vol. 27, 2012, pp.23-47.
[8] G. W. Wei. "Some induced geometric aggregation operators with intuitionistic fuzzy information and their application to group decision making", Applied Soft Computing, Vol. 10, 2010, pp. 423-431.
[9] G. W. Wei, X. F. Zhao, R. Lin. "Some Induced Aggregating Operators with Fuzzy Number Intuitionistic Fuzzy Information and their Applications to Group Decision Making", International Journal of Computational Intelligence Systems, Vol. 3, 2010, pp. 84-95.
[10] D. H. Hong, and C. H. Choi, "Multicriteria fuzzy decision-making problems based on vague set theory", Fuzzy Set and Systems, Vol. 114, 2000, pp. 103-113.
[11] D. F. Li, Y. C. Wang, S. Liu, and F. Shan, "Fractional programming methodology for multi-attribute group decision-making using IFS", Applied Soft Computing, Vol. 9, 2009, pp. 219-225.
[12] G. W. Wei, "Maximizing deviation method for multiple attribute decision making in intuitionistic fuzzy setting", Knowledge-Based Systems, Vol. 21, 2008, pp. 833-836.
[13] G. W. Wei, "GRA method for multiple attribute decision making with incomplete weight information in intuitionistic fuzzy setting", Knowledge-Based Systems, Vol. 23, 2010, pp. 243-247.
[14] G. W. Wei, "Gray relational analysis method for intuitionistic fuzzy multiple attribute decision making", Expert Systems with Appllications, Vol. 38, 2011, pp. 11671-11677.
[15] Z. J. Wang, K. W. Li, and W. Z. Wang, "An approach to multiattribute decision making with interval-valued intuitionistic fuzzy assessments and incomplete weights", Information Sciences, Vol. 179, 2009, pp. 3026-3040.
[16] Z. S. Xu, "Models for multiple attribute decision making with intuitionistic fuzzy information", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 15, 2007, pp. 285-297.

Yujun Luo received B.S. degree from Sichuan Normal University in 1996 and M. S. degree from Chengdu University of Technology in 2002. He currently works as an associate professor at Department of Computer Science and Mathematics, North Sichuan Medical College, Nachong, China. His current research interests include uncertain multiple attribute decision making with fuzzy information and computer applications.

Xianfu Li received B.S. degree from Sichuan Normal University in 1996 and M.S. degree from China West Normal University in 2008. He currently works as an associate professor at Department of Imaging Medicine, North Sichuan Medical College, Nachong, China. His current research interest is digital image processing in three-dimension conformal radiation therapy for tumor.

Ying Yang received M.S. degree from China West Normal University in 2006. She currently works as a lecturer at Department of Computer Science and Mathematics, North Sichuan Medical College, Nachong, China. Her current research interest is computer applications.

Zhenglong Liu received B.S. degree from China West Normal University in 2000 and M.E. degree from Sichuan University in 2008. He currently works as an associate professor at Department of Computer Science and Mathematics, North Sichuan Medical College, Nachong, China. His current research interests include intelligent optimization algorithms, computer applications and mathematical methods in biomedicine.

Review of Intelligent Techniques Applied for Classification and Preprocessingof Medical ImageData

Hota H.S.[1], Shukla S.P.[2] and Gulhare Kajal Kiran[3]

Abstract

Medical image data like ECG, EEG, MRI and CT-scan images are the most important way to diagnose disease of human being in precise way and widely used by the physician. Problem can be clearly identified with the help of these medical images. A robust model can classify the medical image data in better way .In this paper intelligent techniques like neural network and fuzzy logic techniques are investigated for MRI medical image data to identify tumor in human brain. Also need of preprocessing of medical image data is explored. Classification technique has been used extensively in the field of medical imaging. The conventional method in medical science for medical image data classification is done by human inspection which may result misclassification of data sometime this type of problem identification are impractical for large amounts of data and noisy data, a noisy data may be produced due to some technical fault of the machine or by human errors and can lead misclassification of medical image data. We have collected number of papers based on neural network and fuzzy logic along with hybrid technique to explore the efficiency and robustness of the model for brain MRI data. It has been analyzed that intelligent model along with data preprocessing using principal component analysis (PCA) and segmentation may be the competitive techniques in this domain.

Keywords: *Magnetic Resonance Imaging (MRI), Intelligent Techniques Artificial Neural Network (ANN), Fuzzy Logic (FL), Principal component analysis (PCA).*

1. Introduction

Automatic detection of any problem persist in medical image data attracted and motivated the researchers to design and develop a decision support system (DSS) to assist physician in the decision making process[32]. A DSS to support the physician can be developed using various intelligent techniques like Artificial Neural Network, Fuzzy logic and genetic algorithm with hybridization of all these techniques beside this various techniques are also used as a preprocessing of the medical image data using PCA andwavelet transformation. A DSS model developed with the help of intelligent techniques are basically a classifier to classify the data either as a normal data or abnormal data (Medical Image data with abnormality) as shown in Figure1.

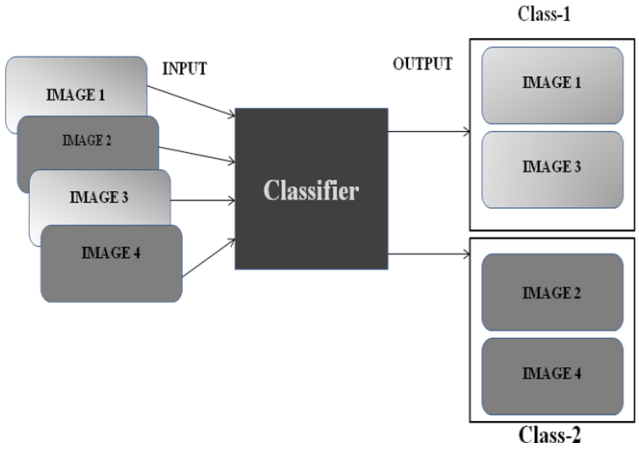


Fig. 1 Image Classification Process

Model receives the data as image 1, image 2, image 3 and image 4 where as image 1 and image 3 belongs to class 1 where as image 2 and image 4 belongs to class 2.A high classification accuracy is required from the model because correct diagnosis of the problem is essential for correct treatment and medication, since it is directly related to a life of human being.

Recent trend of classification of medical image data that is being utilized by many researchers are the intelligent hybrid techniques which can be developed with the help of artificial neural network, fuzzy logic and genetic algorithm.

In this paper we have explores various techniques with their findings for classification of various medical image data. A hybrid model will be beneficial in terms of accuracy by reducing drawbacks of individual techniques. A preprocessing of medical image data is essential due to high dimensionality of features present in image data, we have also explored and analyzed various techniques popularly used for preprocessing and segmentation of image data.

Various medical image data which are considered in this paper to study intelligent model developed by various authors are as follows:

1.1 Magnetic Resonance Imaging (MRI)

Magnetic resonance imaging (MRI) of different parts of the human body like brain and spinal is a safe and painless test that uses a magnetic field and radio waves to produce detailed images. A brain MRI image is very helpful to identify disease related to brain like tumor and the brain stem. MRI uses a powerful magnetic field, radio frequency pulses and a computer to produce detailed pictures of organs, soft tissues, bone and virtually all other internal body structures. MRI can detect a variety of conditions of the brain such as cysts, tumors, bleeding, swelling, developmental and structural abnormalities, infections, inflammatory conditions, or problems with the blood vessels.

1.2 Electroencephalogram (EEG)

Electroencephalogram or EEG is used for measuring electrical activities of the brain. EEG is mainly used for diagnosing seizure disorders, infections, tumors degenerative disorders and metabolic disturbances affecting the brain. EEG testing comes with certain adverse conditions.

2. Research Methodology

Various intelligent techniques are being used by the researchers to classify and segmentation of medical image data especially MRI data to detect abnormalities found in different parts of the human body, this study confined to utilization of these techniques for classification and segmentation of medical image data as special case. These techniques are explained below:

2.1 Fuzzy Logic

Zadeh [14] introduced the fuzzy set theory; a major contribution of fuzzy set theory is its capability of representing vague data. Fuzzy sets and fuzzy logic are powerful mathematical tools for modeling: uncertain systems in industry, nature and humanity, and facilitators for common-sense reasoning in decision making in the absence of complete and precise information. A fuzzy number is characterized by a given interval of real numbers, each with a grade of membership between 0 and 1.Fuzzy logic based clustering technique is frequently utilized for classification of medical image data.

2.2 Artificial Neural Network (ANN)

Artificial neural network (ANN) [11] is an interconnected group of natural or artificial neurons that uses a mathematical or computational model for information processing. Some of the architectures of ANN are explained below:

2.2.1 Multi Layer Neural Network (MLNN)

Multilayer Neural Networks [11] solve the classification problem for non linear sets by employing hidden layers, whose neurons are not directly connected to the output. The additional hidden layers can be interpreted geometrically as additional hyper-planes, which enhance the separation capacity of the network. Multi layer neural network is mostly used for classification of different categories of data .A popularly used MLNN is back propagation network (BPN) with gradient descent.

Back propagation artificial neural network (BPANN) is a neural network technique which is able to train nonlinear data and is based on gradient descent. This network is trained with popular error back propagation algorithm (EBPA). This algorithm has two passes: feed forward phase in which output is calculated and feed backward phase in which the calculated error is propagated back to the network to adjust the weights.

2.2.2 Polynomial Neural Network (PNN)

Polynomial neural networks (PNN) are multilayer perceptrons of neuron-like units which produce high order multivariate polynomial mappings. These are tree structured hierarchical cascades of first-order and second order activation polynomials in the nodes, and input variables passed from the leaves. The activation polynomial outcomes are fed forward to their parent nodes, where partial polynomial models are made.

2.2.3 Radial Basis Function Neural Network (RBFNN)

Radial basis functions [11] are powerful techniques for interpolation in multidimensional space. A RBF is a function which has built into a distance criterion with respect to a center.Radial basis function (RBF) networks are feed-forward networks trained using a supervisedtraining algorithm. It has single hidden layer generally with special type of activation function known as basis functions one can use a suitable basis function like radial basis, polynomial , sigmoid or linear basis function

as per suitability of data pattern. These are also known as kernel type and can be changed to tune the network.

2.3 Hybrid Techniques

To overcome problem of individual techniques hybridization is required, a suitable hybrid technique [8] with combination of two or more intelligent techniques like Neuro-Fuzzy, Neuro-Genetic or Neuro-Fuzzy-Genetic can be utilized. Authors [8] are currently using hybrid techniques for medical image data classification. Some very well known hybrid techniques are explained below:

2.3.1 Neuro-Fuzzy Technique

A Fuzzy Neural Network or Neuro-Fuzzy System [4] is a learning machine that finds the parameters of a fuzzy system by exploiting approximation techniques from neural networks. This means that the main intention of Neuro-Fuzzy approach is to create or improve a fuzzy system automatically by means of neural network methods. A Neuro-Fuzzy system based on an underlying fuzzy system is trained by means of a data-driven learning method derived from Neural Network theory. It can be represented as a set of fuzzy rules at any time of the learning process, i.e. before, during and after. Thus the system might be initialized with or without prior knowledge in terms of fuzzy rules. The learning procedure is constrained to ensure the semantic properties of the underlying fuzzy system. A Neuro-Fuzzy network is a fuzzy inference system in the body of an artificial neural network. Depending on the Fuzzy Inference System (FIS) type, there are several layers that simulate the processes involved in a fuzzy inference like fuzzification, inference, aggregation and defuzzification.

2.3.2 Adaptive Neuro-Fuzzy Inference System (ANFIS)

ANFIS, developed by Jang [13] is an adaptive network incorporates the concept of fuzzy logic into the neural networks, and has been widely used in many applications. ANFIS largely removes the requirement for manual optimization of the fuzzy system parameters. An adaptive network is network of nodes and directional links. Associated with the network is a learning rule - for example back propagation. It's called adaptive because some, or all, of the nodes have parameters which affect the output of the node. These networks are learning a relationship between inputs and outputs. By using a hybrid learning procedure, the proposed ANFIS can construct an input-output mapping based on both human knowledge (in the form of fuzzy if-then rules) and stipulated input-output data pairs.

2.3.3 Neuro-Genetic Technique

The Neuro-Genetic [19] model is a hybrid model which exhibits the characteristics of both ANN and GA. It can be used as the tool for decision making in order to solve the complex nonlinear problems. In this method first we define a network structure with a fixed number of inputs, hidden nodes and outputs. Second we employed the GA in the learning phase of the network, as it is capable to search in a large search space. The hybridization of ANN and GA is able to select the optimal weight sets as well as the bias value for the classification.

3. Process of Medical Image Data Classification

Medical image data classification using intelligent techniques are very important and useful technique to detect or diagnose critical disease like brain tumor. The recent research for medical image classification uses different individual and hybrid techniques. A detail outline of the phases related to this is depicted in Figure 2 and explained as below:

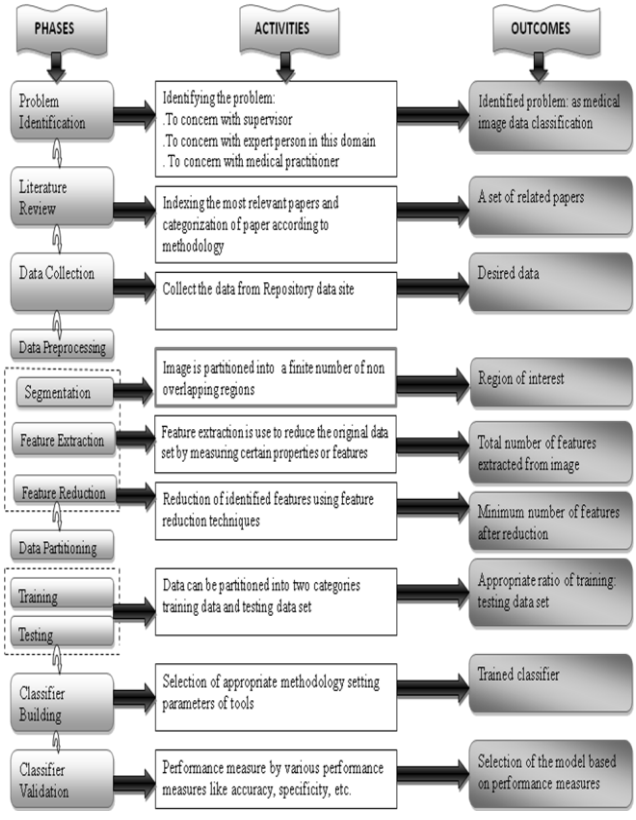


Fig. 2 Phases of medical image data classification

Phase 1: Problem Definition

We need to identify problem to work with there may be various dimension of the work to be done related to medical image data it may be data preprocessing through segmentation or feature selection or may be classification of medical image data or combination of both. Defining problem with clear objective provides strong base for the rest of the process. Classification of medical image data is one of the important areas for research work. In this phase, we need to identify the problem with the help of concern supervisor, expert persons and medical practitioners.

Phase 2: Literature Review

This phase helps us to analyze the research work already done in this domain and to find out new directions of the work to be done which consists collecting, analyzing and indexing relevant papers related to the objective. Number of research papers related to the medical image data used for segmentation and classification to diagnose problem are collected and arranged. Works done by different authors along with their findings are explained in more detail as below:

Artificial Neural Network is one of the most powerful and widely used intelligent technique provides strong feature to classify medical image data [32], there are various Neural Network which can be categorized as supervised and unsupervised neural network model like back propagation, probabilistic neural network, radial basis neural network and self organizing map (SOM) as explained above. MohdFauzi Othman et al., Ehab F. Badran et al., and Vinodkumar et al. [17][7][29] have used different types of artificial neural network like probabilistic neural network which is a type neural network mostly used for classification problems having good classification accuracy ,in these research work principal component analysis (PCA) and wavelet transformation have been used for feature extraction from brain MRI data to identify brain tumor. Probabilistic Neural Network: Another work done by Adams [1] for the brain MRI medical image data with the help of ANN in all the above work classification accuracy was found satisfactory.

Different types of Artificial Neural Network is also combined with various feature extraction techniques like wavelet transform by [20][8][24][23] in which feature extraction is done using discrete wavelet transformation (DWT) while PCA is used as feature reduction, they have compared the performance of two neural network based technique : Feed forward back propagation neural network (FP-NN) and K-nearest neighborhood (K-NN) with discrete wavelet transform as feature extraction .A comparative result of both the techniques shows that K-NN with DWT produces better classification accuracy as compare to FF-NN. Brain tumor classification is also performed [27][16] using EEG image data using wavelet and multi layer neural network, they found that performance of the model in detecting the brain tumor using EEG is very much encouraging. Other authors [2] also studied and implemented artificial neural network using Gaussian decomposition as a preprocessing of MRI data.

On the other hand unsupervised neural network like SOM along with above extraction techniques is utilized by [26] authors have proposed a special case of SOM known as hierarchical SOM (HSOM) for detection and characterization of brain tumor.

Neural network is combined with another intelligent technique fuzzy logic. Authors [6] have used Neuro-Fuzzy technique for classification normal and abnormal brain images.

Image segmentation is one of the broader areas of research in medical image data as a preprocessor in which image is partitioned into a finite number of non overlapping regions with respect to some characteristics to find out region of interest (ROI) before feeding the data to the classifier model. Many authors have worked in this area to segment medical image data before feature extraction and feature reduction. J.K. Singh and et al. [12] have proposed self adaptive RBF network based segmentation of medical images of the brain; images are segmented into three different regions. A fuzzy hopefield neural network is used by [9] for MRI image segmentation and achieved good classification accuracy of the three segmented regions.

Fuzzy C-means (FCM) algorithm is a common clustering algorithm used for segmentation of MRI images .Authors [31] have used FCM with its improved version (IFCM) by introducing two new parameters, these parameters are computed with the help of ANN and genetic algorithm (GA) a similar type of work is done by [22] in which optimization of the parameters is done with the help of particle swarm optimization (PSO) instead of GA. A genetic fuzzy based segmentation is proposed by [15]. A neural model and fuzzy model is compared for brain MRI image segmentation by [4][6] they have used neural network technique :linear vector quantization **(**LVQ**)** and fuzzy logic technique FCM .It is concluded that accuracy of neural classifier is more than fuzzy classifier.

FCM is also used by [21] for brain image segmentation and then ANN is trained with fuzzy back propogationalgorithm.FCM is also combined with artificial ant colony (ABC) algorithm to improve its accuracy .Authors have compared their result with other combination of segmentation techniques like combined FCM with PSO and GA and they proved that FCM with ABC is producing better result compare to others. A fuzzy approach is also used with active surface model **(**ASM**)** by

[3] for infantile brain MRI classification with the help of fuzzy rule base.

A hybrid unsupervised neural network model is proposed for segmentation by [18][25] using SOM and fuzzy adaptive resonance theory (ART).These two techniques have been used in sequential manner first SOM is applied and then output of SOM is presented to fuzzy ART.
Support vector machine (SVM) is based on the statistical learning theory founded by Vapnik [28] .The main idea of SVM is to map multi dimension data to more multi dimension but linear dividable space and a linear classifier can perform classification task.

In recent years Xinyu et al. [30] have used SVM for segmentation of MRI image another authors [10] also used least square SVM for brain MRI slices.

Phase 3: Data Collection

Medical Image data must be collected for training and testing the model proposed for MRI image classification either from some repository sites or from some hospitals. A sufficient number of data is required to train and test the model successfully.

Phase 4: Data Preprocessing

Data preprocessing is an important and essentials for MRI image data due to noise and other raw information inserted in images and due to high dimensionality of data. Data preprocessing may have three sub phases: Segmentation, feature extraction and feature reduction. In segmentation region of interest (ROI) is selected with important parts of image then various features of the segmented image are extracted and at last irrelevant, incomplete, noisy and inconsistent features available in segmented image are eliminated from the image and subset of features are selected through feature reduction techniques. The prime objective of preprocessing is to improve the image data quality by suppressing undesired distortions and enhancing the required image features for further processing.

Phase 5: Data Partitioning

An optimum size of data is partitioned into different partitions: Training and testing. The ratio of partition may be 70%:30% or 60%: 40% respectively for training and testing. Partition size of training and testing data also played crucial role to provide high classification accuracy of medical image data, therefore a suitable partition of training and testing samples are necessary.

Phase 6: Classifier Building

A classifier model can be developed using intelligenttechniques like ANN, Fuzzy logic and optimization algorithm as discussed above. Model will be utilized to classify medical image data either as abnormal case or normal case .We need to train the model developed with the help of training data set .In the traditional classification approach single classification methods like Artificial neural network (ANN) or Fuzzy Logic has been used like, whereas in the recent years hybrid of various intelligent techniques are being used.

Phase 7: Model Validation

To check the robustness of the model testing data set is used. A set of rules will apply to check the validation of classifier and test the proposed modelby various performance measures like accuracy, specificity and sensitivity.

4. Conclusion

Medical image data classification using intelligent techniques is essentials for appropriate decision making process by the physician as a decision support system. Literature shows that due to high dimensionality and noise in the data it is necessary to preprocess before feeding it to the models. Segmentation, feature extraction and feature reduction are the three stages found in all most all the research work. Most of the authors have used PCA and other intelligent techniques for medical data preprocessing .In this review work mostly MRI image of human brain has been considered for experimental purpose. This review also concluded that artificial neural network is a promising technique for medical image data classification however in very few literatures other than ANN techniques have been used by the authors. ANN is also combined with fuzzy logic to develop hybrid classification model for human brain MRI data classification. Classification accuracy achieved in all these research work is satisfactory however by using and integrating some other techniques accuracy can be improved.

References

[1]Adam D.B., Gade S.S. and et.al, “Neural Network Based Brain Tumor Detection using MR images”, International Journal of Computer Science and Communication (IJCSC), Vol.2 , No.2,July, December 2011.
[2]Carlos Arizmendi,Daniel, A.Siena and et.al “Brain Tumour classification using Gaussian Decomposition and neural networks”, AIC(Annual international conference) of the IEEE 2011.

[3]Chuan-Yu Chang, Da-FengZhuang and et.al, "A Fuzzy-Based Learning Vector Quantization Neural Network for Recurrent Nasal Papilloma Detection",IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS, VOL. 54, NO. 12, DECEMBER 2007.
[4] D.Jude Hemanth1, C.KeziSelva, and et.al, "Comparative Analysis of Neural Model and Fuzzy Model for MR Brain Tumor Image Segmentation", IEEE, 2009.
[5]D.JudeHemanth,C.KeziSelvaVijilaandet.al"Application of Neuro-Fuzzy Model for MR Brain Tumor Image", Biomedical Soft Computing and Human Sciences, Vol.16,No.1,pp.95-102, 2009.
[6]Dipali M. Joshi, Dr.N. K. Rana and et.al, "Classification of Brain Cancer Using Artificial Neural Net work", IEEE, 2010.
[7]Ehab F. Badran, EsraaGalal Mahmoud, and et.al, "An Algorithm for Detecting Brain Tumors in MRI Images", IEEE, 2010.
[8]EL-Sayed, A. EL-Dahshan, and et.al, "A Hybrid Technique for Automatic MRI Brain Images Classification", Studia Univ. BABES_BOLYAI, INFORMATICA, Volume LIV, Number 1, 2009.
[9]GholamaliRezai-Rad and Reza ValipourEbrahimi and et.al, "Modified Fuzzy Hopfield Neural Network Using for MRI Image segmentation", Research Publishing Services (RPS), 2006.
[10] H. Selveraj, S.Thamarai, and et.al "Brain MRI Slices classification using least square support vector machine,(ICMED),vol.1,No.1,issue1, 2007.
[11] Haykin, S. (1999) Neural Networks: A Comprehensive Foundation, Prentice Hall
[12] J. K. Sing, D. K. Basu, and et.al, "Self-Adaptive RBF Neural Network-Based Segmentation of Medical Images of the Brain", Proceedings of ICISIP, 2005.
[13] Jang, Sun, Mizutani(1997) Neuro-Fuzzy and Soft Computing-Prentice Hall,pp 335-368, ISBN 0-13-261066-3.
[14]L.A. Zadeh, "Some reflection on soft computing, granular computing and their roles in the conception, design and utilization of information/ intelligent system", Soft computing, vol. 2, 1998.
[15]M. Hasanzadeh, S. Kasaei and et.al, "Multispectral Brain MRI Segmentation Using Genetic Fuzzy Systems", IEEE, 2007.
[16]M. Murugesan and Dr. (Mrs.)R.Sukanesh and et.al, "Automated Detection of Brain Tumor in EEG Signals Using Artificial Neural Networks", International Conference on Advances in Computing, Control, and Telecommunication Technologies, 2009.
[17] MohdFauzi Othman, Mohd Aniffanan and et.al "probabilistic Neural Network forBrain Tumor Classification", Second International conference on Intelligent system (ICIS), Modeling & simulation, IEEE, 2011.
[18Momoyo Ito, Kazuhito Sato and et.al, "Brain Tissues Segmentation for Diagnosis of Alzheimer-Type Dementia", Nuclear Science Symposium Conference Record (NSSCR) 978-1-4673-0120-6/111, IEEE, 2011.
[19N.Benamrane, A. Aribi, L.Kraoula and et. al, "Fuzzy Neural Networks and Genetic Algorithms for Medical Images Interpretation".IEEE Proceedings of the Geometric Modeling and Imaging, 2006.
[20]N. HemaRajini, R.Bhavani , and et.al "Classification of MRI Brain Images using k-Nearest Neighbor and Artificial Neural Network", IEEE-International Conference on Recent Trends in Information Technology, MIT, Anna University, Chennai. June 3-5, 2011, ICRTIT 978-1-4577-0590-8/11, IEEE 2011.
[21]NanditaPradhan and A.K. Sinha and et.al, "Intelligent computing for the analysis of Brain Magnetic Resonance Images", International Conference on Integrated Intelligent Computing, 2010.
[22] NosratallahForghani, MohamadForouzanfarand et.al, "MRI Fuzzy segmentation of Brain Tissue Using IFCM Algorithm with particle swarm optimization", 1-4244-1364-8/07, IEEE, 2007.
[23]Ramakrishnan, Selvanand et.al, "Classification of Brain Tissues Using Multiwavelet Transformation and Probabilistic Neural Network", I.J. of SIMULATION Vol. 7 No. 9 ISSN 1473-804x online, 1473-8031, 2006.
[24]ShahlaNajafi, Mehdi ChehelAmirani and et.al, "A New Approach to MRI Brain Images Classification", 2011.
[25]ToshimitsuOtani, Kazuhito Sato, and et.al, "Segmentation of Head MR Images Using Hybrid Neural Networksof Unsupervised Learning", 978-1-4244-8126-2/10, IEEE, 2010.
[26]V.P.Gladis, PushpaRathi, Dr.S.Palani and et.al,"Detection and Characterization of Brain Tumor Using Segmentation based on HSOM, Wavelet packet feature spaces and ANN",978-1-4244-8679-3/11, IEEE, 2011.
[27] V. SalaiSelvam and S. Shenbagadevi, and et.al, "Brain Tumor Detection using Scalp EEG with Modified Wavelet-ICA and Multi Layer Feed Forward Neural Network",IEEE,September 3, 2011.
[28] Vapnik, V. The Nature of Statistical Learning Theory. New York, NY: Springer-Verlag, 1995.
[29]Vinod Kumar, NiranjanKhandelwal and et.al. "Classification of Brain Tumors using PCA-ANN", 978-1-4673-0126-8/11, IEEE 2011.
[30]Xinyu Du, Yongjie Li, Dezhong Yao and et.al, "A Support Vector Machine Based Algorithm for Magnetic Resonance Image Segmentation", 978-0-7695-3304-9/08, IEEE DOI, 2008.
[31] YounessAliyariGhassabeh, NosratallahForghani and et.al, "MRI Fuzzy Segmentation of Brain Tissue Using IFCM Algorithm with Genetic Algorithm Optimization", IEEE, 2007.
[32]Bhaiya L. and Hota H.S "Diagnosis of Brain Tumor with Artificial Neural Network Using MRI Images : An Investigation " to be published in international conference Sasstartha to be held at RCET ,Durg during 8-9 Feb. 2013

Hota H.S.is currently working as Assistant Professor in Department of Computer Science Guru GhasidasUniversity(GGU), Bilaspur(C.G.), India.
Shukla S.P.is currently working as Professor in Department of Electrical Engineering Bhilai Institute of Technology (BIT), Durg(C.G.), India.
GulhareKajalkiranis currently working as Assistant Professor in Department of Computer Science C.M.D. College, Bilaspur(C.G.), India.

Research on Spatial Estimation of Soil Property Based on Improved RBF Neural Network

Jianbo Xu[1], Quanyuan Tan[2*], Lisheng Song[1], Kai Hao[1], Ke Xiao[1]

[1] College of Infoematics, South China Agricultural University, Guangzhou, 510642, China
[2] Hunan City University, Yiyang, 413000, China

Abstract

To seek optimal network parameters of Radial Basis Function (RBF) Neural Network and improve the accuracy of this method on estimation of soil property space, this study utilizes genetic algorithm to optimize three network parameters of RBF Neural Network including the number of hidden layer nodes, expansion speed and root-mean-square error. Then, based on optimized RBF Neural Network, spatial interpolation is conducted for arable soil property under different sampling scales in the study area. The estimation result is superior to RBF Neural Network method without optimization and geostatistical method in terms of the fitting capacity and interpolation accuracy. Compared with the result of space estimation by RBF Neural Network method without optimization, among the 5 schemes, the forecast errors of RBF Neural Network optimized by genetic algorithm reduce greatly. Mean absolute error (MAE) reduces 0.4868 on the average and root-mean-square error (RMSE) reduces 1.492 on the average. Therefore, RBF Neural Network method optimized by genetic algorithm can gain the information about regional soil property spatial variation more accurately and provides technical support for arable land quality evaluation, accurate farmland management and rational application of fertilizer.

Keywords: *Genetic algorithm, RBF Neural Network, Spatial forecast, Error analysis*

1. Introduction

Soil is the loose surface with certain fertility covering the earth surface on which the plans can grow. Soil is formed through combined actions of parent material, climate, living beings, terrain, time and human factor. It has highly spatial heterogeneity [1]. Nutrient contents of arable soil are different in spatial distribution in different locations due to interactions of physical, chemical and biological processes. This is the specific representation of soil spatial heterogeneity [2]. Full understanding of changes in arable soil nutrients plays a vital role in soil nutrient management, rational application of fertilizer and improvement of farmland management efficiency [3, 4].

Foreign scholars put forward soil spatial variability as early as 1960s. The research methods underwent initial Fisher statistical method, geostatistical method in late 1980s, geographic information technique, neural network and high-accuracy curve modeling. Since geostatistical method was introduced in soil property spatial variation study, it has become the major method. But in some circumstances, some preconditions cannot be met as follows: during use of Kriging interpolation, the study area must be homogeneous; different parts of the landscape different use semi variograms. So, Kriging cannot well describe spatial distribution of soil property with nonlinear characteristics [5,6,7]. Meanwhile, the complexity and peculiarity of geosciences phenomena make it hard to apply theoretical models established under various ideal conditions in practices. The parameters and even the structure of deterministic models need continuously modifying with the changes in the place and time. Thus, to a larges extent, the universality of models is lost [8]. Artificial neural network is an approach to simulate biomechanism by computer. It has strong ability to deal with nonlinear system. In recent years, it has been gradually applied to study of soil property spatial variation [6,9]. JoséA. C. Ulson et al. [7] utilized back propagation network (BP Network) algorithm to train the soil property data collected from the field on the basis of designing a neural network with hidden-layer multi-layer perception, and then conducted spatial interpolation. The forecast accuracy of the result is higher than that of Kriging interpolation. Shen Zhangquan [5,10] compared soil nutrient spatial forecast by generalized regression network, integrated BP Network and Kriging interpolation under three sampling scales through designing three different soil sampling point collection schemes. The result showed in most cases, spatial forecast accuracy of generalized regression network, integrated BP Network was higher than that of Kriging interpolation. Besides, with the decrease in the number of samples, interpolation accuracy of generalized regression network, integrated BP Network showed more superiority. But these studies just established a mapping relation between space coordinates and soil properties, and overlooked other ecological processes at sampling point locations. Moreover, soil property spatial variability is very complex, which makes this method unable to fully reflect spatial

variation characteristics of soil property. Later, Li Qiquan et al. [6] utilized Radial Basis Function (RBF) Neural Network to study on spatial interpolation of soil property with different degrees of variation by RBF Neural Network under the condition of adding adjacent sampling point information as network input and compared RBF Neural Network method only taking space coordinates as network input and Kriging interpolation. The result showed the ability of RBF Neural Network adding adjacent sampling point information input to describe spatial distribution of soil property information improved greatly and could well reflect local variation information of soil property. However, RBF Neural Network has some problems in network topology, width and center confirmation as well as weight calculation from the hidden layer to the output layer, thus imposing great influence on interpolation accuracy. The researches of Chai Jie et al. [11] and Li Yu et al. [12] show genetic algorithm can optimize the weight of RBF Neural Network and network hidden-layer structure. Based on this, Dong Min et al. [13] utilized genetic RBF Neural Network model to optimize the weight of from the hidden layer to the output layer of RBF Network, then adopted optimized RBF Network to carry out spatial interpolation for available zinc in the soil in the study area and then compared it with the interpolations of RBF Network without optimization and Kriging. The result showed the error of the interpolation of genetic RBF Neural Network was small and that the interpolation chart could better reflect practical spatial distribution of available zinc element in the soil. But, there are many parameters needing confirming in RBF Neural Network. Only through optimizing the weight from the hidden layer to the output layer, the improvement effect of forecast accuracy is not obvious.

This study utilizes three network parameters of RBF Neural Network including the number of hidden layer nodes, expansion speed and root-mean-square error to design 5 different soil sampling point layout schemes on the basis of 637 soil samples in line with the thought of gradual improvement of sampling scales. Besides, under different sampling scales, the fitting capacity and estimation accuracy of soil property space of RBF Neural Network optimized by genetic algorithm, Ordinary Kriging as well as RBF Neural Network without optimization are respectively compared, which provides technical support for accurately estimating soil property spatial variability, reducing the number of soil samples and decreasing the cost of soil sampling.

2. Research Methods

2.1 Preprocessing of soil sampling point data

To check the accuracy of interpolation method on spatial interpolation of available phosphorus in soil, the soil sampling points should be first divided. Create Subsets function in Geostatistical Analyst module of ArcGIS software was used to sample. Besides, spatial distribution of soil samples should be even. Through referring to the studies of Lei Nengzhong et al., the modeling scheme of 5 different sampling scales was set up at the interval of 100 sampling points [14]. 500 training samples were drawn from 637 available phosphorus sampling points. The sampling point layout composed of this dataset is Scheme e. Then, based on sampling point layout of Scheme e, 400 sampling points were drawn at random from 500 training samples to form Scheme d. Then, based on Scheme d, 300 sampling points were drawn at random as Scheme c. Scheme d and a can be formed by parity of reasoning. Finally, after Scheme e was formed, 100 sampling points were drawn from 137 soil sampling points as the check sample of interpolation results of all schemes.

2.2 Interpolation of RBF neural network

Radial Basis Function Neural Network (hereinafter referred to as RBF Neural Network) was formed through Broomhead and Lowe [15,16] applying radial basis function raised by Powell (1985) to artificial neural network. Initially, it was used to interpolate data points in a group of multi-dimensional space. The objective of interpolation was to seek a function which could map each vector to corresponding target values.

In the process of spatial interpolation of RBF Neural Network, Gaussian kernel function was adopted in this study as the basis function:

$$\mu_j = \exp(-\frac{(X-C_j)^T(X-C_j)}{2\delta_j^2}), j = 1,2,...,N_h \quad (1)$$

Where, μ_j is the output of nodes at the j^{th} hidden layer; X is output sample; C_j is the central value of Gaussian kernel function; δ_j is a standard constant; N_h is the number of nodes at the hidden layer. The output range of the nodes is between 0 and 1. Besides, the input sample is closer to the center of nodes, and the output value is larger.

This paper adopts Matlab neural network tool kit to realize RBF spatial interpolation. The main steps are as follows:

(1) Data preprocessing

In ArcEngine, geodetic coordinates were transformed to decimal plane rectangular coordinates. The property row

was formed for longitude and latitude coordinates of sampling points and added in ArcGIS. Derivation function of ArcGIS was utilized to derive the file in dbf format. Then, decimal system longitude and decimal system latitude were drawn from soil sampling point coordinated system and put into a new Excel. Then the data in Excel were processed by Matlab.

To prevent excessive variable value of Matlab in operational process and improve learning speed, normalization processing is required for coordinate data. Assume left bottom and top right corner within spatial area coverage of interpolations are $(x_{\min}, y_{\min})$ and $(x_{\max}, y_{\max})$ respectively, and

$$x' = (x - x_{\min})/(x_{\max} - x_{\min}) \quad (2)$$

$$y' = (y - y_{\min})/(y_{\max} - y_{\min}) \quad (3)$$

As well, normalization processing is conducted for main physicochemical index values of soil. Assume the minimum value and the maximum value of a physicochemical index value of soil are $Z_{\min}$ and $Z_{\max}$ respectively, and normalization express is:

$$Z' = (Z - Z_{\min})/(Z_{\max} - Z_{\min}) \quad (4)$$

(2) Generation of interpolation grid point

The grid can be established according to the size by use of Matlab order meshgrid in line with specific conditions of the study area. Denser grid means higher interpolation accuracy, the operating rate reduces exponentially. For grid coordinate points set up, the coordinate value should be calculated under normalization according to coordinate normalization formula.

(3) Seeking 5 data pints nearest to the training point

Ergodic program was adapted to train sampling point dataset. Euclidean distance between other sampling points in the dataset and the training point was calculated. The computational formula is:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (5)$$

d_{ij} means spatial distance between the sampling point i and j; (x_i, y_i) refers to the coordinate of i^{th} sampling point; (x_j, y_j) refers to the coordinate of j^{th} sampling point.

Sort algorithm was used to sequence spatial distance stored in the array. 5 nearest sampling points were selected according to the principle of $d_1 \le d_2 \le d_3 \le d_4 \le d_5$ and all information was stored in the array.

(4) Network establishment and analog simulation

Matlab neural network tool kit function was used to establish RBF Neural Network. Firstly, training dataset array was substituted into RBF Neural Network for training so as to gain non-linear relationship between the property of soil sampling points and adjacent points respectively. Then, grid generated in (2) was substituted into Neural Network for simulation, thus getting soil property value of any grid unit in the whole grid.

(5) Comparison of data recovery and interpolation effect

Recovery operation was implemented for interpolation-generated soil property values and geographical coordinates. Testing data were used to evaluate the accuracy of interpolation model and obtained the value of accuracy evaluation factor under current operation mode. Finally, the data can be written in Excel through Matlab function. Latticed data can be shown in ArcGIS through ArcGIS grid analysis.

2.3 RBF Neural Network interpolation improved by genetic algorithm

Genetic algorithm is a theory and approach with initiative significance jointly studied by a psychology professor in University of Michigan – Holland as well as his colleagues and students in 1975. Such approach was a highly concurrent, random and self-adapting search algorithm developed by referring to natural selection and evolutionary mechanism in the biosphere. This paper adopts genetic algorithm to seek optimized the number of nodes at the hidden layer, expansion speed and root-mean-square error of RBF Neural Network. Binary coding is adopted as gene code system in accordance with roulette model [17] as realization model of genetic algorithm natural selection. In this paper, the main steps to combine genetic algorithm and RBF Neural Network are as follows:

(1) Rewrite interface function of RBF interpolation algorithm; possible value scopes of the number of nodes at the hidden layer, expansion speed and root-mean-square error in RBF Neural Network are revealed in variable form for use.

(2) Solve the maximal length of binary coding necessary in genetic algorithm. In this study, the maximum value of parameter scope subtracts the minimum value of the scope. Then, compare the value with the figure expressed by binary coding so as to solve the shortest binary system length required by parameter scope, i.e. chromosome length.

(3) Produce initial group. An initial group is produced by Random function of Matlab. The length is the large random matrix of the sum of all lengths expressed by binary of chromosome.

(4) Calculate the fitness of each individual in the group. Draw individual gene length information and read binary coding of the gene by sections. Then, utilize transformation relation between binary number and

decimal numeral to calculate segmental gene information to decimal integers or decimals. Establish fitness function of genetic algorithm with the measurement standard of root-mean-square error and mean value error of spatial interpolation of RBF Neural Network model. Interpolation is carried out for the given points through modifying RBF spatial interpolation program and utilizing well-trained network. Then, solve root-mean-square error and mean value error.

(5) Method to gain optimized RBF network structure. In given iterative algebra range, calculate root-mean-square error and mean value error for all groups in each iteration and store them in temporary array. After solving fitness function of each iteration, solve the fitness of the most excellent individual and the genotype of this individual through sort algorithm and record them into external temporary array. The program will record the optimized individuals and their genes (RBF network parameter) of corresponding algebras according to specified iterative algebra, and sequence again and compare the numerical values in this array when finally returning to optimal individuals. The program will ultimately gain the algebra and individual with the best fitness.

(6) Implementation method of natural selection. The process of practical programming and running shows most fitness function values solve previously are between 0 and 1. Therefore, the large array can be produced through the way of generating pseudo-random numbers between 0 and 1 to simulate selection requirements of natural environment so as to retain individuals meeting natural selection conditions and weed out those not adapting environmental requirements.

(7) Implementation method of gene crossover. For individuals retained, every two are selected as parent individuals. Random function Rand is adopted to produce a random number between 0 and 1. If such random number is less than crossover probability, gene crossover operation is carried out. Under the effect of random function, a position in corresponding gene segments of both parents is produced as the cross point. Individual exchange is carried out for the corresponding gene of both parents at the central position of the cross point.

(8) Implementation method of genovariation. For each offspring individual, it is required to judge whether the random values between 0 and 1 produced by random function are less than variation probability specified by the function. If they are less than the variation probability, variation operation is implemented, or else, variation operation is not implemented for individuals. In corresponding gene segment of individuals, specific variation points may be gained through the values between 0 and 1 produced by random function multiplying by the length of the gene segment. Then, variation operation is conducted under binary condition. This study is realized through judging binary values on the gene points. If the value is 1, it changes to 0; if the value is 0, it changes to 1.

In the end, the parameters of the number of nodes at the hidden layer, expansion speed and root-mean-square error gained through genetic algorithm optimization are input into RBF Neural Network. Interpolation grid is generated by use of meshgrid order in Matlab. The dataset of ergodic program is adopted to train sampling points and calculate Euclidean distance between other sampling points in the dataset and the training sample point so as to seek 5 adjacent points (soil features participating in the training). In the interpolation process, nonlinear function between soil property values z and x hide in the network after convergence. The specific expression is unknown. Geographical coordinates are regarded as network input to realize forecast of soil property space at unknown points.

2.4 Evaluation of interpolation accuracy

Spatial interpolation accuracy is tested by authentication dataset. Interpolation accuracy evaluation is conducted through comparing mean absolute error (MAE) and root-mean-square error (RMSE) of the forecasted value of soil property and measured value at the verification point. MAE reflects actual measurement error range of estimated values. The error can be given quantitatively. RMSE reflects the effect of the estimated value and the extreme value of sampling point data. The computational formulas are as follows:

$$MAE = \frac{1}{n}\sum_{k=1}^{n} | \hat{Z}_k - Z_k | \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n} (\hat{Z}_k - Z_i)^2} \quad (7)$$

In the expression, $\hat{Z}_k$ refers to the forecasted data of he points to be tested; Z_k means measured data at the testing point; n means the number of sampling points tested. It can be seen from the expression that two smaller parameters indicates higher accuracy when evaluating interpolation method.

3 Result and analysis

3.1 Parameters of RBF Neural Network interpolation

When spatial interpolation is implemented by use of RBF Neural Network, since optimized parameters of the

network cannot be gained in advance, "test method" is generally adopted for repeated trials. Then, the parameters can be confirmed in line with the whole accuracy of the network. In this study RBF Neural Network parameters based on adjacent points and geographical coordinates use defaults of Matlab software. The number of nodes at the hidden layer is 8; the expansion parameter is 1.0; the error coefficient is 0.001.

3.2 RBF Network Parameter parameters improved by genetic algorithm

RBF Neural Network improved by genetic algorithm can give full play to the characteristic of global optimization of genetic algorithm to seek optimized parameters of RBF Neural Network. The researches of Zhou Ming et al. [18] show the parameters of RBF Neural Network improved by genetic algorithm can be given through experience. Generally, the expansion parameter is 1; the group size is 20-100; the crossover probability is 0.4-0.99; variation probability is 0.0001-0.1; the end algebra is 100-1000. The number of nodes at the hidden layer is gained by trial method. Through trial and estimated empirical values, genetic algorithm network structure and parameters are shown in Table 1; structural parameters of RBF Neural Network improved by genetic algorithm are shown in Table 2.

Table 1 Structure and parameters of genetic algorithm neural network

Scheme	Expansion parameter	Number of individuals	Crossover probability	Variation probability	Genetic algebra
A	1	30	0.7	0.1	100
B	1	30	0.7	0.1	100
C	1	30	0.7	0.2	100
D	1	30	0.7	0.2	100
E	1	30	0.7	0.2	100

Table 2 RBF Neural Network improved by genetic algorithm

Scheme	Number of nodes at hidden layer	Expansion speed	Error coefficient
A	2	0.128014	0.001713
B	2	3.374016	0.078953
C	30	0.411811	0.058685
D	2	3.374016	0.078953
E	1	1.191339	0.020488

To evaluate the fitness capacity of RBF Neural Network improved by genetic algorithm, this study will utilize such method for contrastive analysis of spatial forecast result of soil organic matter under 5 sampling scales with that of RBF Neural Network method without optimization and Ordinary Kriging method through scatter diagram so as to solve regression equation of the forecasted value and measured value of the training sample and the determination coefficient (R^2). The matching degree of the forecasted value and the measured value of the training sample is judged by the determination coefficient. If the determination coefficient approaches 1, this indicates the matching degree is higher and the fitness capacity of interpolation method is stronger. Under five sampling scales, the scatter diagram of forecast results gained by 3 spatial estimation methods are shown in Fig.1-5 (Ordinary Kriging is Ordinary Kriging interpolation method; RBF is the method of RBF Neural Network spatial interpolation without optimization; GARBF is the method of RBF Neural Network spatial interpolation optimized by genetic algorithm).

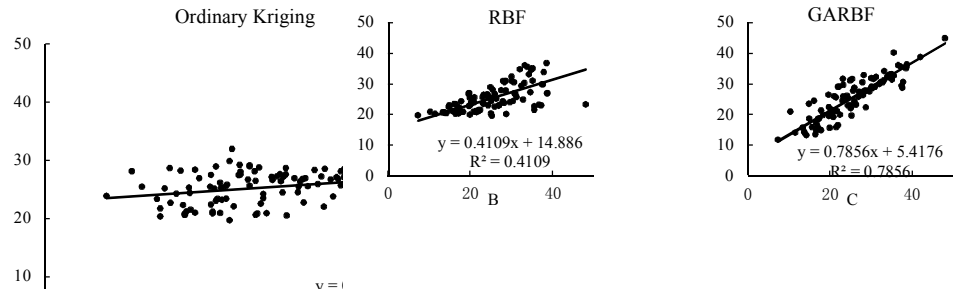


Fig.1 Scatter diagram of measured values and forecasted values of training sample gained by 3 methods in Scheme A

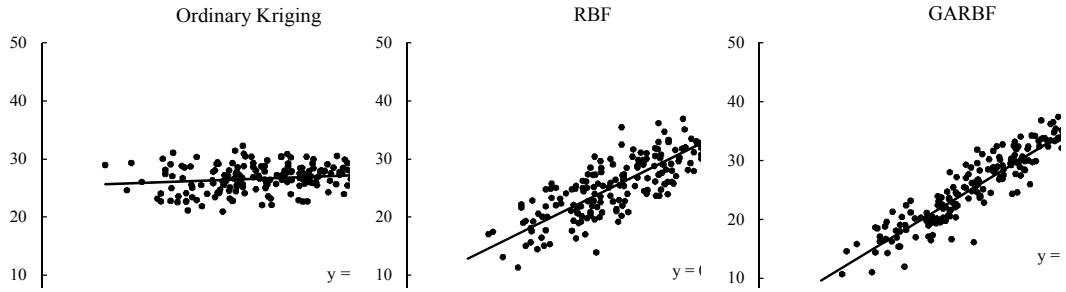


Fig.2 Scatter diagram of measured values and forecasted values of training sample gained by 3 methods in Scheme B

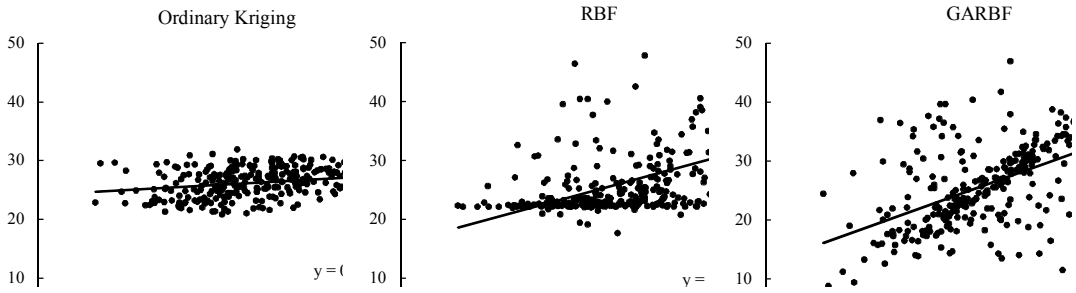


Fig.3 Scatter diagram of measured values and forecasted values of training sample gained by 3 methods in Scheme C

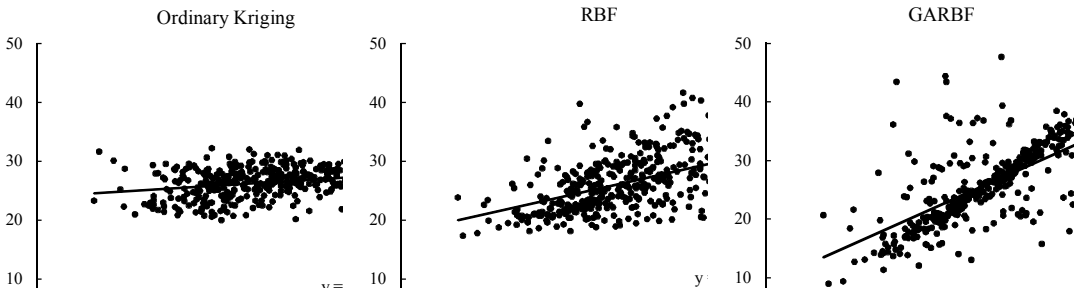


Fig.4 Scatter diagram of measured values and forecasted values of training sample gained by 3 methods in Scheme D

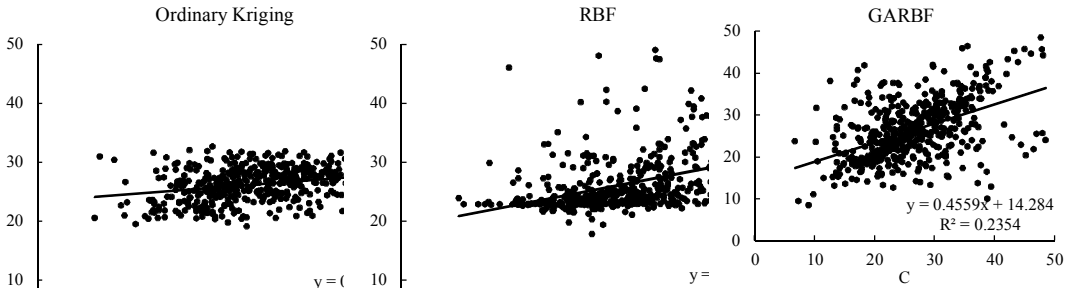


Fig.5 Scatter diagram of measured values and forecasted values of training sample gained by 3 methods in Scheme E

Note: Ordinary Kriging is Ordinary Kriging interpolation method; RBF is RBF Neural Network method based on adjacent points; GARBF is RBF Neural Network method optimized by genetic algorithm.

Table 3 Comparison of approximate errors of training samples in all cases

	Scheme A		Scheme B		Scheme C		Scheme D		Scheme E	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Ordinary Kriging	6.547	8.495	6.056	7.520	5.275	7.099	5.147	6.247	5.187	6.313
RBF	6.221	7.724	5.647	7.086	5.478	6.864	5.383	7.346	6.035	7.507
GARBF	5.673	6.059	5.321	5.835	4.892	5.391	4.903	5.631	4.941	6.152

Note: MAE is mean absolute error; RMSE is root-mean-square error; Ordinary Kriging is Ordinary Kriging method; RBF is RBF Neural Network method based on adjacent points; GARBF is RBF Neural Network method optimized by genetic algorithm.

5. Conclusion and discussion

Since spatial variability of soil property is large, three important preconditions during application of geostatistics and smooth effect of Kriging interpolation cannot be met. These to some extent cause the inaccuracy of expressing abnormal area of soil property, thus reducing reliability of the forecast result [5,13]. Infinite approximation capability of RBF Neural Network can well solve this problem, but some problem still exists in optimization of network parameters. Therefore, in this study three parameters of RBF Neural Network optimized by genetic algorithm including the number of nodes at the hidden layer, expansion speed and RMSE are used to improve the accuracy and reliability of spatial estimation of soil property. The study result shows: spatial interpolation ability of RBF Neural Network optimized by genetic algorithm is superior to RBF Neural Network without optimization and geostatistics. Such superiority is not just reflected in the fitness capacity of RBF Neural Network method. In the aspect of testing MAE of RMSE between the forecasted value and the measured value of the samples, RBF Neural Network optimized by genetic algorithm is obviously less than other two methods. Besides, the superiority is more significant when the quantity of interpolation sampling points is less.

Very complex non-linear relationship exists between soil properties and various influencing factors. Besides, mutational boundary exists among different influencing factors [19]. So, when spatial estimation of soil property is conducted by use of neural network, if relevant factors such as parent material, soil type, planting system, elevation, climate element as well as other soil properties can blended in the network, not only the stability and forecast accuracy of the network can be improved, but also synchronous estimation of multiple properties of soil can be realized. What's more, if soil property estimation process can well comply with geoscience laws, spatial estimation will tend to realer and more reasonable, and can better describe detailed information of soil property vibration.

References

[1]Hua Meng, Wang jian. Soil Physics. Beijing: Publishing house of Beijing Agricultural University, 1992.
[2]Wang Jun, Fu Bojie, Qiu Yang, et al. Spatial distribution patterns of soil nutrients in a small catchment of the Loess Plateau- Kriging method. Geographical Research,2003,22(3):373-380.
[3]Hammond. M W. Comparison of phosphorus and potassium utilization with conventional and variable fertility management. Better Crops,1994,78(4): 22-23.
[4]Franzen. W D.，Hofman. L V.，Halvorson. D A. Sampling for site-specific farming: Topography and nutrient considerations. Better Crops,1996,80(3): 14-18.
[5] Shen Zhangquan, Shi Jiebin, Wang Ke, et al. Study on spatial variety of soil properties by means of generalized

regression neural network[J].Transactions of the Chinese Society of Agricultural Engineering, 41(3):471-475.
[6]Li Qiquan, Wang Changquan, Yue Tianxiang, et al. Error analysis of soil property spatial interpolation with RBF artificial neural network with different input methods[J]. Acta Pedologica Sinca,2008,45(2):360-365.
[7]Lson J.S.A.U.， Silva I.N.， Benez S.R.H. Modeling and identification of fertility maps using artifical neural networks，2000: 2673-2678.
[8]You Shucheng, Yan Tailai. A study on artifical neural network based surface interpolation. Acta Geodaetica etCarto Graphica Sinica,2000,29(1): 30-34.
[9]He Yong, Zhang Shujuan, Fang Hui. Interpolation method of field information based on the artifical neural network. Transactions of the Chinese Society of Agricultural Engineering，2004，20(3): 120-123.
[10]Shen Zhangquan, Shi Jiebin, Wang Ke, et al. Spatial variety of soil properties by BP neural network ensemble[J]. Transactions of the Chinese Society of Agricultural Engineering,2004,20(3): 35-39.
[11]Chai Jie, Jiang Qingyin, Cao Zhikai. Function approximation capability and algorithms of RBF neural network. Pattern Recognition and Artificial Intelligence，2002，15(3): 310-315.
[12]Li Yu,Kong Fanguo. Optimization of neural network based on fuzzy theory and genetic algorithm. Journal of Shanghai University of Engineering Science，2007，21(2): 130-131.
[13]Dong Min, Wang Changquan, Li Bing, et al. Study on soil available zinc with GA-RBF-Neural-Network-Based spatial interpolation method[J]. Acta Pedologica Sinca,2010,47(1): 42-50.
[14] Lei Nengzhong, Wang Xinyuan, Jiang Jingang, et al. Spatial variability of soil nitrogen by BP neural network interpolation[J]. Transaction of the Chinese Society of Agricultural Engineering(Transactions of the CSAE) 2008,24(11):130-134.
[15]Broomhead D.S.， Lowe D.. Multivariable functional interpolation and adaptive networks[J]. . 1988, Volume 2. Proc. IEEE Int. Conf. on Complex Systems，1988(2): 321-355.
[16]D. S. Broomhead D.L. Multivariable functional interpolation and adaptive networks. Proc. IEEE Int. Conf. on Complex Systems，1988，Volume 2: 321-355.
[17] Wang Chengzhang, Yin Bocai, Sun Yanfeng, et al. An improved 3D face-modeling method based on morphable model[J]. Acta Automatica Sinica,2007,33(3):232-239.
[18] Zhou Ming, Sun Quandong. Theory and application of genetic algorithm. Beijing: Publishing house of National Defence Industry,1996.
[19] Shi W, Liu J, Du Z, et al. Surface modelling of soil pH[J]. Geoderma,2009, 150(1/2): 113-119.

Software Process Improvement Framework Based on CMMI Continuous Model Using QFD

Yonghui CAO[1, 2]
1, School of Economics & Management, Henan Institute of Science and Technology, Xin Xiang, 453003 ,China
2, School of Management, Zhejiang University, Hang Zhou,310058 ,China

Abstract
In the rapid technological innovation and changes era, the key to the survival company is the continuous improvement of its process. In this paper, we introduce Software Process Improvement (SPI) and Quality Function Deployment (QFD); and for combining also the staged model and the continuous model in CMMI, the Software Process Improvement framework with CMMI has two parts: 1) Software Process Improvement framework with CMMI staged model based on QFD and 2) SPI framework for CMMI based on QFD continuous model. Finally, we also draw conclusions.
Keywords: *Software Process Improvement, CMMI, QFD*

1. Introduction

Software Process Improvement (SPI) is the modification of current software process methods in many software development organizations. Its aim is to improve the organization's ability to produce better software products (Humphrey, 1990)[1].The Capability Maturity Model Integrated (CMMI) is a SPI models, which came from the Software Engineering Institute. But in the process of improvement, models and standards can not be used in business or other requirements in an company independently.

At present, many international models or standards are developed for Software Process Improvement (SPI). For example, these standards have ISO standard, CMM (Capability Maturity Model), CMMI (Capability Maturity Model Integrated). And CMMI is a SPI models from the Software Engineering Institute. With regard to process and quality improvement, these standards and models have a share in some common consideration. CMMI emphasizes continuous improvement, but the ISO standard emphasizes the minimum criteria with the quality system. It is unfairness to do a judgment with which one is much more better(Caulk, 1994) [2].

During process improvement, these models and standards cannot be applied independently in commerce and other departments in an section. However, in consideration of the more detailed guidance and bigger scope offerred from CMMI, it may be a better choice for some software development organizations (Francois Coallier,1994) [3]. Philosophically, the CMMI is a specific implementation of Total Quality Management (TQM). Drawing upon the works of Deming(1986)[4], the CMMI is the framework by integrating systems and ameliorateing systems and software engineering systems. Process improvements have been shown to increase productivity, quality, and cycle times, and result in organizations more accurately predicting schedules and budgets. CMMI is intended to cover both product and service throughout their life cycle of development, deployment, and maintenance, as well as being extensible to incorporate new bodies of knowledge (Chrissis et al., 2003)[5]. The current four bodies of knowledge supported in the current CMMI, also referred to as disciplines, are systems engineering, software engineering, integrated product and process development, and supplier sourcing (Chrissis et al., 2003)[5].

On engineering improvement systems, look as all the other models and standards, CMMI addresses the question of "what to do it" by departting from "how to do it" to organizations system. Consequently, more measure is required to transform CMMI Practices to a series of activity which are more carefully that can be abided by software engineering improvement systems.

In this research , architecture was shapeed to assist directing business or some other procedure necessary condition in one Commercial enterprise to CMMI key element, and help shape action orientation to fulfil those necessary condition making use of Quality Function Deployment (QFD).

Quality Function Deployment (QFD) has been applied in the world in almost any business and department of precedence customer requirement from 1966. In order to

change the requirement into behaviour and project like technological property and standards. So can set up and transmit the quality merchandise and service through concentrating on accomplishing one identical target of one customer aspiration.

There are three original contributions in the proposed framework, all with the help of QFD. First, commerce and the other necessary condition in one institution or business should apply for target and behaviour in CMM. The link line is set up in order to make the mechanism can predict accurately how CMMI assistes in the mercantilism. Second, Business requirements and software system process necessary condition from different sources are assembled and optimized. Third, QFD is applied to assist change key elements of the company to machining actions by CMMI. Study shows that these observable recordes in the progress of the organizational process.

2. QUALITY FUNCTION DEVELOPMENT

When most of the quality models offer supervise for either the accomplishment of a much more better procedure or the evaluating of the nowaday proceeding, all of them have a share in only one common performance , and the models Scopes "what to do" , and not Scopes "how to do it" to personality corporation. It is satisfactory that have a method to conduct the corporation by a development with action projects in SPI. All of these activities should be on account of the software process necessary condition by correlation resources. Quality Function Deployment (QFD) is an suitable instrument in the convertting from customer requirement into goods. From software process requirements, it is available in providing the objective of originating from action proposes to SPI.

In the late 1960s, Quality Function Deployment was developed by Professor Shigeru Mizuno and Yoji Akao in Japan, and was recommended including the United States and European countries, and other countries of world, in the early 1980s. Setting up the sound from the customer is a method, not only spoken but also not spoken, with regard to one product. There are big difference in QFD and the other quality methodologies, and the most difference is Quality Function Deployment increases values about the product, by way of maximizing the product's positive quality, but the traditional product quality systems target to minimize negative quality about one product [6].At present, QFD has been practically applied to almost every commerce and industry, containing software process development [7][8].

The tools used in QFD are the Seven Management and Planning Tools, which are listed in table 1:

Table 1: Seven Management and Planning Tools

the Seven Management and Planning Tools
1, Relations Diagram
2, Matrix Diagram
3, Tree Diagram
4, Affinity Diagram
5, Activity Network
6, Process Decision Program Chart
7, Matrix Data Analysis Chart

The most significant measure in QFD is the House of Quality (Figure 1). The house of quality is one table which links between the Engineer's sound and the Customer's sound.

There are six large ingredients in The House of Quality:
1. Customer requirements (WHAT's).
Customer requirements is a formal structured summarizing of the necessary condition originates from the customer's declaration.

2. Technical correlation (Roof) matrix.
Technical correlation matrix can be made use of ensuring what technological necessary condition backing or prevent one another in the merchandise's design. It can emphasize reformation chances.

3. Technical requirements (HOW'S).
Technical requirements assemble a structured of concerned and quantifiable manufacture characteristics.

4. Interrelationship matrix.
Interrelationship matrix accounts for one team's of the QFD concept of correlation within customer and technology.

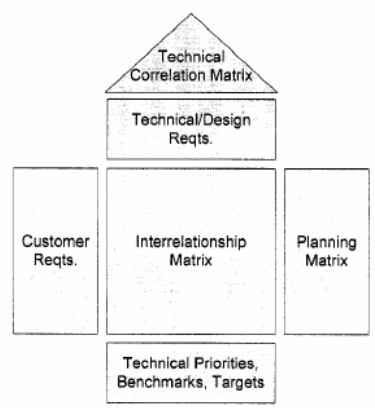


Fig. 1 House of Quality in QFD

What is demonstrated through making use of sign and diagram. How to heap the member in the matrix, includes researches and accordance of one team. Focusing on critical connection, and how to minimize the quantity of necessary condition are available skill to decrease the requirement from resources.

5. Planning matrix.
Planning matrix clarifies customer conceptions surveyed from marketplace. Consistes of correlative significant customer necessary condition, corporation or contender representation when encounter these necessary condition.

6. Technical priorities, benchmarks and targets.
Technical priorities, benchmarks and targets are applied to take notes: The precedence specified to technological necessary condition through the matrix; Estimates of technological capability attained by rival products; The level of hard contained in expanding every necessary condition.

When Professors Akao and Mizuno presented the notion of QFD, this conception was predicated to contain two constituent part: 1) Product Focused QFD or Quality Deployment (QD), 2) Process Focused QFD or Narrow definition QFD [9] [10].The first key element, like the name's suggestion, centers on raising the products quality through becoming customer necessary condition to product property. These have been comprehensively accepted by more and more enterprises and industries in the world. The second key element, centers on raising the quality of procedure, was contrived to confirm that constituent procedure and activity are in subordinated to established criterion, for example as ISO 9000, IS014000, or Some other criterion. Like software corporations, these "narrow definition QFD" can assist them to heighten software optimization procedure by the standards put forward clearly in criterion like ISO 9000, CMM, and so on. But these necessary conditiones have been ignored by most of the QFD followings of the commerce, particularly in the scopes of software system development [11].

3 . SPI FRAMEWORK FOR CMMI CONTINUOUS MODEL BASED ON QFD

QFD is used to help SPI framework with CMMI, which is become more and more popular in the industry. First of all, enterprises and other organizations' demands are mapping to the CMMI process areas and practices. To establish a connection, so that how CMMI helps to do with its business objectives that can be seen clearly by organization. Second, software process requirements from multiples perspective priority, the requirement is more and more stronger affect other demand can get higher priority value. Third, QFD helps convert process of the organization into action through the process area (PAs) and practice in CMMI. Therefore, sorts of action taken is based on how they and two software process requirements and the corresponding practice in CMMI.

How is the framework designed? Always in the way that through the proposed framework plan of action, the process demand can be reflected. Using priority evaluation technique that is introduced in section 4.1, requirements from various angles are related to each other. So each requirement's priority value are adjustment from the influence of other needs assessment.

Two parts were contained by the SPI framework based on CMMI:1) the CMMI staged model of SPI framework and 2) CMMI continuous model of SPI frameworke.

A. SPI framework for CMMI staged model using QFD

Figure 2 shows that the SPI framework for CMMI staged model.

For every four maturity levels, the group requirements are related to target. This goal is based on the process requirements priority. Therefore, if the higher target achieve, the higher importance was get by the overall satisfaction process requirements.

In order to achieve these goals, CMMI staged model has general practice which is divided into four common characteristics and specific practice corresponding "activity implementation" common feature in the CMM. The practice is preferred to their correlation and target.

Therefore, general practice in each common characteristics and specific practice based on prioritized target respectively. Practice aims to achieve higher overall satisfaction goal will get higher importance value. Different action plan is from the general practice in each common characteristics and specific practice. This action helps supporting the more important practical get the higher priority.

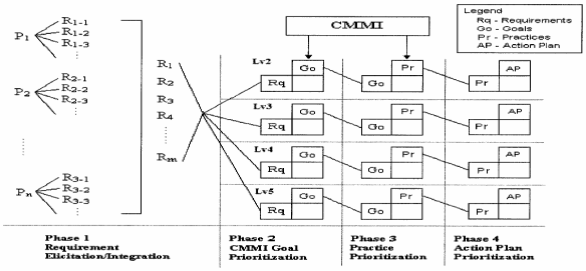


Fig. 2 Software Process Improvement through CMMI Staged Model Using QFD

Therefore, in the CMMI model, the actions follow the operation of the process maturity standard and meet the demand of the process. The higher importance value of the actions could help to achieve higher process demand satisfaction.

Because the CMMI staged model and CMM model are similarity in some extent, four stages of SPI framework based on CMMI staged model shown in figure 2.

In figure 2, the first stage is the identical SPI framework based on CMM. All sorts of views expressed as P1 to Pn. Each sort of view includes more than one need. In perspective 1, software process needs expressed as R1-1, R1-2, and so on. These views software process needs can priority integrated into a single set of requirements based on their relative importance within the organization. In figure 2, the integrated demand for R1 said by Rm, where m is the total number of software process requirements from all angles. Ensure that the priority of needs are comparable to each other from different angle, integrated reflects the needs of correlation from different angles. This phase of the deliverable is a set of priority and integrated software process demands, as input to the next stage.

The second stage to the fourth stage of the SPI framework is used to the CMMI model's part 2 to part 5. The priority and integrated demands from the first stage with all the goals in every four level in the CMMI model with relationship matrix stage. The first goal is used as a fundamental practice. Finally, the first practice transfer into first action plans by House of Quality (HoQ).

In the second stage, the target of all Pas in specific maturity level in "CMMI goal priority," which were choosed first based on demands of all the stage before. This stage can help to realize two important goals. First of all, the organization must follow CMMI standard. Of course, a specific maturity level must to achieve, this process is also meet business and the needed demands.

In phase 2, there is a connection between requirements of organization and goal in the CMMI, and in order to establish the connection, the relationship matrix is used to. The matrix proof that comply with the CMMI standards also helps to meet business and other needs of the organization. The second stage is that the last action plans which is based on priority needs to be priorities, so that if want to gain more resources, the more important action should to do. Priority target, the requirement from the organization can be transformed into practice in the third stage, the final action plan in the final stage. In this way, a group of perform behavior not only to achieve a specific maturity level in CMMI, also in order to meet the needs of the organization's process.

In the third stage, it illustrates "practice priority", including the priority of practice in a particular level of all PAs. All these practice must carry out in order to achieve a specific maturity levels because he CMMI specifications. These practice as a bridge between the demands, the final action, it is necessary to know the practice reflects the needs of the software process. In order to show the connection request and the action plan, which must give first priority to the practice on the basis of the goal of, now, this reflects the demand priority. Some CMMI document show that there is mapping between the target and the practice [12].

In the fourth stage, it illustrate the "action plan development and priority", a group of behavior from the priority practice. These actions should reflect demand integration in the first stage. At the same time, they also need to implement the state in order to achieve a specific CMMI maturity levels. These actions are guided by process improvement. Therefore, more resources should be allocated to these behaviors with high priority.

As shown in the figure above, through will demand from the organization for the action plan through the goal and practice relationships, organizational goals and CMMI maturity level becomes clear.

B. SPI framework for CMMI continuous model using QFD

There is great difference between the SPI framework for CMMI continuous model and other staged framework. However, the same technology related basic priority help QFD used to frame. In the continuous model of CMMI, ability level is assigned to individual disrespect. Different PAs can be on different ability level.

Every PA has two types of goals: the first one is general objectives and the second one is specific objectives. General objectives make the CMMI level institutionalized with a general objective for each level. Specific objective describe practice, which must achieve to meet the process area. The goals should make the general practice and specific practice satisfied. Figure 3 illustrates how to practice and act to distinguish priority SPI framework using QFD in continuous CMMI model. The process requirements are used in both PAs and practice. The first thing to do is to calculation of the priority value disrespect. Then practice is the first two process requirements and disrespect. It depends on the PA a practice is, priority value needs practice is multiplied by the PA priority. Finally, the action of the priority value is computed from practice priority value.

Therefore, as shown in figure 3, PAs are priority based on the process requirements and PAs, helping to achieve higher overall satisfaction process demanding higher importance.

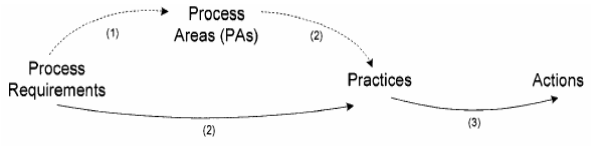


Fig. 3 Priority Calculation in SPI Framework Based on CMMI Continuous Model Using QFD

At the next part, in order to improve the PAs, the different ability level of generic and specific are prioritized. Different priority practice ability level depends on their correlation with the same set of process requirements. As is known that in CMMI continuous model, different PAs has different priority ability level so that the practice does individual PAs. Therefore, in the framework for the CMMI model, practice each level of personal assistant can priority respectively. Practice aims to achieve higher overall satisfaction's key objectives that will get higher importance value. Priority value each PA in the previous stage can be used to calculate the priority of the practice.

Therefore, the Pas, practice, and action reflect process requirements. Both of them follow the operating process capability standard in CMMI and meet the production requirements. The higher importance value to do that the higher process demand satisfaction could achieve.

In figure 4, the first stage is the identical with SPI framework based on CMM. All sorts of views were expressed as P1 to Pn. Each view includes more than one need. Software process needs in perspective 1 expressed as r1-1, r1-2, and so on. These views software process needs can priority based on their relative importance within the organization and integrated into a single set of requirements.

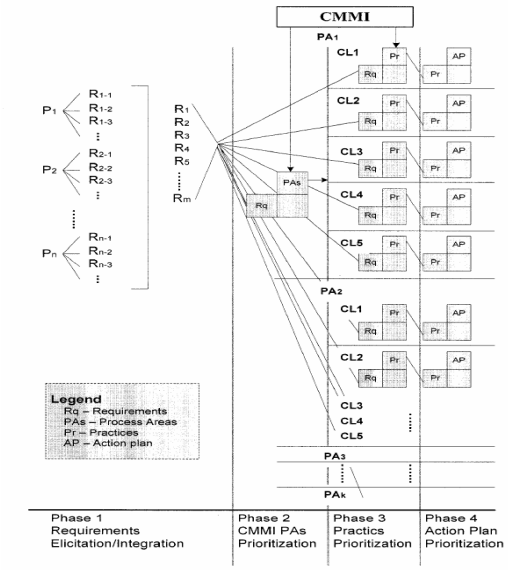


Fig. 4 Software Process Improvement through CMMI Continuous Model Using QFD

In Figure 4, these integrated demands are expressed as from Rl to Rm, where m is the total number of the software process from all angles. In order to ensure that the demands priority from different angles is comparable with each other, the integrated reflect demand correlation from different angles. This part deliverables are a group of priority and integrated software process demands, as input to the next stage.

Then the framework is applied to the Pas in CMMI continuous model from the second stage to the fourth stage. Because of in the CMMI continuous model, different ability level, suitable for different power PAs, the framework of the stage model cannot be applied. And from first stage to the second stage, mapping priority and comprehensive demands in a particular maturity, which are connected together in two stages and appraisal system

level, depends on the ability of target, the practice of connected together, in the third stages using relation matrix. Finally, using the House of Quality, the priority approach changed into priority action plans.

In the second stage, this is "CMMI PA prioritized," based on the demands priority from all the stage before; all PAs are prioritized and selected. This stage can help to realize two important goals.

First of all, the organization should follow CMMI standards. While, organizations need to ensure that by improving process field to higher ability level, this process is also meet business within the organization and other needs.

In Phase 2, in order to establish connection between the requirements organization and each Pas, the relation matrixes were used which illustrate that comply with the CMMI standards also helps to meet business and other needs of the organization.

Second, based on the priorities of demands, the final set of action plans need prioritized, so that more important action to get more resources. The PAs is bridge between demands and the action plan. Priority's personal assistant, the organizations' demands can be changed to practice in the third phase by prioritizing the Pas. The final plan of action to do is in the final stage. In this way, a group can perform a behavior not only to reach a higher level in different PAs, but also to meet organizational process needs.

In the third phase of framework, it mainly discussed "practice priority" which concludes the priority approach, a specific ability level in each PA. According to the CMMI specifications, all these practice ability level in the PA must be carried out, so that the PA could reaches a certain level of performance. However, they may not need the same amount of resources. These practices are bridges between the demands and the final action, and how to put these practice reflects the software process requirements which is need to know. The link between demands and the final action plan is also reflected demand priority.

In the fourth phase of the framework, it mainly discussed "action plan development and priority," groups of action originated from priority practice in different PAs. These actions should reflect demand integration in the first stage. At the same time, they also need to implement state what to be executed in order to achieve a specific ability level of a particular PA. These actions guide SPI. Therefore, more resources should be allocated to these behaviors with high priority.

Pictured above framework, the requirement from the organizations to take action plan goal and practice the connection between the organization's goals and PA ability level become clear.

As is shown in the framework, the demand from the tissue into action plans by goals and practices the linking between the target and the PA ability becomes clear.

4. CONCLUSIONS

This research is to solve this problem that using QFD as a kind of tool to established connection between demands in the organization and action plans in SPI. After carefully look back on some improvement methods, and the Software Engineering Institute (SEI) was selected as the basis of the proposed method. The new framework of SPI is on the basis of mature and development of the research. This new framework, discussed in detail how to arrange and integration demands, how to map demands of the various components, and how to priority action plan. The framework has three goals: 1) mapping process demands, including business demands, by using quality function deployment, 2) developing a new method, based on quality function deployment to the integration and priority demands from various angles (group); and 3) can prioritize SPA that is on the basis of process demands.

Acknowledgments

This work is financially supported by the National Natural Science Foundation of China (Project No. 90718038). Thanks for the help.

References

[1] Humphrey, W. S. Managing the software process Reading, MA: Addison-Wesley ,1990.
[2] Paulk, Mark C. "A Comparison of ISO 9001 and Capability Maturity Model for Software." Technical Report. CMU/SEI-94-TR-12, ESC-TR-94-12, July, 1994.
[3] Francois Coallier, "How ISO 9001 Fits Into the Software World," IEEE Software, Vol. 11, No. 1, January 1994, pp. 98-100.
[4] Deming;W.E.(1986).Out of the crisis. Cambridge MA:MIT Center for Advanced Engineering.
[5] Chrissis, M. B., Konrad, M., Shrum, S. (2003). CMMI: guidelines for process integration and product improvement. Boston, MA: Addison-Wesley Publishing Company.,2003.

[6] Akao, Yoji, ed., Quality Function Deployment: Integrating Customer Requirements into Product Design, Cambridge, MA, Productivity Press, 1990.
[7] Liu X, Inuganti P., Veera C. 2003. An Integration Methodology for Software Quality Function deployment. Final Project Report to the Toshiba Corporation.
[8] Xiaoqing (Frank) Liu, Yan Sun, Praveen Inuganti, Chandra Sekhar Veera, and Yuji Kyoya. "A Methodology for the Tracing of Requirements in Object-Oriented Software Design Process Using Quality Function Deployment," Software Quality Professional Journal, September 2007, Volume 9, Issue 4.
[9] Akao, Yoji, Glenn H. Mazur. "Using QFD to Assure QS9000 Compliance." 4th International Symposium on Quality Function Deployment, Sydney, 1998.
[10] Zultner, R.E. "Quality Function Deployment (QFD) for Software." American Programmer, 1992.
[11] Akao Y., Hayazaki T. "Environmental Management System on ISO 14000 Combined with QFD." Transactions of the Tenth Symposium on QFD. Novi, Michigan. ISBN 1-889477-10-9

Author Yonghui Cao received the MS degree in business management from Zheji**ang** University in 2006. He is currently a doctorate candidate in Zhejiang University. His research interest is in the areas of management information systems.

Research of the Decision-theoretic Intelligent Multi-agent Self-organization System

Yonghui CAO[1,2]
1 School of Management, Zhejiang University
2 School of Economics & Management, Henan Institute of Science and Technology

Abstract

This research will attempt to answer these commonly asked questions from the machine learning literature. The heart of the problem is how the agents will learn the environment independently and then how they will cooperate to establish the common task. In this paper, we explain the structure of Self-organization of the intelligent agents. This agent is designed by a Bayesian network and an influence diagram. We study a multi-agent organization system. At the same time, we also analyze bi-directional learning feature. Finally, we design the system representation of the decision-theoretic intelligent agent, feedback control and adaptive control.

Keywords: *Self-organization System, Decision-theoretic Intelligent, Feedback Control, Adaptive Control*

1. Introduction

As discussed in the literature, several methods are employed in multi-agent learning and organization problem such as temporal difference (TD(λ)), genetic algorithms, and learning classifier systems. The advantages and disadvantages of these methods are also examined in the literature. The main disadvantage of these methods is that they perform badly when the data is not fully observable. Additionally, they do not have the desired bi-directional learning property. We proposed Bayesian networks to ease these problems because they can perform well with the partially observable data and, more importantly, Bayesian networks have the bi-directional learning ability. The following paragraphs will illustrate how Bayesian networks can solve the multi-agent self-organization problem with the help of influence diagrams. Next section, we will explain the structure of an agent, which is designed by a Bayesian network and an influence diagram. Then, we will examine a multi-agent organization system and the bi-directional learning feature of the proposed multi-agent self-organizing system. Finally, we present the system representation of the decision-theoretic intelligent agent design.

2. A Decision-theoretic Intelligent Agent Design

An agent is defined as an entity that can be viewed as perceiving environment through sensors and acting upon that environment through effectors. Therefore, an agent should have sensors and actuators to interact with the environment. On the other hand, an intelligent agent is an agent that reasons with the sensory information and creates optimal actions to satisfy a goal. Therefore, a reasoning system and a decision support system are necessary elements of an intelligent agent. Bayesian networks and influence diagrams can be considered as reasoning systems and decision support systems respectively.

Communication between the agents is also necessary to establish organizational behaviors in a multi-agent self-organizing system. Therefore, an intelligent agent should have sensors, actuators for actions, a Bayesian network, an influence diagram and a communication system.
An intelligent agent has five levels: sensors, belief, preferences, capabilities and actions. In this design, Shohams' agent oriented programming paradigm is followed. According to this paradigm, the mental state of agents can be represented in terms of their belief, capabilities, and preferences. The belief level consists of a Bayesian network (V_A or V_E) and its nodes represent agent's possibly uncertain beliefs about the world. The nodes in V_A represent variables related to the other agents in the system. The nodes in V_E represent the variables related to the agent itself. The preference level is represented as a utility node (U_A and U_E) that expresses the desirability of a world state. The capability level is represented by decision nodes (V_{DA} and V_{DE}) that contain alternative courses of action, which the agent can execute to interact with the world. This is also called

belief, desire, and intention (BDI) architecture in the literature.

Each agent models other agents as an influence diagram by modeling other agents' variables(V_A), utility function(U_A), and decision nodes (V_{DA}). Duryadi and Gmytrasiewicz stated that other agents' models could be learned using influence diagrams. As a modeling representation tool, the influence diagram is able to express an agent's belief, capabilities and preferences, which are required if we want to predict the agent's behavior. Duryadi and Gmytrasiewicz established the learning of other agents' behaviors in the following way: Given an initial model of an agent and a history of its observed behavior, new models can be constructed by refining the parameters of the influence diagram in the initial model. The details of the learning method can be seen.

Agents also need a model of the environment. Bayesian networks can model the environment efficiently. The nodes in V_E model the environment and provide beliefs about the environment. Then, these beliefs are dragged into the utility node U_E .The utility node U_E represents the agent's own preference that is defined by the goal of the multi-agent organization system. The utility U_E is a function of the belief about the environment (V_E), the expected actions of the other agents (A_2), its possibly course of actions (A_1). Figure 1 presents the proposed intelligent agent model.

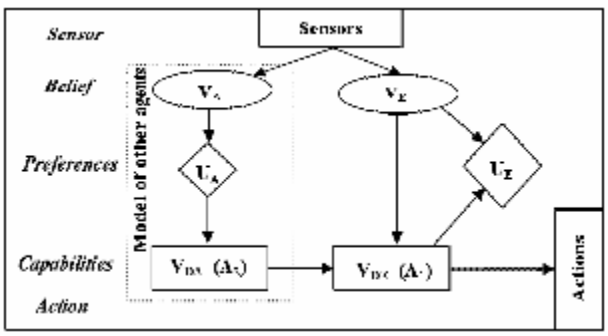


Fig. 1 the structure of an intelligent agent

After establishing the world model and the utility function, the agent needs to take an optimal action according to the principle of maximum expected utility (PMEU). The PMEU lets the agents choose the best action from its set of action (A_1), given the belief about the environment (V_E), and other agents' expected behavior (A_2). Formally, it can be expressed as

$$\max_{a_{1i}} U_E = \max f\{V_E, A_1, A_2\} \quad (1)$$

Where $V_E = \{X_1, X_2, \cdots, X_n\}$, the variables X_i are the nodes of the Bayesian network V_E , $A_1 = \{a_{11}, a_{12}, \cdots, a_{1k}\}$ is the action set of the agent, $A_2 = \{a_{21}, a_{22}, \cdots, a_{2i}\}$ is the expected action set of the other agents. Therefore, an agent takes its actions after evaluating the environment and the other agents. This property will help to obtain self-organization ability of the system. Each agent first check to see if other agents are performing task before it takes its actions to perform the task.

3. Multi-agent Self-organizing System

This section will examine the learning problem when we have more than one agent. The agent described in the previous section is specifically designed for multi-agent systems. In a multi-agent environment, coordination requires an agent to recognize the current status and to model the actions of the other agents to decide on its own next behavior. That's why agents model other agents as well as the environment. A computational difficulty may arise if the number of agents is large in the system because agents model the internal structure of other agents in their network. The Bayesian network in the agent may become so large that the calculation of the conditional probabilities might become difficult. The agents are independent but they take their actions by considering the other agents. Thus, agents take their actions together in coordination. Formally speaking, the agent's utility function U_E depends on the expected actions of other agents (A_1), see Equation (1).

We can explain this ability with an example. Suppose we have two dogs and a sheep, as in the sheepdog problem. Dogs are our agents and their goal is to put the sheep into a barn. Dogs will explore the environment and they will model the environment. In this case, the environment contains another dog, a sheep, and a barn. First, the dogs

will probably locate the sheep. Then, they will make movements to direct the sheep into the barn. If the dogs do not consider (model) each other, they might not be able to put the sheep into the barn since one's action might hinder the other's action. Thus, they need to cooperate and make movements together. If each dog learns the model of the other dog, then they can make movements together to put the sheep into the barn. If there is no coordination, both dogs will probably go behind the sheep and direct it into the barn. If there is coordination between the dogs, while one of them goes behind the sheep, the other may move back and forth so that the sheep will not escape as shown in Figure 2.

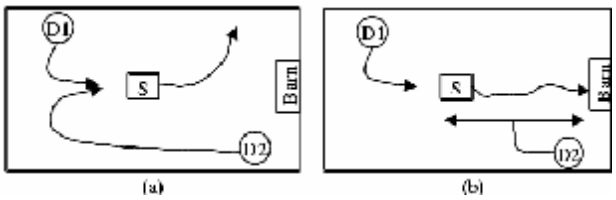


Fig. 2 Multi -agent behavior without coordination (a) and with coordination (b)

A multi-agent self-organization system with two agents can be seen in Figure 3. The multi-agent system is designed by using the agents, shown in Figure 1.

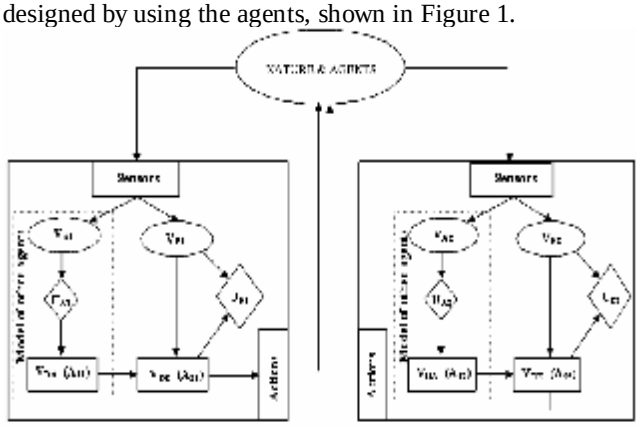


Fig. 3 Multi-agent self-organizing scheme with two agents

In summary, agents will fire actions to change the environment as well as to organize themselves. Self-organization will happen eventually because each agent takes its actions considering other agents' behaviors in the environment. This property will make our system a multi-agent self-organizing system. In the proposed learning system, an agent learns the environment using the sensory data, and modifying its world model (Bayesian Network) accordingly. Then, an agent calculates the expected state of the environment using the world model and creates actions to change the environment. Thus, the learning structure is bi-directional because the agent interacts with nature and the world model in both directions.

4. Bi-directional Learning

As stated earlier, bi-directionality is the most important feature of an intelligent learning system because it combines the supervised learning method and unsupervised learning method and facilitates them at the same time. That is why a Bayesian network is chosen to construct the learning system. Figure 3 shows the learning model of the proposed system. The proposed system has four directed edges among nature, the learning system, and the world model: evidence, action, adaptation, and expectation.

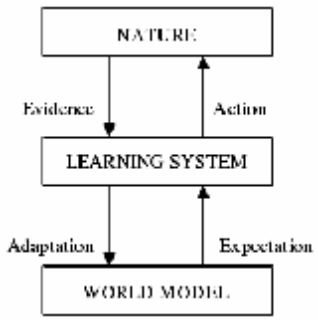


Fig. 4 The learning model of the proposed system

The learning system collects evidence through sensors. Then, it creates optimal actions to change the environment according to the objective (utility). These two steps are represented by Evidence and Action edges in Figure 4. On the other hand, the learning system adapts the world model (Bayesian network) using the evidence from the environment. In other words, adaptation is the parameterization of the BN utilizing the evidence. Then, the learning system calculates the expected state of the environment using the world model. Last two steps are represented by Adaptation and Expectation edges in the Figure 4. Evidence and action edges represent unsupervised learning while adaptation and expectation edges represent supervised learning. This justifies that the proposed learning system is bi-directional since supervised and unsupervised learning schemes are employed simultaneously.

5 . System Representation of the Decision-theoretic Intelligent Agent System

The decision-theoretic intelligent agent system has adaptive learning ability with feedback from the environment. The agent starts with a limited knowledge of the plant (environment), then it explores (samples) the plant to learn the plant's parameters. After it learns about the plant, it takes its actions accordingly. The agent first estimates the plant's behavior using the previous observation, and then takes its action according to the estimation. The plant, then responds to the agent's action with an output. The output of the plant in this stage is used as feedback to update the plant parameters in the predictor (BN). Figure 5 shows the decision theoretic-intelligent agent learning system in a block diagram.

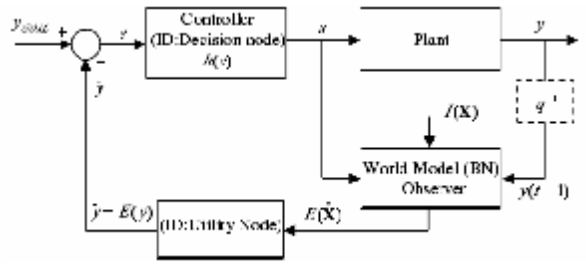


Fig. 5 System Block representation of the intelligent agent system

In Figure 5, $I(X)$ represents the initial state of the plant, $E(\hat{X})$ is the expected value of the state, $E(y)$ is the expected value of the plant output, and y_{GOAL} is the desired plant (system) output. The symbol q^{-1} represents one unit delay. The controller (ID) applies controls to the plant to provide a certain plant output because the controller creates the control according to the error between the expected value of the plant output and the reference. The reference is the desired output to be provided by the plant. The observer (BN) models the plant by using the plant's input/outputs. After a control is applied to the plant, the plant output is used in the next step to update the plant model. Thus, there is a time delay between the control and the output of the plant. The controller creates the control using a priori knowledge about the plant (environment).

The decision theoretic intelligent agent system (DTAS) has potential use in feedback control and adaptive control because it uses the plant's output as a feedback and modifies the controller and the observer accordingly.

5.1 Feedback Control

In the literature, there are two main types of feedback control, namely output feedback and state feedback. Output feedback is performed by a path (loop) from the output back to the controller as shown in Figure 6.

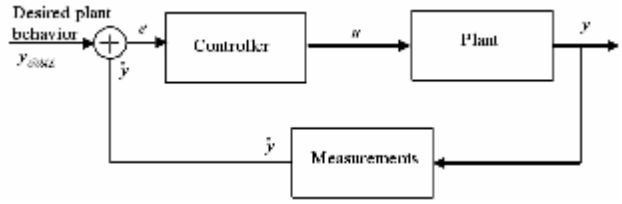


Fig. 6 Output feedback control

The equations for the system in Figure 6 can be given as:

$$e = \hat{y} - y_{GOAL} \quad (2)$$

$$u = f(e) \quad (3)$$

$$y = g(u) \quad (4)$$

Now, let us compare the system equations in the feedback control system and the decision-theoretic intelligent agent system. In the DTAS, the output of the plant, y, also depends on the control input, u. Let us compare the control signal u in both systems.

$$u_{DTAS} = h(e) \Leftrightarrow u_{FEEDBACK} = f(e) \quad (5)$$

If we choose the functions h and f to be equal, then the controllers will give the same control u with the same error e. Let us compare the errors in both systems. In the DTAS, the error is the difference between the desired output and the expected value of the plant output provided by the predictor. This is very similar to the feedback control system but the expected value of the plant output replaces the measured plant output. These two values are equivalent only if the predictor estimates the output of the plant well enough. In the DTAS, it is shown that the predictor estimates the plant output well enough when there is sufficient data from the plant's input/output. Therefore, the expected value in the DTAS is equivalent to the measured value of the plant output in a feedback control system. The following equations summarize the discussion.

$$e = y_{GOAL} - E(y) \quad (6)$$

$$E(y) \cong \hat{y} \quad (7)$$

$$e = y_{GOAL} - \hat{y} \quad (8)$$

From Equations (6), (7), and (8), we may conclude that the DTAS exhibits feedback control properties.

Another type of feedback control is state feedback control. In state feedback control, the state variables are sensed and fed back to the input through appropriate gains. If there is direct access to the state variables, the state variables can be easily measured and fed back to the input. If there is no direct access to the state variables, then an observer may be employed to perform the estimation of the state variables. Figure 7 illustrates a state feedback control system with an observer.

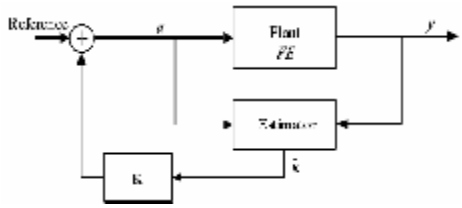


Fig. 7 A control system with the state feedback

In Figure 3, the block denoted by FE is the plant. The estimator predicts the state variables of the plant. The estimated state variables are fed to the input with a gain K. Then, the control signal becomes the following:

$$u = r + K\hat{X} \quad (9)$$

Thus, the control is a function of estimated state variables and the reference input. Let us compare the controls in both systems. In the DTAS, the control is defined as

$$u = f(e) \quad (10)$$

where $e = y_{GOAL} - \hat{y}$ .The term $\hat{y}$ represents the estimated output of the plant. The term $\hat{y}$ is a function of the estimated state variables because it is calculated by the utility function of the system. Therefore, we can represent $\hat{y}$ with the following equation.

$$\hat{y} = \hat{C}\hat{X} \quad (11)$$

where the vector $\hat{X}$ is the estimated state vector and the matrix $\hat{C}$ is the transformation matrix between the states and the output. Thus, the control can be rewritten as follows:

$$u = f(y_{GOAL} - \hat{y}) \quad (12)$$

$$u(X) = f(y_{GOAL} - \hat{C}\hat{X}) \quad (13)$$

Let us assume that the function f is a linear function with the following form.

$$f(x) = A \cdot x \quad (14)$$

$$u = A \cdot (y_{GOAL} - \hat{C}\hat{X}) = A \cdot y_{GOAL} - A \cdot \hat{C}\hat{X} \quad (15)$$

Let $K = -A \cdot \hat{C}$, and $r = A \cdot y_{GOAL}$, then the control becomes

$$u = r + K \cdot \hat{X} \quad (16)$$

As seen in Equation (16), the control signal in the DTAS can be interpreted as the control signal in the state feedback control. This concludes the analysis of how the DTAS corresponds to a feedback control system. It can be concluded that the DTAS will have the inherent advantages of feedback control. The following section investigates the adaptive control capabilities of the DTAS.

5.2 Adaptive Control

The term adaptive control covers a set of methods that provide a systematic approach for automatic adjustment of the controllers in real time, in order to achieve or to maintain a desired level of performance of the control system when the parameters of the plant dynamic model are unknown and/or change in time. A block diagram presenting a basic configuration of an adaptive control system is shown in Figure8.

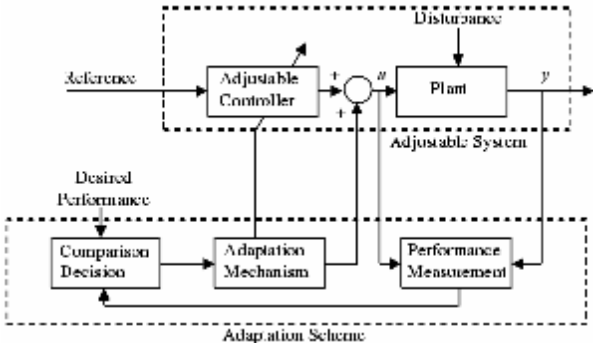


Fig. 8 A basic adaptive control system

The following definition provides an adaptive control system given in Figure 4. An adaptive control system calculates a certain performance index (IP) of the control system using the measured inputs, the states, the outputs, and the known disturbances. From the comparison of the performance index and a set of given ones, the adaptation mechanism modifies the parameters of the adjustable controller and/or generates an auxiliary control signal in order to maintain the performance index of the control system close to the set of given ones (i.e., within the set of acceptable ones).

An adaptive control system will monitor the performance of the system in the presence of parameter disturbances in addition to a feedback controller with adjustable parameters acting as a supplementary loop upon the adjustable parameters of the controller.

There are three types of adaptive control schemes in the literature: open loop adaptive control, direct adaptive control, and indirect adaptive control. In open loop adaptive control, the adaptation mechanism is a simple look-up table stored in the computer that gives the controller parameters for a given set of environment measurements. In the literature, this is also called gain-scheduling.

Direct adaptive control is based on the observation that the difference between the output of the plant and the output of the reference model (called plant-model error) is a measure of the difference between the real and the desired performance. The reference model is a realization of the system with desired performance. This information is used by the adaptation mechanism (called parameter adaptation) to directly adjust the parameters of the controller in real-time in order to force (asymptotically) the plant model-error to zero. This scheme corresponds to the use of Model Reference Adaptive Systems (MRAS) for the purpose of a general concept called Model Reference Adaptive System (MRAS) for the purpose of control. The indirect adaptive control was originally introduced by Kalman.

In an indirect adaptive control system, shown in Figure 9, the basic idea is that a suitable controller can be designed on line if a model of the plant is estimated on line from the available input-output measurements. The scheme is called indirect because the adaptation of the controller parameters is performed in two stages:

1 .On-line estimation of the plant parameters (e.g. Bayesian network construction)

2. On-line computation of the controller parameters based on the current estimated plant model (e.g. Influence Diagrams-making decisions)

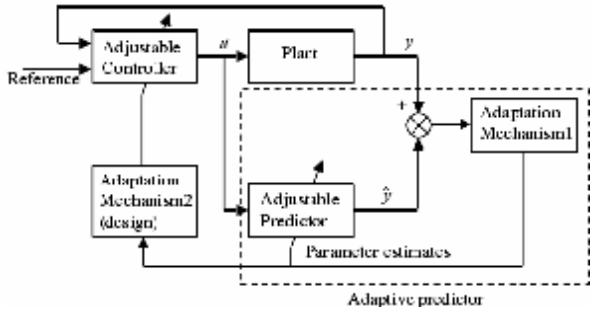


Fig. 9 Indirect adaptive control system

The main goal is to create an adjustable predictor for the plant output and compare the predicted output with the measured output. The error between the plant output and the predicted output (called prediction error or plant-model error) is used by a parameter adaptation algorithm which at each sampling instant will adjust the parameters of the adjustable predictor in order to minimize the prediction error in the sense of a certain criterion.

There are two options given to effectively implement an indirect adaptive control strategy. The choice is related to a certain extent to the ratio between the computation time and the sampling period.

Strategy 1:①Sample the plant output;②Update the plant model parameters;③Compute the controller parameters based on the new plant model parameter estimates;④ Compute the control signal;⑤Apply the control signal;⑥ Wait for the next sample.

In this strategy, there is a delay between $u(t)$ and $y(t)$. This delay should be smaller than the sampling period.

Strategy 2: ① Sample the plant output; ② Compute the control signal based on the controller parameters computed during the previous sampling periods;③Apply the control signal;④Update the plant model parameters; ⑤Compute the controller parameters based on the new plant model parameter estimates; ⑥ Wait for the next sample.

In the second strategy, the delay between $u(t)$ and $y(t)$ is smaller than in the previous case. In this strategy, a priori parameter estimation is performed since we apply the control without updating the plant parameters.

In the above paragraphs, a general definition of an adaptive control system is provided. A greater importance is given to indirect adaptive control systems because the decision-theoretic agent system (DTAS) has the properties of an indirect adaptive control system. The DTAS has the same steps as the indirect adaptive control system.

Additionally, the learning strategy in DTAS is very similar to the second strategy of the indirect adaptive control system.

The first step, the on-line estimation of the plant model parameters, is performed by structuring a Bayesian network and calculating its parameters in the DTAS. The online Bayesian network learning is performed to model the plant. The second step, the online computation of the controller parameters, is performed by a decision system (influence diagrams).

As shown in Figure 5, there are two adaptation mechanisms in the indirect adaptive control. The first adaptation mechanism corresponds to the online Bayesian network learning in the DTAS. The second adaptation mechanism corresponds to the utility node in the influence diagram part of the decision-theoretic intelligent agent because it determines which action will be fired in the decision node. The adjustable predictor corresponds to the Bayesian network in the DTAS. Finally, the adjustable controller corresponds to the decision nodes in the influence diagram in the DTAS.

Now, the indirect adaptive control system can be redrawn by using the decision-theoretic intelligent agent components, shown in Figure10.

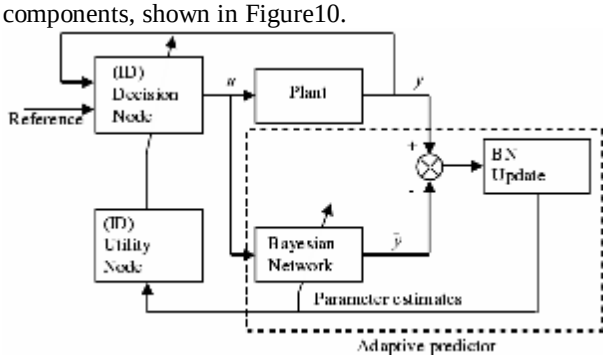


Fig. 10 Indirect adaptive control representation of the DTAS

Consequently, the online Bayesian learning determines the plant model structure and parameter estimation; and, the influence diagram determines the controller parameters. Therefore, it can be concluded that the decision-theoretic intelligent agent system implements an indirect adaptive control system.

6 Conclusions

Self-organization of the intelligent agents is accomplished because each agent models other agents by observing their behavior. Agents have belief, not only about environment, but also about other agents. To study the proposed intelligent agent's learning and self-organizing abilities, in this paper, we explain the structure of an agent, which is designed by a Bayesian network and an influence diagram, and then examine a multi-agent organization system and the bi-directional learning feature of the proposed multi-agent self-organizing system. We present the system representation of the decision-theoretic intelligent agent design. The decision-theoretic intelligent agent system has adaptive learning ability with feedback from the environment. The agent starts with a limited knowledge of the plant (environment), then it explores (samples) the plant to learn the plant's parameters. After it learns about the plant, it takes its actions accordingly.

References

[1] C. Boutilier, "Planning, learning and coordination in multi-agent decision processes," in Sixth conference on Theoretical Aspects of Rationality and Knowledge, The Netherlands, 1996.
[2] C. Claus, "Dynamics of multi-agent reinforcement learning in Cooperative multi-agent systems," Ph.D. Dissertation, Univ. of British Colombia, Canada, 1997.
[3] C. Gerber, "Evolution-based self-adaption as an expression for the autonomy degree in multi-agent societies," in Proceedings of the IEEE Joint Conference on the Science and Technology of Intelligent Systems, Gaithersburg, MD, pp. 741-746, September 1998.
[4] D. Suryadi and P. J. Gmytrasiewicz, "Learning models of other agents using influence diagrams," in Proceedings of User Modeling: The Seventh International Conference, Springer Wien, New York, 1999, to apper.
[5] J. Pearl, "Constraint-propagation approach to probabilistic reasoning," in L. M. Kanal and J. Lemmer (Eds.), Uncertainty in Artificial Intelligence, North-Holland, Amsterdam, pp. 357-288, 1986.
[6] J. Pearl, Probabilistic Reasohihg in Intelligent Systems: Networks of Plausible Inference. San Mateo, CA: Morgan Kaufmann, 1988.
[7] J. Pearl, "A probabilistic calculus of actions," in Proceedings of the Tenth Conference oh Uncertainty in AI (UAI-94), San Mateo, CA: Morgan Kaufmann, 1994.
[8] N. Friedman, K. Murphy, and S. Russell, "Learning the structure of dynamic probabilistic networks," in G.F. Cooper and S. Moral (Eds.), Proceedings of Fourteenth Conference oh Uncertainty in Artificial Intelligence (UAI '98), San Francisco, CA: Morgan Kaufmann, 1998.
[9] S. L. Lauritzen and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," Journal of the Royal Statistical Society, Series B, vol. 50(2), pp. 157-224, 1988.
[10] S. Non and P. J. Gmytrasiewic, "Coordination and belief update in a distributed anti-air environment," in Proceedings of the 31S' Hawaii International Conference on System

Sciences, vol. V, pp. 142-145, Los Alamitos, CA: IEEE Computer Society, January 1998.
[11] S. Sen and M. Sekaran, "Multi-agent coordination with learning classifier systems," in Proceedings of the IJCAI Workshop oh Adaptation and Learning in Multi-agent Systems, Montreal, pp. 84-89, 1995.
[12] S. Russell and P. Norvig, Artificial Intelligence: A modern Approach, New Jersey: Prentice Hall, 1995.
[13] T. Malsch and I. Schulz-Schafer, "Generalized media of interaction and inter-agent coordination," in Socially Intelligent Agents-Papers from the 1997 AAAI Fall Symposium, Technical Report FS-97-02, AAAI, 1997.
[14] Y. Shoham, Agent-oriented programming," Artificial intelligence, vol. 60(1), pp. 51-92, 1993.

Author Yonghui CAO received the MS degree in business management from Zhejinag University in 2006. He is currently a doctorate candidate in Zhejiang University. His research interest is in the areas of management information systems.

Refactoring Model of Legacy Software in Smart Grid based on Cloned Codes Detection

Fanqi Meng[1], Zhaoyang Qu[2] and Xiaoli Guo[3]

[1] School of Information Engineering, Northeast Dianli University,Jilin, Jilin 132012, China

[2] School of Information Engineering, Northeast Dianli University, Jilin, Jilin 132012, China

[3] School of Information Engineering, Northeast Dianli University, Jilin, Jilin 132012, China

Abstract

The construction of smart grid relies on the development of many new software systems, whereas it would be very expensive and time-consuming if these new software systems are completely developed anew. Since the existence of many legacy software systems in the former power grid, the problem may be solved well supposing that those legacy software systems are reused reasonably and efficiently in the construction of smart grid. In view of this situation, a refactoring model of legacy software is proposed. The model is based on reverse engineering and its kernel is cloned codes detection and components extraction. Firstly, the cloned codes in the scanned source code of the legacy software will be detected by means of CCFinder. Secondly, the abstract syntax trees of the functions which include the cloned codes will be created. Thirdly, the degree of variation between the functions which include the cloned codes belonging to the same clone set will be calculated according to their abstract syntax trees, and then some functions whose similarities of abstract syntax trees are in the allowed range will be combined. Finally, the combined functions and other frequently invoked functions will be encapsulated in a new class (or a DLL file), and all of these classes (or DLL files) will be reused as components in the development of new software systems of the smart grid.

Keywords: *Smart Grid, Legacy System, Code Clone, Refactoring*

1. Introduction

Although the term “smart grid” has been used since at least 2005 [1], it still hasn’t a uniform definition in the entire world. However, all of the countries consider smart grid as the inevitable trend of the development of electrical grid, so that smart grid has another name called “electrical grid 2.0”. A smart grid is an advanced electrical grid that can gather and process the information by using computers and other technology. The gathered and processed information has widely resources, ranging from the behaviors of suppliers and consumers to the status of devices running in smart grid. The processing of the collection and computation of the information should be in an automated fashion, in order to enhance the efficiency, reliability, economics, and sustainability of the supply of electricity [2]. The above background and the features (especially for efficiency and reliability) of smart grid decide that its construction relies on the development of many new software systems (such as monitoring software, controlling software, marketing software, etc.) to gather and process the information. However, it must be very expensive and time-consuming if these new software systems are totally developed anew. The cost and deadline of the task must be considered under the circumstances. Thus, an efficiency software development mode for smart grid is imperative.

Legacy software usually is a large-scale and complex software system which has run for a long time (more than 20 years) [3]. Since the development language of legacy software mostly is the third or early programming language (such as ASM, COBOL or Turbo C etc.), and the development framework of legacy software has been outdated, the legacy software is hardly to be maintained and evolved. Even if it is no longer used, legacy software may continue to impact the organization due to its historical role. Most functions in legacy software are stability and credibility in processing the existing business, thus the method that reuses these functions in developing new software systems which can handle both existing business and emerging business has been adopted by many programmers. The smart grid commonly has many legacy software systems which had been used by former electrical grid. Using them efficiently in the construction of a smart grid is potentially the solution to the above problem (expensive and time-consuming). So the study of the method to efficiently use legacy software is significant to the construction of smart grid.

Refactoring is a programming technique for optimizing the structure or pattern of an existing body of code by altering its internal nonfunctional attributes without changing its external behavior[4][5][6][7]. By applying a series of "refactorings", the software can obtain some advantages,

including improved code readability and reduced complexity to improve the maintainability of the source code, as well as a more expressive internal architecture or object model to improve extensibility. The code refactoring of legacy software is one of basic methods to achieve efficient reuse. This paper presents a study of how to reuse the legacy software by means of refactoring. The remainder of the paper is organized as follows. Section 2 gives an overview of the related work in the area. The refactoring model of legacy software is proposed in Section 3. And its kernel processes, cloned codes detection and components extraction are presented in Section 4. Section 5 analyzes the results obtained in a number of experiments and Section 6 outlines the conclusions and future work.

2. Related work

Software reuse is still a popular research issue in the field of software engineering. The modernization of legacy software is an important research direction of software reuse. The evolution of software systems can be divided into three types: maintenance, modernization and replacement. Maintenance can only meet the small changes of the requirements by correcting and enhancing the functions of the software system. Replacement has high risk and will cost long time to develop new system. Replacement will not happen unless the system can't be maintained or modernized. So the modernization of legacy software is now regarded as the most feasible method in software reuse. There already exist several modernization methods to deal with legacy systems, for reusing them in the development of new systems. These methods mainly fall into three categories: Redevelopment, Wrapping and Migration.

2.1 Redevelopment

Redevelopment is a high-risk and low-reuse method, almost abandons whole codes of the legacy systems. Since the methods of this kind realize the functions of the legacy systems in the new system via programming anew, they are usually used to eliminate the structural flaws of legacy systems. CORUM (Common Object-based Re-engineering Unified Model), CORUMⅡ, MARMI-RE and OSET are all the methods belonging to redevelopment[8][9][10].

2.2 Migration

Wrapping methods can be classified into three types: UI-based wrapping (UI, user interface), data-based wrapping and function-based wrapping, according to the wrapped contents. UI-based wrapping reuses the UIs of the legacy system in the new system by interface mapping. Data-based wrapping includes the means, such as DB (data base) gate, XML and data copy etc. Data-based wrapping inherits the data structure of the legacy system, so the data of the legacy system can be used in the new system. Function-based wrapping uses component wrapping, object wrapping and gate wrapping etc. to realize the reusing of the service logic. These wrapping methods can reuse legacy systems for a short time, but they will increase difficulties in the maintenance and management of the new system.

2.3 Migration

The migration of legacy systems usually divides into two types: component-based migration and system-based migration. Component-based migration classifies the legacy system into independent components, and then migrates the components singly. System-based migration integrates the whole legacy system and its data into the new system. Representative methods and models of migration are Chicken Little, Butterfly, SGF and AGRIP etc. Migration merely suits for small-scale legacy systems, since it is more possible for losing information if the scale of the legacy system is larger.

3. Refactoring model

The refactoring of legacy software is a process to reengineering the old software system by component technology. This process can be roughly divided into two steps: The first step is reverse engineering. Reverse engineering is the process of analyzing a subject system to create representations of the system at a higher level of abstraction [11]. It can also be seen as going backwards through the development cycle. Reverse engineering often involves taking computer program apart and analyzing its workings in detail to be used in maintenance, or to try to make a new program that does the same thing without using or simply duplicating (without understanding) the original. The second step is forward engineering. Forward engineering has the process similar to conventional development of software. It follows the flow: requirements analysis, outline design, detailed design, testing and modification. Figure 1 shows the refactoring model that is built to reengineer the legacy software in smart grid based on component extraction, update and reuse.

In this model, firstly, Requirement Change leads to the Architecture Readjustment of legacy software system; Then, Architecture Readjustment needs Component Update to provide new components; Finally, Component Update helps Software Refactoring coming true. On the stage of requirement analysis, according to the change of requirement, Requirements Analysis Engineers increase

new requirements or delete useless requirements based on the result of Requirement Analysis that comes from the reverse engineering of legacy software.

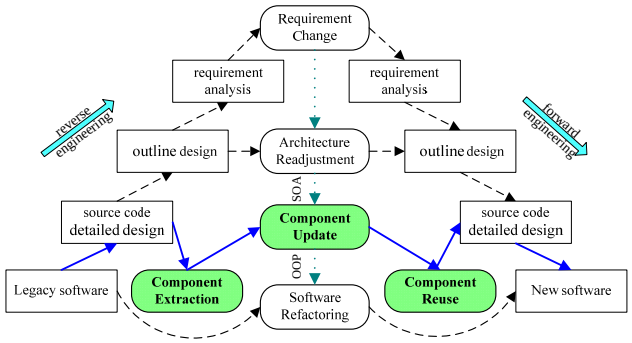


Fig. 1 The refactoring model of legacy software in Smart Grid

On the stage of outline design, engineers readjust the architecture of legacy software to Service Oriented Architecture (SOA). On the stage of detailed design, programmers update component, including abandoning useless components and regaining new components, in order to make the components compatible with the demand of Architecture Readjustment and Object-Oriented Programming (OOP). Usually, three ways can be used to gain the needed components. The first one is to purchase from others; the second one is to renew development by yourself; the last one is to extract components from the source code of legacy software. The model we proposed adopts the third method to get components, so the Component Extraction in the model is both the beginning of refactoring process and the kernel method of the refactoring model.

4. Components extraction

The refactoring of legacy software is also known as Software Systems Modernization. Software Systems Modernization using SOAs and Web Services represents a valuable option for extending the lifetime of mission-critical legacy systems [12]. Components play an important role in SOA. Software engineers regard components as part of the starting platform for service-orientation. Actually, the refactoring model uses component-based software engineering (CBSE) as the forward engineering method. Figure 2 shows the refactoring process from component perspective.

4.1 Cloned codes detection

Copying code fragments and then reuse by passing with or without minor modifications or adaptations are common activities in software development. This type of reuse approach of existing code is called code cloning and the passed code fragment (with or without modifications) is called a clone of the original. For instance, Baker has found that on large systems between 13% - 20% of source code can be cloned code. For an object-oriented COBOL system, the rate of duplicated code is found even much higher, about 50% [13].

Although code clones may adversely affect the software systems' quality, especially their maintainability and comprehensibility, the cloned code in legacy software are potentially most valuable code to be refactored into new components. One piece of code is cloned more times, and then it has more reuse value. So we should detect cloned code before refactoring. In addition, cloned code detection will compress the length of source code in legacy software, and reduce the work load in component extraction.

We have used CCFinder to detect the cloned code in legacy software. CCFinder is a token-based cloned code detection tool [14]. The work principle of CCFinder is followed: First, each line of source code is divided into tokens by a lexer and the tokens of all source code are then concatenated into a single token sequence. The token sequence is then transformed. After that, each identifier related to types, variables, and constants is replaced with a special token. A suffix-tree based sub-string matching algorithm is then used to find the similar sub-sequences on the transformed token sequence where the similar sub-sequence pairs are returned as clone pairs/clone classes. Once the clone pair/clone class information is obtained with respect to the token-sequence(s), a mapping is required for obtaining the clone pair/clone class information with respect to the original source code. Figure 3 shows the detection interface of CCFinder 10.2.5.0 (download from http://www.ccfinder.net/)

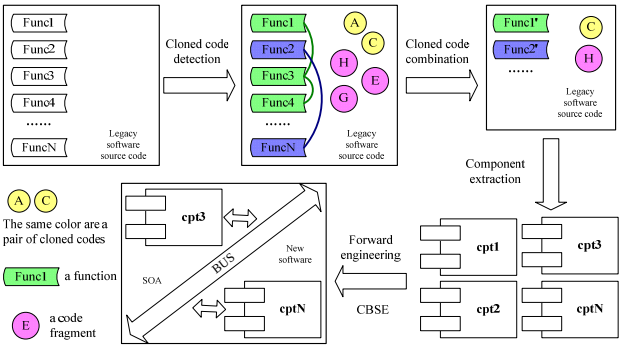


Fig. 2 The process of component extraction

4.2 Abstract syntax trees creation

Abstract syntax tree is a production generated after the lexical analysis and parsing of source code. Abstract

syntax tree fully reflects the grammatical structure of the source code, and its leaves represent identifier or constant etc. Figure 4 shows an abstract syntax tree of a code segment.

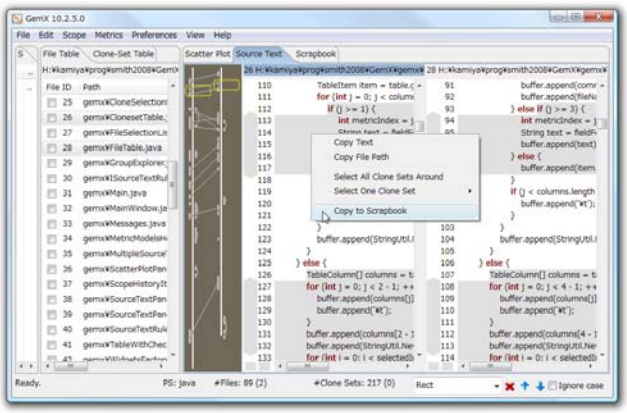


Fig. 3 The detection interface of CCFinder

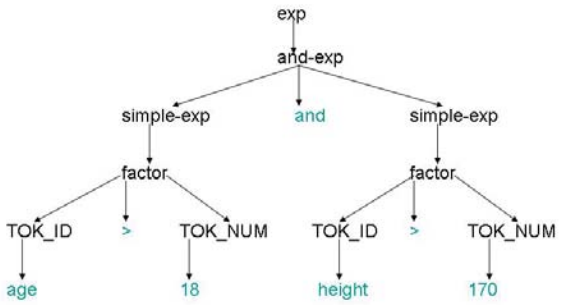


Fig. 4 An example of abstract syntax tree

Function is the basic unit in the third generation programming languages which are the main tools used in software development 20 years ago. Therefore the emphasis of reusing legacy software is functional refactoring for reuse. According to the refactoring model, the functions include cloned code which will be compared for calculating the similarity. Generally, the two functions with similar syntax structure are probably the same. So the abstract syntax tree of the functions includes cloned code which should be built after code clone detection. Various tools for building abstract syntax tree can be downloaded easily from Internet, for example The GNU Compiler Collection and JavaCC.

4.3 Differences degree calculation

The functions may be very different even though they have cloned code belonging to the same clone set. The relationship between the function and the cloned code probably has two cases. In **the first case**, the function is totally cloned (mainly exists between different legacy software). In **the second case**, the function includes the cloned code. Even though some large cloned code may also include functions, but the larger cloned code can finally be divided into the functions totally cloned and the small cloned code pieces included by functions. In addition, the relationship between cloned codes which are detected by CCFinder also has three cases. In **case one**, the two cloned codes are the same. In **case two**, they are merely different in some identifiers' name. In **case three**, they may be minor different in variable types or syntax. All above cases happened because the detective method of CCFinder is token-based.

4.3.1 Totally cloned

If the function is totally cloned according to the detection result presented by CCFinder (it still has three cases discussed above), then we should calculate its difference degree with other related cloned functions by traversing abstract syntax tree twice. We directly compare the value of the node of the abstract syntax trees with the functions in the first traversing. If the result shows that the abstract syntax trees are the same, it means that the two functions are the same (note it as **Type A**). If not, we change all customer identifiers into $ when traversing the abstract syntax trees. If the result shows the same, the two functions are merely different in some identifiers' names (note it as **Type B**); else they are different in variable types or others (note it as **Type C**).

4.3.2 Partly cloned

If the function is partly cloned, it means that the function includes cloned code in its body. We traverse the abstract syntax tree of the function with changing customer identifiers into $. Those functions which include the cloned code belonging to the same clone set will be compared with traversing results. We adopt Levenshtein Distance (or Edit Distance) for the compare method. The algorithm of the calculation of Levenshtein Distance between the two string fp1 and fp2 is shown as follows [15]. The different degree between two functions can be gotten via calculating the expression: ***DD= (matrix (len1, len2)/max (len1, len2))*100%***. The bigger the value of DD is, the more different the two functions are. We can use a threshold value to decide whether the functions are similar. If the value of DD is below the threshold value, note the two functions as **Type D**, else noted them as **Type E**.

4.4 Cloned functions combination

A cloned function is a function with cloned code. The combination of cloned functions can be divided into five cases according to the types of cloned function.

```
1: len1 ← strlen(fp1)
2: len2 ← strlen(fp2)
3: initialize_two_dimensional_matrix(matrix, len1, len2)
4: for i = 0 → len1 do
5:    for j = 0 → len2 do
6:       if fp1[i] = fp[j] then
7:          cost = 0
8:       else
9:          cost = 1
10:      end if
11:      matrix[i, j] = min(matrix[i−1, j]+1, matrix[i, j−
         1] + 1, matrix[i − 1, j − 1] + cost)
12:   end for
13: end for
14: return matrix(len1, len2)
```

Fig. 5 The algorithm of Levenshtein Distance calculation [15]

Case 1, the functions which will be combined are Type A. In this case, all of the functions are the same, so we select one of them and add it into function base.

Case 2, the functions which will be combined are Type B. In this case, all of the functions are very similar except individual identifier's name, so we select the shortest one for saving space and add it into function base.

Case 3, the functions which will be combined are Type C. In this case, the differences between the functions are in types of variable or in other aspects, so we select the longest one for retaining enough information and add it into function base.

Case 4, the functions which will be combined are Type D. In this case, even though the difference degree is lower than a preset threshold value, the functions are more different with each other than Type A, Type B and Type C. so we should flexibly adopt various existent refactoring method to combine the functions. The combined function will be added into function base.

Case 5, the functions are Type E. Since the similarity is too low, the functions are not recommended to combine. All of the functions are respectively added into function base.

All of the functions in the function base have a value which denotes invoked times in legacy software. The value is an important reference for component extraction. In Case 1 to Case 4, the invoked time of a combined function is the sum of invoked times of all related cloned functions.

4.5 Component extraction

A component may be a software package, a Web service, or a module that encapsulates a set of related functions (or data). Programmers can use these functions which have been stored in function base as various forms. For example, some functions can be encapsulated into a new class as function members by tiny modification, or assembling some functions to generate a DLL files. In addition, several tools have been used in extracting components, such as CodeMiner, CARE (Computer-Aided Reuse Engineering) and PATRicia (Program Analysis Tool for Reuse) [16].

5. Analysis and result

We did some experiments about cloned code detection which is the basic work in the model. We selected the former versions of Cook, Snns, Weltab and Postgresql as our experimental subjects. The four applications were both written in C or C++. We used CCFinder to detect cloned codes hidden in the four applications. Before the detection, the value of Minimum Clone Length was set at 120 and the value of Minimum TKS was set at 30. These values were more suitable for two reasons: firstly, the codes will be no much reuse value if its length is shorter than 40 characters and 10 TKS; secondly, the cloned codes usually do not have a length longer than 200 characters and 50 TKS. The metric results are shown as Table 1, Table 2 and Table 3.

Table 1: File metrics

Name	***Min.***	***Max.***	***Average***
LEN	1	33586	1000.28
CLN	0	11	0.128866
NBR	0	11	0.181701
RSA	0	1	0.056779
RSI	0	0.51	0.008532
CVR	0	1	0.064539
RNR	0.024	1	0.900082

The meaning of the Names in the tables can be found at http://www.ccfinder.net/doc/10.2/en /tutorial-gemx.html

Table 2: Clone set metrics

Name	***Min.***	***Max.***	***Average***
LEN	123	3220	732.071
POP	2	12	3.38571
NIF	1	12	2.85714
RAD	0	6	1.2
RNR	0.52	0.996	0.789677
TKS	30	53	34.5143

LOOP	0	17	3.35714
COND	2	48	14.4714
McCabe	3	58	17.8286

Table 3: Line-based metrics

Name	*Total*	*Min.*	*Max.*	*Average*
LOC	432005	2	11618	278.354
SLOC	217849	0	5562	140.367
CLOC	12750	0	603	8.21521
CVRL	-	0	1	0.058527

From Table 3, we know that the four applications have 432,005 lines in their source files, and 217,849 lines including at least one token. In other words, 217,849 lines are executable codes. 12,750 lines are cloned codes. The ratio of the lines including cloned codes is 0.058527. The ratio is lower than common case because of the bigger preset value for Minimum Clone Length and Minimum TKS. According to these results, we found some cloned functions which can be treated as reusable component candidates. Figure 6 shows a pair of Type A cloned functions, named *next_token*. They are found in different files of the same application (hba.c and miscinit.c in postgresql). Figure 7 shows a pair of Type B cloned functions, named *TEST_JE_Backprop* and *TEST_JE_BackpropMomentum*. They are detected in the same file of same application (learn_f.c in snns). Besides the different function name, the two functions have another difference in line 5698 and 5743 (if condition, <3, <5). Figure 8 shows a pair of Type B cloned functions found in different files of different applications (Lex.yyz.c in snns and bootscanner.c in postgresql). More cloned functions are Type D or Type E, but we didn't further study these partly cloned functions in our experiments.

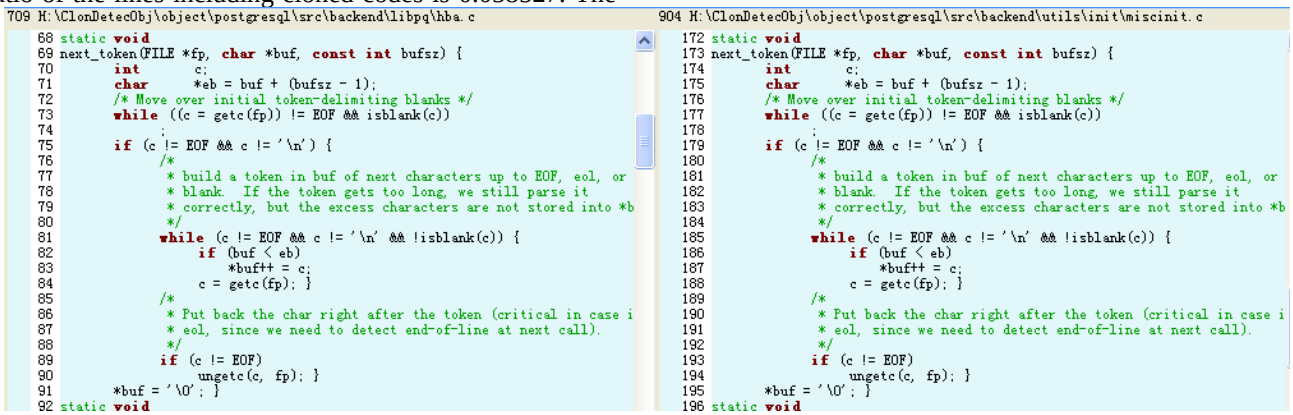


Fig. 6 An example of Type A cloned functions from different files of the same applications

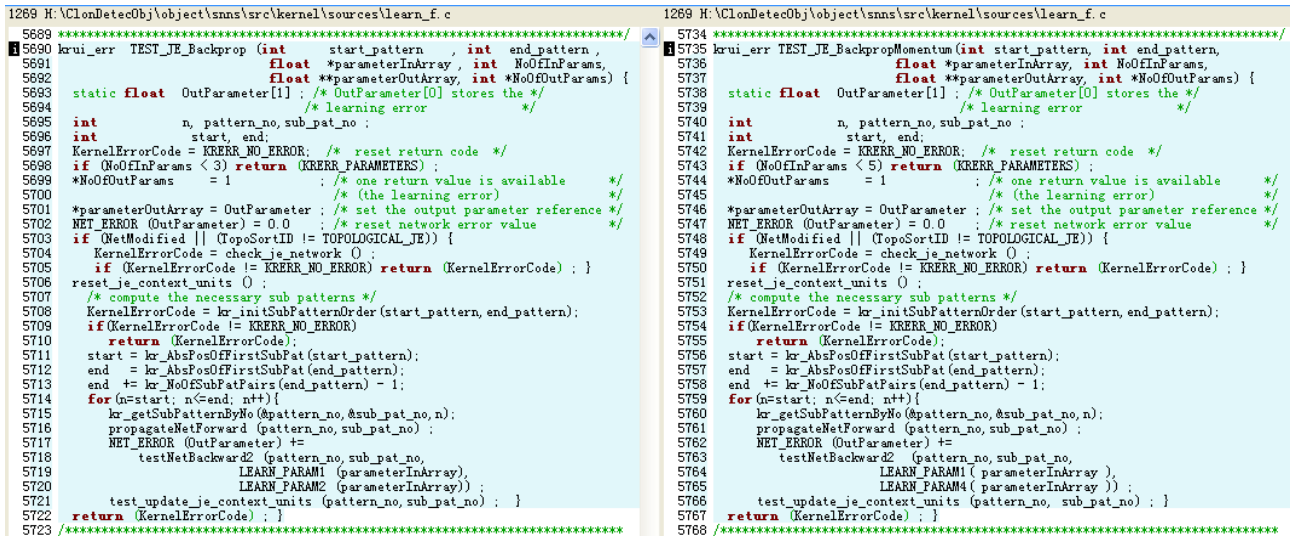


Fig. 7 An example of Type B cloned functions from a same file of a same application

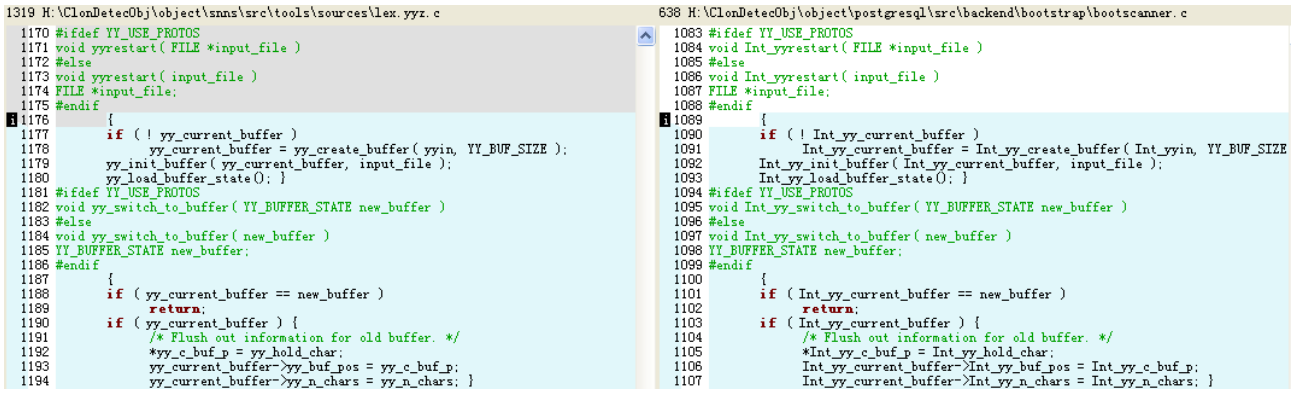


Fig. 8 An example of Type B cloned functions from different files of different applications

6. Conclusions

The legacy software that was used by former electrical grid may help the construction of smart grid in efficiencies and costs, but it depends on whether the legacy software can be reused. In order to reuse the legacy software efficiently, we proposed a refactoring model based on cloned code detection. By detecting cloned code, we can firstly reduce the candidates for component extraction, thereby lower the complexity; secondly, the remained cloned functions after function combination are more valuable for components generation, thus enhance the reliability of the refactoring. However, the result shows that the valuable cloned functions are not too much in legacy software, so the method of this model should be used as a subsidiary method in refactoring large-scale legacy software.

Acknowledgments

This work was funded in part by Natural Science Foundation of China (Grant Number 51077010) and Natural Science Foundation of JiLin province of China (Grant Number 20101517).

References

[1] S. Massoud Amin, Bruce F. Wollenberg, “Toward a smart grid”, IEEE P&E Magazine, vol. 3, no. 5, 2005, pp. 34 – 41.

[2] Amrita Dey, Nabendu Chaki, Sugata Sanyal, "Modeling Smart Grid using Generalized Stochastic Petri Net", JCIT, AICIT, vol. 6, no. 11, 2011, pp. 104 – 114.

[3] Wen-Shin Hsu, Jiann-I Pan, Hua Hu, "Exercise Prescription Monitoring System: Using Sensor Network and Service-Oriented Architecture", IJMIA, AICIT, vol. 2, no. 2, 2012, pp. 44 – 55.

[4] Nien-Lin Hsueh, Peng-Hua Chu, "A Pattern-based Refactoring Approach for Multi-core System Design", IJACT, AICIT, vol. 3, no. 9, 2011, pp. 196 -209.

[5] Ibrahim, Safwat M.”Identification of Nominated Classes for Software Refactoring Using Object-Oriented Cohesion Metrics”, International Journal of Computer Science Issues, v 9, n 2 2-2, p 68-76, 2012

[6] Ananda Rao, A. “Identifying clusters of concepts in a low cohesive class for extract class refactoring using metrics supplemented agglomerative clustering technique”, International Journal of Computer Science Issues, v 8, n 5 5-2, p 185-194, 2011.

[7] Arora, Madhulika. “Refactoring, way for software maintenance”, International Journal of Computer Science Issues, v 8, n 2, p 565-570, 2011

[8] Woods Steven, O'Brien Liam, Lin Tao, Gallagher Keith, “An architecture for interoperable program understanding tools”, 6th International Workshop on Program Comprehension, 1998, pp. 54 – 63.

[9] Rick Kazman, Steven G. Woods, S. Jeromy Carrière, “Requirements for Integrating Software Architecture and Reengineering Models: CORUM II”, WCRE '98 Proceedings of the Working Conference on Reverse Engineering, 1998, pp. 154 – 163.

[10]Eun Sook Cho, Jung Eun Cha and Young Jong Yang, “MARMI-RE: A Method and Tools for Legacy System Modernization”, Lecture Notes in Computer Science, Vol. 3647, 2006, pp. 42-57.

[11]Chikofsky, E. J.; Cross, J. H, "Reverse engineering and design recovery: A taxonomy", IEEE Software, vol. 7, no. 1, 1990, pp. 13–17.

[12]Gerardo Canfora, Anna Rita Fasolino. “A wrapping approach for migrating legacy systeminteractive functionalities to Service Oriented Architectures”, The Journal of Systems and Software,vol. 81, 2008, pp. 463 – 480.

[13]Dongxiang Cai, Miryung Kim, “An Empirical Study of Long-Lived Code Clones”, International Conference on Fundamental Approaches to Software Engineering, 2011, pp. 432–446.

[14]Toshihiro Kamiya, Shinji Kusumoto, Katsuro Inoue, “CCFinder: A Multilinguistic Token-Based Code Clone Detection System for Large Scale Source Code”, Transactions on Software Engineering, Vol. 28, no.7, 2002,pp. 654- 670.

[15]Wu Zhou, Yajin Zhou, Xuxian Jiang, "Detecting Repackaged Smartphone Applications inThird-Party Android Marketplaces", In Proceedings of the 2nd ACM Conference on Data and Application Security and Privacy,2012,pp.120-130.

[16]MF Dunn, JC Knight, "Automating the detection of reusable parts in existing software", In Proceedings of the 15th international conference, 1993, pp.10-19.

Fanqi Meng is a member of ACM, and received his M.S. degree in computer science and technology from Northeast Dianlil University, Jilin, China, in 2010. He is now a lecturer in the School of Information Engineering, Northeast Dianlil University. His main research interest is cloned code detection and refactoring.

Zhangyang Qu received his Ph.D. degree in power system automation from North China Electric Power University, Bejing, China, in 2010. He is now a professor and tutor of postgraduates in the School of Information Engineering, Northeast Dianlil University. His main research interest is power system information processing.

Xiaoli Guo received her M.S. degree in computer science and technology from Changchun University of Science and Technology, Jilin, China, in 2006. She is now a professor and tutor of postgraduates in the School of Information Engineering, Northeast Dianlil University. Her main research interest is computer education.

Rolling Bearing Diagnosis Based on LMD and Neural Network

Baoshan Huang[1,2], Wei Xu[3*] and Xinfeng Zou[4]

[1] National Key Laboratory of Vehicular Transmission, Beijing Institute of Technology, Beijing, China
[2] Beijing Institute of Technology, Zhuhai, China
[3,4] University of Macau, Macao SAR

Abstract

Inner ring pitting, the outer indentation and rolling element wear are typical faults of rolling bearing. In order to diagnose these faults rapidly and accurately, the paper proposes a novel diagnosis method of rolling bearing based on the energy characteristics of PF component and neural network by the vibration signal of local mean decomposition(Local mean decomposition, LMD). The vibration signal is decomposed into several PF components by the local mean decomposition, the calculated energy characteristics of the PF component are inputted to the neural network to identify the type of rolling bearing faults. At the same time, the genetic algorithm is introduced to optimize the structure parameters of neural network, which improves diagnostic rate and accuracy of faults. The results show that this method has a higher diagnosis and recognition rate for the typical faults of rolling bearing.

Keywords: *Rolling bearing, LMD, Genetic algorithms, Neural network, Fault diagnosis.*

1. Introduction

Rolling bearing is a critical part of the transmission gear. Failures like inner ring pitting, outer ring creasing and rolling elements wear, etc. can be resulted from wear, fatigue, corrosion, overload and so on, while the equipment is operating. The measured vibration acceleration signal of the roller bearing is a kind of typical non-stationary signal which reflects weak energy of status information, this will bring some trouble to fault diagnose because of its. Therefore, knowing how to extract fault information characters from non-stationary vibration signals is very important for Fault Diagnosis of Rolling Bearings. Up to now, the main approaches to process non-stationary signal include Wigner Distribution, short-time Fourier transform, wavelet transform, EMD and LMD, etc. But all of them have their own limitation. For example, cross terms appear when analysis multicomponent signal by using Wigner Distribution, the time-frequency window of short-time Fourier transform is fixed; although the time-frequency window of wavelet transform is variable, but it is also mechanical lattice type division of time-frequency plane, the same as Fourier transform, so essentially, it's not a kind of self-adaptive signal process approach;

Problems like over envelope, owe envelop, mode confusion, end effect, IMF criterion and no fast algorithm, etc. and the unexplainable negative frequency will be produced by using Hilbert transform to get analytic signal and compute instantaneous frequency. Recently, a new approach of self-adaptive time-frequency analysis which is called Lockalmean decomposition (LMD) is proposed by Jonathan S.Smith. LMD represent a complicated multicomponent signal as the sum of several production functions (PF). Each PF component is the product of one envelope signal and one pure FM signal, and the complete time-frequency distribution (TFD) of original signal is combination of instantaneous amplitudes and instantaneous frequencies of all PF components, the characteristic information of original signal can be achieved more precisely and effectively, so LMD is an ideal method to process multicomponent AM and FM signal, to extract energy of PF component as characteristic. This paper presents an approach of combining neural network with energy characteristic of PF component extracted by using LMD to diagnose faulty rolling bearing. First, decompose the vibration signal by LMD to get energy characteristic of PF component as the input of neural network, and optimize the neural network structure parameters by applying genetic algorithm to improve the faults recognition speed and accuracy. Comparing the diagnostic result of rolling bearing work in well condition, work with inner ring pitting, work with outer ring indentation and work with rolling body wear, it shows the approach of combining neural network with energy characteristic of PF component proposed by this paper has the advantages of faster diagnosis speed and higher accuracy, and at the same time, it shows this method is available for classic fault diagnosis of bearings in gearboxes.

2. LMD method

Essentially, LMD is a kind of method that isolates pure FM signal and envelop signal from original signal, next for loop processes PF components (these components have physical significance) which are products of the pure FM signal and envelop signal until all PF components are extracted, then we can get the time-frequency distribution of the original signal. The original signal $x(t)$ is the sum of all the PF components and, that is:

$$x(t) = \sum_{p=1}^{k} PF_p(t) + u_k(t) \quad (1)$$

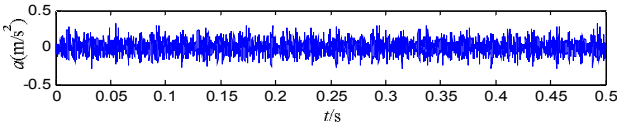


Fig. 1 Vibration signal of a rolling bearing with inner race fault

Fig. 1 shows the vibration acceleration signal of a rolling bearing with inner ring fault. The result of decomposing the signal using LMD method is shown in Fig. 2, the complicated multiple-component AM and FM signal have been decomposed to simple component AM and FM signal. The different characteristic components can be reflected by the relationship between the PF components and the corresponding components of the signal.

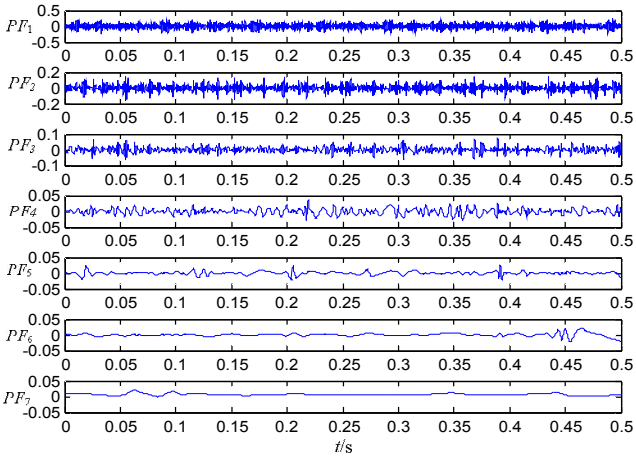


Fig. 2 LMD decomposition result of the vibration signal of the rolling bearing with inner ring fault

Using LMD method to decompose the signals from 4 different kinds of rolling bearings: proper one, one with inner ring pitting, one with outer ring indentation and one with rolling element abrasion. Then compute the energy of each in the top 7 layers after decomposing as the characteristic vector, the results shown in the following table 1:

Table 1 Character vector of normal, inner race pitting, rolling body abrasion and outer race indentation

	PF1	*PF2*	*PF3*	*PF4*	*PF5*	*PF6*	*PF7*
Normal bearing	0.128	0.382	0.212	0.125	0.107	0.234	0.129
Inner ring pitting	0.292	0.300	0.114	0.047	0.013	0.016	0.029
Rolling element wear	0.070	0.045	0.052	0.058	0.017	0.006	0.005
Outer ring indentation	0.083	0.126	0.086	0.033	0.007	0.056	0.065

As shown in the table above, the energy characters of PF components derived from different faults by using LMD method are different, therefore, LMD is an effective way to decompose signal, and the energy characters of PF components can be input into the neural network.

3. The Fault Diagnosis Model based on optimized Neural Network

3.1 The model structure of Neural Network (NN)

BP network has the merits such as parallel processing and distributed storage, and it is one of the most widely used Neural Network (NN) [8, 9] in practical application. The common structure of BP network is constituted of input layer, hidden layer and output layer, with each layers connected by a weighting value.

(1) The number of input-layer nodes

The selection of input layers node has a direct bearing on the whole structure and the output of the network. The number of nodes should not too many and not too few. The whole network structure will be no viable in the area of fault recognition with fewer nodes. On the other hand, more nodes will increase the complex of network which will cause in network running slowly. In this paper, through LMD decomposing, the four types of fault signals including normal bearing, inner ring pitting, rolling element wear and outer ring indentation will respectively choose the energy of each layer as the characteristic vectors from the earliest seven layers PF1, PF2, PF3, PF4,

PF5, PF6 and PF7. Therefore, the number of output layer nodes is m =7, namely (x1, x2…… x7,).

(2) The number of hidden-layer nodes
Select three layers as the BP network model. There is not a unified formula in choosing the number of hidden-layer nodes at present, so it can choose an empirical formula to decide the number of hidden layer nodes by the experiences of forefathers. The formula as follow:

$$T = 2m + 1 \quad (2)$$

T is the number of hidden layer nodes; m is the number of input layer nodes. As a result, T =15.

(3) The number of output-layer nodes
Identify the four kinds of fault signals, "The inner ring pitting", "The outer indentation", "The rolling element wear" and "Normal bearing". Since the ideal output result could be identified directly according to the fault signals, the output-layer nodes are four types of fault signals (y1, y2, y3, y4). This paper uses binary encoding format as fault outputs. The type of output sample can be judged by the corresponding category which has a maximum node within the real network output. The desired outputs are shown in table 2 below.

In conclusion, the BP neural network structural is (7, 15, 4), Wij is the weight and bi is the threshold with input layer and hidden layer respectively, Wjk is the weight and bk is the threshold with hidden layer and output layer respectively, i =1,2……7, j =1,2……15, k =1,2,3,4, Fig. 3 is the three-layer model.

Table 2: Desired output of Bearing

Bearing Type	***Desired output vector***
Normal bearing	(1 0 0 0)
Inner ring pitting	(0 1 0 0)
Rolling element wear	(0 0 1 0)
Outer indentation	(0 0 0 1)

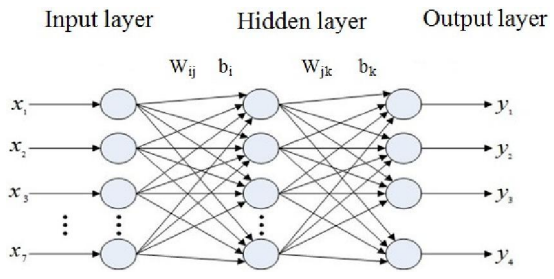


Fig. 3 The three-layer model of the BP neural network structural

3.2 Genetic algorithm optimizing BP neural network (GA-BP)

BP network Genetic algorithm is a kind of global search method algorithm based on the reference the natural selection biology evolution process and the mechanism of nature genetics [10, 11]. By simulating the process of natural evolution to search for the optimum solution, regard the argument of solving problems as gene, transform the problem parameters which need to be optimized into the coded string, and make a suitable selection with coded string by fitness function and a series of genetic manipulation, to retain the individual which has the highest fitness. Apply GA to the optimization of BP neural network structural parameters. Specific steps are as follows:

According to BP, construct the initial population by setting each parameter of GA and encoding, sequencing and building chromosomes with network Wij and bi;
Apply the inverse of E which is the BP error sum of squares as the individual fitness function, evaluate the quality of link weight and threshold, abandon the lower adaptive value of weight and threshold and retain the higher weight and threshold, select the individual which has a higher adaptive value pass on to the next generation. The function of network error sum of squares is

$$E = \sum_{k}^{p} (T_k - Z_k) \quad (3)$$

The fitness function is as follow,

$$F(k) = 1/E \quad (4)$$

T_k is the desired output, and Z_k is the actual output of network, k =1,2,3,4;

Judge E, the biggest of fitness of individual in the populations, whether satisfied with the accuracy requirements. If the accuracy requirements are met, continue the evolutionary process. Otherwise, implement the fourth step, until the condition is met;

The genetic operators include choice, crossover and mutation can be utilized to optimize the current population, produce the next generation and then turn to the step 2;

Output the initial weight and threshold of BP;

Calculate the error of output and estimate whether satisfied with the accuracy requirements. If the condition is met, then end the practice. Otherwise, continue to fix the BP weight and threshold until the accuracy requirements are met.

4. PF Component and Neural Network Optimization in the Application of Rolling Bearings

In order to verify the effectiveness of the fault diagnosis method based on PF energy characteristics and neural network optimization that proposed in this paper, we conducted the experiments at integrated fault simulation experiment platform of Spectra Quest's company, use the SCX-1000 data acquisition system to collect data, adopted QTH8-YD65 piezoelectric acceleration sensor which is mounted on the horizontal and vertical directions of the rolling bearing pedestal's both ends to measure the vibration signals. We simulated the four fault states of normal bearing, inner ring pitting, outer ring indentation, rolling element wear. For the acquired signal, first using LMD to decompose the original signal, then take the energy of each layer after decomposing as the input of the neural network optimization, last output four diagnostic results of normal bearing, inner race corrosive pitting, outer race indentation, rolling element wear. The fault diagnosis model based on the PF energy characteristics and neural network optimization is shown in Figure 4.

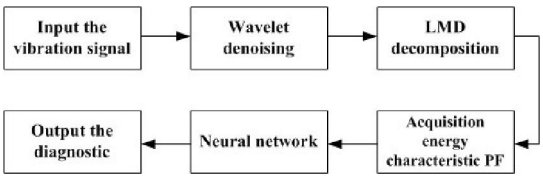


Fig. 4 The fault diagnosis model based on the PF energy characteristics and neural network optimization

Because the useful information in the layer would be less along with the LMD decomposition, to reduce the complexity of computing, we took the decomposing PF of previous 7 layers, extracted 50 samples in each fault, composed the 200*7 sample set. The characteristic parameters we obtained are presented in Table 3:

Table 3 Part of the characteristic parameter's values

SAMPLE	*PF1*	*PF2*	*PF3*	*PF4*	*PF5*	*PF6*	*PF7*
x1	*0.12849*	*0.38254*	*0.21228*	*0.12538*	*0.10799*	*0.23454*	*0.12971*
x2	*0.1422*	*0.2847*	*0.32626*	*0.38693*	*0.41058*	*0.021541*	*0.041392*
x3	*0.29246*	*0.30039*	*0.04763*	*0.013058*	*0.016375*	*0.016375*	*0.029089*
x4	*0.41113*	*0.17747*	*0.12231*	*0.064193*	*0.020069*	*0.021797*	*0.0014103*
x5	*0.070903*	*0.044931*	*0.052882*	*0.058284*	*0.017178*	*0.0065585*	*0.005993*
x6	*0.070992*	*0.094987*	*0.13547*	*0.055548*	*0.0156*	*0.0020125*	*0.00022*
x7	*0.08323*	*0.1262*	*0.086161*	*0.033388*	*0.0076132*	*0.056062*	*0.065736*
x8	*0.14744*	*0.14172*	*0.12501*	*0.05144*	*0.026*	*0.0046917*	*0.0019385*

BP neural network training is built on the basis of large quantity of fault signal samples, this article set 40 samples of each fault as the training sample, the other 10 samples as the test sample. For the same fault sample, we use BP neural network after optimization and BP neural network to conduct the fault diagnostic identification. The results are shown in Table 4:

Table 4 Part of antifriction bearing identification results of BP and GA-BP algorithm

Signal type	*algorithm*	*Output of the network*				*Target output*
Normal	*GA-BP*	1.0024	-0.00067	-0.00048	-0.0021	1 0 0 0
	BP	1.0000	-0.000132	0.013277	-0.0132	
Inner ring pitting	*GA-BP*	0.01317	0.9807	0.02232	-0.0213	0 1 0 0
	BP	0.11752	0.97026	-0.0196	-0.06822	
Rolling element wear	*GA-BP*	0.005102	-0.00089	1.0042	-0.00916	0 0 1 0
	BP	-0.000133	-0.000378	0.85117	0.14879	

Outer indentation	*GA-BP*	-0.000435	0.009101	-0.00345	0.99361	0 0 0 1
	BP	0.00338	-0.00489	-0.09911	1.1006	

The accuracy rate of GA-BP and BP algorithm is shown in Table 5:

Table 5 Bearing fault diagnosis testing result

Bearing type	***algorithm***	***Test sample size***	***Diagnosis result***		***The accuracy rate of diagnosis***
			correct	***error***	
Normal	*GA-BP*	10	10	0	100%
	BP		9	1	90%
Inner ring pitting	*GA-BP*	10	9	1	90%
	BP		9	1	90%
Rolling element wear	*GA-BP*	10	7	3	70%
	BP		5	5	50%
Outer indentation	*GA-BP*	10	8	2	80%
	BP		7	3	70%
Total	*GA-BP*	40	34	6	85%
	BP		30	10	75%

It can be seen from the table, the precision and accuracy rate of BP neural network are improved after optimization. It has been proved that BP neural network after optimization can be used in the application of antifriction bearing identification.

As can be seen from the above comprehensive comparison, through optimizing BP neural network's structural parameters, we not only improved the efficiency of network training significantly, but also greatly improved the accuracy rate of BP neural network used in antifriction bearing fault identification, provided a practical and feasible method for antifriction bearing fault identification.

5. Conclusion

LMD is an adaptive signal decomposition method, it can decompose the complicated multi-component as the aggregation of the finite instantaneous frequency PF component which has the physical meaning. We separately extracted the PF component energy characteristics of normal bearing, inner ring pitting, rolling element wear, outer ring indentation, the result indicate that characteristic value extracted by the LMD method has a significant difference.

Experiment simulation proved that the bearing fault diagnosis method of genetic algorithm optimized BP neural network improved the accuracy and failure rate compared with the fault diagnosis method of BP neural network.

The proposed fault diagnosis method based on PF energy characteristic and optimized neural network, provide a new method to achieve the high efficiency and high precision in antifriction bearing fault diagnosis

Acknowledgments

This work was supported by General Assembly Department about Electromagnetic compound stageless transmission (62201020204).

References

[1] Classen T, MecklenbraukerW. The aliasing problem in discrete-time Wigner distribution [J]. IEEE Transactions on A-coustics, Speech, and Signal Processing, 1983, 31 (5):1067-1072.
[2] Lee JH, Kim J, KimH J. Development of enhanced Wigner-Ville distribution function [J]. Mechanical Systems and Signal Processing, 2001, 13(2): 367-398.
[3] Mallat S. A theory for multi-resolution decomposition, the wavelet representation [J]. IEEE Trans. P. A. M. I. 1989, 11(7): 674-689.
[4] Cheng Junsheng, Yu Dejie, Yang Yu. Energy operator demodulating approach based on EMD andits application inme-chanical fault diagnosis [J]. Chinese Journal of Mechanical Engineering,2004,40(8): 115-118.

[5] Smith J S. The localmean decomposition and its applicationto EEG perception data [J]. Journal of the Royal Society In-terface, 2005, 2(5): 443-454.

Baoshan Huang, associate professor of Beijing Institute of Technology, Zhuhai. He obtained his MS, PhD in Mechanical Design and Theory at South China University of Technology (SCUT) in 2006. He worked on a number of projects. Such as, State 863 projects and Natural Science Foundation of China (NSFC).

Wei Xu is a PhD candidate, has been on the University of Macau since 2010 and specializes in information systems, his research focus on RFID & IOT innovative technology applications in manufacturing processes, business process management and enterprise modeling and simulation. He received MS, Control Theory Engineering and BS, Electronic Information Engineering from SHU and HAUST.
*corresponding author Jerryxw@live.com

Xinfeng Zou is a master student in the Electromechanical Engineering of the Faculty of Science and Technology at University of Macau. His current research interests are in the areas of intelligent management, the matching relationship of work-in-progress (WIP) and worker.

Research on the Classification of Reviewers in Online Auction

Li-cheng Ren [1], Ming Wu [2] , Jin-tao Lu [3]

[1] School of Economic and Management, Tai Yuan University of Science and Technology, P.R.China,

[2] School of Economic and Management, Tai Yuan University of Science and Technology, P.R.China,

[3] School of Management, Northwestern Polytechnical University, P.R.China

Abstract

Online reviews are more influential than expert reviews towards marketing communication. Starting from the network reputation, this article classified online reviewers through the study of consumer reviews motivation and dominant willingness. The research deemed that the online-reviewers should be classified into Dominant egoist, Dominant altruist, Robust egoist, Robust altruist, Accommodated egoist, Accommodated altruist, Avoidant egoist and Avoidant altruist, and these four types enrich online users' behavior characteristics theory. Different marketing strategies for various online users' motivation were also proposed, which may help the website operators to understand their users, to choose the communication channels of EWOM and design internet marketing strategy.

Keywords: *Online reviews; customer word-of-mouth; Word-of-mouth motivation; Dominant willingness.*

1. Introduction

HITWISE released that mobile phone buyers over the age of 18 in the United States were affected by the online media significantly. Based on it, IResearch found that nearly 61% of them were influenced by other users' evaluation and introduction; while about 30% consumers' purchase decisions were affected by the blogs. With the development of online review systems, word-of-mouth sites and virtual communities, online reviews are more influential than expert reviews and become an important foundation for consumers' purchase decisions. Jianyuan Yan researched the relationship between review content and its usefulness, which revealed that the deeper and more objective of the content review was, the higher of the usefulness would be [1].

In many document researches, most scholars had committed to the IWOM motivation research [2], review comment [1], and there were also a small amount of scholars researching online consumers' classification criteria [3], while few researches on the classification of online reviewers. With the basis and proceeding from the perspective of the online reviews, this article analyzes online reviewers' behavior motivation and discussed their motivation and behavior rules to understand the characteristics of user behavior, and provided reference for website operators to market strategically.

2. Comments research

2.1. Word of mouth and online reviews

The traditional word-of-mouth is mainly among non-commercial individuals who communicate the products or the companies. It is of great influence and often occurs in strong ties such as the crowd of relatives and friends. The traditional word-of-mouth has been playing an important role in consumption, and has influenced the consumers' purchase decisions [4], as well as views of the post-purchase product [5]. In the United States, more than 90% of the respondents claimed that they based on recommendations of friends before deciding to purchase a product or service [6].

Along with the development of communication technology and modern network, the concept of electronic word-of-mouth, which assaulted and improved the traditional word-of-mouth concept, has been proposed. Consumers' communication relating to information on the Internet is called "Internet consumers Exchange" [7]. Through a variety of network channels (such as forums, blogs, etc.) consumers and other Internet users share information about the product or service. This form of information is known as "electronic reputation", which is EWOM (Electronic Word-Of-Mouth) [8]. EWOM can overcome the limitations of the traditional reputation [9]. And there is an abundance of online consumers' reviews than traditional reviews in the offline world [10]. Whether "Internet consumers" communication or electronic word-of-mouth, both of which can be seen as different periods of network development form.

The paper argues that the electronic word-of-mouth is positive or negative assessments which are published on the web and disseminated to others about products, services, or the company. It is online reviews essentially. Similar to the traditional WOM, online consumer reviews serve as a recommendation of the relevant products or sellers [11].

2.2. Source of online reviews

With the proliferation of e-commerce and the increasing numbers of product reviews, consumers are more and more relying on them for target search. Being compared with traditional WOM, online reviews are presented in the form of text, so information and expressions are becoming richer and more diverse. In the process of online reviews, positive WOM will deliver the pleasure and satisfaction of consumers experience and negative WOM can allow consumers to vent chagrin, resentment and anger. Comments propagation of results and potential decision-making will be affected by emotional attitude and way of expression.

Online consumers are the important information source of online reviews as well as important assessors [12]. The online-reviewers are disseminators of reputation, which can be roughly divided into three categories, which are marketers, individuals with direct experience and individuals with indirect experience [13]. WOM becomes another fast and convenient channel for marketers to promote their products.

2.3. Opinion Leaders

Opinion leaders are one of the important factors of online reviews WOM. "Opinion leaders" can largely promote the speed and range of information dissemination. What's more, the information is more likely to trust and have greater influence [14].

The research of opinion leaders was originated from a decision-making mechanism of the referendum made by Lazarsfeld in 1948. The study proposed the "two-level mobile communication theory", which thought the mass media did not "flow" to the general audience directly, but was through the middle part of the opinion leaders, the mode is mass communication----opinion Leaders----general audience [15].Although scholars have different definitions over opinion leaders, they basically thought that opinion leaders are the "activists" who often provide information, opinions, and comments for others and influence others.

Many papers pointed out that the opinion leaders have fields [16] and have some common characteristics, such as a wide range of social and information channels, compassion and responsibility [17]. Yung-Ming Li established a review mining model to evaluate the influence of online-reviewers, and used this model to find the opinion leaders for carrying out the network marketing [18]. Feng Li etc. retrieved to the author and the reader, then analyzed the relationship between them through the blogs content, and identified opinion leaders in order to implement an effective marketing strategy [19].Obviously, opinion leaders have a large influence in the promotion of information [14], and play important roles in WOM and marketing [19].

3. Classification of online-reviewers

3.1. Classification of online-reviewers based on the motivations

In the process of information communication, the disseminators share valuable information and establish a mutually beneficial relationship with the recipient, prompting the receiver to feedback out of responsibility, which leads to consumer information transmission behaviors[20].Most scholars have researched traditional WOM in term of motivation in previous researches. Thurau etc. (2004) proposed that the classification is also applicable to electronic WOM motivation behavior [21].This paper discusses and summarize the online reviews motivation, shown in Table 1.

Table 1.Motivation research on word-of mouth behavior

Scholar	***WOM motivation***
Dichter E. (1966)	Product/ Self / Other / Massage Involvement
Sundaram, Mitra, Webster (1998)	Altruism , Product involvement , Self-enhancement, Helping the company,
Gianfranco (2004)	Happy, Help, Responsibility
Hennig-Thurau (2004)	Caring for others, Social interaction needs, Economic incentives, Expression of positive emotions
Peddibhotla , Subramani (2007)	Self-directed, Other-directed
Y. Tong, X. Wang, H.H. Teo (2007)	Information feedback, Help others
Jason Y. C. Ho,	Express individuality, Altruism

Melanie Dempsey (2010)	
Jun Yan, Yinbo Jiang, Yaping Jiang (2011)	Emotional share, Support / punish businesses, Community flourished, Improve service, Rewarded, Enhance image
Christy M. K. Cheung (2012)	Enhance reputation, Sense of belonging, Help others

According to the documents, most scholars studied in term of review motivation, but they did not classify them. In the process of behavior, motivation dominates the direction and intension of the behavior. Jun Yan, Yinbo Jiang and others divided motivation into 9 species, which could be classified as egoistic and altruistic categories [22]. Christy etc. extracted several main motivations from the view of social psychology, and proved that reputation, helping others and sense of belonging significantly related to EWOM motivation [2]. Improving the reputation, sharing information to help others and enjoying the process is a commitment, responsibility or a sense of belonging to them, and these all affect consumers EWOM willingness. Based on aforementioned discussion, this paper divides online reviews motivation into two types of egoism and altruism.

Table2. Reputation of behavior motive

Scholar	***egoistic type***	***Altruistic type***
Engel, Blackweli, Miniard (1966)	Self Involvement	Product-Involvement Other-Involvement Massage-Involvement
Sundaram Mitra, Webster (1998)	Self-enhancement, Vengeance	Altruism, Product involvement, Helping the company
Gianfranco (2004)	Happy	Help, Responsibility
Hennig-Thurau (2004)	Social interaction needs, Economic incentives, Expression of positive emotions	Caring for others
Peddibhot, Subramani (2007)	Self-directed	Other-directed
Y.Tong, X. Wang, H.H. Teo (2007)	-----	Information feedback, Help others
Jason Y. C. Ho,Melanie Dempsey (2010)	Express individuality	Altruism
Jun Yan, Yinbo Jiang, Yaping Jiang (2011)	Emotional share, Rewarded, Enhance image	Support/punish businesses, Community flourished, Improve service
Christy M.K.Cheung (2012)	Enhance reputation, Sense of belonging	Help others

3.1.1. Egoistic type

A motive is considered egoistic if the ultimate goal is to increase the actor's own welfare [23]. Individuals are deemed as egoistic when they aim at tangible or intangible returns after sharing information with others. Being rational, people try to look for returns (e.g. pay, prizes, reputation, and recognition) by maximizing their benefits and minimizing their cost during information exchange process with others [24]. This perspective has been widely adopted in many EWOM communication publications [21, 25]. People share and contribute their knowledge because they want to gain an informal recognition and establish themselves as experts [26]. The typical one is opinion leaders.

For U.S. consumers, encouraging usage of online reviews and providing an Internet shopping site with an online review component may be an effective way to maximize the effect of online reviews [27]. Online reviews can be used as an indirect network marketing communication tool that plays a recommended role in the relevant product or seller[11], so businesses will take measures to incentive online shoppers to add some positive WOM comments, from which online shoppers could get benefits, such as discounts, reward, and better service [21].Consumers will cover up the product defect, and give a positive evaluation for immediate benefits, even if the product does not reach a satisfactory level. This type of online reviewers is egoistic. In addition, as it is mentioned above, enhancing self-image [22], improving reputation, expression of personality [28], the demand for social interaction [21], or willingness to express their pleasant consumer experience [19], all of which are egoistic factors for consumers to make online reviews , no matter the factor is active or passive.

3.1.2. Altruistic type

Altruism is motivation with the ultimate goal of increasing the welfare of one or more individuals other than oneself [23]. Individuals acting on altruistic goals are willing to volunteer themselves to contribute their knowledge to online consumer reviews without expecting direct rewards in return. For example, consumers may share purchasing experience just because others have a need for it [29]. The online reviewers committing evaluation is considered to be an online knowledge sharing behavior in the reputation system [30].Reviewers provide advice to others in online communities, blogs or forums[31], which aim at helping others make the right purchase decision or preventing them encountering a similar situation from their experience of failure [32].This type of online-reviewers' motivation is out of concern for others, or supporting a positive or negative opposition which is to vent their dissatisfaction with the product or service and publicizing negative WOM [31].

3.2. Classification of online -reviewers based on the dominant intention

Domination refers to the process and behavior which disposable subjects influenced on objectives in accordance with the given conditions and objectives. Domination wishes mean that direct others to commit adhering to their own thing. Different personalities have different domination willingness. Therefore, domination willingness is divided into 4 levels: the dominant type is someone who has a strong desire; robust type is someone who has the domination wishes, but lacks mobility; accommodated type is someone who has the lower domination willingness; and the avoidant type is someone who has little control willingness.

Table3. Domination willingness and behavioural characteristics

Domination willingness	***Behavior characteristics***
Dominant Type	Unique; Interested in the success and achievement; Desire for influence and gain respect;
Robust Type	Calm ; good at thinking and analysis; judgment; possessed of definite views
Accommo-dated Type	Friendly; focus so much on the ideas of others and enjoy being with others
Avoidant Type	Do not take the initiative to share information; poor expression; evasive when cope with stress

4. Classification of online reviews based on consumer motivation and domination willingness

The same behavior may have different motivations, in other words, a variety of motives can be manifested by the same kind of behavior; different behaviors can also be motivated by the same or similar motives. The motivations varied in the same individual, some motives were in dominating position, and some were in subordinate position. As the discussion shown above, online-reviewers have two dominant motivations---egoistic and altruistic motivations. The article discusses the cross-effect between consumer reviews motivation and domination willingness, and classified online -reviewers into dominant egoist, dominant altruist, robust egoist, robust altruist, accommodated egoist, accommodated altruist, avoidant egoist and avoidant altruist. As shown in Figure1.

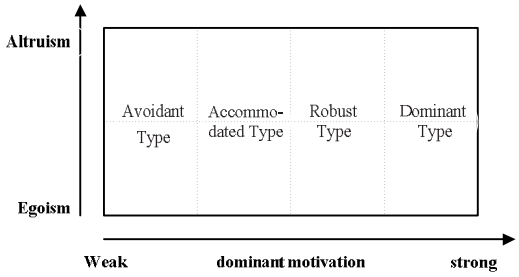


Fig1. Motivation behaviour pattern

4.1. Dominant egoist and dominant altruist

Dominant individuals do not want to be ordered or restricted. They will use all appropriate means to be in dominant position, such as argument, rank which is a location of power. The dominant egoists are interested in success and achievement; they hope to get others' recognition and respect, and take the initiative to fight to earn a reputation, or enhance the image. The dominant altruist will criticize discontent, when they will persuade and dominate others to comply with their comments and suggestions. Opinion leaders are the typical representatives, who always get more mass media information than other groups, have thorough understanding and comprehensive evaluation, and can make a secondary diffusion of information.

4.2. Robust egoist and robust altruist

Robust individuals are calm, like thinking, and have the judgment and strong-minded. They often observe first when they face things around, and would like to find out the context of the matter, then involve in. They would like to use their own judgment to distinguish right from wrong, and comment with concern over justice and fairness.

When they encounter negative information, they will be opposed resolutely, and will not to take the initiative to change other people's views and opinions. Egoistic and altruistic motivation of robust individual is not obvious enough. A robust individual's behavior will turn into dominant type or the other, showing the egoistic or altruistic motives under certain conditions.

4.3. Accommodated egoist and accommodated altruist

Accommodated individuals are friendly, paying much more attention on others, and enjoy being with others, and they are generous or have benevolent gesture. They will ignore their own interests to meet the concern of others; they still can obey the commands of others, or succumb to the other people's point of view in order to find a part to meet both favorable and acceptable ways even if they are not reluctant.

For instance, in the evaluation of the online shopping, the sellers often kept contacting the buyers who did not give decent comments, the buyers felt so boring that they modify the bad into praise comment [33]. Reviewers of the characteristics belong to the accommodated egoistic type; they are more cautious and will conform to the interests of others in order to avoid trouble. In addition, the main content of Chinese traditional culture includes the fate, moderation and humility. In such a cultural context, consumers believe that fate and tend to be outer control, which makes accommodated altruist attribute the failure to uncontrollable factors such as environment, luck. If accommodated altruists get a service failure or defects of the commodity, they adhere to the idea of a forward-looking, and give their buying advice to support corporate businesses, expecting products and services would be improved [21], or through their own positive publicity to help companies increase sales [32] expecting to do a little to help correct it and are willing to maintain long-term contact [34].

4.4. Avoidant egoist and avoidant altruist

Avoidant individuals are silent, not good at expression and have a strong self-protection consciousness. When facing the pressure, they will avoid argument and be evasive. Reviewers of this type only browse the information, and less share information and with more features of the egoistic in order to protect their own interests. When not infringed, they remain silent. If the messages are negative, avoidant individuals do not do negative WOM publicity, which is negative altruism.

4.5. Marketing Implications

Online reviews can help consumers to make the purchase decision [35] and represent the market performance of the product, and are often used as indicators of product awareness [36] ， which are playing an increasingly powerful influence on consumer decision-making [37]. Because of this, the research of psychological behavior is particularly important in the comment process. Different types of online shoppers have different dominant wishes and comments motivations, website operators only according to the different types of people to carry out network marketing to make the marketing effectiveness maximization.

Because the WOM behavior affected by emotion sharing and self-improvement, website operators should provide some special services to meet the needs of reviewers who are egoists with dominant behavior (such as opinion leaders) or altruistic person. Specific measures may include: open related product forum and discussion group on the platform，which not only meet consumers' desire to express their views and share information but also make the enterprise understand consumer issues and give timely resolution.

Individuals of robust behavioral characteristics possess definite views but lack of mobility. In this regard, the network operators should improve information sharing mode at community platform, creating a good atmosphere for the community and establishing an incentive mechanism so that individuals can participate in.

In order to help others, reward / punish businesses, accommodated individuals will tolerate others sometimes expecting the operators improvement, but it will easily lead to the occurrence of negative WOM and the actual loss of consumers in the long run. Therefore, website operators should remedy the failed transaction actively, reshape consumer satisfaction, and increase customer loyalty.

For avoiding troubles and pressure, avoidant individuals with characteristics of motivation less share information or experience who belong to the potential of online-reviewers. This type of online consumers is a waste of resources. In this regard, website operators should create a community platform of communication atmosphere, such as the establishment of independent club on the Internet and channels of communication. In the club, players can design their own interesting topics that can attract and encourage such consumer reviewers.

5. Conclusion

How to manage consumers' WOM and carry out WOM marketing has become a problem that can not be ignored. Online reviews provide a powerful inexpensive and influential channel and publicity tool for marketing communications. Marketers will take the opportunity to use online reviews for the publicity and promotion of product or service. As a source of information and WOM publicity for online reviews, consumers play a crucial role in the decision-making of the potential shoppers.

Starting from the network reputation and discussing consumer reviews motivation with domination willingness, the article deems that the online-reviewers should be classified into Dominant egoist, Dominant altruist, Robust egoist, Robust altruist, Accommodated egoist, Accommodated altruist, Avoidant egoist and Avoidant altruist. And the article puts forward different marketing strategies for different motivated behaviors of online-reviewers, which may be helpful for web site operators to understand WOM motivation that consumer participated in and know how to choose WOM communication channels and design marketing strategies.

References

[1]Yan Jianyuan, Zhang Li, Zhang Lei, An Empirical Study of the Impact of Review Content on Online Reviews Helpfulness in E-commerce, Information Science, vol. 30, no. 5, pp. 713-716, 2012.

[2] Christy M. K. Cheung, Matthew K. O. Lee. What Drives Consumers to Spread Electronic Word of Mouth in Online Consumer-Opinion Platforms, Decision Support Systems, vol. 53, no.1, pp. 218-225, 2012.

[3] Wang Haiping, Research Review of Online Consumers Typology, East China Economic Management, vol. 25, no. 3, pp. 147-150, 2011.

[4]Engel, J. F., Blackwell, R. D. and Kegerreis,R. J. How information is used to adopt an innovation, Journal of Advertising Research, vol. 9, no. 4, pp. 3-8, 1969.

[5] Bone, P. F. Word-of-mouth effects on short-term and long-term product judgments, Journal of Business Research, vol. 32, no.3, pp. 213-223, 1995.

[6] AC Nielson. Trust in Advertising: A Global Nielsen Consumer Report, Nielsen Media Research, New York 2007.

[7] Stauss ,B. Using new media for customer interaction: A challenge for relations hip marketing [M].In T.Henning - Thurau& U.Hansen (Eds.)，Relationship Marketing，Berlin: Springer, pp.233- 253,2000.

[8]Hanson，W.A. Principles of Internet Marketing. South-western College Publishing，Cincinnati, pp.211-215, 2000.

[9]D. Godes, and D. Mayzlin, Using Online Conversations to Study Word of Mouth Communication, Marketing Science, vol.23, no. 4, pp. 545–560, 2004.

[10]Jumin Lee , Do-Hyung Park , Ingoo Han ,The effect of negative online consumer reviews on product attitude: An information processing view, Electronic Commerce Research and Applications , vol.7, pp. 341–352, 2008.

[11]Y. Yubo Chen ， Jinhong Xie , Online consumer review: a new element of marketing communications mix, Working Paper Management Science, vol. 54, no. 3, pp. 477-491, 2008.

[12]Jinhyung Cho ,Kwiseok Kwon ,Yongtae Park.Q-rater: A collaborative reputation system based on source credibility theory. Expert Systems with Applications, vol.36, pp.3751–3760, 2009.

[13]Wang Rong-lin, Wang Yu-qi, Study on the dissemination mechanism of internet consumer's mouth word, Science-technology and management, vol. 12, no. 6, pp. 108-111, 2012.

[14]Luo Xiaoguang, Xi Lulu, On the Opinion Leader of Customer Word-of-Mouth Communication Network Based on the Social Network Analysis Approach, Management Review, vol. 24, no. 1, pp. 75-81, 2012.

[15]Lazarsfield P., Berelson B., Gaudet, H. The People's Choice, Columbia University Press, New York, 1948.

[16]Hellevik O., Bjorklund T. Opinion Leadership and Political Extremism, International Journal of Public Opinion Research, vol. 3, no. 2, pp. 157-181, 1991.

[17]Beth Harben, Soyoung Kim. Attitude Towards Fashion Advertisements with Political Content: Impacts of Opinion Leadership and Perception of Advertisement Message, International Journal of Consumer Studies, vol. 32, no. 1, pp. 88-98, 2008.

[18]Yung-Ming Li，Chia-Hao Lin, Cheng-Yang Lai. Identifying Influential Reviewers for Word-of-Mouth Marketing, Electronic Commerce Research and Applications,vol.9, no.4, pp. 294-304, 2010.

[19]Feng Li,Timon C. Du. Who is Talking? An Ontology-Based Opinion Leader Identification Framework for Word-of-Mouth Marketing in Online Social Blogs, Decision Support Systems, vol. 51, no. 1, pp. 190-197, 2011.

[20]H. Elizabeth, Melanie Wall endorf. Motives Underlying Marketing Information Acquisition and Knowledge Transfer, Advertising, vol. 11, no. 3, pp. 25-31, 1982.

[21]Hennig-Thurau T, Gwinner K.P., Walsh G., Gremler D. Electronic Word-of-Mouth via Consumer-Opinion Platforms: What Motivates Consumers to Articulate Themselves on the Internet? Interactive Marketing, vol. 18, no. 1, pp. 38-52, 2004.

[22]Yan Jun, Jiang Yinbo, Chang Yaping, Relationship between Motives and Behavior of eWord-of-Mouth, Management Review,vol.23, no.12, pp.84-89, 2011.

[23] C.D. Batson, Why act for the public goods? Four answers, Personality and Social Psychology .Society and Natural Resources. vol.20, no.5, pp. 603–6101, 1994.

[24]K.R. Lakhani, E. Von Hipper, How open source software works:'free'user-to-user assistance. Research Policy, vol. 32, no. 6, pp.923-942, 2003.

[25]Y. Tong, X. Wang, H.H. Teo. Understanding the intention of information contribution to online feedback systems from social exchange and motivation crowding perspectives , Proceedings of Hawaii International Conference on System Sciences, Hilton Waikoloa Village, Big Island, 2007.

[26]M.M. Wasko, S. Faraj, Why should I share? Examining social capital and knowledge contribution in electronic networks of practice? MIS Quarterly, vol. 29, no. 1, pp.35–57, 2005.

[27]Cheol Park ,Thae Min Lee .Antecedents of Online Reviews ' Usage and Purchase Influence: An Empirical Comparison of US .Journal of Interactive Marketing, vol. 23, no. 4, pp. 332–340, 2009.

[28]Jason Y. C. Ho,Melanie Dempsey. Viral Marketing: Motivations to Forward Online Content. Journal of Business Research，vol.63, no.9, pp.1000-1006, 2010.

[29]Peter Kollock, Marc Smith. The economies of online cooperation: gifts, and public goods in cyberspace, Communities in Cyberspace. Routledge, New York, pp. 220–239,1999.

[30] Dellarocas，C. The digitization of word of mouth：Promise and challenges of online feedback mechanisms, Management Science, vol. 49, no. 30, pp. 1407-1424, 2003.

[31]Inge M. etc. Never Eat in that Restaurant, I did! : Exploring Why People Engage in Negative Word-of-Mouth Communication. Psychology and Marketing, vol. 24, no. 8, pp. 661-680, 2007.

[32]Sundaram，DS. ，Mitra,K. ，&Webster,C. "Word-of-mouth communications: A motivational analysis" ，advances in Consumer Research, vol.25, pp. 527-53, 1998.

[33]Hu Zunrang, Yuan Miao, Wang Zhiyi, The unreasonable Talking of C2C e-commerce credit evaluation system, e-business journal, vol. 7 , pp. 29-33, 2010.
[34]Matos，C．A．d. , Rossi，C．A．v．”Word-of-mouth communications in marketing：a meta-analytic review of the antecedents and moderators”, Academy of Marketing Science,vol.36, pp. 578-596, 2008.
[35] Sajjad Nazir, Arsalan Tayyab, Aziz Sajid, Haroon ur Rashid , Irum Javed, “How Online Shopping Is Affecting Consumers Buying Behavior in Pakistan?” International Journal of Computer Science Issues, Vol. 9, No. 1, pp: 1694-0814, 2012
[36] Chevalier, J. A., & Mayzlin, D., The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research, vol. 43, no. 3, pp. 345-354, 2006.
[37] Gerzema, J., & D’ Antonio, M. Spend shift: How the postcrisis values revolution is changing the way we buy, sell and live, San Francisco: Jossey-Bass, pp. 126-128, 2011.

Li-cheng Ren is now a professor and the Dean of the School of Economics and Management in Taiyuan University of Science and Technology. He is also a member of Shanxi Soft Science, whose research interests cover E-commerce and Information Management.

Ming Wu received the BBM degree from Inner Mongolia Agricultural University in 2006 and is a postgraduate student in Taiyuan University of Science and Technology, who’s the corresponding author and major in Management Science and Engineering. Her current research interests include E-commerce.

Jin-tao Lu received the B.E. degree from Lanzhou University of Technology in 2006 and an M.E. degree from Taiyuan University of Science and Technology in 2012. He is now a Ph.D candidate in Northwestern Polytechnical University, whose major is Management Science and Engineering. His current research interests include System Engineering.

The Study on the Application of Business Intelligence in Manufacturing: A Review

Ernie Mazuin Mohd Yusof[1], Mohd Shahizan Othman[2], Yuhanis Omar[3] and Ahmad Rizal Mohd Yusof[4]

[1,2] Faculty of Computer Science and Information System, University Technology Malaysia
Johor Bahru, 81310, Malaysia

[3] Faculty of Information System, University Kuala Lumpur
Kuala Lumpur, 50250, Malaysia

[4] Institute of Occidental Studies (IKON), University Kebangsaan Malaysia
Bangi, 43600, Malaysia

Abstract

A manufacturing based organization operates in an environment where a fast and effective decision is needed. This is to ensure that the output is met with customer compliance. There exists manufacturing systems that collect the operational data and the data turns out to be in a high volume due to the state of the art of the abundant manufacturing operational data. Having a lot of data without the tool to analyze and extracting valuable information from it, increases the amount of time spent by employees focusing on the data itself. This eventually leads to a delay in a decision making process, resulting in a delay of products delivery to customer. To fill in this gap, a Business Intelligence (BI) implementation will be reviewed, with the aim to execute the right action at the right time or in other words, to improve the decision making process of an organization.

Keywords: *Business Intelligence, Manufacturing, Visual Representation.*

1. Introduction

The manufacturing industry may be the main resources for profit for a certain country. It is one of the major business activities. As the competition rises and customers become more demanding, the world has started to find a way to sustain and increase their profit. Because of the business states and environments which have now become globalized, there is a need to have a fast decision based on the updated information. The growth in the manufacturing sector has supported the world economy positively [25]. Growth in 2010 was revised from 4.3% to 4.5%, while in 2011 it was revised from 3.8% to 3.9%.

In Malaysia, sales in the manufacturing sector went up to 8.5% from the year 2009 to November 2010 [24]. The growth is seen rapidly high in the area of computer peripherals and electronics manufacturing industry. The computer peripherals and electronic product manufacturing company produces computers, computer peripherals, communications equipment and other electronic products. Examples of the products are printers, scanners, fax machines and so on. These products are used in homes and businesses, as well as in government and military sectors. The focus to synchronize business with the manufacturing unit of the manufacturing operations is needed as the segment has increased globally for more value-added chain [21]. Even though the computerized systems in the manufacturing companies for higher productivity, quality and lower production costs produce large volumes of data, the valuable knowledge might be hidden in it [14]. Having a lot of data does not guarantee that the most critical information is being attended. In a manufacturing based organization, a fast and quick decision is very much needed to ensure that the in house operation corresponds to the customer needs. The problem that arises in a shop floor control with this abundance of data is that, decisions are difficult to make in real-time by the status of the shop floor [16]. Two technologies are seen to improve the knowledge available to decision makers. They are the Business Intelligence (BI) and Knowledge Management [29]. The BI systems are chosen since they are becoming increasingly more critical to the daily operation of organizations [27].

2. Manufacturing Processes and Problems

A manufacturing organization consists of many processes initiating from customer orders until the delivery of products to customers [10]. The process flow in a manufacturing company is as shown in Figure 1. Being the general flow of the manufacturing organization, it might vary from one organization to another.

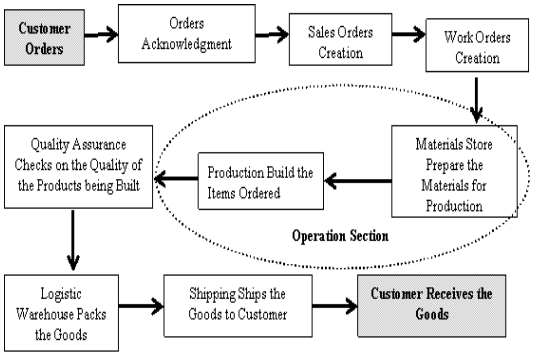


Fig. 1 Manufacturing Process Flow.

The problem seen in the operation section is that, whether the products have been completely built by the production folks or not, they are unknown [4]. The whole process stays invisible to others as there is no real-time information, unless we go down to the production floor itself and check the status ourselves. The problem seen in the Operation Section or also called the shop floor and production here is, the urgent customer orders are often overlooked. In other words, the priority of the orders in accordance to its delivery schedule is not being monitored and carried out. Employees tend to pick a simple order and item (that does not have so many materials to build for example) to fulfil. In addition, if ever exists an order which requires further attention, even though remarks are put in the list, this order is often neglected. Rarely will it be reviewed back by the production employees after the remarks have been updated. This results in the delay of delivery of that item, eventually affects the on time delivery performance of the organization. Moreover, in the program management side of the organization, a frequent follow up with the operational staffs has to be made to push them to fulfil the top priority orders.

In addition, many manufacturing organizations struggle with issues like the overall enterprise processes and information visualizations are limited, and also, manual forms and unstructured data not readily integrated or understood in relation to other data and systems [23]. Data are recorded from nearly all of the processes in the organization like the scheduling, assembly, material planning and control and many others. However, to make use of the collected data turns out to be an issue [12].

There exist several systems to serve the purpose of monitoring the shop floor activities like the Manufacturing Execution System (MES), Enterprise Resource Planning (ERP), Manufacturing Resource Planning (MRP) and Supply Chain Management (SCM) [15, 22]. However, those systems are lacking of analytical and historical data aggregation features that are needed for an organization to build up its value by executing intelligent business processes.

BI is said to overcome those problems as its implementations in the manufacturing industry, particularly the electronics and computer peripherals section will be reviewed here.

More and more manufacturing enterprises hope to take advantage of BI to transform the abundant data into information and knowledge to acquire competitive edge [7]. Without having to dig the valuable information from tedious reports and spreadsheets, BI application has the ability to foresee the future, like monthly delivery requirements, single and real-time operational data view and important information consolidation and presentation in high level [19]. A survey from Gartner and Forrester shows that majority of the firms are interested in investing the BI systems [5]. In the context of a widespread data analysis, BI is used to generate information that is decisive for appropriate actions to be taken [5].

3. Business Intelligence in Manufacturing

BI is defined as the method of converting data into information and subsequently to knowledge [18]. The types of knowledge obtained are about the customer requirements and decisions, organizational performance in the industry and the global trends. Another definition of BI, particularly the BI systems is, BI systems put together the gathering and storage of data and knowledge management with analytical tools to present a ready-for-action and complicated information to the planners and decision makers [28]. This is to assist them to obtain the right information at the right time, location and form.

Cindi Howson defines BI as a set of technologies and procedures that permit people at all levels of an organization to access and analyze data [6]. It permits people at all levels of an organization to access, interact with, and analyze data to manage the business, improve performance, discover opportunities and operate efficiently [6].

In this paper, Business Intelligence is defined as, information obtained to aid the decision making process of a business segment through the transformation of the existing data. The information is presented visually to give the intended users a clear guidance for a smooth decision making process and most importantly, an accurate and fairly fast decision.

The BI has been widely used nowadays in the manufacturing industry, to solve organizational issues from the business perspective, especially in decision making to maintain the company's competitiveness. As shown in Figure 2, a research by the Ventana Research on the BI applications has come out with the most of the respondents coming from the Services and Manufacturing industries.

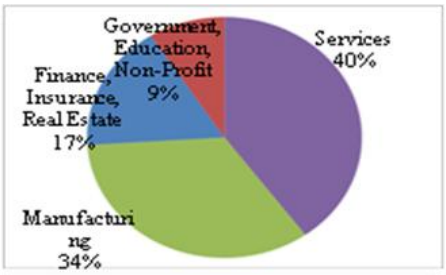


Fig. 2 Type of Industry with the most Participants of BI Demographic Survey. (Source: Ventana Research). 2006.

Since the production section of a manufacturing company plays a very important role where the operation runs, the BI is commonly applied in this area of an organization. Figure 3 shows that the Best-in-Class organization applies BI in the operation section.

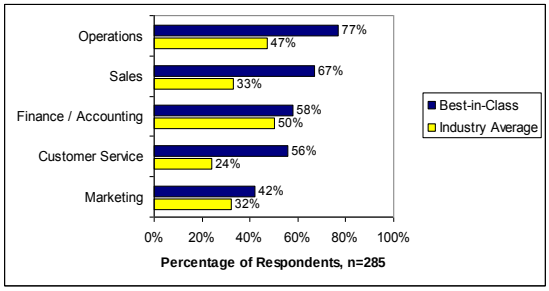


Fig 3. Statistics on the Usage of Business Intelligence Applications in Operations for Best-in-Class Organizations (Source: Abedeen Group). 2009.

Thus, this study will focus on the application of BI in the operation section of a manufacturing company. The next section will analyze the previous research pertaining to the application of BI in the operation or production department of different segments of manufacturing organizations.

4. Previous Studies

An elaboration of the previous studies related to the application of BI in manufacturing organizations in the operation or production site will be discussed in this sub topic as Table 1 shows.

From Table 1, nine paper works from different manufacturing sectors will be analyzed. The classification is done according to the manufacturing sector, problems, BI solution for the problems and the results obtained from the BI tools applied.

All the nine researchers who studied on the BI application in the manufacturing company applied it in the Production or Operation section of the organizations. There are different areas of manufacturing where the studies had been done, which are semiconductor, cement, chemical, faucet, electronics, general manufacturing enterprise, plastics and chemistry and automotive.

The studies show problems related to business data and execution of the organization. The most common problem is the reports inconsistencies and difficulties. This eventually imposed a delay in decision making process. Other than that, the lack of visibility of certain business activities are also the major concerns for the manufacturing firms. The need to increase the production output is also among the common challenge for the manufacturing organizations to implement the BI.

With the major problems faced by the manufacturing organizations, researchers have come out with different types of BI framework. Majority of the studies focused on developing frameworks that are doing the integration of the existing systems with the BI services. The second popular BI framework for solving the problems of business execution for manufacturing organizations is the dashboard. There is also the web based tool implemented.

Above all, it is the benefit of the BI applications that all of the manufacturing companies are looking for. The most obvious results of BI applications that benefit them is the ability to see the performance of certain business process in real-time or the visual representation of data in an informative manner. A number of manufacturing organizations also experienced higher productivity while reducing the manufacturing cost. The benefit of improving the customer related activities is also gained.

Thus, it can be concluded from the previous studies that the manufacturing industry indeed did implement the BI applications in order to boost up its growth. In its highly competitive market, where the manufacturing organizations are facing with a large volume of data, BI is seen to be the best solution for all. In order to establish a BI framework, researcher must focus on the visual representation of data that shows the performance of

organization's operation. The integration of existing manufacturing systems with BI tools also should be taken into consideration, as well as having the web and portal for the business process.

5. Conclusions

This paper reviews the various applications of BI for the improvement of an organization performance in the manufacturing industry. In all the case studies, BI applications helped the organizations to overcome most of the problems they had, particularly in relation to the information overload while there is a need to extract a valuable information from the data. Without having enough visualization and information, it is time consuming for the management and employees in general to plan future steps and path forwards to run the operation smoothly, subsequently lead to the remarkable poor on time delivery performance, higher production cost, poor production planning, etc. This paper proves that BI should not be neglected nowadays, if we have a lot of data but could not answer the question of what is important in the data.

With the reviews being discussed, this paper opens up extensive research for the implementation of BI in the manufacturing industry. Further study will be done, in which it is expected to help the decision makers make full use of their business information, in the sense that data is turned into a useful information and knowledge. In the next case study, it is hoping that the BI framework to be designed is able to benefit the frontline and operational employees of the manufacturing company, by helping the organization improves its on time delivery performance consistently.

Table 1: The Application of Business Intelligence in Manufacturing

Researcher	Manufacturing Sector	Area in Organization	Problems	BI Solution	Results
A.L. Azevedo and J.P. Sousa, 2000	Semiconductor	Production and Operation	• Order prioritization is only by date • Unlimited capacity assumption • Time-consuming plan regeneration	• Decision Support System – Business Systems and Manufacturing Execution Systems integration	• Customer orders management in real-time in a distributed environment. • Delivery dates are determined based on capacity check, thus improve the due date calculation efficiency, precision and reliability.
Russell Barr, Fayyaz Hussain and James Sommers, 2005	Cement	Operation & Finance	• Information is shared by e-mail with excel spreadsheet attached leads to data inconsistency • E-mail sent is from different time frames	• Real-time Performance Dashboard	• 3% reduction in operation costs. • 5% increase in production
Gang Xiong, Timo R. Nyberg and Feiyue Wang, 2010	Chemical	Production & Global	• No common visibility among departments – inconsistent decision making • Low production output due to no real-time response ability to manufacturing disruptions and demand changes • High maintenance cause due to no real-time between production plan and execution	• XMII (Manufacturing Integration and Intelligence)	• 3% - 5% reduction in manufacturing costs • 8% - 10% increase in production yield • Increase customer responsiveness

Juhani Heilala, Matti Maantila, Jari Montonen, Jarkko Sillanpaa, Paula Jarvinen, Tero Jokinen and Sauli Kivikunnas, 2010	Faucet	Production	• Manufacturing simulation data is updated only once or very rare • Simulation analysis produces many tables, lists and reports – difficult and time consuming for decision makers to locate the information	• Simulation-based Decision Support System focusing on visualization.	• Capable to see the potential bottlenecks or other production problems to take corrective actions • Pro active planning and problem solving for production • Benefit for production operators: Early information for upcoming work • Benefit for production engineers: Planning changes or new systems
Anil B. Jambekar and Karol I. Pelc, 2006	Electronics Measuring Instruments	Production, Finance, Competitors and Customers	• No monitoring systems to adapt to industrial operational condition. • No preparation for managers for potential increased production sale. • Serious needs to increase sales and expand business.	• Managerial Dashboard	• Ability to monitor the firm's operation performance • Managers benefit it by able to identify technical and managerial knowledge to prepare for a large scale manufacturing
G R Gangadharan and Sundaravalli N Swami, 2004	Electrical and Electronics Components	Production, Store and Sales	• Difficulty to forecast sales, production and distribution • Poor service and high inventory level • Reporting systems are hard to use, inflexible and outdated	• Data Mart, Data Tracker, Reporting and Web Integration	• Boosted up the company's revenue by 36% • Information that used to take hours or days to report is available instantaneously – in sales, forecasting, production, planning, order tracking, profit analysis and ad-hoc reporting

Cheng Yuan and Li Zhigang, 2010	General Manufacturing Enterprise	Production involving Technology, Planning, Dispatcher, Manufacturing, Store and Logistics.	• The application of traditional BI is separated with business process execution • There is a need to convert business-relevant data into analytic information systematically	• Process-oriented Business Intelligence – Integrating BI services with business processes like production, planning, procurement, store etc.	• Close monitoring on key performance indicators by engineers – Improve overall technical process control • Planners can establish reasonable production plan where orders are available analytically • Supervisors can monitor production schedule, improve resources management – minimize cost and maximize production output • Operators able to use production equipment effectively and arrange tasks reasonably
Leo Sennott and Jorge Willemsen, 2009	Semiconductor	Production	• There is a need for the company to improve product and process yield with thin profit margin • Different data sources come from different facilities – A need for data integration	• Dashboard (Desktop Status) • Web-based analysis tool (Parameter Viewer) • Portal (Skyworks Data Portal and Rapid Prototype Line Portal)	• Yield monitoring capability • Improve product performance activities • Real-time production build status • Real-time visibility into various plant manufacturing operations • Provide real-time knowledge to improve the company's competitiveness
Margarete T. Koch, Henning Baars, Heiner Lasi and Hans-Georg Kemper, 2010	• Plastics and Chemistry • Automotive	Production / Manufacturing Operation	Plastic and Chemistry: • No Overall Equipment Effectiveness-indicator • No daily reports • No features for process analysis • No business oriented analysis in MES Automotive: • No package cycle analysis • Insufficient integration with non-production related systems	• Operational BI - Integrating Manufacturing Execution Systems with BI	• Increase business performance by integrating complete processes and enriching technical indicators with economic data • Machine and production data can be used to do combined analysis. e.g. To analyze the production related choices on customer and financial side

References

[1] Accreditation Commission for Programs in Hospitality Administration. (n.d.). Handbook of accreditation. Retrieved from http://www.acpha-cahm.org/forms/acpha/acphandbook04.pdf

[2] A. L. Azevedo, and J. P. Sousa, "A Component-based Approach to Support Order Planning In A Distributed Manufacturing Enterprise", Journal of Materials Processing Technology, Vol. 107, No. 1-3, 2000. pp. 431-438.
[3] A. B. Jambekar, and K. I. Pelc, "A Model of Knowledge Processes In a Manufacturing Company", Journal of Manufacturing Technology Management, Vol. 17, No. 3, pp. 315-331
[4] B. Hameed, J. Minguez, M. Wörner, P. Hollstein, S. Zor, S. Silcher, F. Dürr, and K. Rothermel, "The Smart Real-Time Factory as a Product Service System", in Proceedings of the 3rd CIRP International Conference on Industrial Product Service Systems, Technische Universität Braunschweig, Braunschweig, Germany, 2011, pp. 326-331.
[5] B. S. Sahay, and J. Ranjan, "Real Time Business Intelligence In Supply Chain Analytics", Information Management & Computer Security, Vol. 16, No. 1, pp. 28-48.
[6] C. Howson, Successful Business Intelligence: Secrets to Making BI a Killer App, USA: The McGraw-Hill Companies, 2008.
[7] C. Yuan, and L. Zhigang, "The Research & Application of Process-oriented Business Intelligence in Manufacturing Industry", in International Conference on Management and Service Science, 2010, pp. 1-4.
[8] G. Xiong, T. R. Nyberg, and F. Wang, "Real-time Manufacturing Integration and Intelligence Solution Applied in Global Process Industry", in Service Operations and Logistics and Informatics (SOLI), IEEE International Conference, 2010, pp. 270-275.
[9] G. R. Gangadharan, and S. N. Swami, "Business Intelligence Systems: Design and Implementation Strategies", in 26th International Conference of Information Technology Interfaces (ITI), 2004, Vol. 1, pp. 139-144.
[10] H. P. Wiendahl, H. A, ElMaraghy, P. Nyhuis, M.F. Zäh, H. Wiendahl, N. Duffie and M. Brieke, "Changeable Manufacturing - Classification, Design and Operation", CIRP Annals Manufacturing Technology, Vol. 56, No. 2, 2007, pp. 783-809.
[11] IDC Research. Worldwide Business Intelligence Tools 2005 Vendor Shares, 2006, USA, IDC #202603.
[12] J. A. Harding, M. Shahbaz, and A. Kusiak, "Data Mining in Manufacturing: A Review", Journal of Manufacturing Science and Engineering, Vol. 128, 2006, pp. 969 – 976.
[13] J. Heilala, M. Maantila, J. Montonen, J. Sillanpaa, P. Jarvinen, T. Jokinen, and S. Kivikunnas, "Developing Simulation-Based Decision Support Systems for Customer Driven Manufacturing Operation Planning", in Proceedings of the 2010 Winter Simulation Conference, pp. 3363-3375.
[14] J. Jenkole, P. Kralj, N. Lavrac, and A. Sluga, "A Data Mining Experiment on Manufacturing Shop Floor Data", in Proceedings of 40th CIRP International Manufacturing Systems Seminar, 2007.
[15] J. Ranjan, "Role of Business Intelligence in Supply Chain Management", Global Journal of e-Business & Knowledge Management, Vol. 5, No. 1, 2009, pp. 1- 7.
[16] J. Shin, S. Park, C. Ju, and H. Cho, "CORBA-based Integration Framework for Distributed Shop Floor Control", Computers & Industrial Engineering, Vol. 45, 2003, pp. 457–474.
[17] L. Sennott, and J. Willemsen, "Web-Based Business Intelligence for Semiconductor Manufacturing", in International Conference on Compound Semiconductor Manufacturing Technology, 2009.
[18] M. Golfarelli, S. Rizzi, and I. Cella, "Beyond Data Warehousing: What's Next In Business Intelligence?" in DOLAP '04, Washington DC, 2004.
[19] M. Kristiansen, R. Young and P. Ittycheria, "The New View: Dashboards Show Pipeline Enterprise In Real Time", Pipeline & Gas Journal, 2008, http://www.pgjonline.com
[20] M. Lewis and N. Slack, Operations Management: Critical Perspectives on Business and Management, London: Routledge, 2003.
[21] M. J. Shaw, "Information-Based Manufacturing with the Web", The International Journal of Flexible Manufacturing Systems, Vol. 12, 2000, pp. 115–129.
[22] M. T. Koch, H. Baars, H. Lasi, and H. G. Kemper, (2010). "Manufacturing Execution Systems and Business Intelligence for Production Environments" in Proceedings of the Sixteenth Americas Conference on Information Systems, 2010.
[23] Microsoft Dynamics™ AX, "Build a Competitive Edge for Manufacturing Plant Operations", 2006, White Paper.
[24] Ministry of International Trade and Industry, MITI Weekly Bulletin, Kuala Lumpur (Malaysia): Weekly Bulletin, 2011.
[25] Organization of the Petroleum Exporting Countries, Monthly Oil Market Report, Vienna, Austria: Issued 17 January 2011.
[26] R. Barr, F. Hussain, and J. Sommers, (2005). "Real Time Modeling for Financial and Performance Management", in Cement Industry Technical Conference, 2005, pp. 43-51.
[27] R. T. Herschel and N. E. Jones, "Knowledge Management and Business Intelligence: The Importance of Integration", Journal or Knowledge Management, Vol. 9, 2005, No. 4, pp. 45-55.
[28] S. Negash, and P. Gray, (2003). "Business Intelligence", in Americas Conference on Information Systems (AMCIS), 2003.
[29] W. F. Cody, J. T. Kreulen, V. Krishna and W. S. Spangler, "The Integration of Business Intelligence and Knowledge Management", IBM Systems Journal, Vol 41, No. 4, 2002, pp. 697.

Ernie Mazuin Mohd Yusof received her B. Eng. Degree in Computer and Information Systems Engineering from the International Islamic University Malaysia in 1999. She is currently working as a Senior Program Executive in an electronics manufacturing company. She was holding a Senior Engineer post before. She is also currently taking a Master of Science (Computer Science) in University Technology Malaysia. Her research interests cover the business intelligence, visual representation and decision making tool.

Mohd Shahizan Othman received his BSc in Computer Science with a major in Information Systems from Universitiy Technology Malaysia, in 1998. Then he earned MSc in Information Technology from the Universiti Kebangsaan

Malaysia (UKM), Malaysia. Soon after, he graduated for his PhD in Web Information Extraction, Information Retrieval and Machine Learning from UKM. He is currently a senior lecturer at the Faculty of Computer Science and Information Systems, UTM, since 2001. His research interests are in information extraction and information retrieval on the web, web data mining, content management and machine learning.

Yuhanis Omar is a lecturer of Information System Department, Malaysian Institute of Information Technology in Universiti Kuala Lumpur. She is now pursuing her Ph.D degree in Information Science at Universiti Kebangsaan Malaysia on 'e-Train : the Effectiveness of Engagement Environment in Educational Portal Assessment Module'. Her research interests are in e-Learning and Software Engineering. Previously she has heavily involved in the development of Geographical Information System (GIS), Management Information Systems and educational portal.

Ahmad Rizal Mohd Yusof received his B.IT in Industrial Computing from Universiti Kebangsaan Malaysia, in 2000. Then he received M.IT in Computer Science from the Universiti Kebangsaan Malaysia (UKM) in 2003. He received his PhD in Knowledge Management from UKM in 2009. He is currently a senior lecturer at the Institute of Occidental Studies (IKON), UKM. His research interests are in the Knowledge Management, Formal Methods, JAVA and C++ Programming Language.

Problems in Software Quality Assurance and Reasons

Mohammed Alshammri

Faculty of Engineering and Information Technology , University of Technology, Sydney
Sydney, NSW 2007, Australia

Abstract

This paper is aimed at highlighting the problems which has been faced by the project managers as well as the companies regarding the quality assurance. It has been seen that people do not pay much attention towards the quality assurance issues and thus eventually end up with wasting their money as well as time. That's why it is important to make sure that the project meets the quality requirements.

Keywords: *software quality, Quality Assurance, software problem.*

1. Introduction

Software Quality Assurance (SQA) aims at monitoring the software engineering processes to ensure quality of the software. Quality has also been added in the triple constraints of the software development which includes time, scope and cost. Now it is known as quadruple constraint as quality has also been added to it. Multiple testing standards are available which have their own pros and cons. The researches for the testing standards are ongoing and are also highly important for the software development companies as this helps them in choosing the right standard for their company which suits their software requirements and fulfill their needs.

Research should be conducted to evaluate these testing standards and provide them with more improvements and amendments as this can help the companies for developing more quality software. The software companies are itself the stakeholders along with the clients and the sponsors. The sponsors and the clients want their software to meet all the four constraints which are time, cost, scope and quality. If the software does not fulfill the client's quality criteria then it is of no use and eventually ends up producing nothing. The software industry is the host for software quality assurance.

2. Background

Most people simply accept the poor quality software from the Information Technology products. So what if your computer crashes twice a day? You simply backup your files. So what if you are not able to log in to your corporate intranet or internet? Just try after sometime when there is less traffic. Is this the solution? No! This shows the bad quality software which lacks in providing the features.

2.1 Timing Difference Quality Issue

In 1981, a small timing difference caused by a computer program change created a 1 in 67 chance that the space shuttle's five on-board computers would not synchronize. This error caused a launch abort. This real life example shows that how much it is important to ensure the quality of the project. (Dong, 1984)

2.2 Fatal Doses of Radiations

In 1986, two hospital patients died after receiving fatal doses of radiations from a Therac 25 machine. A software problem caused the machine to ignore calibration data. This does not seem like a little problem. This shows how a little problem in software can cause major problem. Similarly Britain's Coast Guard was unable to use its computers for several hours in May 2004 after being hit by the Sasser Virus. So it is highly necessary to maintain the quality of the software in order to avoid future problems. (Jones, 2011)

3. Problem Statement

The problem which has been faced in most of the software projects is due to the lack of quality. The developers do not pay much attention towards the quality assurance of the projects and thus face the difficulties in the future. The project either results as a failure or create much problems in future which eventually lead to a bigger problem.

4. Factors Influencing Success of the Projects

Many factors influence the success of the project. Cost, time and quality are the factors that influence the success of the project. If these triple constraints of the project are met then it is way easy to make or declare a project successful.

To construct the customer relation management system it is highly important to first analyze the issues the business is facing right now at the current time and then gather all the requirements for the new system. It is impossible to build a system if its requirements are not well known.

5. High Profile Projects' Failure – Reason "Lack of Quality"

Below are some high profile projects examples which lead to the project failure because of lack of project quality. One important thing which is to be noted that quality does not merely means that the project works fine.

There are a number of factors like the meeting the satisfaction needs of the clients, fulfillment of the requirements which include functional and non-functional requirements and also other important factors like long term maintenance. There are a number of bench marks across which the project is tested in order to see whether it fulfills the needs or not. (Muneo Kitajima, 2012). It is true that developing a project is easy but developing it within the budget, scope and requirements is different. As all these factors combine to form a complete quality project.

5.1 Huntington Bancshares, Inc.

A very best example of this high profile project is the Huntington Bancshares, Inc. (Schwalbe, The Importance of project phases and management reviews)The CIO of this company Joe Gottron said that there were "four or five very intense moments" when everyone was seeing this project as a major failure just because of its complexity.

5.1.1 Problems faced by Huntington Bancshares, Inc.

- The project developers decided and selected a technology which was complex enough but later it was discovered that it is not applicable in the real world thus resulting in a lot of wastage of time and money.
- They never prioritized the user requirements and a lot of trouble was created at the time of project completion. The problems should be categorized in the following categories:
    - Must have
    - Should have
    - Could have
    - Would have

5.2 Jordan Telecom (JT)

Another very interesting case is of Jordan Telecom (JT). It is Jordon's only telecom operator. They were having a lot of problems regarding the project management criteria. As it was not possible for them to set the results according to the client's benchmarks but later it turned out to be a success.

They managed a number of ways later to process the model in an effective way. JT developed a three lines of processes based on the size of the project. (Douz_Korned)

JT used the models of project development and also found ways on how to develop a customized design for the project and its development.

5.2.1 Problems faced by Jordan Telecom (JT)

- The managers did not consult with the clients and stakeholders which then lead to problems like the system does not produce the expected results and cause a lot time and money wastage.(www.bleuphish.com)
- The managers or developers assumed the requirements. The developers should ask the customers about what they are expecting from the system. Mostly the clients are not well aware of what they want from the system. The developer should help the client in conveying the requirements of the system in an effective way. They should ask the client about the requirements instead of assuming the requirements.
- The developers started to solve the before even knowing what exactly the problem is.

6. Reasons for Benchmark Failure

There are a number of benchmarks developed to maintain the quality of the projects. These benchmarks are actually the quality standards which help in understanding the current status of the project in terms of quality. Theses benchmarks make sure that the project is according to the quality standards or not.

At times these quality standards are also unable to check the standard of quality of the project. There are a number of reasons for the benchmark failure. Project managers play an important role in ensuring that the project is up to the mark and fulfills all the quality measurements.

6.1 Non-Termination of Failure Projects

It is the project manager's duties to terminate such projects which are most of times in lose but some managers' start to invest more money and resources in order to prove that the project is a success.
The project managers should view all the past and future outcomes of the project in order to decide whether to continue the project or terminate it. If the project managers have these three psychological factors which are information biasing, reinforcement and self justification then it is more likely to continue the project and make it a success.

Some project manager's terminate the projects because they don't want to expose their mistakes in front of others or they have lost all the hope. Lack of managerial support and insecurity of job even forces the managers to terminate the projects instead of investing more. It merely depends upon the managers' nature. If they have strong leadership qualities then even after great failure they continue with the project and eventually make it a success.

6.2 Lack of Vision

It is not necessary that the project vision remains the same from the very start. Sometimes the managers are not even sure about what they are going to develop. The project sometimes takes turn and to assure its success the managers need to take decisions which in turn results in change in the vision of the project.

Sometimes in order to fulfill the demands of the stakeholders, clients and sponsors it becomes really important to change the vision of the project in order to meet their goals and satisfy their needs.

The stakeholders are at times so interested in the success or progress of the company instead of the organization thus resulting in changed project vision to attain the company's specific goals instead of the overall goal for which the project was basically developed. The changed vision statement should be communicated to the team members in a very good way with some strong facts and figures. (Patil, 2011)

7. Conclusions

Now-a-days software developers are not paying much attention towards the quality of the software products. This attitude could lead to them to the road of failure and would eventually result in the wastage of time and money both. In order to highlight the importance of quality assurance there should be a separate department to carry out the quality assurance methodologies to each information technology projects.

References

[1] Dong, C. (1984). Failure mode and effects analysis based on fuzzy utility cost estimation. International Journal of Quality & Reliability Management.

[2] Douz_Korned. Chapter 2. In Douz_Korned, IT Project Management**.**

[3] Jones, R. H.-A. (2011). The Quality Assurance Journal.

[4] Muneo Kitajima, H. T. (2012). Journal of Quality Assurance in Hospitality & Tourism.

[5] Patil, A. (2011). International Journal of Quality Assurance in Engineering and Technology Education (IJQAETE).

[6] Patil, A. (2011). International Journal of Quality Assurance in Engineering and Technology Education (IJQAETE), http://www.igi-global.com/journal/international-journal-quality-assurance-engineering/41026

[7] Schwalbe, K. The Importaqnce of project phases and management reviews.

[8] In K. Schwalbe, Information Technology Project Management (p. 62).

[9] www.bleuphish.com. (n.d.). Adaptive Development Life cycle. Retrieved April 2012, from http://www.bleuphish.com/adaptive_product_develpment_lifecycle.html

Policy-Based Support for Mobile Grid Services

Tariq Alwada'n[1], Thair khdour[2], Helge Janicke[3], Abdulsalam Alarabeyyat[2], Abdel Rahman Alkharabsheh[1]

[1]Faculty of Technology, The World Islamic Sciences and Education University
Amman, Jordan
[1]{taiq.alwadan,ar.karabsheh}@wise.edu.jo

[2]Prince Abdulah Ben Ghazi Faculty of Information Technology, Al- Balqa Applied University
Amman, Jordan
[2]{khdour,alarabeyat}@bau.edu.jo

[3]Faculty of Technology, De Montfort University
Leicester, UK
[3]{heljanic}@dmu.ac.uk

Abstract

In a multi-organization environment like the GRID, each institute might want to apply some boundaries on how its resources are being utilized by other institutes. A disagreement between the multi-Virtual Organizations (VOs) might happen in the security aspect for the policy framework. Mobile Grid Services has given the ability to move jobs, data and application software from nodes to nodes during jobs' execution in the grid environment. It has also solved some of the lack in finding suitable resources for the jobs. To facilitate the ability to support mobile resource sharing between multiple heterogeneous VOs, an authorization policy management framework is needed to support authorization for heterogeneous authorization systems. Traditional authorization policy management frameworks act well in authorization policy for a single VO where the contributing hosts grant the permission to follow a global authorization system. However most of policy management tools do not provide a clear support for sharing mobile resources between multiple heterogeneous VOs. To solve this problem, we present a dynamic and heterogeneous policy management framework that can give a clear policy definition about the ability to move jobs, data and application software from nodes to nodes during jobs' execution in the grid environment. We introduce an architecture for policy based resource management in the case of mobile sharing, and a scenario that explain the advantages of mobility mechanism and the role of policy in the grid systems. To check the performance of this architecture, a set of experiments had conducted. The results, analysis and the overheads estimation are presented in this journal.

Keywords: Grid Computing, Mobility, Policy.

1 INTRODUCTION

Due to the advances in communication technology and global system of interconnected computer networks (internet), grid computing appears as a result of a combination of multi-network computer system to develop a wide range and heterogeneous system used to solve scientific or industrial problems [1]. A grid is a system that should have the ability to organize resources[1] which are not under the subject of centralized domain, utilize protocols and interfaces and supply high quality of service [2]. Thus, the major advantage of grid computing is the capability to organize and share resources [3], [4]. As a result of such technology many challenges, such as finding suitable resources and reducing number of rejected jobs, stand in front of developing it. There are a lot of contributions to solve some of these challenges, for example: the mobility has solved some of the lack in finding the suitable resources for the jobs, but not a lot of attentions was given to the policy (aspect of security and privacy) in this solution.

Mobility is the ability to migrate or relocate jobs, data and application software among grid nodes. These migrations depend on the grid's users and the grid's nodes policies. Mobility facilitates the accomplishment of requirements for grid jobs as well as grid users. It also assists grid evolution, improves performance of operating applications by relocating data to the target host, therefore reducing the communication consumption and solving the load balancing issues. David G. Rosado et.al [5] described the mobility as "In the purview of Grid and Mobile Computing, Mobile Grid is an heir of the Grid, which addresses mobility issues, with the added elements of supporting mobile users and resources in a seamless, transparent, secure and efficient way [[6], [7], [8]]". Mobility can be divided into the following category: computer, personal and computational mobility. In personal mobility; the grid's user can do the job at sites remote from the actual physical hardware without having to move jobs around with them. They can launch a job in one site and move it to an other place in the world no matter what the machine type, such as web-based email accounts. Meanwhile the computer mobility is interested in moving actual computer hardware parts from one place to another, for example relocating PCs notebook and other PC parts. The last one, this paper is interested in, is the computational mobility. This type of mobility deals with the movement of software between nodes [9], [10]. Computational mobility may also be known as a control migration, data migration,

1. Resources refere to management and computing resources. For example: computer, software applications, etc.

link and object migration [11]. This type of migration allows the data and codes to migrate and execute on various systems across the network. Also it offers moving execution control and the ability to connect software elements at runtime while migrating from one system to another and back to the original system again.

Security is an essential element in grid computing. One of the important issue that research into grid environment tries to solve is how to keep distributed resources from unauthorized users and at the same time allowing the sharing of resources and the accountability for resource handling. Every resource applies its own security policy that may result in the refusal of requests for utilizing of its resources. Because of the fact that there are a lot of elements, like users and resources, contributing to the grid, security has become a critical aspect in checking the element trying to use a service (authentication), and in verifying whether this element is allowed or not, to use the service (authorization). Securing the grid therefore is vital to give confidence to both grid users and resource providers.

Policies are groups of regulations, standards and practices written by the administrators of resources about how their resources or jobs can be handled and used. Policies decide the way that a specific job should be accomplished, how security is applied in a domain and how an organization organizes, secures and distributes their resources. Depending on the Globus Toolkit [12], before the job submission, there should be many steps for authenticating the users who ask to use resources [13], [14]. However, after the authentication, there are no further resource access restrictions on how to use the resources.

The rest of the paper will be organized as follows. The next section describes the component of our grid architecture and describes each component in a separated section. In section three, we give an explanation of the grid portal as part of the grid architecture and its advantages. Section four presents our resource broker and its architecture including the suggested framework for the mobile grid policy services and a scenario that explains the advantage of mobility mechanism and the role of policy server in it.The following section introduces our simulation followed by validating this simulation and presents the results. In the last section we discuss future possibilities and conclude the paper.

2 Architecture Structure and Components

Grids depend on enhanced software that guarantees seamless communication between components nodes. It uses an effective mechanism which determines the suitable policy(s) that should be applied to achieve the best way to utilize resources in a way that guarantee privacy and security for both grid users and grid resources.

Figure(1) shows our proposed architecture. It applies Client/Server architecture since this architecture is the most favorable type in heterogeneous environments [15]. Client/Server network includes clients and servers who operate on the proper hardware and software for their jobs. There are two forms of client/server architecture; two and three-tier (multi-tier). Our architecture employs the last model which compromises of the client (grid portal) as the first tier, the resource broker as second tier and grid nodes as third tier. The following describes the functions for each one of them.

3 GRID PORTAL

A grid portal or grid interface is a virtual computing resource performing an interface on behalf of grid users to approach the grid. A portal has many features such as hiding the complexity of the grid from users via a simple interface which facilitates the classification of grid job necessities.

4 Resource Broker

The Resource Broker is one of the major grid elements. It performs significant functions in building a valuable grid environment by arranging user jobs onto grid resources to reach particular accomplishment targets, like cutting communication delays, raising the resource exploitation, reliability and distributing jobs across resources without depending on a particular resource. The main job for the broker is to discover and choose suitable resources for jobs by sending jobs input files to the resources, monitoring jobs and sending outputs to users.The resource broker presented in this paper is based on the mobility framework and isolates the user from the grid's middleware.

4.1 Resource Broker Architecture

The resource broker accepts job requirements from the portal and looks for appropriate resources that can fit these requirements. First it asks for all information about the available resources from the information service and the data information stored in the replica catalogue. Then it chooses the resources that can fit the job requirements and asks the grid policy agent about policies for those resources. According to that, the resource broker's architecture compromises of three components indices: information service, the replica catalogue and the grid policy agent.

4.2 Information service

Information service is a crucial element in grid computing. It is a directory service holding data about all the grid resources and the entire grid activated jobs operating on those resources. This information can be either dynamic or static information. The last one is for the hardware conditions and the operating system, while dynamic information related to the resources available time, the job presently running, type of application software, disk space and policies. In order to advertise their information the resource broker communicates to both resources and the information service to ask for this information.

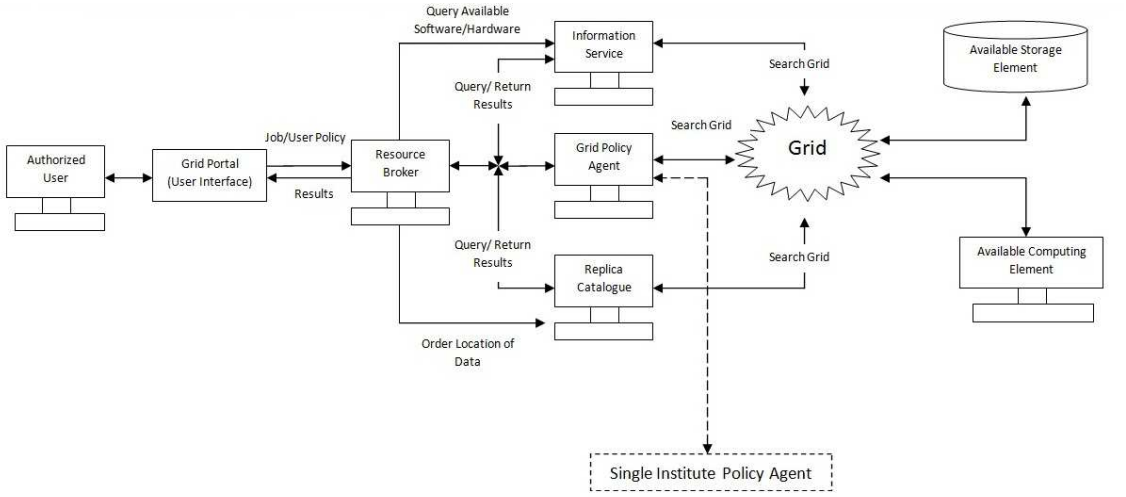


Fig. 1. Grid Architecture

4.3 Replica Catalogue

This is also an important component for the grid, because it presents information which helps in accessing the stored data in the grid. It determine the places of data in the grid, updates data resources and maps logical file names to the actual physical places on grid resources. In order to use the data on the grid the resource broker communicates with a replica catalogue to ask for information about data place and the access control needed to utlize this data.

4.4 Grid Policy Agent

The grid policy agent contains all the policies information about all resources in the grid. Each institute should have as a minimum one policy manager (agent) that has the capability to access the policy database or policy information for that institute. All policy agents (PAs) in all domains in the grid should be registered with the grid policy agent and should send their policy information (e.g. policy framework) or any changes or updated data about their policies to the grid policy agent[16]. Grid administrator can specify the policies for units participated in the grid but it does not have any policy managers (agents) that can use it directly. As an alternative, a grid policy agent operates as a proxy for the policy agents which run at each of the different institutes.

Our resource broker is differentiated from others by adding the mobility feature as a new characteristic for resource brokers. Mobility is the ability to move physical or virtual computational resources (software code, data, portable notebook PC's, running objects and mobile agents) from one site to another through a local or wide network. Mobility is a wide idea used in distributed computing. Advantages in related to services that have the capability to migrate between nodes such as increase resource utilization, enhance the organization between services and presented resources[8].

Figure(2) shows the architecture of our mobile policy agent and its components. The main job now for the institute policy agents is to merge the policies from the organization administrator and the policies from the global grid to obtain the efficient set of policies for resources belonging to that institute. The efficient set of policies are the ones applied by the policy agents attached to each resource assigned to that institute in the grid. The following describe each one of them.

4.4.1 Data/Application Software Agent

This agent is responsible for the data and application software movments. Our grid architecture lets application software and/or data to migrate from one node to another in the grid system so as to adapt the resources needed to fit the job requirements. If the resource fits the job hardware conditions and the time available, but does not have the needed application software or data needed for the job(s), the resource broker will look in the grid for the nodes that have this data/application software by checking the replica catalogue and information service and putting these nodes into a new list. Each node will be checked, one by one, by asking the mobile policy agent to decide whether or not the data/application's software policy in these nodes allows their movements or allow copying the needed software from them. The Data/Application Software Agent will check the policies for the nodes that contain the required data or application software and return the results to the resource broker. If one of the nodes does support the mobility feature for data/application software, the resource broker will copy or move (migrate) properly and send it to the resource that fits the job hardware and time requirements along with its policy. If all the nodes' policies do not support the data/application software mobility, the broker will inform the user that the grid cannot run the job.

4.4.2 Job Agent

This agent is responsible for checking the grid users' policies. Our grid architecture lets the job and its execution state to migrate from one resource to another and restart on the new one in order to fit the job conditions and requirement. If the resource that fits the job hardware conditions is occupied at the time needed, our resource broker will evacuate this resource by moving the presently operating job to other resources (if

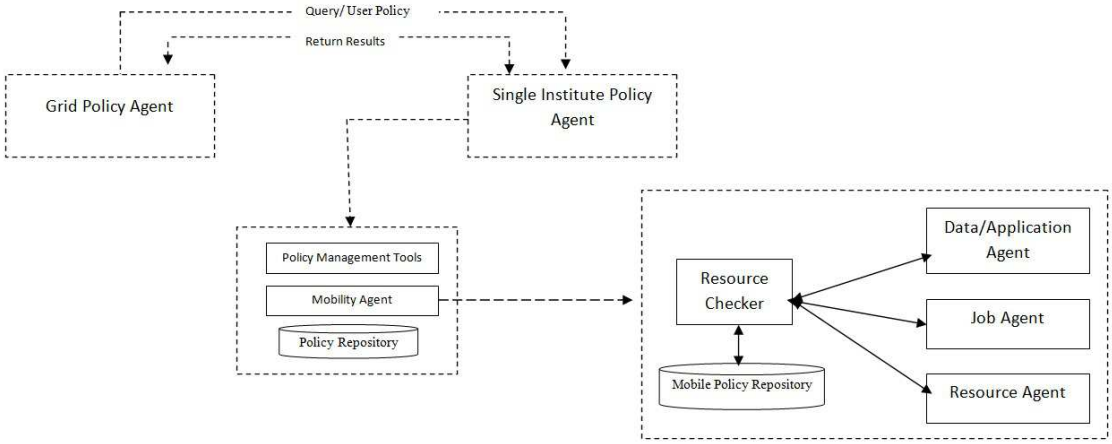


Fig. 2. Mobile agent Architecture

they exist and have the job requirements). This can be done by looking for jobs that are running on this required resource and acquire the needed information about them from the replica catalogue and the information service. If the job conditions can be fulfilled by other resources the resource broker will ask the mobile policy agent if the currently running job(s) is allowed to be migrated to another resources. The Job Agent in the mobile policy agent will check whether or not the grid user's policy allows migrating the running job to the new resource and returning the results to the resource broker. If the policy allows this kind of migration, then the resource broker will relocate these jobs to the new target resource(s) and transfer the new job to the vacated resource which can fulfil its conditions.

4.4.3 Resource Agent

This agent checks whether or not the resources' policies allow the migration for jobs, data and application software between various resources. Our grid architecture allows jobs, data and application software to move from one node to another in the grid system in order to acclimatize the resources needed to fit the job needs. If the resource that meets the job requirements is currently busy and there is a need to migirate to other resources, or there is a need for data or application software migiration, the resource broker will ask the mobile policy agent to check the policy aspect in these situations. The Resource Agent in the mobile policy agent will determine whether or not the current resource's policy allows the job migration from its node to the destination resource, or if the destination resource can accept jobs from the original resource. In both cases, it will inform the resource broker about the results. In the case of data/application software migration the resource agent in the mobile policy agent will determine if the addition or migrating of data/application software policies are allowed this type of action in the current resource and the destination resources. If they do not, the broker will inform the user that the grid cannot operate the job. If they do, the broker will apply the migration between those resources.

4.4.4 Resource Checker

As soon as the mobile policy agent makes its decisions about any possible migration(s) either for jobs, data or application's software, it stores indexes for these decision using the resource checker and stores these indexes in the policy repository prior to submitting the decision's results to the resource broker. The aim of these indexes is to track any changes or updates in the target policy(s) and inform the resource broker about them. This helps in enhancing the mobile policy agent performance and throughputs by returning to these indexes for any new requests from the resource broker instead of going for the whole checking operation again.

After the authorized grid users submit their jobs to the core of the grid system (resource broker), it asks the Grid Information Services (GIS) and Replica Catalogue about the free resources in the grid. Later, it sends this information along with the related policies (Users policies) to the Grid Policy Server which forward it to the Single Institute Policy Server to make the final Policy decisions, then it sends the results back again to Grid Policy Server. The Grid Policy Server sends the results to the resource broker to enforce the policy results in its decisions[17].

As a result the mobility has created a new environment that can solve these cases. In order to apply the mobility, the policies for the elements in Figure(3) should be checked before any migrations can take place. In our model mobile policy agent plays a significant role to achieve these requirements. The mobile policy agent checks the policies for each element in the three levels (Figure3). Each element in these levels is consider as a Policy Decision Point (PDP) where the policy decision is taken place and forward this decision to its related agent. If the policies in (level 1) have given the green light for the migration, the mobile policy agent checks the policy for the target domain (Level 2) to see if that domain is allowed to have all (or any) of the elements in (Level 1). If so, the next step should be checking the node which is going to be the new host for the elements in (Level 1). By checking the policies for both elements in (Level 1) and node's policy in (Level 3) mobile policy agent takes the decision if that node allows to have the immigrant element in (Level 1) or not.

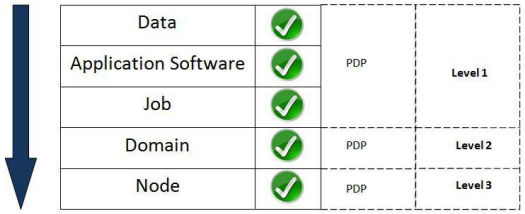


Fig. 3. Policy's Level

4.5 A Mobile Policy Agent Example

The following scenario explains the advantage of mobility mechanism and the role of policy in it within grid systems. It is divided into three sections.

- **First section: Grid Resources Specifications**
  The grid contains five nodes; each node has different conditions and specifications. These specifications are: hardware, domain, application software, data and policies. Each node is responsible for defining its own policy. Also it contains the running jobs, if presented, as shown in Tables (1), (2) and (3).
- **The second section: Jobs Requirements**
  There are five jobs which need to be executed by the grid resources. The requirements needed to accomplish the jobs include hardware, software, input, output, dmoanin and policies, as shown in Tables (4), (5) and (6). Grid users are responsible for defining their policies when submitting their jobs to the grid.
- **The third section: Fits the Jobs Requirement to Grid Resources**
  The resource broker is responsible for locating the optimal resource that can meet the job requirements and scheduling the jobs into grid resources with respect to the policies. All of these issues will be illustrated in the following. It is also shown in Figures (4) and (5).

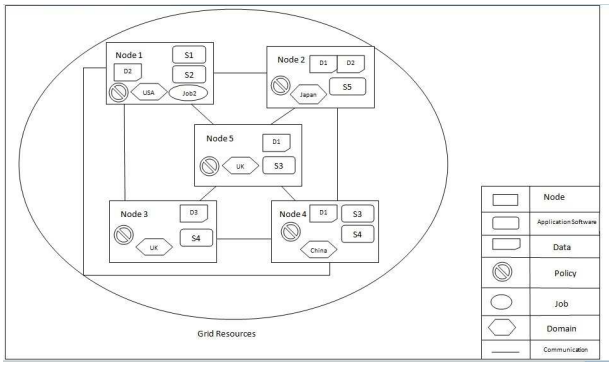


Fig. 4. Grid Resources (Infrastructure)

- **Job Migration**
  From Tables (4, 5 and 6) it can be noticed that Job1 requirements fit the Node1 specification in Tables (1, 2, 3), but Node1's policy, Table (3), is to allow only a single job to run at any time (exclusive execution), so there is a need to migrate the existing job (Job 2) on Node1 to another node that fits Job2 requirements. The resource broker looks for this substitute node and finds Node4 and Node5; but Node4 domain is in China which is against the policy of Job2 and Node1 policy. Therefore, the resource broker sends job1 to Node1 and move Job2 together with its status (memory image) to Node5 for execution.
- **Data Migration (case 1)**
  In Tables (1, 2, 3), Job3's requirements fit Node3's specifications in Tables (4, 5 and 6), but Node3 does not contain data (D2); this data is available in Node1 and Node2, Node1 policy is to allow movement of this data as well as Node3's data requirements, while Node2 is not. The resource broker will therefore send a message to Node3 telling it to take data (D2) along with its policy from Node1 and execute Job3.
- **Data Migration (case 2)**
  In Tables (1, 2, 3), Job4's requirements fit Node4's specifications in Tables (4, 5 and 6), but Node4 does not contain data (D2); which is available in Node2 and Node3(after migration). Node2 policy is not to allow movement of data to China domain, but the policy in Node3 allows this kind of movements, but the data in Node3 was moved originally from Node1 which its policy does not allow to move data to China domain. Therefore, the resource broker will send a message to User4 which says that the grid is unable to execute Job4, because the needed data is unavailable.

TABLE 2
Grid Nodes Application/Data Specifications.

Node Name	Application				Data	
			Requirement			
	File	Version	CPU Speed	Disk Space	File Name	Size
Node 1	S1	9.2	0.5	200	D2	1000
	S2	1.0	0.5	700		
Node 2	S5	2	1.0	300	D1/D2	900
Node 3	S4	1.0	1.0	500	D3	700
Node 4	S3	1.0	1.0	900	-	D1
	S4	5.0	1.0	250		
Node 5	S3	1.0	1.0	500	D1	700

TABLE 4
Node Specification Requirements.

User Name	Job Name	Node Specification			
		CPU		Memory	
		Speed	Count	RAM	S/D
U1	Job 1	1	1	1024	-
U2	Job 2	1	1	2048	-
U3	Job 3	2	2	2048	-
U4	Job 4	1	1	2048	-
U5	Job 5	2	2	2048	-

TABLE 1
Grid Nodes Hardware Specifications.

Node Name	Hardware				Domain
	CPU		Memory		
	Speed	Count	RAM	Shared or Disturbuted	
Node 1	1	1	1024	-	USA
Node 2	1	1	2048	-	Japan
Node 3	2	2	2048	D	UK
Node 4	1	1	2048	-	China
Node 5	1	1	2048	S	UK

TABLE 3
Grid Nodes Policy Specifications and Running Jobs

Node Name	Policy					Jobs Running
	Exclusive Execution	Move Data	Move Application	Move Job	Restricted (Domain/ User/ Job)	
N1	Yes	Yes	Yes	Yes	China/U4	Job2
N2	No	No	Yes	Yes	China	-
N3	Yes	Yes	No	Yes	Non	-
N4	Yes	Yes	Yes	Yes	Non	-
N5	Yes	Yes	Yes	Yes	Non	-

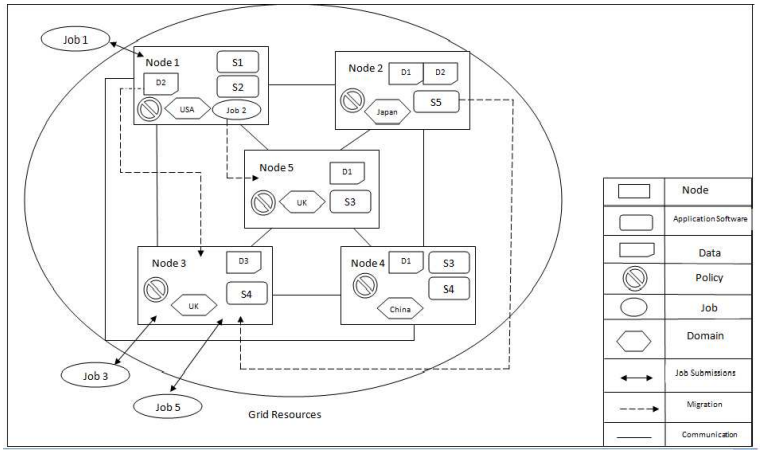


Fig. 5. Grid Resources after Mobility

TABLE 5
Application Software Requirements.

User Name	Job Name	Application Software		Data
		Name	Version	
U1	Job 1	S2	1	-
U2	Job 2	S3	1	D1
U3	Job 3	S4	1	D2
U4	Job 4	S4	5	D2
U5	Job 5	S5	1.1	D3

TABLE 6
Job Domain/Policy.

User Name	Job Name	Policy			Domain
		Exclusive Execution	Move Job	Restricted (Domain/ User)	
U1	Job 1	No	Yes	Non	China
U2	Job 2	No	Yes	China	UK
U3	Job 3	No	No	Non	USA
U4	Job 4	No	No	Non	USA
U5	Job 5	No	No	Non	UK

- **Application Software Migration**
  In the previous Tables, it can be seen that Job5's requirements fit Node3's specifications. But Node3 does not have application software (S5). Node2 does, however, and its policy is to allow this application software as well as node3's application software requirements. The resource broker will therefore send Job5 with a message to Node3 telling it to take application software (S5) from Node2 and execute Job5.

5 SIMULATION

In our simulation design we build a heterogeneous grid environment which has an unlimited number of resources in a fully connected topology. These nodes have the ability to migrate data, application software and jobs between them. The migration depends on the grid's policy, resources' policies, and grid users' policies. We have developed a Java User Interface that can simplify our work by creating a grid environment, configuring its nodes by each with its own application software, data, policies, hardware specifications and node names and finally sending jobs to the grid system.
The grid system has been simulated by using Jade simulator, which is a software framework fully implemented in Java language and allows agents to execute tasks defined according to the agent policy. When running the simulation, the main portal interface turns up. It composes all the functions needed to configure a new grid with all of its elements as shown in Figure(6).
In this interface, the grid administrator will be able to create a new grid environment by choosing grid name, configuring grid nodes and sending jobs to the grid system. The interface will then directly pass all this information to the Jade simulator to create them.

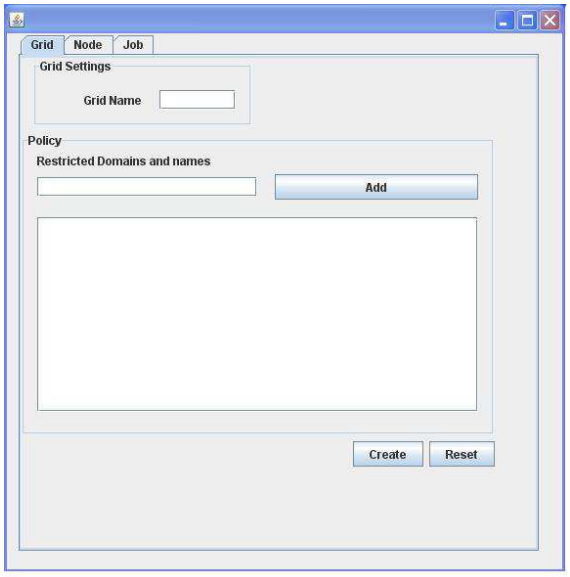


Fig. 6. Main Simulation Interface

5.1 Grid Configuration

Using the Interface in Figure(6) the grid administrator can create a new grid environment by choosing the grid name and any domains and/or users who are not allow to work under this grid. This can be done by entering the names of these domains and/or users in the Restricted Domain and names filed. This filed is going to be under the grid policy section which will be translated into XML file once the administrator clicks on the create botton, and to be sent later to Jade simulator to create the grid environment under this policy.

5.2 Node Configuration

Our interface can simulate the nodes by configuring their specifications. This is can be done by specifying their names, grid names, domain name, number of jobs that can be processed at the same time, hardware specifications, application software, data and policies. As shown in Figure(7).

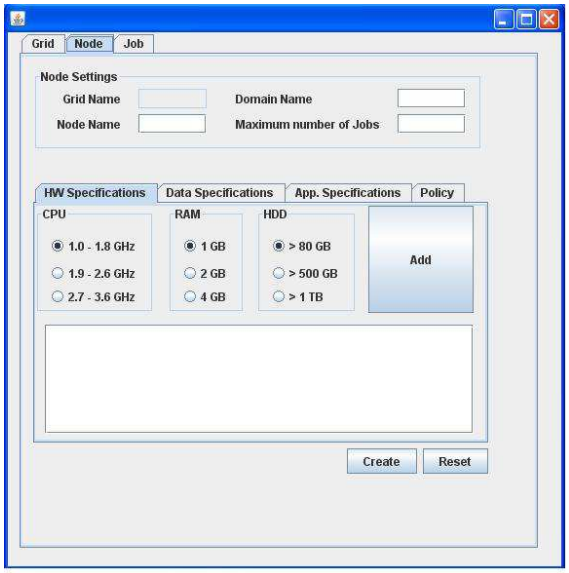


Fig. 7. Node Configuration Interface

After determining the node's name, the administrator can choose the domain name for that node. This domain name helps in sorting out the nodes into groups, which in turn will help in making the policy decision later. The administrator also has the ability to determine how many jobs each node can handle at the same time. The other fields are described as follow:

5.2.1 Node Hardware Specifications

In this step the administrator can determine the node's hardware specifications. These specifications include CPU speed, memory size and hard disk space. By choosing these specifications the system will represent the choices by an agent which can be understood by the Jade simulator and in the same time store them into the hardware section in an XML file which represents the overall node's specification. As shown in Figure(7).

5.2.2 Node Data Specifications

The interface in this part helps the administrator to configure the data in the node by determining data name and the policy for this single data. If the administrator chooses a Moving feature when creating a new data, the policy for this single data will allow this data to be moved wherever it is allowed to be moved. Otherwise it will prevent it from moving from its original node. The Copy feature has the same job but it will perform a copy action for the data instead of moving.
After finishing creating node's data, the system will represent each single data by an agent which can be understood by the Jade simulator and at the same time store the policy for this single data into the data section in the XML file which represents the overall node's specification.

5.2.3 Node Application Software Specifications

The interface in this section helps the administrator to configure the application software's in the node by determining application name and the policy for this single application. If the administrator chooses a Moving feature when creating new application software, the policy for this single application will allow this application to be moved wherever it is allowed to be moved. Otherwise it will prevent it from moving from its original node. The Copy feature has the same job but it will perform a copy action for the application instead of moving.
After creating node's application software, the system will represent each single application by an agent which can be understood by the Jade simulator and at the same time store the policy for this single application into the application software section in the XML file which represents the overall node's specification.

5.2.4 Node Policy Specifications

The interface related to the policy specifications helps the administrator to configure the policy in the node by determining any restricted domain(s) or user(s) whom they are not allowed to work under this node. Also in this section the administrator can determine whether this node is allowed to execute two jobs (or more) at the same time or not. This feature can be applied using the Exclusive choice in the policy. By choosing this option, the node is not allowed to execute more than one job at the same time.
After creating the node's policy, the system will store the policy for this node into the policy section in the XML file which represents the overall node's specification.

5.3 Job Configuration

After configuring the grid with its components by the administrator, this environment will be ready to receive jobs which have been submitted by the authorized users via the grid interface. Figure(8) shows this interface which will help the users to describe their jobs requirements in a simple way. These requirements will then be converted to a language that can be understood by the Jade simulator in the system, and at the same time it will be stored in an XML file that describes the jobs with its policies.
Our interface can simulate the jobs by configuring their

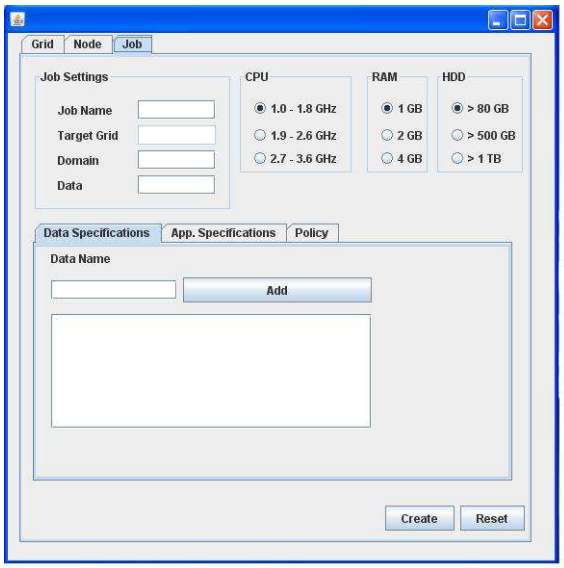


Fig. 8. Job Configuration Interface

requirements. This is can be done by specifying their names, grid names, domain name, any data attached with the job, hardware specifications, application software, data and policies. As shown in Figure(8).
After determining the job's name, the administrator can choose the domain name for that job. This domain name helps in sorting out the jobs into groups, which will help in sending the jobs to the appropriate domain later. The administrator also has the ability to attach a specific data along with the job to be processed during the execution time to fulfil the job requirements. The other fields are described as follow:

5.3.1 Job Hardware Specifications

In this step the administrator can determine the job's hardware specifications. These specifications include CPU speed, memory size and hard disk space. By choosing these specifications the Interface will represent the choosen ones by an agent which can be understood by the Jade simulator and at the same time store them into the hardware section in an XML file which represents the overall job's specification. As shown in Figure(8).

5.3.2 Job Data Specifications

By using Data interface section the administrator can configure the data needed to process the job by the grid nodes. By determining this name the system will add this data to the job requirements which will be sent later to the Jade simulator to find the suitable node(s) that owns this data. At the same time the system will store the name of this data in XML file that describes the job requirements.

5.3.3 Job Application Software Specifications

By using Application Software interface section the administrator can configure the application software(s) needed to process the job by the grid nodes. By determining this name(s) the system will add it to the job requirements which will be

sent later to the Jade simulator to find the suitable node(s) that owns this application(s). At the same time the system will store the name of this application(s) in XML file that describes the job requirements.

5.3.4 Job Policy Specifications

This interface helps the administrator to configure the job's policy before submitting it to the grid environment. The administrator or the grid users can determine any restricted domain(s) or user(s) who are not allowed to handle their jobs. Also in this section the administrator can determine whether this job is allowed to execute with other jobs at the same time or not. This feature can be applied using the Exclusive choice in the policy. By choosing this option, the job is not allowed to execute with other jobs. The Moving feature allows the job to be moved wherever it is allowed to be moved. Otherwise it will prevent it from moving from one node to another. By choosing this feature the system will store the mobility feature in the policy section in the XML file that presents the job's specifications.
After creating the job's policy, the system will store the policy for this job into the policy section in the XML file which represents the overall job's specification.

6 VALIDATION

We applied various grid environments in our simulation with numerous nodes and jobs. Each one of them with different hardware specifications, data, application software and polices. Our aim is to simulate and analyze the effect of the policy on the resource mobility (jobs, data and application software) operations in the grid environment.
Our program allows job, data and application software to migrate from one node to another in the grid environment. The aim of the simulation is to present the effect of the policies on the number of rejected jobs and number of nodes used in the grid during these migrations. To accomplish this aim, we constructed a grid environment that contains 20 nodes; each node has distinctive (or similar) hardware, application software and data specifications from others. Afterwoards we sent 30 jobs sequentially to this environment. Then we applied the job and resource mobility within the grid according to the following scenarios and configurations:

- Case 1: No mobility. In this stage we configured the policies for all of the jobs, data and application software not to be allowed to migrate within the grid along with preventing grid's nodes to accept migration resources between them. We then sent 30 jobs sequentially to the grid with distinctive (or similar) hardware, application software and data needed to accomplish these jobs. Figures (9 and 10) show the effect of polices on number of rejected jobs and number of the overall nodes used in the grid in the case of no mobility.
- Case 2: Partial Mobility (%25). In this stage we configured quarter of the policies for the jobs, nodes, data and application software to be allowed to migrate within the grid. Also we configured quarter of the grid's nodes polices to accept resource migration between them. We then sent 30 jobs sequentially to the grid with distinctive (or similar) hardware, application software and data needed to accomplish these jobs. Figures (9 and 10) show the effect of polices on number of rejected jobs and number of the overall nodes used in the grid in the case of Partial Mobility (%25).
- Case 3: Partial Mobility (%50). In this stage we configured half of the policies for the jobs, nodes, data and application software to be allowed to migrate within the grid. Also we configured half of the grid's nodes polices to accept resource migration between them. We then sent 30 jobs sequentially to the grid with distinctive (or similar) hardware, application software and data needed to accomplish these jobs. Figures (9 and 10) show the effect of polices on number of rejected jobs and number of the overall nodes used in the grid in the case of Partial Mobility (%50).

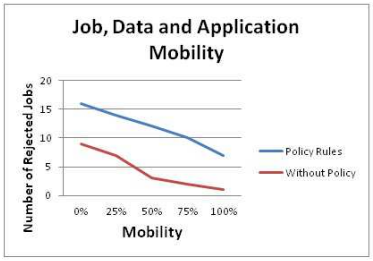


Fig. 9. Rejected Jobs with Job, Data and Application Software Mobility

- Case 4: Partial Mobility (%75). In this stage we configured (%75) of the policies for the jobs, nodes, data and application software to be allowed to migrate within the grid. Also we configured (%75) of the grid's nodes polices to accept resource migration between them. We then sent 30 jobs sequentially to the grid with distinctive (or similar) hardware, application software and data needed to accomplish these jobs. Figures (9 and 10) show the effect of polices on number of rejected jobs and number of the overall nodes used in the grid in the case of Partial Mobility (%75).
- Case 5: Full Mobility. In this stage we configured all of the policies for the jobs, nodes, data and application software to be allowed to migrate within the grid. Also we configured all of the grid's nodes polices to accept resource migration between them. We then sent 30 jobs sequentially to the grid with distinctive (or similar) hardware, application software and data needed to accomplish these jobs. Figures (9 and 10) show the effect of polices on number of rejected jobs and number of the overall nodes used in the grid in the case of full Mobility.

6.1 Rejected jobs

Once the grid is not able to accept a job due to the short of the job requirements (hardware, data and application software) amount of rejected jobs will rise. The mobility has solved this problem by migrating data or application software needed by

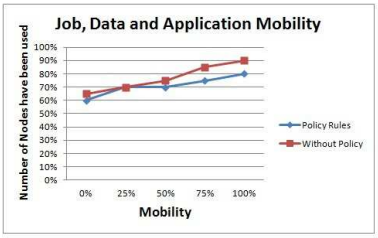


Fig. 10. The overall Used Nodes with Job,Data and Application Software Mobility

the new jobs or even evacuated the required node to fit the new jobs if necessary. Our results show the effect of the mobility on number of rejected jobs when it's applied. This means that the grid can fit a node to execute the job, consequently; the number of rejected jobs in the grid will be reduced. Nevertheless; that reducing is affected by the grid policy, node policy and job policies. It can be seen clearly that when applying these policies, the number of rejected jobs is less than the situation when the policies are not applied. In other words, the number of rejected jobs when applying the policies is less than without polices, but applying the policies over mobility gives the grid, grid's nodes and grid's user's the ability and the privacy to have control over their data, application software's and jobs.

6.2 The Overall Used Nodes

In the normal case, not all the grid's nodes own data or application software needed by all the users' jobs. In this case it can be seen clearly that some nodes are not utilized due to the short in these resources, and most of the jobs are sent to a specific nodes because of the fact that these nodes own the needed resources required by most of the jobs rather than the other nodes. In this case the gird finds itself in a situation not to accept a job, at some point, due to the short of the job requirements (hardware, data and application software) although it owns un-utilized nodes in its system. The mobility has solved this problem by migrating (or copying) data or application software needed by the new jobs, or even evacuated the required node to fit the new jobs if necessary. In this case the mobility helps in distributing user's jobs to most of the grid's nodes and that's will help in load balancing and reducing number of rejected jobs. Our results (Figure(10)) show the effect of the mobility on number of nodes used by the grid to fulfill the grid user's jobs. As a result, when applying the mobility solution, more nodes had been used than the situation without mobility.

7 CONCLUSIONS AND FUTURE WORK

We have presented in this paper a new dynamic policy management framework for mobile grid services that has the capability to deal with policies of multiple virtual organizations. Mobility assists grid evolution, improves performance of operating applications by migrating data to the execution host and therefore reduces the communication consumption and solves the load balancing problems. The other advantage of this architecture is taking the policies of the external users of the grid into account when making policy decisions. We presented our simulation for the grid environment in the case of applying grid policy, nodes' policies and users policies over mobility and finally presented the evaluation and the results of our simulation. Based on our contributions, our future work is concerned with conducting more experiments on our simulation to see the effect of the policies over mobility on the load balancing in the grid and how we can extend our simulation on the other simulators in the future.

REFERENCES

[1] I. Foster and K. Kesselman. The grid: Blueprint for a future computing infrastructure. In *Morgan Kaufmann in Computer Architecture and Design*, 1999.

[2] Alex Galis, Bernhard Plattner, Jonathan M. Smith, Spyros G. Denazis, Eckhard Moeller, Hui Guo, Cornel Klein, Joan Serrat, Jan Laarhuis, George T. Karetsos, and Chris Todd. A flexible ip active networks architecture. In *Proceedings of the Second International Working Conference on Active Networks*, pages 1–15, London, UK, 2000. Springer-Verlag.

[3] Rajkumar Buyya, David Abramson, and Jonathan Giddy. A case for economy grid architecture for service-oriented grid computing. In *Proceedings of the 15th International Parallel & Distributed Processing Symposium*, IPDPS '01, pages 83–, Washington, DC, USA, 2001. IEEE Computer Society.

[4] Zsolt N. Németh and Vaidy Sunderam. A formal framework for defining grid systems. In *Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid*, CCGRID '02, pages 202–, Washington, DC, USA, 2002. IEEE Computer Society.

[5] D.G. Rosado, E. Fernandez-Medina, J. Lopez, and M. Piattini. Developing a secure mobile grid system through a uml extension. *Journal of Universal Computer Science*, 16(17):2333–2352, 2010.

[6] Tao Guan, Ed Zaluska, and David De Roure. A grid service infrastructure for mobile devices. In *Proceedings of the First International Conference on Semantics, Knowledge and Grid*, SKG '05, pages 42–, Washington, DC, USA, 2005. IEEE Computer Society.

[7] Hassan Jameel, Umar Kalim, Ali Sajjad, Sungyoung Lee, and Taewoong Jeon. Mobile-to-grid middleware: Bridging the gap between mobile and grid environments. In *EGC'05*, pages 932–941, 2005.

[8] J.H. Park. Usf-pas : Study on core security technologies for ubiquitous security framework. *Journal of Universal Computer Science*, 15(5):1065–1080, 2009.

[9] P. Nixon T. Walsh and S. Dobson. Review of mobility systems. In *TCD Computer Science Technical Report*, 2000.

[10] Philip W.L. Fong. Viewer's discretion: Host security in mobile code systems, 1998. School of Computing Science, Simon Fraser University.

[11] Luca Cardelli. Secure internet programming. chapter Abstractions for mobile computations, pages 51–94. Springer-Verlag, London, UK, 1999.

[12] G. Alliance. Globus toolkits.

[13] Luis Ferreira, Viktors Berstis, Jonathan Armstrong, Mike Kendzierski, Andreas Neukoetter, MasanobuTakagi, Richard Bing, Adeeb Amir, Ryo Murakawa, Olegario Hernandez, James Magowan, and Norbert Bieberstein. *Introduction to grid computing with globus*. IBM Corp., Riverton, NJ, USA, first edition, 2003.

[14] Rampure. Vishal Wu. Jin, Leangsuksun. Chokchai Box and Ong. Hong. Policy-based access control framework for grid computing. In *Proceedings of the Sixth IEEE International Symposium on Cluster Computing and the Grid*, CCGRID '06, pages 391–394, Washington, DC, USA, 2006. IEEE Computer Society.

[15] Coulouris, Jean Dollimore, and Tim Kindberg. *Distributed Systems: Concepts and Design (4th Edition) (International Computer Science)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[16] Tariq Alwadan, Helge Janicke, Omer Aldabbas, and Mai Alfawair. New framework for policy support for mobile grid services. In *To appear in proceedings of the 6th International Conference on Risks and Security of Internet and Systems (CRISIS2011)*, 2011.

[17] Tariq Alwadan, Helge Janicke, Omer Aldabbas, and Hamza Aldabbas. New framework for dynamic policy management in grid environments. In *Recent Trends in Wireless and Mobile Networks, Third International Conferences, WiMo 2011 and CoNeCo 2011,*, volume 162, pages 297–304. Springer, 2011.

Phishing Detection Taxonomy for Mobile Device

Cik Feresa Mohd Foozy[1], Rabiah Ahmad[2] and Mohd Faizal Abdollah[3]

[1,2,3] Center for Advanced Computing Technology, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), Karung Berkunci No. 1752 Pejabat Pos Durian Tunggal, 76109 Melaka Malaysia

Abstract

Phishing is one of the social engineering attacks and currently hit on mobile devices. Based on security report by Lookout [1], 30% of Lookout users clicking on an unsafe link per year by using mobile device. Few phishing detection techniques have been applied on mobile device. However, review on phishing detection technique on the detection technique redundant is still need. This paper addresses the current trend phishing detection for mobile device and identifies significant criterion to improve phishing detection techniques on mobile device. Thus, existing research on phishing detection technique for computer and mobile device will be compared and analysed. Hence, outcome of the analysis becomes a guideline in proposing generic phishing detection taxonomy for mobile device.

Keywords: Mobile, Phishing, Security, Social Engineering, Taxonomy

1. Introduction

A mobile device is defined as a very small, lightweight device that provides functionality like a laptop computer [2]. Examples of mobile devices are Palm and other PDAs, tablet PC and smart mobile[3]. Mobile devices have become so popular for business and personal use because of their features such as portability and long battery life.

However, the rapid growth of mobile devices has contributed to security problem due to its functionality connect to Internet. According to a website *mysecurecyberspace.com*[4], one example of security problem in mobile device is the security threat such as mobile banking password, contact number, photo and others. Moreover, Symantex [5] also has reported the increasing intrusion and identity theft on mobile phones.

The purpose of phishing attack is to steal valuable information such as credit card and social security numbers, user IDs and passwords [6]. According to Boodae [7], mobile device users are three times more likely to enter a web-based phishing attack than desktop users. Since the reported shows the increasing phishing attack on mobile device, many studies has been done to detect the phishing attack. Thus, to get clear view what is criterion for modern mobile device phishing detection, taxonomy for mobile device will be proposed.

This paper proposes phishing detection taxonomy on mobile device and structured into four sections as follows: In Section 2, describes the related work in phishing. In Section 3 also discusses, the methodology used to identify the categories of phishing attacks and detection techniques in mobile devices. Moreover, section 4 conducts the analysis on detection technique. Finally, in Section 5, the proposed taxonomy and future work for the next research are presented.

2. Related Work

Current studies in phishing are much focused on specific phishing detection technique for mobile devices and most of the studies are more on verifying the validity of the computer website. Therefore, this section will review the phishing attack and detection for desktop and mobile device.

2.1 Taxonomy

Taxonomy can be defined as a simple classification to categories into the specific groups [8]. There are few techniques to mitigate phishing attack such as filtering on browser and toolbar, anti-virus, anti-phishing and through education and training.

2.2 Phishing

Phishing term has been introduced in early 1990's by America Online (AOL) because of the stolen data happened on that time. Since, financial lost and stolen data can be happened in mobile device, few phishing detection techniques on mobile device that have been proposed are filtering on browser and toolbar, anti-virus, anti-phishing and via education and training. There are some advantages and disadvantages of each techniques mention above. For example, one of the limitations of anti-phishing is the signatures phishing need to be

updated frequently. However, the advantages, it is widely used in industries and easy to be updated.

Maggi et al. [9] has done a research on phishing of voice channel and found the phishing attack can be categories into traditional and modern type. Email method is identified as traditional way to attack and for modern phishing attack, instant message, social network and phone system phishing.

Crain et al.[10] classified the phishing defense method into technical and educational method. Example of technical defense method are browser toolbar, email verification, anti-phishing. There are several limitations of defense method on toolbar and anti-phishing such as frequent update the phishing signature, unsecure connection between server client and user can switch off the anti-phishing on client [11]. Moreover, for training and educational methods it requires times for human to adapt with new process.

In addition, Kumaruguru et. al [12] listed several defense method through education and training but it becomes a problem when the worker leaves the organization. Moreover, the company must frequently train their for phishing awareness. However, a study by Alnajim and Munro [13] categories the defense mechanism into technical and training techniques. The defense mechanism consists of anti-phishing for email, web, IQ test, class assessment and tool bar.

There are few detection techniques that have been proposed to overcome the phishing issues in education and technical part. However, this paper review detection technique and focus on technical based solution for phishing attack on mobile device.

2.3 Phishing Detection

Phishing launch the attack through browser and email[14]. Additionally, for mobile device, phishing can attack via bluetooth, SMS, Voice Over IP, mobile application and mobile browser.

One of phishing detection technique on mobile device are using content-based filtering[15]. Moreover, there are more detection techniques that will be discuss in this section later.

Phishing detection technique is a research area that can help to reduce the effect of phishing attack on mobile device. Commonly, phishing attack will attack website, client application, visual and images. G. Xiang, et al. [16] listed two phishing detection techniques such as blacklist and feature-based. Moreover, J. a. Huh and H. Kim [17] listed three types of detection techniques such as blacklist, whitelist and heuristic. In addition, Zhang et al.[18], listed blacklist and heuristic as common phishing detection. Moreover, Chhabra [19] listed three email detection techniques such as blacklist, whitelist and graylist. Table 1 shows the summary desktop and wired phishing detection techniques that has been discusses in the literature.

However, phishing solutions for desktop and wired computers are not suitable for wireless and mobile devices due to processing, power and storage limitations[20],[21]. Thus, phishing detection must be lightweight and high accuracy in detecting phishing attack on mobile device.

2.4 Common Phishing Detection Technique on Mobile Device

The possibilities of lost and risks for mobile device are increasing when the device is connected to the network[22]. This shows the phishing detection for mobile device is still a significant research area to be improved since phishing attack has revolutionized the strategies into mobile device.

In addition, mobile operating systems and browsers not have secure application[23]. Losing money and stealing data such as password, contact number, account number and etc. can be occurred if mobile application and website are interacting with each other.

Approximately 1 in 20 users will click on a phishing link every year on Android devices[1], since phishing detection for mobile device is different from wired computers, developing taxonomy for phishing attack and detection techniques is needed in order to propose an overview for suitable phishing detection technique for mobile device.

Dunham [21], identify the phishing attack on mobile devices into Bluetooth phishing, Short Message Service (SMS) phishing and Voice over IP Phishing or known as vishing.

Example Bluetooth phishing attack has been discuss by [21], the Bluetooth phishing attack works when user connect to the Wi-Fi hotspot. Attacker can steal the data when the user connects to the Wi-Fi.

Figure 1 shows the example of SMiShing attack in Malay language and this attack can be used as a strategy to trick mobile phone user to transfer money to their bank account.

Fig. 1 Example SMiShing attack

Based on the findings, common phishing detections for mobile device are SMS phishing detection, voice call phishing detection and mobile web browser phishing detection. Thus, below are common detection techniques for mobile device that has been review:

i. Content Based Filtering:

This technique has shown a successfully detection phishing attack on email. J. W. Yoon, et al.[15] implement this technique with challenge-response scheme. The combinations of these techniques are needed to improve the traditional spam filtering detection technique on mobile device since the content-based filtering alone is less efficient. Moreover, content-Based filtering can be divided into rule based and statistic based [24].

ii. Blacklist:

Blacklist is a method that need human to verification. Since this technique have very low False Positive(FP), it is widely applied in the industries as anti-phishing in toolbar. If user enter the blacklist website, a warning will be appeared. However, this method is not suitable to detect new phishing attack[16]. This technique is also not efficient in update and verify the phishing attack database globally [17],[25]. In addition this technique have less capablities to protect users [26]. Moreover, this technique also has been implemented in fraud telephony(vishing)[27] and SMS filtering by [28].

iii. Whitelist:

Whitelisting method is different from blacklist-based, this technique need to maintain all website in the cyber world. The limitation of this technique is impossible to cover all website [25]. This technique has been implemented by [28] to detect SMS phishing.

3. Methodology

The aim of this paper is to classify phishing attack and identify the defense technique on mobile device. The data in this study were retrieved from various databases such as ACM Digital Library, SpringerLink, IEEE Xplore, ScienceDirect, Google, Google Scholar, and Yahoo.

Using these databases, a statistical analysis on every selected article about phishing detection and filtering was done to propose taxonomy of phishing detection on mobile device.

Step 1: Analysis phishing attack on mobile device category

Step 2: Analysis of phishing attack classification and detection techniques.

The purpose of the first step is to identify the phishing attack on mobile device. This paper discusses the preliminary analysis on phishing attack by identifying the common attack on mobile device and finally to propose taxonomy of phishing detection attack on mobile device. Figure 2 shows the overview of the analysis process.

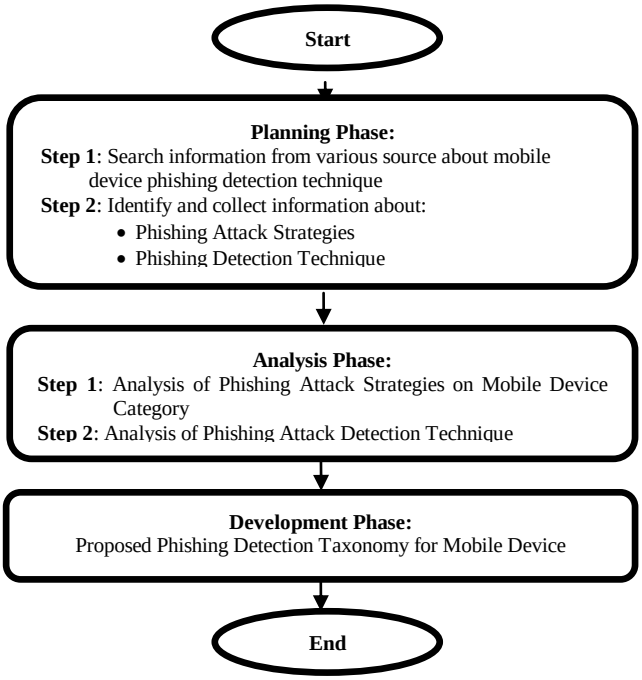


Fig. 2 Overview the analysis process

4. Analysis and Findings

According to the [29], phishing is a type of technical-based social engineering. This attack is dividing into traditional and modern attack where email is traditional attack and phone system phishing such as mobile device is a type of modern phishing attack. There are three types of modern phishing strategies on mobile device such as SMS, Voice Call and Bluetooth.

4.1 Taxonomy Elements for Mobile Device Phishing

The development phishing taxonomy is to provide basic understanding on phishing attack and detection concept on mobile device. Main elements of proposed taxonomy are consisting of the Attack Strategies and Phishing Detection Techniques.

In the proposed taxonomy, all these elements will be applied to build phishing detection taxonomy for mobile device. This can be as alternative to understand the components to build a framework of phishing detection for mobile device. Figure 3, shows the main elements in intrusion detection taxonomy.

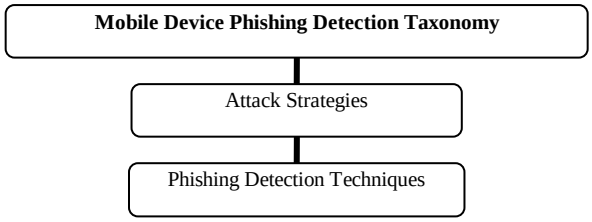


Fig. 3 Main elements of intrusion detection taxonomy

4.2 Mobile Device Phishing Attack Strategies

Table I listed few study by [21], [30] and [23] on mobile device phishing attack strategies. All listed phishing type is relevant to be included in the taxonomy since it has been discussed by the researchers.

Table 1. Analysis of phishing attack on mobile device (item found=√)

References / Phishing Attack Strategies	[21]	[30]	[23]
Bluetooth Phishing	√		
SMS Phishing	√	√	
Vishing	√		
Mobile Web Application Phishing			√

4.3 Mobile Device Phishing Detection Techniques

In addition, Table II listed the phishing detection technique that has been studies by the researchers below. This shows, the content based filtering and blacklist is widely used to detect phishing on desktop and it has been applied to detect phishing attack on mobile device.

Table 2. Analysis of phishing attack on mobile device (item found=√)

References / Phishing Detection Techniques	[15]	[24]	[27]	[28]	[31]	[32]
Content Based	√	√				
Blacklist			√	√		
Whitelist				√		
Hotspot					√	
Gaussian Mixture Model						√

4.4 Mobile Device Phishing Detection Techniques Based on Attack Strategies

In addition, Table III listed the relation between Attack Strategies and Phishing detection Techniques. These explain that each attack strategies have different phishing detection techniques. Blacklist detection techniques have more than two occurrences which applied at SMS, Vishing and Mobile Application. For whitelist, it also have more than one occurrence and been applied at SMS and web application.

Table 3. Analysis on mobile device phishing detection techniques based on attack strategies (item found=√)

Attack Strategies / Phishing Detection Techniques	Bluetooth	SMS	Vishing	Mobile Web/ Application
Content Based		√		
Blacklist		√	√	√
Whitelist		√		√
Hotspot Wireless Defense Tool	√			
Gaussian Mixture Model			√	

5. Result

As a result from the analysis section, the elements that will be includes in our taxonomy are the Attack Strategies and the Phishing Detection Techniques. Since, phishing attacks are widely discussed in many areas, it is less consideration for mobile device phishing detection technique.

This section discussed the outcome of the analysis which is about Mobile Device Phishing Attack Strategies Taxonomy, Mobile Device Phishing Detection Techniques and Mobile Device Phishing Detection Techniques Taxonomy.

5.1 Mobile Device Phishing Attack Strategies Taxonomy

From the Table I, [21], [30] and [23] has listed several attack strategies on mobile device and taxonomy are illustrate as Figure 4.

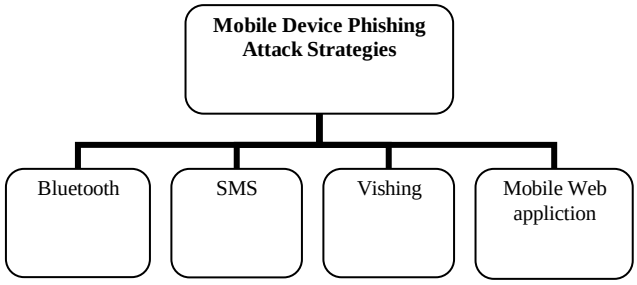


Fig. 4 Mobile Device Phishing Attack Strategies Taxonomy

5.2 Mobile Device Phishing Detection Techniques

Phishing detection techniques on mobile device also have been compared in order to develop phishing taxonomy and the advantages and disadvantages of each method are summarizes in the Table IV.

There are six detection techniques that has been review such as Content Based, Blacklist, Whitelist, Hotspot, Gaussian Mixture Model and Graylist. According to the table, Content based detection technique has been review by [15] and [24].

For blacklist detection attack, this techniques has been applied by [27] and [28]. Whitelist is also a detection techniques that has been aplied by few researchers to detect phishing attack on SMS and Mobile Web. This method has been discussed by [28]. This techniques need to collct the data of trusted senders and only can detect phishing from the known sanders.

Moreover, bluetooth phishing attack can be defense by using wireless hotspot defense tools[32]. The advantages of this tools it can detect wireless attack and it can check any changes on ESSID, MAC address of the access point, MAC address of the default gateway and radical signal strength fluctuations on the network. However, this tools need and expert to monitor and understanding the changes happened on the network.

This technique has been applied by [31], the result shows this method is effective to detect vishing attack on mobile device and can identifies lies and true statements.

Table 4. Summary of phishing attack detection techniques for mobile device

Phishing Attack Strategies	Technique	Advantage	Disadvantage
• **SMiShing**	Content Based Filtering	• Flexible	• Less efficient
• **SMiShing** • **Vishing** • **Mobile Web**	Blacklist	• Low False Positive • Effective detection known phishing URL	• Not suitable to detect new attack • Less efficiency in updating and verify the attack in database • Inefficient to protect user from phishing attack
• **SMiShing** • **Mobile Web**	Whitelist	• Have list of trusted senders	• Detect phishing from known sender
• **Blueetooth**	Wireless Hotspot Defense Tools	• Check any changes on MAC address and etc.	• Need expert to review and monitor the network.
• **Vishing**	Gaussian Mixture Model	• Identifies lies and true statements	• To assign pattern into lies and true voice pattern

5.3 Mobile Device Phishing Detection Techniques Taxonomy

The mobile device phishing detection taxonomy has been developed as Figure 5. The taxonomy shows the general view on phishing attack strategies and phishing attack detection techniques for mobile device.

As a social engineering based attack, phishing need to be countermeasure and various defense solutions has been introduced to detect phishing but not many alternatives solutions to detect phishing attack on mobile devices. Since, mobile device is one of modern technology that essential today, this attack has evolving their strategies into mobile device.

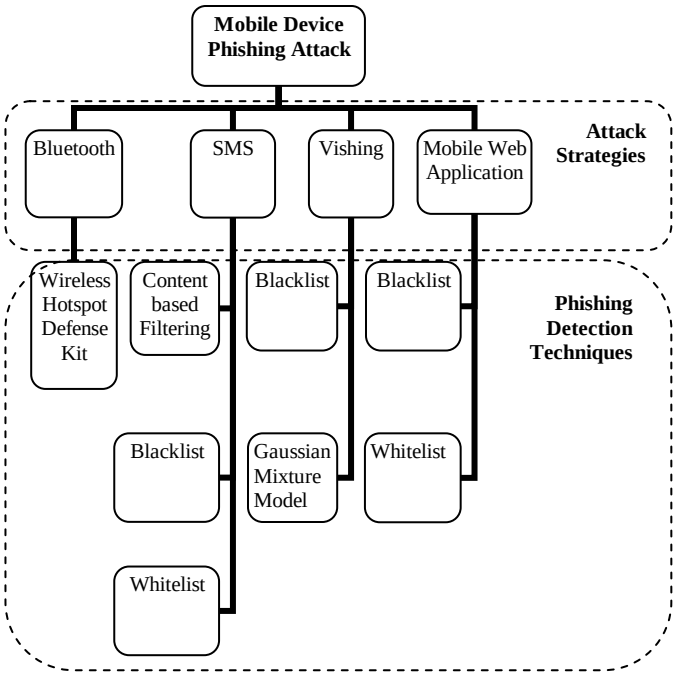


Fig. 5 Mobile device phishing detection attack taxonomy

6. Conclusion

According to[32], there is no anti-phishing solution dedicated to mobile device. Thus, a review on mobile device phishing detection technique will help to develop phishing detection taxonomy because taxonomy can help user to have understanding about the specific topic.

This paper discusses mobile device phishing attack and the develop taxonomy has classifies the mobile device phishing attack strategies into several phishing attack. This paper is a preliminary study for future work and it contributes ideas on how to identify mobile device phishing attack.

Acknowledgments

The authors would like to thank University Tun Hussein Onn Malaysia(UTHM) and Ministry of Higher Education Malaysia for supporting this research.

References

[1] I. Lookout, "Lookout Mobile Threat Report August 2011," 2011.

[2] H. Wen-Chen, *et al.*, "Mobile Data Protection Using Handheld Usage Context Matching," in *Mobile Data Management: Systems, Services and Middleware, 2009. MDM '09. Tenth International Conference on*, 2009, pp. 594-599.

[3] J. Stonemetz, *et al.*, "Handheld Devices Anesthesia Informatics," ed: Springer New York, 2009, pp. 409-424.

[4] MySecureCyberspace.com. (2011, The Trend of Tablet PCs. Available: http://www.mysecurecyberspace.com/articles/features/the-trend-of-tablet-pcs.html

[5] S. Corporation, "Symantec Intelligence Report: July 2011," 2011.

[6] Microsoft. (2011, 8th June). *Email and web scams: How to help protect yourself*. Available: http://www.microsoft.com/security/online-privacy/phishing-scams.aspx

[7] M. Boodae, "Mobile Users Three Times More Vulnerable to Phishing Attacks," in *Trusteer* vol. 2012, ed, 2011.

[8] P. Rich, "The Organizational Taxonomy: Definition and Design," *The Academy of Management Review*, vol. 17, pp. 758-781, 1992.

[9] F. Maggi, *et al.*, "A social-engineering-centric data collection initiative to study phishing," presented at the Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, Salzburg, Austria, 2011.

[10] J. Crain, *et al.*, "Fighting Phishing with Trusted Email," in *Availability, Reliability, and Security, 2010. ARES '10 International Conference on*, 2010, pp. 462-467.

[11] S. Abu-Nimeh and S. Nair, "Bypassing Security Toolbars and Phishing Filters via DNS Poisoning," in *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE*, 2008, pp. 1-6.

[12] P. Kumaraguru, *et al.*, "Protecting people from phishing: the design and evaluation of an embedded training email system," presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, San Jose, California, USA, 2007.

[13] A. Alnajim and M. Munro, "An evaluation of users' tips effectiveness for Phishing websites detection," in *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*, 2008, pp. 63-68.

[14] P. Soni, *et al.*, "A phishing analysis of web based systems," presented at the Proceedings of the 2011 International Conference on Communication, Computing; Security, Rourkela, Odisha, India, 2011.

[15] J. W. Yoon, *et al.*, "Hybrid spam filtering for mobile communication," *Computers & Security*, vol. 29, pp. 446-459, 2010.

[16] G. Xiang, *et al.*, "CANTINA+: A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites," *ACM Trans. Inf. Syst. Secur.*, vol. 14, pp. 1-28, 2011.

[17] J. a. Huh and H. Kim, "Phishing Detection with Popular Search Engines: Simple and Effective Foundations and Practice of Security." vol. 6888, J. Garcia-Alfaro and P. Lafourcade, Eds., ed: Springer Berlin / Heidelberg, 2012, pp. 194-207.

[18] Y. Zhang, *et al.*, "Cantina: a content-based approach to detecting phishing web sites," presented at the Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada, 2007.

[19] S. Chhabra, "Fighting Spam, Phishing and Email Fraud," Master of Science in Computer Science, UNIVERSITY OF CALIFORNIA RIVERSIDE, 2005.

[20] S. Abu-Nimeh, *et al.*, "Distributed Phishing Detection by Applying Variable Selection Using Bayesian Additive Regression Trees," in *Communications, 2009. ICC '09. IEEE International Conference on*, 2009, pp. 1-5.

[21] K. Dunham, "Chapter 6 - Phishing, SMishing, and Vishing," in *Mobile Malware Attacks and Defense*, D. Ken, Ed., ed Boston: Syngress, 2009, pp. 125-196.

[22] J. Networks, "Malicious Mobile Threats Report 2010/2011," 2011.

[23] A.P. Felt and D. Wagner, "Phishing on Mobile Devices," 2011.

[24] H. Peizhou, *et al.*, "A Novel Method for Filtering Group Sending Short Message Spam," in *Convergence and Hybrid Information Technology, 2008. ICHIT '08. International Conference on*, 2008, pp. 60-65.

[25] Y. Cao, *et al.*, "Anti-phishing based on automated individual white-list," presented at the Proceedings of the 4th ACM workshop on Digital identity management, Alexandria, Virginia, USA, 2008.

[26] S. Sheng, Wardman, B., Warner, G., Cranor, L., Hong, J., & Zhang, C, "An empirical analysis of phishing blacklists," *6th Annual Conference on Email and Anti-Spam (CEAS), Mountain View, CA.*, 2009.

[27] Devinder Singh, *et al.*, "Telephony Fraud Prevention," US Patent, 2011.

[28] T. M. Mahmoud and A. M. Mahfouz, "SMS Spam Filtering Technique Based on Artificial Immune System," *IJCSI International Journal of Computer Science Issues*, vol. 9, 2012.

[29] R. A. Cik Feresa Mohd Foozy, Mohd Faizal Abdollah, Robiah Yusof and Mohd Zaki Mas'ud, "Generic Taxonomy of Social Engineering Attack," *Malaysian Technical Universities International Conference on Engineering & Technology*, 2011.

[30] O. Salem, *et al.*, "Awareness Program and AI based Tool to Reduce Risk of Phishing Attacks," in

Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on, 2010, pp. 1418-1423.

[31] J. H. Chang and K. H. Lee, "Voice phishing detection technique based on minimum classification error method incorporating codec parameters," *Signal Processing, IET*, vol. 4, pp. 502-509, 2010.

[32] Saeed Abu-Nimeh and S. Nair, "Phishing Attacks in a Mobile Environment."

Cik Feresa Mohd Foozy is currently working with Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia. Feresa holds a Master's degree in Computer Science(Information Security) from Universiti Teknologi Malaysia, Malaysia and a Bachelor's degree in Information Technology and Multimedia from Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia. She is currently pursuing her PhD at the Universiti Teknikal Malaysia Melaka, Malaysia.

Rabiah Ahmad is an Associate Professor at the Faculty of Information Technology and Communication, Universiti Teknikal Malaysia Melaka,Malaysia. She received her PhD in Information Studies (health informatics) from the University of Sheffield, UK, and M.Sc. (information security) from the Royal Holloway University of London, UK. Her research interests include healthcare system security and information security architecture. She has delivered papers at various health informatics and information security conferences at national as well as international levels. She has also published papers in accredited national/international journals. Besides that, she also serves as a reviewer for various conferences and journals.

Mohd Faizal Abdollah is a Senior Lecturer at the Faculty of Information Technology and Communication, Universiti Teknikal Malaysia Melaka,Malaysia. He received his PhD in Computer and Network Security from Universiti Teknikal Malaysia Melaka,Malaysia, and M.Sc. (Computer Science) from the University Kebangsaan Malaysia. His research interests include network and mobile security and network monitoring. He has delivered papers at various network security conferences at national as well as international levels. He has also published papers in accredited national/international journals. Besides that, he also serves as a reviewer for various conferences and journals

Security Aspects of Sensor Networks

[1]Mohd Muntjir, [2]Mohd Rahul, [3]Mohammad Asadullah
College of Computers and Information Technology
Taif University, Taif, Saudi Arabia

Abstract

Sensor networks are amassed wireless networks of small, low-cost sensors that collect and propagate environmental data. The emerging field of wireless sensor networks integrates sensing, computation, and communication into a single device. The power of wireless sensor networks verifies in the capability to deploy huge numbers of small nodes that collaborates and configure them. Wireless sensor networks simplify monitoring and handling of physical environments from remote locations with best accuracy. Security protocols associated to sensor network are analyzed in this paper.

Keywords: *Application areas, system evaluation metrics, sensor nodes, security protocols.*

1. Introduction

A sensor network is an integration of a large number of sensor nodes that are obtusely deployed either inside the anomaly or very close to it. Random deployment in inaccessible domain or disaster relief operations of sensors is done. Sensor nodes are assumable with an onboard processor. Instead of sending the raw data to the nodes incumbent for the fusion, they use their processing capacity to locally carry out simple computations and broadcast only the required and fractionally processed data [1].

Sensors associated into structures, machinery, and the environment, conjugated with the efficient delivery of sensed information, could provide extraordinary benefits to society. Potential benefits append: minor catastrophic failures, conservation of natural resources, elaborated manufacturing fertility, improved emergency response and enhanced homeland security. However, barriers to the outspread use of
sensors in structures and machines remain. Bunches of lead wires and fiber optic "tails" are subject to wreckage and connector failures. Long wire bundles personify a expressing installation and long term preservation cost, limiting the number of sensors that may be dispose and accordingly reducing the total quality of the data revealed. Wireless sensor networks can discard these costs, easing installation and dismissing connectors. The ideal wireless sensor is networked and supplying, consumes very little power, is smart and software programmable, capable of fast data possession, reliable and genuine over the long term, costs short to take and install, and requires no real conservation. Selecting the ideal sensors and wireless communications link requires knowledge of the application and obstacle definition. Battery life, sensor update rates, and size are all extensive design deliberation. Examples of low data rate sensors combine temperature, humidity, and maximize strain captured peacefully. Examples of high data rate sensors combine strain, acceleration, and vibration. The way of wireless sensor networks is based on a simple equation;

Sensing + CPU + Radio = Thousands of potential applications

As soon as the people distinguish the capabilities of a wireless sensor network, hundreds of applications buck to mind. It looks like a genuine combination of modern technology. A wireless sensor network (WSN) extensively consists of a base station (or "gateway") that can communicate with a number of wireless sensors via a radio link. Data is poised at the wireless sensor node, compressed and send to the gateway straightly or, if required, uses other wireless sensor nodes to forward data to the gateway. After this the transmitted data is then given to the system by the gateway connection. The total aim of this chapter is to given a brief technical introduction to wireless sensor networks and existent a few applications in which wireless sensor networks are enabling.

2. WIRELESS SENSOR NETWORK ARCHITECURE

There are lots of different topologies for radio communications networks. A concise discussion of the network topologies that apply to wireless sensor networks are defined below.

A. Star Network (Single Point-to-Multipoint):

The star topology is useful in WSN, mainly in the development of Wireless Body Area Networks (WBAN).In this topology, a central node has the responsibility of the allocation with the medical sensors and the communication outside the BAN. The advantage of this type of network for wireless sensor networks is in its clarity and the ability to keep the remote node‟s power desolation to a minimum. It also allows for low suspension communications between the remote node and the base station. The deprivation of such a network is that the base station must be within radio transmission range of all the individual nodes and is not as fit as other networks in view of its dependency on a single node to conduct the network. The star network topology is as usual used in Body Area Networks (BAN), also called Body Sensor Network (BSN), where sensors are allocated on the body of a specimen.

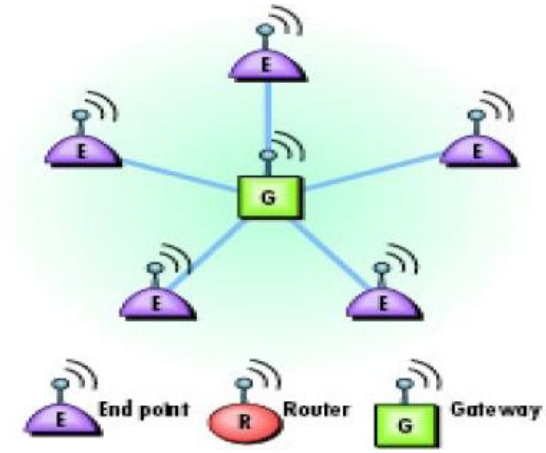


Fig. 1: Star Network

B. Mesh Network:

A mesh network allocates for any node in the network to broadcast to any other node in the network that is within its radio transmission domain. This network topology has the favor of redundancy and reliability. If a particular node fails, a remote node still can telecast to any other node in its range, can forward the message to the desired location. In this way, the range of the network is not limited by the range in between single nodes; it can simply be enlarged by adding more nodes to the system.

The disadvantage of this type of network is in power depletion for the nodes that implement the multichip communications are extensively higher than for the nodes that don‟t have this potential, generally limiting the battery life. Furthermore, as the number of communication mingle to destination increases; the time to redeem the message also increases, primarily if low power operation of the nodes is a requirement. The current enlargements in WSN are also centralizing on mesh network topology because it admits for the communication between devices without a central node for routing using a mesh of nodes. This feature discards the central failure, and contributes self-healing and self-organization.

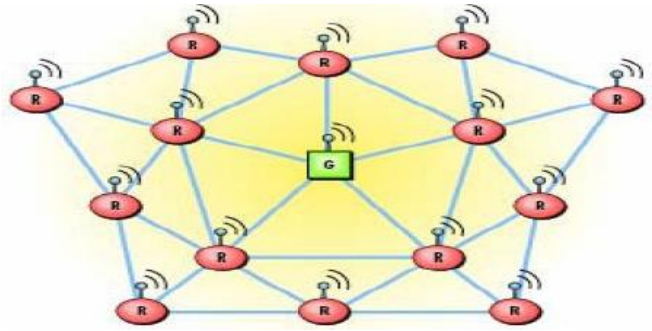


Fig. 2: Mesh Network

3. NETWORK APPLICATION AREAS

A. Applications Classification

It could be ordered into two categories: event detection (ED) and spatial process estimation (SPE). In event detection sensors are expanded to expose an event such as fire in a forest, a quake, etc. [2–3]. Signal processing within tools is very simple; each device has to compare the systematic quantity with a given inception and to send the binary information to the sink(s).

The density of nodes must assure that the event is detected and delivered to the sink(s) with a applicable probability of success while cultivating a low probability of wrong alarm. The detection of the phenomenon of interest (POI) could be finished in a decentralized (or dispensed) way.

In SPE the WSN intent at estimating a given physical phenomenon, that can be modeled as a bi-dimensional random process. It evaluates the all behavior of the spatial process based on the samples taken by sensors that are commonly placed in random positions [3–4].

4. SYSTEM EVALUATION METRICS

Meanwhile the key evaluation metrics for wireless sensor networks are full time, coverage, cost and ease of categorization, response time, physical accuracy, security, and effective sample rate.

A. Lifetime

Energy is the bounded factor for the lifetime of a sensor network. Each node must be designed to conduct its local supply of energy. Nodes can be evidently powered or self-powered.

B. Coverage

Multi-hop networking protocols elaborate the power consumption of the nodes, which may decline the network lifetime.

C. Cost and ease of deployment

Wireless sensor network must construct itself. All through the lifetime of a deployment, nodes may be dislocated or large physical objects may be arranged so that they interfere with the communication between two nodes.

In an actual deployment, a fragment of the total energy budget must be devoted to system maintenance and verification. The generation of characteristic and reconfiguration traffic deflates the network lifetime. This can also decreases the efficient sample rate.

D. Response Time

Although low power operation, nodes must be proficient of having immediate, high-priority messages communicated across the network as swiftly as possible response time must be as low as possible.

Network lifetime can be increased by having nodes only compelled their radios for limited periods of time but it reduces the responsiveness of system.

E. Temporal Accuracy

The network must be capable of constructing and maintaining a global time base that can be used to sequentially order fragments and events. In a distributed system, energy must be distributed to maintain this expanded clock.

The time synchronization information must be constantly communicated between nodes. The frequency of the synchronization messages is dependent on the aspired accuracy of the time clock.

F. Security

Encryption and cryptographic authentication are used for security but it charged both power and network bandwidth [8-9]. The extra computation must be achieved to encrypt and decrypt data and extra authentication bits must be transmitted with each and every packet.

G. Effective Sample Rate

Effective sample rate is denoted as the sample rate that sensor data can be taken at each and every sensor and communicated to an acquisition point in a data collection network.

In a data collection tree, a node must hold the data of all of its descendants. Network bit rates combined with maximum network size end up smashing the effective per node sample rate of the complete system [10].

Distinct forms of spatial and temporal compression can be used to reduce the communication bandwidth demanded while maintaining the same active sampling rate. Local storage can be used to collect and store data at a high sample rate for limited periods of time. The data can then be downloaded over the multi-hop network as bandwidth grants.

5. SENSOR NODES

Miniaturization, low power and low cost composed are likely the most exacting technical problem for sensor nodes.

A. Device Classes

Two forms of device classes are Commodity devices and Custom built nodes from commercially-available electronics segments.

1) Commodity Devices

Commercially available commodity devices are used to frame prototypical sensor network algorithms and functions. Commodity devices include laptop computers, PDAs, mobile phones, cameras.

Many commodity devices afford regulated wired and wireless interfaces and application protocols that allow using the device‟s serviceability without a extreme programming act.

2) COTS Sensor Nodes

CTOS abbreviates for the custom-built sensor nodes. COTS nodes are fabricated from several commercially off-the shelf (COTS) electronic components. COTS node expansion provides an adaptable, commonly applicable sensor node.

A classical setup consists of an RF transceiver and antenna, one or more sensors, as well as a battery and power regulating circuitry collected around a general-purpose processor.

Those processors are often 8-bit microcontrollers having internal memory, remains with some additional external memory.

3) Sensor-Node Systems-on-a-Chip

The research groups have recently oppressed the development of whole sensor-node systems-on-a-chip (SOC). Such designs collaborate most (if not all) sensor-node subsystems on a single die or multiple dies in one package. This integrates microcontrollers and memories but also novel sensor designs as well as wireless receivers and transmitters. Examples of sensor-node SOCs are Smart Dust, the Spec Mote, and SNAP.

B. Sensor-Node Components

1) Processors

Sensor node designs have 8-bit RISC microcontroller as their major processor. The microcontroller may also have to handle a simplistic RF radio.

The computational power of 8- bit microcontrollers is often as limited as to perform complex tasks, some sensor nodes designs use 16 or even 32-bit microcontroller, or they have additional ASICs, DSPs, or FPGAs.

2) Memories

Sensor nodes are generally based on microcontrollers that generally have Harvard architecture.

Maximum novel microcontroller designs feature constructed data and instruction memories, but do not have a memory management unit (MMU) and thus cannot enhance memory security. Mostly the sensor-node designs to add external data memory or nonvolatile memory just like as FLASH-ROM.

Microcontrollers used in COTS sensor nodes accommodate between 8 and 512 Kbytes of non-volatile program memory and up to 4 Kbytes of volatile SRAM. Memory absorbs a significant fraction of the chip; the die area is a commanding cost factor in chip design.

3) Wireless Communication Subsystems

Numerous sensor networks utilize radio frequency (RF) communication; even light and sound have also been engaged as physical communication medium. Sensors, Sensor Boards and Sensor Interface:

All the sensor node constructs are Application-specific, General-purpose node design are minor with external interfaces. The external interface grants to connect different sensors or actuators precisely or to attach a preconfigured sensor board.

The sensors-node constructs are for visible light, infrared, audio, pressure, temperature, acceleration, position (e.g., GPS).

Fewer common sensor types integrate hygrometers, barometers, magnetometers, oxygen saturation sensors, and heart-rate sensors. Simple analog sensors are inspected by the processor via an analog-to-digital converter (ADC).

6. SECURITY PROTOCOLS IN SENSOR NETWORKS

A. Key Management

Key management is foremost to assure purity of sensor data and protected communication through cryptographic techniques random key pre-distribution and localized encryption and authentication protocol. RKP (Random Key Pre-distribution) RKP schemes have many variants.

Eschenauer and Gligor [16] propose a key pre-distribution scheme that commits on probabilistic key allocation among nodes within the sensor network.

These system works by circulating; a key ring to each and every participating node in the sensor network before formation. RKP scheme is distributed into three states: first one is key setup and next is shared-key discovery, and third one is path-key establishment. And thus it contributes the key revocation phase.

• *Key Setup*

Each node‟s key ring responds of a number of randomly chosen keys from a big pool of keys developed offline.

The purpose of key setup phase is to confirm that a limited number of keys are accessible to probabilistically create a common key between two or more sensors during shared key discovery phase.

• *Shared-key Discovery*

Each node telecast a key identifier list, and compares the list of identities collected to the keys in their key chains.

• *Path-key Establishment*

A node tries to connect through intermediate nodes that already have a link established through the preceding phase.

• *Key Revocation*

An arbitrated sensor node can be reason for a lot of damage to the network. So retraction of a compromised node is very useful in key distribution scheme.

When a node is compromised by an antagonist, the key ring must be deleted. Each and every neighbor should delete the key of a compromised node from their key circle. LEAP (Localized Encryption and Authentication Protocol) developed by Zhu et al. (2003) as a key management protocol for sensor networks [14]. Featherweight, energy sufficient operation and robustness and survivability are the major design targets of this protocol.

A standard implementation of LEAP (LEAP+) was configured on the Berkeley Mica2 motes [15]. RC5 is uses by the protocol for encryption and CBC-MAC for authentication.

Four different keying mechanisms provided by LEAP:

1) Individual Keys, 2) Group Keys, 3) Cluster Keys and 4) Pair wise Shared Keys.

The Individual Key is a unique key that each and every node shares with the base station. This permits for private communication between the base station and individual nodes, useful for important instructions or keying material etc.

The Group Key is a publically shared key that is adopted by the base station for sending encrypted messages to the whole sensor network (or Group). This may be treated to send queries or interests, or to generate a mission to the nodes of the network.

A Cluster Key is identical but is shared between a node and its neighbors. This is generally employed for securing private broadcast messages (routing information or enabling passive participation) [13].

A Pair wise Shared Key is a key which every node shares with each of its current neighbors.

These keys are used under this scheme for protected communications that demand privacy or source authentication. It could also use this key to disperse a Cluster Key, for example. The use of these keys includes reserved participation.

B. Cryptography & Authentication

TINYSEC Karlof et al. (2004) designed the replacement for the deficient SNEP, known as TinySec and a "Link Layer Security Architecture for Wireless Sensor Networks" [13]. This affords services as like access control, message integrity and confidentiality and scalability.

Access control and integrity are assumed through authentication and confidentiality through encryption. Semantic security is acquired through the use of a different initialization vector (IV) for each invocation of the encryption algorithm. TinySec grants for two specific variants:

TinySec-Auth, affords for authentication only, and the second, TinySec-AE, affords both authentication and encryption. For TinySec-Auth, the whole packet is authenticated using a MAC, but the charged data is not encrypted; although using authenticated encryption, TinySec encrypts the charged data and then authenticates the packet with a MAC.

The Security Protocols for Sensor Networks (SPINS) [11] activity possess of two main threads of work: an encryption protocol for Smart Dust motes called Secure Network Encryption Protocol (SNEP) and a telecast authentication protocol that is called micro-Timed Efficient streaming learnt Authentication (TESLA).

In SPINS, each sensor node contributes a different master key with the base station. On the other hand the keys required by the SNEP and the TESLA protocols are copied from this master key.

1) SNEP is based on Cipher Block Chaining implemented in the Counter mode (CBC-CTR), with the sense that the initial value of the counter in the sender and receiver is the same.

To achieve authenticated telecasts, TESLA uses a time-released key chain and gives authenticated cascading telecast, and SNEP (Secure Network Encryption Protocol) that provides data confidentiality and two edge data authentication, and data freshness with low overhead.

In Sensor Network Encryption Protocol (SNEP) the encrypted data has the following format: $E = \{D\}(Kencr,C)$, where D is the data and encryption key is $Kencr$ and the counter is C. The MAC is $M = MAC(Kmac,C|E)$.

The both keys $Kencr$ and $Kmac$ are derived from the master secret key K. The whole message that A sends to B is: $A\_B : \{D\}(Kencr,C),MAC(Kmac,C|\{D\}(Kencr,C))$. SNEP has attributes like Semantic security and Data authentication, Replay protection, Weak freshness and Low communication overhead.

2) µTESLA compete asymmetry through the delayed disclosure of symmetric keys and serves as the telecast authentication service of SNEP.

µTESLA requires that the base station and the nodes be closely time synchronized and each node knows an upper bound on the biggest error for synchronization.

The base station calculates a MAC on the packet with a key that is secret at that point in time. When a node gets a packet, it can assure that the base station did not yet display the corresponding MAC key, using its closely synchronized clock, maximum synchronization error and the time at which the keys are to be revealed.

The node reserves the packet in a buffer and aware that the MAC key is only known to the base station, and that no opponent could have fixed some packets during the transmission. When the keys are to be displayed, the base station telecasts the key to each and every receiver.

The receiver can then authenticate the righteousness of the key and use it to authenticate the packet in the buffer [11].

Each MAC key from the keys is a member of a key chain that has been created by a one way function F. According to generate this chain, the sender elects the end key, Kn, of

the chain at random and applies F regularly to compute all other keys:

Ki = F(Ki+1)

Utilizing the SNEP building block, each node can smoothly dispose time synchronization and deliver an authenticated key from the key chain for the "commitment in a protected and authenticated manner" [12].

7. CONCLUSION

This paper will help the person to know in detail about the sensor network and about the security protocols for WSNs: RKP, LEAP, TinySec, SPINS used in sensor network.

The above work emphasis our preliminary work related to the security protocols. Open-source implementations of the protocols are in the process of being made available for work. LEAP includes collapses and what we do claim is that LEAP is a very good solution. The LEAP protocols are shortly available for the industry needs otherwise they can grow into a best solution. Currently, as for wireless sensor networks, TinySec is a very important expanded protocol for data link security. In this paper work a smooth key update scheme for TinySec is given based on the weight synchronization model.

References

[1] Akyildiz, I.; Su, W.; Sankarasubramaniam, Y.; Cayirci, E. A survey on sensor networks. IEEE Commun. Mag. 2002, 40, 102–114.

[2] Lucchi, M.; Giorgetti, A.; Chiani, M. Cooperative Diversity in Wireless Sensor Networks. In Proceedings of WPMC'05, Aalborg, Denmark, 2005, pp. 1738–1742.

[3] Toriumi, S.; Sei, Y.; Shinichi, H. Energy-efficient Event Detection in3DWireless Sensor Networks. In Proceedings of IEEE IFIP Wireless Days, Dubai, United Arab Emirates, 2008.

[4] Behroozi, H.; Alajaji, F.; Linder, T. Mathematical Evaluation of Environmental Monitoring Estimation Error through Energy-Efficient Wireless Sensor Networks. In Proceedings of ISIT, Toronto, Canada, 2008.

[5] Perrig, A., et al., SPINS: Security protocols for sensor networks. Proceedings of MOBICOM, 2001, 2002.

[6] Rivest, R., The RC5 Encryption Algorithm. 1994: Fast Software Encryption. p.86-96.

[7] Doherty, L., Algorithms for Position and Data Recovery in Wireless Sensor Networks. UC Berkeley EECS Masters Report, 2000.

[8] A. Perrig, R. Szewczyk, V.Wen, D. Culler, and J. D. Tygar. SPINS: Security Protocols for Sensor Networks. Wireless Networks Journal (WINET), 8(5):521.534, September 2002.

[9] Deng, J., Han, R., Mishra, S. (2004) „Intrusion Tolerance and Anti- Traffic Analysis Strategies for Wireless Sensor Networks", The International Conference on Dependable Systems and Networks, 1 July, 2004, Florence, Italy.

[10] Karlof, C., Sastry, N., Wagner, D. (2004) „TinySec: A Link Layer Security Architecture for Wireless Sensor Networks", Proceedings of the 2nd International Conference on Embedded Networked Sensor Systems, Baltimore, MD, USA, 03 – 05 November 2004, New York, NY, USA: ACM Press, 162 – 175.

[11] Zhu, S., Setia, S., Jajodia, S. (2003) „LEAP: Efficient Security Mechanisms for Large-Scale Distributed Sensor Networks", CCS '03, Washington D.C., USA, 27 – 31 October 2003, New York, USA: ACM Press, 62-72.

[12] Zhu, S., Setia, S., Jajodia, S. (2006) „LEAP+: Efficient Security Mechanisms for Large-Scale Distributed Sensor Networks", ACM Transactions on Sensor Networks TOSN, 2(4), 500-528.

[13] L. Eschenauer and V.D. Gligor, "A key management scheme for distributed sensor networksm" In Proceddings of the 9th ACM conference on Computer and communications security, pp. 41-47, Nov. 2002.

Towards a Graph-Based Approach for Web Services Composition

Chaker BEN MAHMOUD[1], Fathia BETTAHAR[2], Hajer ABDERRAHIM[3] and Houda SAIDI[4]

[1]ENIG, University of Gabès
Gabès, Tunisia

[2]ISIMG, University of Gabès
Gabès, Tunisia

[3]ISIMG, University of Gabès
Gabès, Tunisia

[4]ISIMG, University of Gabès
Gabès, Tunisia

Abstract

Nowadays, Web services (WS) remain a main actor in the implementation of distributed applications. They represent a new promising paradigm for the development, deployment and integration of Internet applications. The aim of Web services composition is to use the skills of several departments to resolve any problem that cannot be solved individually. The result of this composition is a compound of Web services that define how they will be used. In this paper, we propose an approach for automatic web services composition based on the concepts of directed graphs for the representation and description of Web services, and the ordering of web services compound execution. In this context, the user query, defined by a set of inputs and outputs, can be viewed as a directed graph composed of Web services.

Keywords:*Web Services, Automatic composition, WSDL, service-oriented architecture, Theory of graphs.*

1. Introduction

Web services provide a new way to develop distributed and dynamic applications. They are considered to be a good solution for interoperability during data exchange between heterogeneous applications within an organization. One of the most important advantages of the Web service is reuse. In fact, Web services are conceptually limited to relatively simple features which are modeled by a collection of operations.

The use and composition of Web services to solve problems remains a difficult task to achieve, however. Web services composition refers to the process of creating a composite service with new functionality from existing relatively simple Web services. This process includes discovery, integration and execution of Web services in a specific order to meet an identified need.

Intense research activities have been conducted in this area in order to achieve correct web services composition. In fact, composition of Web services is not just simple grouping, but rather a composition in which the web services are ordered according to the relations between their semantics. These are usually provided by different organizations independently from any execution context. Since each organization has its own working rules, Web services should be treated as strictly autonomous units.

2. Related work

Several approaches have been developed for Web services composition. Only few used the concept of graph theory. In what follows, we present three approaches closely related to the one presented in this paper.

Elmaghraoui *et al.* presented in [1] a solution for optimizing the computational effort in Web services composition. This approach is based on graph theory. It consists in modeling the relationship between the involved semantic Web services in a directed graph, and calculating the shortest path by using an extended version of the Floyd-Warshall algorithm. This optimization approach is based on two pillars: i) the first is defining the semantic relationships between the available Web services using an directed graph called Service Composition Graph (SCG), and ii) the second is applying a graph search algorithm to calculate the shortest paths between all nodes. Finally, the results of the algorithm are stored in a matrix called the Shortest Predecessor Matrix (SPM).

Hashemian *et al.* [2] has created composite Web services using a graph search algorithm based on input/output dependencies between Web services. In fact, the composite Web services are presented as a dependency graph built using input-ouptut requirements of available elementary web services. A dependency graph G = (V, E) contains information about the existing Web services in the repository as well as their input/output. The set of nodes V represents the actions or statements on inputs/outputs included in the list of inputs/ outputs. There is a directed edge from node v_x to node v_y in the graph (where v_x, $v_y \in$ V) if and only if there exists at least one dependency $v_x \rightarrow v_y$ in the list of dependencies between inputs and outputs. Each edge in E is a set that contains all web services in the

repository having that dependency in one of their dependency sets. This algorithm resolved the composition problem in two steps: i) find Web services that can potentially participate in the composition, ii) find the composition based on these Web services. The author considered the dependencies between input and output parameters without considering the semantic functions, so they cannot guarantee that the generated composite service correctly provides the requested functionality.

Samuel *et al.* presented in [3] a composition technique based on weighted planning graph in which the composition can be found in polynomial dynamic time and in heterogeneous environment. The author conceived the composition problem as a problem of generating the required outputs from the given inputs. Therefore, the order of actions is not important, except that he supposed that the inputs arrive before outputs. Many information systems fall into this category where the inputs and outputs can be retrieved from different WSDL files. It uses a special graph structure called dependency graph to construct an index of available web services and their input/output information. This graph can be considered as a model for the repository specification, because it is accessible by the composition planner in response to a request. The planning graph is a layered directed graph. The vertices can be of two types. The first is the collection of propositions (pre-and post-conditions of Web services), called P and the second is the actions (set of Web services), called A. The edges connect one layer to another. The quality attributes can be assigned on the edges as weights. After the construction of the planning graph for composition, it applies the non-functional quality parameters to find a better composition scheme.

3. Model of Web services composition-based graph

The automatic composition of Web services is a complex task. Indeed, the use of Web services is limited, firstly because they perform a specific task. On the other hand, the structure and the availability of these services are not stable because of the exponential evolution of the Web. In this context, we propose a system for automatic composition of Web services. To achieve this goal, we consider that a solution for composition must be engaged from the Web services discovery to the interaction with users. From a user perspective, once the query is defined, the system commits to identify Web services necessary to satisfy the problem. In order to achieve these objectives, we propose a graph-based model for automatically composing Web services.

3.1 Principle

We assume that Web services are defined by a set of inputs and outputs describing their semantics. The user defines his needs in terms of inputs/outputs with a textual description. These needs must be validated with respect to the domain ontology (concept or term). The proposed system selects the Web service that is the most adapted to the given user inputs and uses the results to continue to meet the goal (all outputs). The resulting Web service can then be viewed as a Web services-based graph constructed according to input/output similarity. This approach covers the following features: service composition, discovery, execution and publication of the composite service.

3.2 Web service modeling

Each Web service contains different operations and is defined by its name, parameter, and state of the world in which it operates. In this work, we assume that each Web service consists in a single operation. For sake of simplicity, we use operation and Web service terms interchangeably.

We propose the following formalism:

WebService (Parameters, State-of-the-world) *(1)*

This representation allows us to introduce a Web service as an entity that is fully defined by its parameters ("PARAMETERS ") and the state of the world ("State-of-the-world ") where it acts. Parameters are represented by the inputs and outputs of Web service, and the state of the world is represented by its preconditions and its effects.

WebService (Inputs, Outputs, Pre-conditions, effects) *(2)*

Where, Inputs, Outputs $\subset$ Parameters and Pre-Conditions, Effects $\subset$ State-of-the-world.

This second representation was used to introduce a web service as an entity capable of producing one or more concrete results based on inputs/outputs requirements. Pre-conditions provide information about the state in which the world must come before the invocation of a service. The effects indicate the state of the world after the invocation of the service.

3.2.1 Conditions on web service's inputs and outputs

The inputs and outputs of a Web service should be able identifiable by a concept defined within a well-established ontology. Therefore, the different parameters of a Web service can be represented by instances of concepts belonging to different ontologies.

For example, the web service "Find_Doctor" uses a single input parameter (denoted City_Name) represented by the concept of "CITY", belonging to the ontology CNAMOnto.

So "Find_Doctor" requires an instance of the concept "CITY". The only output parameter (Doctor) is represented by an instance of the concept DOCTOR present in the ontology CNAMOnto. "Find_Doctor" will return an instance of the concept "Doctor" if the preconditions are

validated. So the inputs and outputs of a web service are clearly defined in terms of concepts of a specialized ontology and for the attributes (e.g.CIN of any person or other).

3.2.2 Conditions on the state of the world (preconditions and effects) of the web service

In order to facilitate reasoning about preconditions and effects, we present them using the first-order predicate logic. Indeed, the preconditions and effects of web services are used to estimate the state of the world in a given situation. It is therefore essential to be able to reason with these world estimators.

So we adopt the following formalism to define the state of the world (preconditions and effects) of a web service:

$$PreCondition(WS,PC1,..,PCn) \leftarrow Valid(PC1) \wedge ... \wedge Valid(PCn) \quad (3)$$

$$Effect\ (WebService,E1, ...,En) \leftarrow E1 \wedge ... \wedge En \quad (4)$$

In fact, a web service is defined by its parameters, but also the states of the world through which it passes. It is therefore necessary to integrate the pre-conditions and effects in the definition of a web service. For example, the web service "Find_Doctor" said two predicates P1 (pre-conditions on the service) and E1 (effects on the service).

$$P1\ (Find\_Doctor, CITY) \leftarrow Exist\ (DOCTOR) \wedge Valid(CITY)$$
$$E1\ (Find\_Doctor, DOCTOR) \leftarrow List\ (DOCTOR)$$

3.3 Operation of the automatic web services composition

3.3.1 Composition module

In this phase, we proceed to the automatic composition of web services. We focus on the operational aspects of web services and we concentrate particularly on the input parameters and output web services. An interaction graph of web services is represented as an oriented graph in which the vertices represent the set of web services and links materialize the flow of information between two web services.

To represent the interactions between a set of web services in the form of oriented graph, vertices can be defined using levels of details (parameter, service).

In an oriented graph, whose vertices are web services, links represent the common parameters (input / output) that allow web services interact.

Let A be a web service described by WSA (I_A, O_A, Pre_A, Ef_A), where I_A is the set of inputs, O_A the set of outputs, Pre_A refers to preconditions and Ef_A refers the effects.

In order to create a link between a source web service A described by (I_A, O_A, Pre_A, Ef_A) and a target Web service described by B (I_B, O_B, Pre_B, Ef_B), the number of output parameters O_A of web service A must be greater than or equal to the number of input parameters I_B of web service B. In this context, two cases may arise: complete relation or partial relation:

- *Complete relation*: if and only if, for each input parameter of the target web service B, there is an output parameter similar in the web service A.
- *Partial relation*: there are at least one output parameter of the source web service similar to an input parameter of the target web service

Process of building the graph composition

In our approach to composition, the user request passes through several processing stages before constructing the graph composition. The resolution of this problem of composition identifies the resolution of a goal (outputs) described by the user's query.

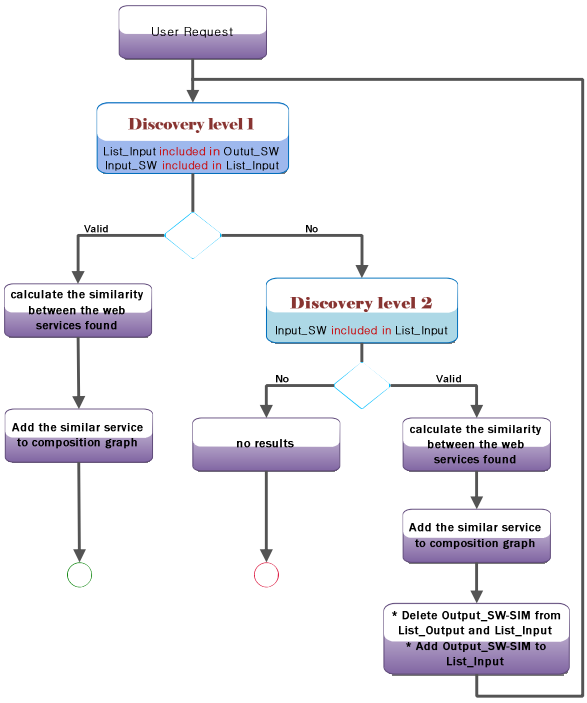


Fig. 1 Process of building graph composition

Example:
Let a set of web services declared as follows:
WS1 ({a,b},{c,d,f},{P1},{EF1,EF2}) ,
WS2 ({c},{m,k},{P2},{∅}) ,
WS3 ({w,m},{t},{P3,P4},{EF3}) ,
WS4 ({k,d,i},{p},{P5},{EF4}) ,
WS5 ({f},{i,g},{P6},{EF5}) ,
WS6 ({h,g,n},{y,q},{P7},{EF5}) ,
WS7 ({a},{f},{P8},{EF}) ,
WS8 ({t},{z,g},{P9},{∅})
And the request of the user is defined as follows:
ReqUti ({a,b,w},{t,p})

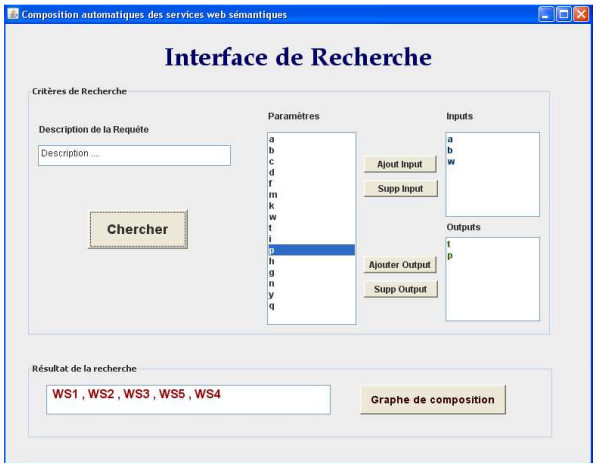


Fig. 2 Search Interface

The graph composition result in the execution of the user query is as follows:

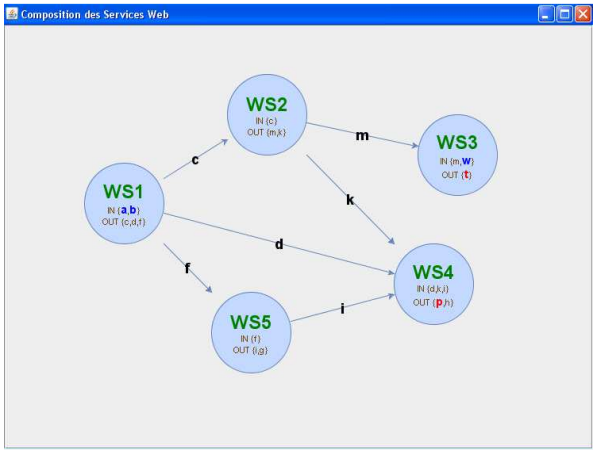


Fig. 3 Result of the composition (composition graph)

3.3.2 Discovery module

This module searches for a list of web services that meets the requirements expressed by the composition module.

The discovery module uses the properties of web services in order to find the ones that best respond to a query. In the discovery process, our module seeks the similarities between the query definition parameters and the web services ones published in registries.

The similarity calculation has a great influence on the search of web services for discovery and composition. For the discovery, the similarity calculation is based on textual descriptions, on the input and output parameters, and the state of the world of web services. For the composition, the similarity calculation is applied to the output parameters of the first service compared and on the input parameters of the second service compared.

Moreover, according to the nature of objects to compare, the similarity can be broken down into syntactic similarity or semantic similarity. Note that the matching semantics can be used on syntactic descriptions by enriching the descriptions for the treatment. Various solutions have been proposed in the literature like the use of tools such as lexical database WordNet [4] or methods such as latent semantic analysis [5].

Syntactic similarity

Syntactic similarity compares parameters from their respective orthographies. Two distinct similarities can be distinguished, the approximate similarity and the equal one.

- The equal similarity uses the strict syntactic equivalence. Two objects are said to be similar if and only if they have the same orthography.
- The approximate similarity uses distance functions d(x,y) to quantify the similarity between two character strings x and y. If the distance between two objects is above a certain threshold, these objects are said similar.

Semantic similarity

For comparing the outputs of a request to the outputs of a published service, four degrees of matching are used [6]:

- *Exact matching*: select a web service if he corresponds exactly at the request (request = Service) that is to say, the inputs and outputs of the request are equal to the inputs and outputs of the web service.
- *Plug-in matching*: returns a web service if he includes a request (request < Service) that is to say, the input of the request includes the inputs of the web service and outputs of the request are subsumed by the outputs of the web service. In this case, the web service is a set that generalizes the request.
- *Subsumes matching*: returns a Web service, if he included in a query (request > Service). In this case, the service does not completely satisfy the request. This service may be used to achieve partially the purpose of the request. One or more additional services may need to be used to meet all the goals of the user.
- *Fail matching*: returns false if no match between the query and service.

In terms of satisfaction of the request, the semantic matching degrees can be ordered according to a scale of preference as follows:

$$Exact > Plugin > Subsumes > Fail$$

In our approach, the discovery of web services is to find links and semantic correspondences between the parameters of the request with Web Services. This

discovery is essentially based on the parameters and the state of the world of web services.

In this context, the similarity can be divided into two parts: parameters similarity (Input and Output) and similarity of state of the world (Pre-condition and Effect).

The system measures the similarity of the parameters (input and output) by attributing a score for each mode of matching: Exact (score=3), Plug-In (score=2), Subsumes (score=1), Fail (score=0). Then, it assigns a score according to the valid states of the world of web services.

Therefore, the matching between the request and a set of Web services can be measured quantitatively. The service has a high similarity score represents the service the most accurate for the request.

The following equation generalizes the comparison between the proposed request by the composition module and web services:

$$Sim\,(Req,SW) = Sim_{In/Out}\,(Req,SW) + Sim_{Pre/Effet}\,(SW) \qquad (5)$$

Where

$$Sim_{In/Out}(Req,SW) = \begin{cases} 3\ (Exact) & if\ InReq = InSW \\ & and\ OutReq = OutSW \\ 2\ (plug\text{-}in) & if\ InReq \supset InSW \\ & and\ OutReq \subset OutSW \\ 1\ (Subsumes) & if OutReq \supset OutSW \\ 0\ (Fail) & if\ OutReq <> OutSW \end{cases} \qquad (6)$$

$$Sim_{Pre/Effet}\,(SW) = \begin{cases} 2 & if\ Valid(Pre\text{-}Condition) \wedge Valid(Effet) \\ 1 & if\ Valid(Pre\text{-}Condition) \vee Valid(Effet) \\ 0 & if\ \neg Valid(Pre\text{-}Condition) \wedge \neg Valid(Effet) \end{cases} \qquad (7)$$

With: InReq denotes the inputs of the request; OutReq denotes the outputs of the request; InSW denotes the inputs of the web service; OutSW denotes the outputs of the web service.

The architecture of this module is illustrated in the diagram below:

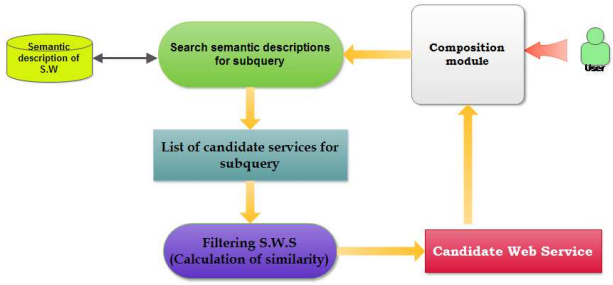


Fig. 3 Architecture of discovery module

Discovery algorithms

Let:
SW: a set of existing Web services in the directory.
SReq: the sub query.
SWF: selected Web service

Algorithm 1:Module Discovery

Input: SW, SReq
Output: SWF
Taux_sim←0
For each S **in** SW **do**
If Similarité (SReq,S) >Taux_**sim Then**
Taux_sim←Similarité (SReq,S)
SWF ← S
End if
End for
Return (SWF)
End.

Algorithm 2: Semantic Similarity

Input : SW, SReq
Output :Taux_Sim
If EntReq = EntSW **and** SortReq = SortSW **Then**
Taux_Sim← 3
Else if EntReq ⊃ EntSW **and** SortReq ⊂ SortSW **Then**
Taux_Sim← 2
Else if SortReq ⊃ SortSW**Then**
Taux_Sim← 1
Else
Taux_Sim← 0
End if
If Pre-ConditionSW = vrai **and** EffetSW = vrai **Then**
Taux_Sim← Taux_Sim+2
Else if Pre-ConditionSW = vrai **or** EffetSW = vrai **Then**
Taux_Sim←Taux_Sim+ 1
End if
Return (Taux_Sim)
End.

4. Conclusion

In this paper, we proposed an approach for automated Web services composition based on directed graphs theory. In this approach, we proposed a formal method for describing Web services, then selecting and ordering the ones which satisfy the required inputs and outputs for the compound Web service.

In future work, we will be working on a module for the verification and validation of composition with respect to user needs. We will also be working on a model of semantic representation of web services in order to assess the degree of similarity or the possibility of interaction between Web services.

References

[1] Hajar Elmaghraoui, Imane Zaoui, Dalila Chiadmi and Laila Benhlima, "Graph based E-Government web service composition," in International Journal of Computer Science Issues (IJCSI), Vol. 8, Issue 5, No 1, pp. 103–110, September 2011.

[2] Seyyed Vahid Hashemian and Farhad Mavaddat, "A Graph-Based Approach to Web Services Composition,"In Proceedings of the 2005 Symposium on Applications and the Internet (SAINT'05).

[3] S. Justin Samuel and T. Sasipraba , "AN APPROACH FOR GRAPH BASEDPLANNING AND QUALITY DRIVENCOMPOSITION OF WEB SERVICES," in Indian Journal of Computer Science and Engineering (IJCSE), Vol. 2, No. 5, p672-679, Oct-Nov 2011.

[4] Seog-Chan Oh, Dongwon Lee, and Soundar R.T. Kumara, "Effective Web Services Composition in diverse and large-scale services networks,"in IEEE TRANSACTIONS ON SERVICES COMPUTING, VOL. 1, NO. 1, JANUARY-MARCH 2008.

[5] Jiangang Ma and Yanchun Zhang , "Web Services Discovery Based on Latent Semantic Approach,"in IEEE International Conference on Web Services, 2008, pp. 740- 747.

[6] M. Paolucci, T. Kawamura, T. R. Payne, and K. Sycara, "Semantic Matching of Web Services Capabilities," in International Semantic Web Conference, 2002, pp. 333-347.

A Calculus for Non Repudiation Protocols

Abdesselam Redouane

Department of Computer Science and Engineering, College of Engineering and Computing
Al Ghurair University, Dubai, UAE

Abstract

We describe a calculus that is specific to non-repudiation protocols. The calculus uses the correspondence assertion of Woo and Lam, that is, if there is a non-repudiation of receipt there should be a corresponding non-repudiation of origin. The main contribution of this work lies in the way we model input and output and hence captures non-repudiation properties. The calculus is a subset of the Pi calculus. The basic constructs are modified in order to handle properties of non-repudiation. We offer a formal syntax and an operational semantics of the calculus. We show the usefulness of the calculus by describing Zhou optimistic protocol.

Keywords: *Non repudiation protocols, Pi calculus, operational semantics.*

1. Introduction

One of the main concerns in e-business, in all its different forms such as B2B, B2C, E2C, is fair exchange of services. In simple terms this is concerned how to ensure fairness between parties. In that, there is no denial by one of the entities of having participated in all or part of an electronic transaction. For example, suppose that a business A instructs its bank to carry out some money transfer to a particular account. The bank executes the instruction requested by A. Later, A denies that he has sent a message for debiting money to that particular account. To avoid such denials the following non-repudiation services are required:

- Non-Repudiation of Origin (NRO) is intended to protect against the originator rejection or denial of having sent a message to the recipient.
- Non-Repudiation of Receipt (NRR) is intended to protect against the recipient rejection or denial of having received the message from the originator.

Non-repudiation protocols rely, usually, on a Trusted Third Party (TTP). All the parties involved in the transacting process trust the TTP. Any dispute will be resolved via this TTP. The trend in these protocols is that they try to minimise its use during a protocol run. Protocols, which do not respect this issue, however, will end up with a bottleneck problem. There are protocols, which eliminate the use of TTP altogether. The approach adopted in this latter case is a probabilistic one [1], [2]. An intensive survey of fair non repudiation protocols can be found in [3].

We describe a calculus which is specific to non-repudiation protocols. We are interested in the more general protocols and which involve the use of a TTP. The calculus is a sub set of the Pi calculus enriched with some primitives to handle non-repudiation properties. We use the technique of Woo and Lam [4] of the correspondence assertion in the sense that for every received NRR there must exists a corresponding NRO. The main contribution of this work lies in the way we model input and output and hence captures non-repudiation properties.

The sequel is organised as follows. In the next section we describe the calculus along with a brief introduction to the Pi calculus. In this section we provide the syntax and an operational semantics of the calculus. Section 3 illustrates the use of the calculus with an example. Related work is given in section 4 while section 5 concludes the paper.

2. The Calculus

The calculus is a subset of the Pi calculus [5] where we modified some of the primitive constructs to handle non-repudiation of origin and non-repudiation of receipt.

2.1. Pi Calculus Overview

The Pi calculus is in essence a process algebra where processes interact by sending data and channel names. The basic computational step is the transfer of a communication link between two processes. The following example illustrates this idea [6]. We have a client which wants to use the printer. The access to the printer is via the server. We have two channel of communication a and b. The channel a is used as an output channel from the server and the printer. The b channel can be used either direction

between the server and the client. Fig. 1 below shows this scenario.

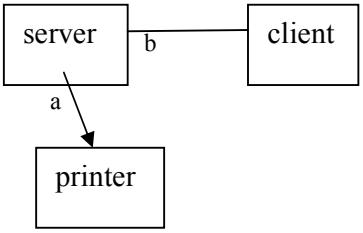


Fig.1: Before interaction between the server and client

There are three processes: S for the server, C for the client and P for the printer. There are two channels a and b as we mentioned earlier. This state can be written in the Pi calculus as follows:

$$\underline{b}a.S \mid b(m).\underline{m}d.C \qquad (1)$$

This expression state that the server will send the link a through the channel b and then behave like S. The client C is using the channel b as input where m is a place holder for the input received. The received input, which is the channel a, is then used as an output channel to send data d. The symbol | is used to mean parallel composition, that is, the two processes S and C are running in parallel and communicating via the channel b.

After the interaction between the processes S and C we have the following expression:

$$S \mid \underline{a}d.C \qquad (2)$$

That is, the channel a is being used by C to send its data to the printer.

Combining the two expressions (1) and (2) the interaction between the server and client can be formulated as follows:

$$\underline{b}a.S \mid b(m).\underline{m}d.C \longrightarrow S \mid \underline{a}d.C$$

Fig. 2 below shows now the new channel between the client and the printer.

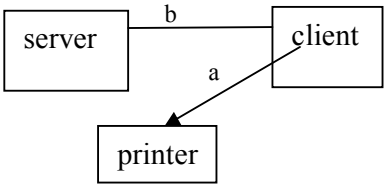


Fig.2: After interaction between the server and client

It should be noted that there are many variants of the Pi calculus which deal with specific area like the SPI calculus [7] and the Ambient calculus [8]. SPI calculus is used for modeling and analysing security protocols and the Ambient calculus is used to model and analyse mobile code.

The Pi calculus, as it is, cannot model non-repudiation of origin (NRO) and non-repudiation of receipt (NRR). It would be simpler; however, to modify the calculus to handle these specific issues elegantly rather than to model these with Pi core primitives and end up with what it might be a cumbersome description.

In order to make the calculus simple we use a biadic calculus rather than a polyadic one. We believe this will suffice to describe non-repudiation protocols, as it is usually the case that in this type of protocols there are two major elements of interest: the message and the non-repudiation service.

The framework in which the calculus should operate is that the evidence of non-repudiation, especially NRR, is generated by the protocol automatically rather than by the user. To this end, we use digital signatures to offer these services, as it is customary in these types of protocols. To accomplish this need, we suppose the availability, to a protocol, the followings: the participating agent identification and his private key. In addition, the generated signature makes use of the notion of a session of a protocol run. Thus, a digital signature is a tuple of the form: (typ, Id, K, α).

Where:

Typ: is the type of the signature: {nro, nrr, sub, con}
Id: the identification of the agent
K: private key of the agent
α: the session of a protocol run

We distinguish two types of digital signatures: the ones, which require non-repudiation of services, and the ones, which do not require such services.

The verification of a received digital signature represented by a NRO or NRR is assumed to be possible by each agent participating in a protocol. The type of communication between agents is asynchronous as we anticipate that an agent who performed a send/receive will not stay idle waiting for a response from the other agent.

2.2 Syntax

Let *N* be a set of names denoting communication actions and variables. Let τ be an internal action capable of being executed by any agent if he wishes to. Let *A* be a set of agent names, *T* a set of TTPs names and *D* a set of digital signatures. As stated earlier, two types of digital signatures are envisaged: those which require non-repudiation of services on one hand and those which they don't require on the other hand. We let the first type ranges over {nro, nrr} and the second type ranges over {con, sub}. The syntax is summarised in Table 1.

Table 1: Syntax

a,b … x,y,z …	Names *N*
T, U, V	TTP agents *T*
no, nr, con, sub	digital signatures *D*
A, B ::= Agents *A*	
0(null)	
|a(x,y).A (input)	
|a(x,y).A (output)	
| A || B (parallel composition)	
| rec(X).A (recursion)	

An informal explanation of the different operators might be useful.

- **0** is the null agent that does nothing.
- a(x,y).A is the input on the channel a to be bind to x and y, and then behave like agent A. Note that y is place holder for a digital signature.
- a(x,y).A is the output that put x and y on the channel a and then behaves like A. Note that y may be a digital signature.
- A || B is the parallel composition.
- rec(X).A is the recursion to allow infinite call to the task accomplished by an agent. This expression binds free occurrences of X in A.

2.2 Operational Semantics

The operational semantics is explained below. Note that not all the symbols are there.

τ τ.A → A
Inp$_{nro}$ a(x,y).A → A[m/x, nro/y] → a(0, nrr) →A
Inp a(x,y).A → A[m/x,0/y]
Out$_{nro}$ a(m,nro).A → A → a(0,x).A
Out$_{nrr}$ a(m,nrr).A → A
Out a(m).A → A

PAR $\dfrac{A \rightarrow A'}{A \,||\, B \rightarrow A' \,||\, B}$

COM $\dfrac{A \rightarrow A' \quad B \rightarrow B'}{A||B \rightarrow A'||B'}$

In the following we comment on these rules and how they should be interpreted.

The Input Rule with an NRO (Inp$_{nro}$)
This action is responsible for the guarantee of non repudiation of receipt and is actually formed in the following steps:

- Get action from the channel, which receive all the input in this case two parameters.
- The input parameters are substituted in their place holder, i.e. x and y respectively
- An output action is generated automatically on the same channel with the first parameter empty and the second parameter is the nrr of the recipient.
- The agent A will, then, continue performing his duties.

It should be noted that the rule is circular-free because once the originator receive the NRR he will not trigger another NRR for the recipient. Of course, the originator is able to see that the non-repudiation service received is a response of his earlier NRO.

It will be noticed from the definition of this rule is that we adopt a style of an early semantics where the substitutions occurs once they have been received and then the process evolves to another state contrary to a late semantics one.

The Input Rule without an NRO (Inp)
This rule is needed if non-repudiation is a not a must. This case may be of interest in a normal

communication between agents or where the NRO and NRR are not required.

In this action the second parameter is empty. Once the recipient detect that the second parameter is empty there is no need to continue, but rather it is obligatory to stop, with his non-repudiation activities.

The Output Rule with NRO (Out_{nro})
The rule is for initiating a non-repudiation handshake. The originator starts by forming his nro and sent it to the recipient. As the rule suggests the originator has to wait on the same channel to get his nrr. It should be noted that this channel will not be used for other communication activities while it is in this status.

The Output Rule with NRR (Out_{nrr})
The rule is for responding to a received NRO. It should be noted that this rule is triggered automatically after an input has been made which contains an NRO

The Output Rule without NRO or NRR (Out)
This rule allows an agent to perform regular communication with another agent where there is no need for non-repudiation services. It is the symmetric counterpart of the Inp rule without non-repudiation services.

The parallel Rule (PAR)
This rule defines the behaviour of the parallel action and it is self-explanatory.

The Communication Rule (COM)
This is the main communication between agents running in parallel and willing to communicate on a common channel. Note that the agents can communicate using non-repudiation services or without them, that is, in a regular communication.

2.4 Bisimulation

In this subsection we define a bisimulation method between processes. The purpose is to be able to make judgment whether two processes are equivalent. This result will be useful, for instance, to verify that an implementation meets its specification.

Two agents A and B are bisimular is that for each transition from A to be matched by a transition from B and vice-versa, leading again to equivalent derivatives A' and B'.

As it has been stated earlier in the calculus rules that we adopted an early semantics, therefore, in the definition of the bisimulation we use an early bisimulation style. A binary symmetric process relation S is bisimulation if (A,B) e S implies:

(i) if A → A' with an input action a(x,y) then for all (z,p) ∃ B': B → B' with the input action a(x,y) and (A'[z/x],B'[p/y]) ε S

(ii) (ii) if A → A' with an action different from an input then ∃ B': B → B' and (A',B') ε S

A and B are bisimular written A ~ B if (A,B) e S for some bisimulation S.

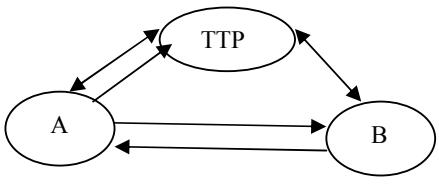
3. Example

In order to illustrate the calculus in practice we specify the optimistic protocol of Zhou [9] (Zhou, 1996). We follow the usual routine in this type of formalism. That is, we provide a specification of the protocol and an implementation. All this encoding is in the calculus. The final step is to proof that the implementation and the specification are bisimular. If this case holds we conclude that the protocol indeed guarantees non-repudiation properties.

3.1 Protocol Description

The main idea of the protocol is to minimise the use of the TTP. For that the originator starts by making a commitment to the recipient by sending the message encrypted. Note, however, that the key is not sent with the message. The originator, then, lodges the key with the TTP. Part of the non-repudiation is that the recipient must retrieve the key from the TTP and the originator, as well, has to get a confirmation from the TTP about the key. Hence the originator must retrieve this confirmation from the TTP.

3.2 Protocol Diagram and its Standard Notation



The protocol in standard notation is as follows:
Message 1. A → B: f_EOO,B,L,C,EOO
Message 2. B → A: f_EOR,A,L,EOR

Message 3. A➔TTP: f_SUB,B,L,K,SUB_K
Message 4. B ⇔ TTP: f_CON,A,B,L,K,CON_K
Message 5. A ⇔ TTP: f_CON,A,B,L,K,CON_K

3.3 Protocol Encoding

We map each send and receive between two agents as one process. That is, between A and B and between any agent (A, B) and the TTP. We have four processes in total, which should be performed in sequential order.

ZhouProtocImp = A1.A2.A3.A4

Where:

A1 = a(m,no).A || a(x,y).B
A2 = a(m,sub).A || a(x,y).T
A3 = a(m,con).T || a(x,y).B
A4 = a(m,con).T || a(x,y).A

On the other hand we need a specification for the protocol which we leave as a future work. The final task then is to show that ZhouProtocImpl is bisimilar to ZhouProtocSpec.

4. Related Work

It should be noted that Schneider [10] has used CSP for the analysis of the above protocol where the proof has been made by hand.

Kremer [11] verified non-repudiation, with a TTP, using a game based model that uses the model of alternating transition systems (ATS) and alternating time temporal logic (ATL) [12].

Zhou work on non repudiation also uses a TTP in his protocols and uses belief logic SVO [13] to verify non-repudiation protocols.

Formal analyses have been also used by Shmatikov [14] and [15] to study fair exchange protocols.

Zhang [16] uses labelled colored Petri nets to model and analyse non repudiation services in a distributed system.

5. Conclusion

We have described a calculus that is useful in the description of non-repudiation protocols. Its syntax and operational semantics have been described.

As a future work, we intend to complete the verification of Zhou optimistic protocol stated in the example. Another area of investigation is an implementation of this calculus in order to take it from a paper and a pencil work to machine automation. To this end, a tool will be useful, in that, given a protocol description, will decide if the protocol is satisfying non-repudiation properties or not.

References

[1] O. Markowitch, Y. Roggeman, “Probabilistic non repudiation without trusted third party”, in *Second* Conference on Security in Communication Networks, 1999.
[2] A. Aldini, R. Gorrieri, “Security analysis of a probabilistic non repudiation protocol”, in PAPM-PROMIV, LCNS 2399, 2002 Springer Verlag.
[3] S. Kremer, O. Markowitch,J. Zhou, “An Intensive Survey of Fair Non-Repudiation Protocols”, in Elsevier Science, 2002.
[4] T. Woo, S. Lam, “A semantic model for authentication protocols”, in IEEE Symposium on Security and Privacy, 1993.
[5] R. Milner, Communicating and Mobile Systems: the π-Calculus, Cambridge University Press, 1999.
[6] Joachim Pi: An introduction to the calculus. pp: 479 - 543 Handbook of Process Algebra, 2001, Elsvier Science, 2001.
[7] M. Abadi, A. D. Gordon: A Calculus for Cryptographic Protocols: The spi Calculus. Information and Computation, Vol. 148, Issue 1, pp: 1-70, 1999.
[8] L. Cardelli, A.D. Gordon. "Mobile Ambients". in proceedings of the First international Conference on Foundations of Software Science and Computation Structure, Lecture Notes in Computer Science (Springer-Verlag), 1378, pp: 140–155, 1998.
[9] J. Zhou, D. Gollmann, “A fair non repudiation Protocol”, in IEEE Computer Society Symposium on Research in Security and Privacy, 1996.
[10] S. Schneider, “Formal Analysis of a Non-Repudiation Protocol”, in Proceeding of the 11th IEEE Computer Security Foundations Workshop, 1998.
[11] S. Kremer, J. Raskin, “A game-based verification of non-repudiation and fair exchange protocols”, Journal of Computer Security, 2003.
[12] R. Alur, T. Henzinger, O. Kupferman, “Alternating time temporal logic”, in Proceeding of the 38th Annual Symposium on Foundation of Computer Science, IEEE Computer Societry Press, 1997.
[13] J. Zhou, D. Gollmann, D., “Towards verification of non repudiation protocols”, in Proceeding of International Refinement Workshop and Formal Methods, Spring Verlag, 1998.
[14] V. Shmatikov, J. Mitchell, “Analysis of abuse-free contract signing”, in Financial Cryptography, LCNS 1962, Spring Verlag, 2000.

[15] V. Shmatikov, J. Mitchell, “Finite state of two contract signing protocols”, Theoretical Computer Science, 2002.
[16] H. Zheng, Y. Yue Du, S. Yu, ”Modeling non repudiation in distributed systems”, Information Technology Journal, 2008.

Dr. Abdesselam Redouane is currently with the college of engineering and computing, Al Ghurair University, Dubai. He received his PhD in Computer Science from Manchester University, UK. His research interest lies in the area of the application of software engineering techniques to new emerging technologies like web and mobile applications. He is also interested in computer security and especially in access control. He is an associate editor of the International Engineering Letter.

Virtual Reality: An Efficient Way in GIS Classroom Teaching

Jiangfan Feng

**College of Computer Science and Technology, Chongqing University of Posts and Telecommunications
Chongqing, 400065, China**

Abstract

Although geographic information system (GIS) education has been spread widely, it becomes increasingly apparent that two-dimensional maps cannot be precisely present multidimensional and dynamic spatial phenomena. It seems that merging GIS and Virtual Reality (VR) is a way to deal with these issues in terms of GIS classroom teaching. Virtual learning environment is the simulation of teaching method, thinking model, cognition manner and control means in the actual learning environment. This paper introduces virtual reality technology and the necessity of applying in GIS education and instruction, which explain the basic method and achieving a way of VR technology applying in the GIS classroom teaching with the instance of VR. The purpose of this paper is to analyze the application of virtual realistic learning environment for GIS education and establish a classroom teaching model accordingly.

Keywords: *Geographic information system, Education, Virtual Reality, Problem.*

1. Introduction

It's well-known that making decisions based on geography is natural to human thinking. For example, where shall we go, what will it shall be, or what shall we do when we get there are applying for the simple event of going to the cinema. By understanding geography and people's relationship to location, we can make informed decisions about the way we live on the earth. Geographic information system (GIS) is such a technology tool for comprehending geography and making intelligent decisions.

GIS organizes geographic data so that a person reading a map can select the data necessary for a specific project or task. A thematic map has a table of contents that allows the reader to add layers of information to a base map of real-world locations. For instance, a social analyst might use the base map of the province, and select datasets from the National Bureau of Statistics of China to add data layers to a map that shows residents' education levels, ages, and employment status. With an ability to combine a variety of datasets in an infinite number of ways, GIS is a useful tool for nearly every field of knowledge of archaeology to zoology.

When students are learning and using GIS, they develop analysis and critical thinking skills, regardless of their field. GIS is a learning platform for conceptual modeling. Students also learn technical skills that will help them in their future employment. Spatial thinking skills acquired in the classroom deepen their understanding of the relationships that exist in the world and the complex problems facing society today.

On the other hand, the potential of VR technology for supporting education is widely recognized. Several programs designed to introduce large numbers of students and teachers with the technology have been established, a number of academic institutions have developed research programs to investigate key issues, and some public schools are evaluating the technology. It has already seen everyday use in an estimated twenty or more public schools and colleges, and many more have been involved in evaluation or research efforts [1, 2].

From the view of education, VR is based on a complete teaching environment learner-centered. The learner is able to control the side of the target environment to observe or study a bit, that the learner is an active observer. The way reflects a kind of new teaching mode, and work with teachers to constitute a new teaching system, students and teachers with the VR system linked to student learning in a VR environment, by sensing devices to operate directly on the virtual environment, teachers, and students learning through the center console of the system and make the appropriate instruction. Its greatest feature is the students along with their own way in the VR environment to learn. VR is able to provide students with a new observation point self-centered. The characteristics of each student

access and adjust the three-dimensional data in the real world, and thus constitute a virtual environment. From the view of information system, it is to accomplish the teaching, feedback control and inspection functions, individualized learning real implementation and experience in the VR system.

2. Related Work

2.1 Teacher Education Programs of VR

VR has been already used in a variety of educational, training, and entertainment settings [3]. The highly visual and interactive nature of VR has been proven to be useful in understanding complex 3D structures and for training in visual tasks [4]. Recognition of this has led to increasing interest in developing VR-based applications for higher education and training.

There are many programs that provide the type of education for teachers regarding the use of VR technology, such as VRRV/Nebraska, Educators' VR Series, QuickTime VR(QTVR), VR in the schools, and virtual education - science and math of Texas (VESAMOTEX), and VR Concentration, M.A. in Education [5].

Virtual reality in education is a leap forward in the development of educational technology. It created a "self-learning" environment, by substituting the traditional way of learning about a new approach through the information environment interaction. VR provides a vivid and realistic learning environment for the students, further, it provides unlimited virtual experience in a wide range of subject areas, in order to accelerate and consolidate the process of learning. Students feel more convincing than the purposeless, abstract teaching, it takes the initiative to go to the inculcation of interactive and passive nature of the difference. Virtual experiments using virtual reality technology create a variety of virtual laboratories, which have advantages of low cost, low risk and limited venues.

2.2 Higher Education in GIS

The role of higher education is to assist students in becoming effective thinkers with the knowledge and skills that will lead them toward becoming meaningful contributors to society. Geographic Information Systems in higher education provide an integrated solution to assist faculty and students with their educational goals.

GIS is no longer just for geography departments. By putting information in the context of geography, it can also be applied across several fields of study to enhance learning and teaching. GIS can give students the skills they need for careers in health, marketing, environmental studies, engineering, natural resource management and, of course, geography.

Although many GIS have been successfully implemented for storage, management, analysis and presentation of spatial data, it becomes increasingly apparent that two-dimensional maps cannot be precisely present multidimensional and dynamic spatial phenomena. Moreover, there is a growing need towards accessing spatial data not only by cartographers and surveyors but also by other users, including naturalists. It seems that merging GIS and multimedia is a way to deal with these issues [6, 7].

Further, education of naturalists is a field where integration of multimedia and GIS can bring enormous benefits. Students will learn faster and more efficiently, using tools that they are likely to meet in their future jobs. In addition, it will be possible to individualize learning and tune it to particular preferences of each student. In this model, a teacher becomes a guide rather than a repository of facts. It is the computer that takes on a role of "an infinitely patient teacher."

In addition, there is a need for teaching professionals to apply the new technology in their fields of expertise. It calls for on-site and just-in-time training. Multimedia GIS could be a very useful tool for such a task. Specialized, tailored courses could be delivered and learnt at a suitable time and adjustable pace. Assuming that some of the trainees would know computer and GIS technologies one can expect the learning process to be fast. And for those new to computers, and GIS in particular, it should be easier to acquire the new knowledge.

Researchers are also demonstrating some really efficient ways that Second Life can benefit the GIS learner. The University of Texas, Arlington, has created a kiosk for its GIS users to get help with the software and even individual teaching or research applications. The University of Illinois has a site that displays GIS-derived maps of the state for people to examine, shown as Fig.1.

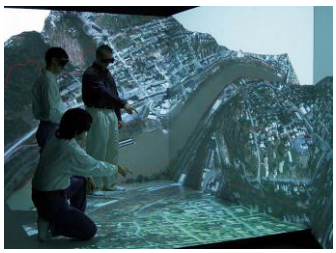


Fig. 1 Visualizing information using the FLEX system.

3. Cognitive learning in virtual reality

3.1 The Process of Learning

Characteristics, the process of learning for medical students, who account for a large proportion of the closed-loop interaction (learner-centered interaction), has obvious advantages in the process of learning. The virtual reality learning environment in which students can take advantage of the process of learning the feedback generated by the various stages of learning self-regulation and control. It's very favorable in the process of learning.

3.2 Learning of Motion Perception Skills

Learning goal is some movement perception skills. For example, the training of geographic procedures, this type of learning skills is important for students to participate in the control link, at the same time, in view of the GIS particularity, feedback is necessary for this type of learning.

3.3 The spatial learning and rehearsal

Virtual reality allows medical students to rehearse certain operations, such as a specific surgical internship example, in brain surgery in the learning process; the students were asked to understand the structure of the brain of a spatial location, shape and contains the tumor or foreign body. This kind of study is very important for medical students in the establishment of human organ structures and spatial location, as well as the brain's ability to form a human "concept map".

3.4 Conceptual Learning

When using virtual reality systems to aid in the understanding of a phenomenon, the two features, both teachers, designers or learners are very important, first of all, with other forms of learning, an effective initiative to explore, rather than passively observed, are very important. Secondly, we must get strong conceptual knowledge; students need different manifestations of the same content. These concepts in the virtual reality experience should be by means of a more abstract representation. For example, in the experience of clinical signs should be by way of illustration, the test data and the language to describe the experience of the same phenomenon at the same time, understand the correlation between the different manifestations of the students are also very important.

4. Teaching Model

Teaching model is shown in Figure 3

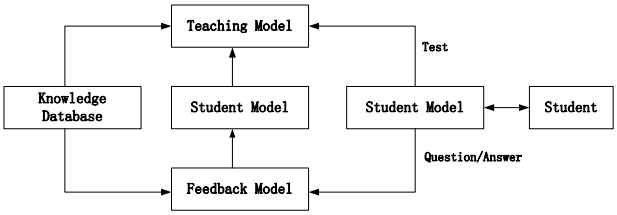


Fig. 2 The using of visualizing information.

Knowledge database includes knowledge of this course (text, 3D database, graphics, etc.), the module can be used to answer students' questions, and provide for the detection, diagnosis, and feedback module factual knowledge.

Feedback module operating or answer the detection student use of diagnostic rules, based on the reaction of the students to determine what knowledge students have mastered what is the error of the students, all of this information are reflected in the student model.

The student model is a record of the students' understanding of the subject knowledge, but also records the student's learning history, for example, the success rate, academic records, test results.

Teaching modules include courses systemic thinking environment, control environment, cognitive, and its role is to select the information according to the student model, a teaching strategy, decide what kind of intervention in the next step of the process of learning when and where. Correct in the teaching process, students' cognitive abilities, they continue to do so, student learning time, cognitive ability evaluation to amend Finally, gradually approaching the accurate value.

5. Case Study

5.1 The Simulation of Environmental Process

Fig. 3, below, is from a model developed as a demonstration of the proposals for the dispersion of air pollution. Embedded within the model is the facility to view the 'before' and 'after' scenes in which some stages of pollution situation are represented to give a third dimension.

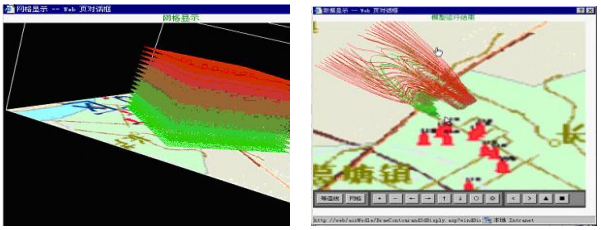


Fig. 3 The simulation of environmental process

The results make the teaching of these new media content to get more real and visual effect, and the technology is able to establish a virtual learning environment for students, which can be called at any time the content of interest.

5.2 Virtual Campus

Fig. 4, below, shows academic buildings, digital library, playground scene is selected as a pilot area, all components of the scene at the specified location on the scene, and the establishment of a virtual campus environment model, dynamic virtual scene roaming. The virtual campus roaming system, as long as the user through the keyboard operation random roaming can be a stroll in the sports arena, on the steps of the building, in a small river viewing.

Fig. 4 The simulation of campus

With the emphasis now placed on high-impact practices, teachers face a sometimes daunting task of developing and offering engaging, impact learning experiences for their students. VR can serve as an excellent tool to enhance such lessons and complement high-impact experiences in a variety of fields and disciplines.

5.3 Geographic Process Simulation

Geographic process models have been increasingly featured in the next generation geographic information science (system), as a method for phenomena simulation and mechanism analysis of the physical environment and its live activities, thereby driving conventional GIS based on data manipulation in the world of dynamic and computational processes.

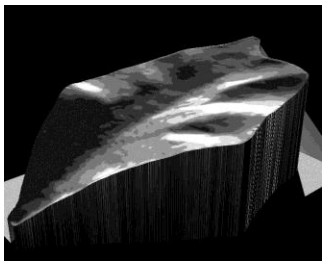


Fig. 5 The simulation of soil erosion

Fig. 5, above, is from a simulation model of soil erosion by reason of rainfall. Rainfall is one of the main agents driving the soil erosion on sloping cropland. VR helps the teachers and students to find out the law of soil erosion under different rainfall process.

6. Conclusions

In the current transition from an industrial society to an information society, traditional instructional approaches based on the use of textbooks in classrooms have been called into question. Instead of memorizing facts, more emphasis is being placed on the high-level thinking skills needed to construct and apply knowledge. Students must learn to locate, interpret, and creatively combine information, and to isolate, define, and solve problems. Additionally, education is no longer seen as something limited to a classroom or to a certain period in a person's life. Instead, education will be life long and must meet the needs of a flexible workforce.

VR as a two-way communication tool offers considerable potential particularly in the area of GIS education. The results clearly show that teachers should do everything possible to give students the ability to incorporate acknowledged good practices such as providing multiple representations and placing at least some instruction under the learner's control. While these latter attributes are not

unique to VR technology, the technology does facilitate their use more than many traditional educational practices [8, 9].

References

[1] Huber M., 1994. Multimedia enhances GIS applications. GIS World, August.
[2] Jacobson R., 1994. Virtual worlds capture spatial reality. GIS World, December.
[3] Vince J. Virtual Reality Systems. Reading, Mass Addison-Wesley Publishing Co1995.
[4] Wei B D, Li L, Hu F. "Research on intelligent GIS services in ubiquitous environment", Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), vol. 21, n. 5, 2009, pp.1-6.
[5] Brown, D.J., S.V.G. Cobb, and R.M Eastgate. 1995. Learning in Virtual Environments(LIVE)." In Virtual Reality Applications, pp. 245-252. Academic Press: San Diego,CA.
[6] United Nations (2003) United Nations Decade of Education for Sustainable Development, Resolution Adopted by the General Assembly, A/RES/57/254, Fifty-Seventh Session, 21st February, 2003,United Nations Documentation.
[7] Kneebone, R. (2003) 'Simulation in surgical training: educational issues and practical implications', Medical Education 37 (3), pp. 267–277.
[8] Bell, J.T. and H.S. Fogler. 1996. "Preliminary Testing of a Virtual Reality Based Educational Module for Safety and Hazard Evaluation Training." In Proc. American Society for Engineering Education Annual Conference, Indiana Sectional Meeting, Peoria, IL.
[9] Application of Virtual Reality as an Educational Tool." In Proc. American Society for Engineering Education Annual Conference, Session 2513, June, Anaheim, CA.

JiangFan Feng He received his B.S. degree from Southwest Agricultural University, and his Ph.D. degree from Nanjing Normal University, in 2002 and 2007. He works as associate professor of Chongqing University of Posts and Telecommunications. His main research area include spatial information integration and multimedia geographical information system.

Semantic Description of Web Services

Thabet Slimani
CS Department, Taif University, P.O.Box 888, 21974, KSA

Abstract
The tasks of semantic web service (discovery, selection, composition, and execution) are supposed to enable seamless interoperation between systems, whereby human intervention is kept at a minimum. In the field of Web service description research, the exploitation of descriptions of services through semantics is a better support for the life-cycle of Web services. The large number of developed ontologies, languages of representations, and integrated frameworks supporting the discovery, composition and invocation of services is a good indicator that research in the field of Semantic Web Services (SWS) has been considerably active. We provide in this paper a detailed classification of the approaches and solutions, indicating their core characteristics and objectives required and provide indicators for the interested reader to follow up further insights and details about these solutions and related software.
Keywords: *SWS, SWS description, top-down approaches, bottom-up approaches, RESTful services.*

1. Introduction

SWS research has as an objective to combines the services with the aim to achieve given goals. Based on goal descriptions and descriptions of available services, a complex service yielding the desired result is composed automatically. SWS research represents a new line of research on service descriptions and their exploitation. The annotation of services with a description using a formal ontology to express their precise mathematical meaning represents the basic idea of services description in the context of the Semantic Web.

The use of semantics is very useful to enables rich support for handling services. Furthermore, the use of ontologies to annotate services allows a higher degree of automation (describes the services in more formal detail).

The main goal of Semantic Web Services approaches is the automation of service discovery and service composition in a SOA [1].

In the last decade, several approaches have been proposed in the literature and these approaches differ in terms of the formalizations and implementations (Ontology language syntaxes) and in terms of the paradigms proposed for employing these in practice.

This paper is dedicated to provide an overview of these approaches, expressing their classification in terms of commonalities and differences. It provides an understanding of the technical foundation on which they are built. These techniques are classified from a range of research areas including Top-down, Bottom-up and Restful Approaches.

This paper does also provide some grounding that could help the reader perform a more detailed analysis of the different approaches which relies on the required objectives. We provide a little detailed comparison between some approaches because this would require addressing them from the perspective of some tasks supported with Semantic Web Services descriptions (i.e., discovery, invocation, composition, etc) and would also require taking into account the frameworks and developed applications.

The remainder of this paper is organized as follows. Section 2 introduces some principles for Semantic Web Service approaches and present in brief the vast popular of those that have been proposed over the years classified into top-down, bottom-up, and Restful approaches. In Section 3 we provide some information whereupon one could make a more efficient comparison and specified evaluation. This section also provides an organized perspective over the state of the art in Semantic Web Service approaches that can better help understand the evolution of the field. Finally, section 4 provides a conclusion and perspectives for future works.

2. Classification of semantic Description of Web Services

The existence of interoperable set of technologies for communication is required for Internet-scale distributed computing. There are currently two major alternative directions in these technologies, named “WS-*” and “REST”. The WS-* set of specifications uses the messaging paradigm and specialized service interfaces, with standardized infrastructure protocols (e.g. for security, transactions etc.). The REST direction relies on the architectural style of the World Wide Web and it views Web services as sets of resources accessible through the uniform interface of HTTP. WS-* technologies are mostly

deployed within enterprises (and behind firewalls), while the public Web is an increasingly large repository of RESTful services.

Web services in the semantic web are enhanced using rich description languages based on Description Logics (DLs) such as the Web Ontology Language (OWL). However, web services that have been enhanced with formal semantic descriptions is the definition of semantic web services. We distinguish two tested and validated approaches for WS-* technologies in addition to the approach based on REST technologies: Top-Down and Bottom-Up approaches for semantic web services. Top-down approaches are related to the development of semantic web services and are based on the definition of high-level ontologies providing expressive frameworks for describing Web services. On the other hand, bottom-up models, have been adopted an incremental approach that includes semantics to existing Web services standards by adding specific extensions which connects the syntactic definitions to their semantic annotations. Furthermore, the bottom-up approach represents an extension of existing standards and technologies including semantic annotations rather than the entirely services modeling based on ontologies.

If the technical or engineering point of view of a system or an organization seems to be clear and well proved through the history of technology dissemination, then the "top down" strategy is adopted: when all parameters are defined in detailed manner, before implementation, then systems operation works out best. This is the conceptual model for any top down strategy and as application it may be applied to e-government interoperability. As example of e-government application, a powerful administrative organization can be located at the top of hierarchy (e.g. a national government or its agency) and advises the interoperability methods and resources to be applied by all the actors on lower levels, supplements may be made on lower levels respectively.

The bottom up strategy is adopted if everyone concerned bring in his/ her requirements and specifications, and we will find a solution for achieving interoperability within the network which is acceptable for all involved, based on these requirements. For example, if local administrative organizations publish their services interfaces and use his/her individual ontologies, then some joint or mutual service should resolve some technical, syntactic and semantic differences as much as possible. As example of e-government application, administrative organizations can be located at the bottom of the hierarchy which recommend and share interoperability methods and resources from their point of view; and furthermore, the centralized direction is only accepted when there is agreement on all lower levels.

As a Web service domain, we consider both commercial and governmental Web services. A case study based on analysis of 493 commercial and 96 governmental Web service operations has been conducted in the work of Kungas and Matskin., 2006 [2] and the result of the analysis of the interaction between commercial and governmental Web services turned out that while ontologies enhance the usage of the commercial Web services, they have no significant impact on the governmental Web services. However, ontologies facilitate automation of semantic integration of commercial Web services with governmental ones. Based on this analysis, we say that TOP-Down approaches are useful when we faced with commercial Web services use.
Additionally, this idea is confirmed by the work presented in [3] which says that the existence of a web services description in a machine-understandable fashion is expected to have a great impact in areas of e-Commerce and Enterprise Application Integration (EAI).
In the remainder of this section, several languages have been presented and classified.

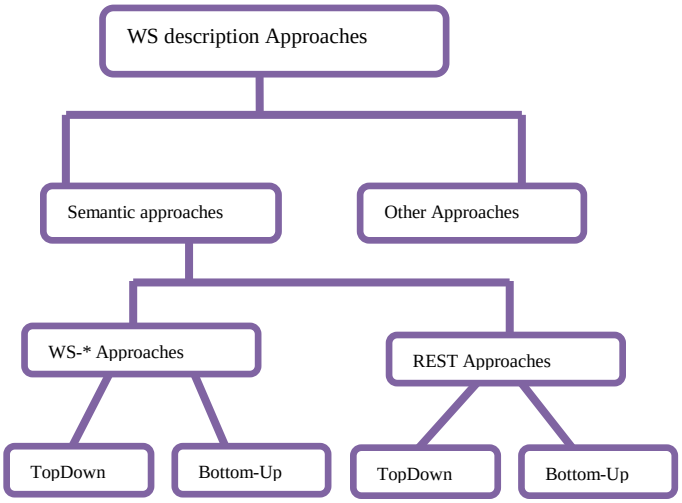


Fig. 1 WS description approaches taxonomy.

2.1 Approaches Using WS-* Technologies

2.1.1 Top-Down Approaches

The term Top-Down means that semantic web services are written directly in a formal language and don't have any dependence to any non-semantic web services. All semantic web services technologies should be able to connect with non-semantic web services (called grounding)

in order to enhance any web service system development. The ability to build new SWS with no relation to the classic web services technologies is the needed features that should characterize this approach. Several languages have been presented for Top-Down approaches:

- **OWL-S** [4] it is mainly a North American development effort, based on the OWL ontology language. The OWL Services (OWL-S) ontology defines an OWL ontology composed by a set of essential vocabularies to describe the "semantics" of Web services. This semantics includes the definitions of the capabilities, requirements, internal structure and interactions details with the service.
- **WSMO** [5] [6] it is a project developed within EU-funded projects (Sekt, DIP, Knowledge Web, ASG and SUPER projects) based on the WSML [6] ontology language. WSMO is a framework for Semantic Web Services that represents a top-down model identifying semantics of web service that uses the WSML (Web Service Modeling Language) language for describing domain-specific semantic models. The description of functional capabilities of services using logical expressions as preconditions, assumptions, postconditions and effects are required by WSMO.
- **SWSL**: SWSL is used to specify the semantics of web services concepts and descriptions as well as individual web services. It includes two sublanguages: ***SWSL-FOL*** is based on first-order logic (FOL) and is designed primarily to express the formal characterization (ontology) of Web service concepts. ***SWSL-Rules*** is based on the logic-programming (or "rules") paradigm and is designed to support the actual language for service specification that use the service ontology in reasoning and execution environments based on that paradigm.
- **DIANE**: DIANE is a framework that allows the automation of the discovery, composition, binding and invocation of services [7]. The framework is based on DIANE Service Description (DSD) and a specialized ontology language for describing service elements called DIANE Elements. DIANE elements exploit the notions of attributes, and reuse the clean separation between schema and instances promoted by description logics. Furthermore, special constructs are included in DIANE elements to describe service such as declarative and fuzzy set as well as variables.
- **SWSO**: The Semantic Web Services Ontology (SWSO) is a part of SWSL language [8], which includes formal conceptual definitions and individual web services. The definition of semantics of the theoretic model of the ontology of SWSO is based on the description of the ontology services, and the description of a first-order logic (FOL) axiomatization (FLOWS - the First-order Logic Ontology for Web Services). The aim of the created service descriptions enable automated discovery, composition, and verification, as well as the creation of declarative descriptions of a Web service that can be mapped to executable specifications.
- **COWS**: The Core Ontology of Web Services (COWS) is based on the Core Ontology of Software Components [9]. To enable extensibility and facilitate reuse, the fundamental concepts of COWS are separated in core ontology. The Core Ontology of Software Components is based on fundamental concepts and associations like software, data, users, policies and so on.
- **MSM**: Minimal Service Model (MSM) introduced together with hRESTS [10] is a simple RDF vocabulary covering what can essentially be considered the core of WSDL. It defines basically *Services* characterized by a number of *Operations* which have an *Input*, an *Output*, and *Faults*. Furthermore, MSM has subsequently been used as a means to integrate heterogeneous services (i.e., WSDLs and Web APIs). The combination between MSM and WSMO-Lite can provides a common framework covering the largest common denominator of the most used SWS formalisms on the Web. With this combination generic publication and discovery machinery has been developed that supports SAWSDL, WSMO-Lite, hRESTS/MicroWSMO, and OWL-S services [11].
- **ServONT**: is an ontology-based hybrid approach designed to improve the effectiveness and the efficiency of service discovery. For this matter, additional semantics is associated to the service ontology **ServOnt**, which organizes services at different levels of abstraction by means of semantic relationships that can be fruitfully exploited to support service discovery. Starting from the bottom layer, we distinguish between *Concrete Services*, *Abstract Services* and *Service Categories*, organized into *Concrete*, *Abstract* and *Category* layer, respectively [12].
- **SSWAP**: Simple Semantic Web Architecture and Protocol (SSWAP) is the driving technology for the iPlant Semantic Web Program[1]. It combines

[1] http://sswap.info

Web service functionality with an extensible semantic framework to satisfy the conditions for high throughput integration [13]. SSWAP utilizes OWL ontologies to describe the features and capabilities of Web services and standard HTTP methods to invoke the services. The architecture of SSWAP is based on five basic concepts Provider, Resource, Graph, Subject, and Object. The Provider organization is the owner and the publisher of resources. The web pages, ontologies and databases, represent the resources which are used to describe services offered on the Web**.**

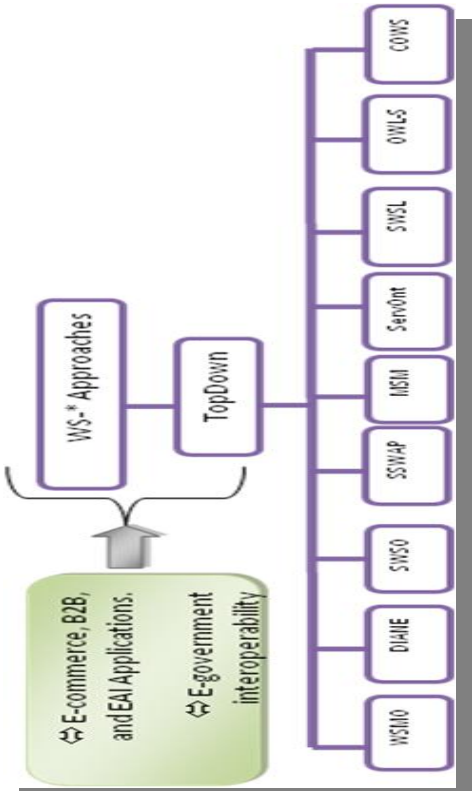


Fig.2 Top-Down WS-* Approaches

In order to illustrate a scenario of an application that adopts a top-down approach, we will briefly describe an application scenario based on [14]. Let us imagine a "Virtual Traveling Agency" (VTA for short) which is a platform providing eTourism services. These services can cover information services concerned with tourism such as events and sights in different areas and services that support booking of flights, hotels, rental cars, etc. By applying Semantic Web Services, a VTA can invoke Web services provided by several eTourism suppliers and aggregate them into new customer services in a semi-automatic fashion.

2.1.2 Bottom-Up Approaches

The aim of annotating Web Services is to add clarity in the Web Service definitions and also to allow the Web Service to be read by machines. This machine-readability increases the power of the SWS by adding the understanding of what the web Service is doing and the ability to interpret the messages that are interchanged. Semantic annotations of web service are used to automate service discovery, composition, mediation, and monitoring. We can state several approaches actually finished:

- **WSDL-S** [15]: WSDL-S specification is a W3C member submission that defines annotations to WSDL documents. The approach of semantic annotation consists in directly annotating the WSDL with semantic information. Semantic annotations that reference concepts in an ontology define the meaning of the inputs, outputs, preconditions and effects of the operations described in a service interface.
- **SAWSDL**[1] [16]: SAWSDL is s a W3C proposed recommendation where the semantic annotations use an extended attributes called modelReference so that relationships between WSDL components and concepts in another semantic model (e.g. ontology) are handled. Hence, the separation of semantic annotation mechanism from the representation of the semantic descriptions makes SAWSDL an approach independent of the semantic representation language. As a result, developer's community has more flexibility to select their favorite semantic representation language, to reuse semantic domain models and annotate descriptions using multiple ontologies.

The described approaches present a main advantage of preparing annotation directly in the WSDL XML Schema. Other advantage is that these specifications are independent to ontology language. Both languages have the necessary development tools and are operational to model and run SWS.

- **METEOR-S**: METEOR-S is an effort to create Semantic Web processes, at the LSDIS lab, University of Georgia. METEOR-S is a framework for semi-automatically marking up web service descriptions with ontologies. It contains an algorithms development to match and annotate WSDL files with relevant ontologies. It provides a mechanism to add data, functional and QoS semantics to WSDL files [17].

[1] http://www.w3.org/TR/sawsdl/

- **MWSAF (METEOR-S Web Service Annotation Framework)** is a semantic web based graphical tool that enables you to annotate existing Web service descriptions with ontologies. It facilitates the parsing of WSDL files and ontologies. This enables the user to annotate Web service descriptions semi-automatically. MWSAF was formerly known as **SAWS (Semantic Annotation of Web Services)**. MWSAF offers various features for programmers looking to create Semantic Web services. It provides: a) a fast and easy method for annotating WSDL files with single or multiple ontologies, b) an intuitive graphical environment for viewing WSDL files as well as ontologies, c) support for RDF-S , DAML+OIL and OWL based ontologies and d) a good solution for selecting the correct domain ontology for annotation from several ontologies.

- **USDL:** Universal Service-Semantic Description Language "is a language for formally describing the semantics of Web services" [18]. The USDL common basis that understands the meaning of services is based on OWL and the use of WordNet. The first attempt of USDL is to capture the semantics of web-services in a universal, yet decidable manner [18]. *USDL is designed* based on two languages: WSDL and OWL and defines a generic class called Concept, which is used to define the semantics of messages parts. The *USDL* Concept class denotes the conceptual objects constructed from the OWL WordNet ontology.

- **ServFace**: The ServFace project [19] aims at creating a model-driven service engineering methodology for an integrated development process for service-based applications[1]. The aims of this approach is to add UI-related annotations to service descriptions, notably WSDLs, in order to better support the development of user interface and to build interactive service-based applications. This project includes the creation of new algorithms for the composition of annotated services to build interactive service based applications based on the user interface annotations.
- **GPO/PSAM**: The General Process Ontology (GPO) and the Process Semantic Annotation Model (PSAM) [20] define business process annotations. The GPO/PSAM approach has been developed into a complete and systematic semantic annotation framework and defines four perspectives: basic description of process models (profile annotation), process modeling languages (meta-model annotation), process models (model annotation) and the purpose of the process models (goal annotation). Profile annotations are basic process description and include the following groups: administrative (e.g., creator, publisher), descriptive (e.g., title, category), technical (modeling language), preservation (documentation) and use (e.g., used in). Meta-model annotations include typical business process constructs such as: **Activity, Actorrole, Input, Output, Merge, Join,** and others. Model annotations use process modeling ontology as metadata to annotate the semantics of constructs in a modeling language. Goal annotations are used to specify aims of business process activities with distinction on *local* and *global* goals.

- **QuASAR [2] / ISPIDER:** The goal of Quality Assurance of Semantic Annotations for Services (QuASAR) [21] is to support the full life-cycle of Web service annotations and to ensure trustworthiness and accuracy of annotations. QuASAR / ISPIDER approach explores the potential uses of an additional source of information about semantic annotations: namely, repositories of trusted data-driven workflows. A workflow is a network of service operations, connected together by data links describing how the outputs of the operations are to be fed into the inputs of others. If a workflow is known to generate sensible results, then it must be the case that the operation parameters that are connected within the workflow are compatible with one another (to some degree). Semantic annotations have been proposed as a means of providing richer information about the behavior of Web services to potential users [21]. Three proposed ontologies of terms used in of service annotation[3]: *Domain ontology*, *Representation ontology* and *Extend ontology*. Domain ontology represents service annotations from similar a domain (e.g., biomedical services and others) that describes common concepts relevant within a given domain. The description of the representation format of service parameters is obtained by the Representation ontology. Extend ontology describes scopes of values of service parameters.

[1] http://www.servface.eu/

[2] http://img.cs.manchester.ac.uk/quasar/
[3] http://img.cs.manchester.ac.uk/quasar/

Information about scopes of values helps to detect incompatibilities between well formed services.

- **BPEL4SWS** :

  BPEL4SWS [22] is a language for Semantic Web Service orchestration based on Business Process Execution Language (BPEL). BPEL is an orchestration language that defines business processes interacting with external entities through web service operations using WSDL. BPEL4SWS extends BPEL and enables the definition of process logic independently from WSDL specific details. It is useful for orchestration of both Web services and Semantic Web Services. Semantic annotations can be attached to any part of BPEL4SWS descriptions. It allows the functionality descriptions or requirements of activities of a process semantically using SWS frameworks such as WSMO or OWL-S instead of using WSDL. BPEL4SWS also makes use of the SAWSDL standard for handling data lifting and lowering and enables bridging the gap between XML data and ontologies and enables semantic service discovery using appropriate middleware such as SEE during runtime.

- **YASA4WSDL**: Yet Another Semantic Annotation (YASA) for WSDL [23] proposes an extension of SAWSDL. YASA4WSDL includes two types of ontologies: The first one is a *Technical Ontology* containing concepts for ontologies describing service concepts (interface, input, output) and ontologies describing non functional concepts of services (ex. QOS attributes). The second type is a *Domain Ontology* that covers a business domain. YASA claims that introducing *serviceConcept* attribute makes SAWSDL descriptions more expressive and allows to explicitly capturing information on service pre-, post-conditions and effects. The separation of semantic annotation mechanism from the representation of the semantic descriptions makes SAWSDL an approach independent of the used semantic representation language.

- **WSMO-Lite:** Has been created due to a need for lightweight service ontology which would directly build on the newest W3C standards and allow bottom-up modeling of services. WSMO-Lite adopts the WSMO model and makes its semantics lighter and allows the use of any ontology language with RDF syntax. WSMO-Lite only defines semantics for the information model, functional and nonfunctional descriptions (as WSMO Service does) and only implicit behavior semantics.

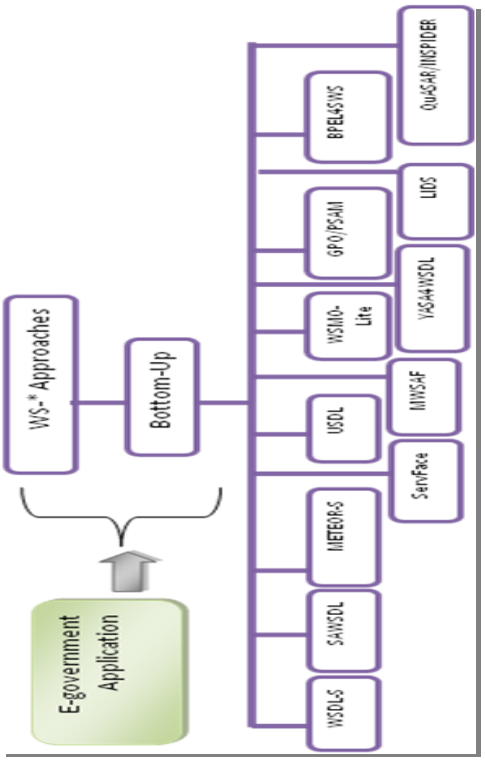


Fig. 3 WS-* Bottom-up Approaches Taxonomy.

- **LIDS:** Linked Data Services (LIDS) [24] denote the integration of data providing services and linked data and represents a lightweight service description model where service inputs and outputs are specified using SPARQL graph patterns. It focuses on the integration of existing data services exposed with Linked Data principles through Web APIs. Furthermore, the Web standards such as HTTP, RDF and SPARQL represent the base of LIDS. In addition to its accessibility over HTTP protocol, LIDS consume and produce RDF triples. LIDS can be directly used by Linked Data consumers and any requirement for data lifting.

2.2 Approaches Using REST Technologies

2.2.1 Top-Down Approaches

RESTful services are currently facing similar limitations to those identified for traditional Web service technologies and present even further difficulties, such as the lack of machine-processable service descriptions. Traditional Web service technologies have a somewhat longer history of research on semantic descriptions and annotation approaches; research in the area of semantic RESTful services is newer and therefore relatively limited. In order to address these challenges and to enable the wider adoption of RESTful service technologie, the following approaches have been developed.

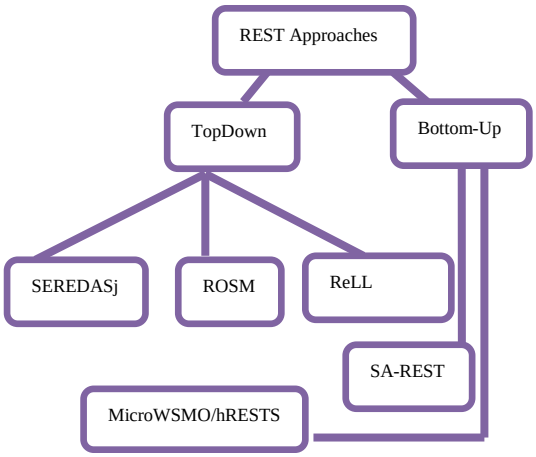


Fig. 4 **RESTful** Approaches Taxonomy.

- **ROSM:** The Resource-Oriented **Service Model (ROSM)**[1] ontology is a lightweight approach to the structural description of resource-oriented (RESTful) services. The use of ROSM enables the annotation of resources included in a service. Furthermore, a resource can be described as a part of collections and accompanied with addresses (URIs) intended for access and manipulation. A resource can be organized in collections, allowing the capture of an arbitrary number of resources and attaching service semantics to them following the SAWSDL approach.

- **SEREDASj**: stands for SEmantic REstful DAta Services, while the "j" should high-light that the approach is based on JSON (this leaves the door open for other data formats). SEREDASj semantically describe RESTful Data Services which in consequence leads to a mechanism to transform the data provided by such services to semantic resources. This aims to contribute to the availability of more semantic datasets [25].

- **Rell**: ReLL [26], the Resource Linking Language, does exactly the opposite. ReLL is a language to describe RESTful services with the aim to transform their exposed data to RDF and thus allowing harvesting already existing Web resources. Currently ReLL does not support any modification of the described re-sources, i.e., at the moment it supports only HTTP GET operations. This clearly restricts the possible use cases of ReLL at this point in time.

2.2.2 Bottom-Up Approaches

We consider the following bottom-up approaches:

- **MicroWSMO/hRESTS**:MicroWSMO [10] is a formalism for the semantic description of RESTful services, which is based on adapting the SAWSDL approach that adds sematic annotations. MicroWSMO uses microformats for adding semantic annotation ro service properties on top of HTML service documentation, by relying on hRESTS (HTML for RESTful Services) (Kopecky et al. 2008) that introduces the service model structure (service, operations, input, output) that allows the descriptions machine-processable. hRESTS enables the annotation of service operations, inputs and outputs, HTTP methods and labels, by inserting HTML tags within the HTML. MicroWSMO enables the identification of RESTful services and brings them to a level where they can be more easily discovered, composed and invoked.

- **SA-REST**: Semantic Annotations for REST (SA-REST) [27] is an open, standards-based approach which adds semantic annotations to RESTful services and Web APIs [15]. SA-REST defines three basic properties that can be used to non-intrusively annotate HTML/XHTML documents, typically to embed ontological meta-data[2]: The

[1] F. F. and N. B. D3.4.6 MicroWSMO v2 – Defining the second version of MicroWSMO as a systematic approach for rich tagging. Soa4all project deliverable.

[2] http://www.w3.org/Submission/2010/SUBM-SA-REST-20100405/

domain-rel property that provide domain information descriptions for a resource. The main objective of this annotation is to provide coarse grained categorizations of the HTML elements. The *sem-rel* property, which refers to the popular rel tag, and used to capture the semantics of a link within an HTML document. This kind of annotation is supposed to be used only within an anchor element (<a>). Finally, the *sem-class* property can be used to single entity annotation within a resource.

3. Comparison of SWS Approaches Functionalities

In the previous section we have briefly introduced the different approaches proposed in the literature, providing a basic description and pointers for the interested reader. In this section, we provide a comparison between these models in terms of their goal of development, their representation language, their conceptual influences and the year they were proposed in. This comparison is located in the Tab 1.

The lack of freely offered services and the acquisition of service descriptions, or the complexity of this task is the major limitation of the efforts described before. Some recent efforts aim to resolve this problem by reducing the complexity of the models and the acquisition task, by using simple RDF(S) vocabularies and Linked Data. These recent approaches present some promising results that could certainly be beneficial for the SWS paradigm.

OWL-S and WSMO is fully edged semantic framework, but WSDL and SAWSDL lack the support for semantic description. The most mature and commonly used in service discovery and composition approaches is OWL-S. But OWL-S presents some drawbacks as stated in [28]:
The process model of OWL-S is neither an orchestration model nor a choreography model. Moreover, OWL-S views Web service description does not consider asynchronous communication, and take into account only synchronous communication. The process model of WSMO offers both an orchestration and choreography view, but the orchestration view is rather primitive and the WSMO Choreography model contains transition rules which represent only local constraints. Furthermore, WSMO hasn't been around as long as OWL-S.

None of the approaches described in the Table 1 provide a complete solution according to the dimensions illustrated, but interestingly WSMO shows complementary strengths because it allows several goals (Discovery, Composition, Invocation, Orchestration and Mediation) and partially SWSO which not allows only the Mediation process.

Additionally the characteristic of UPML, OWL-S, DIANE, GPO and SWSO is very interesting being given that they allows functional, non-functional, informational and behavioral descriptions.

4. Conclusions

Research on SWS has produced several conceptual models, languages, architectures and algorithms that express the potential of these technologies for the Web and organizations. In this paper we have provided an initial description of these works according a number of dimensions. This paper is a first step that presents a breadth of the field, principally in terms of the tasks that could be supported by means of SWS descriptions, allowing a good classification and a comparison in the field. The use of SWS on the Web is unusual and it looks like that the intelligent techniques of the Web that act to the users profit remains as indicated by the reputation of publicly available Web APIS and RESTful services.

It is required to use the domain ontologies, the services taxonomies and in some cases to include complicated logical expressions, in order to create a rich semantic description of a Web service.

References

[1] *S.McIlraith, T. Son, H.* Zeng. Semantic web services. *Intelligent Systems*, 2001, 16(2), pp. 46–53.

[2] P. Kungas; M. Matskin. Web Services roadmap: the Semantic Web perspective, Proceedings of the International Conference on Internet and Web Applications and Services, 2006.

[3] R. Dumitru, K. Uwe, L. Holger, d.B. Jos, L. Rubén, S. Michael, P. Axel, F.ier . Cristina, B. Christoph, and F. Dieter: *Web Service Modeling Ontology*. Applied Ontology, 1(1):77-106, 2005, IOS Press.

[4] D.Martin , M.Burstein, G. Denker, J. Hobbs, L. Kagal, O. Lassila, D. McDermott , S. McIlraith, M. Paolucci, B. Parsia, T. Payne; M. Sabou, E. Sirin, M. Solanki, N. Srinivasan, and K. Sycara,. DAML-S (and OWLS) 0.9 draft release, 2003.

[5] D. Fensel, M. Kifer, de Bruijn J. and Domingue J. Web service modeling ontology (wsmo) submission, w3c member submission, 2005.

[6] J. De Bruijn, C. Bussler, , J. Domingue, D. Fensel, M. Hepp, U. Keller, M. Kifer, B.K onig-Ries, J. Kopecky, R. en Lara , H. Lausen, E. Oren, A. Polleres, D. Roman, J. Scicluna, and M. Stollberg. Web service modeling ontology (WSMO). W3C Member Submission, June 2005.

[13] D. Gessler , G. Schiltz, G. May, S. Avraham, C. Avraham, D. Grant and R. Grant. SSWAP: A Simple Semantic Web Architecture and Protocol for semantic web services. *BMC Bioinformatics*, 10(1):309, 2009.

[14] M. Stollberg, H. Lausen, A. Polleres and R. Polleres. *WSMO Use Case Modeling and Testing*. WSMO Deliverable D3.2, WSMO Working Draft, 2004, latest version available at http://www.wsmo.org/TR/d3/d3.2/.

[15] R. Akkiraju, J. Farrell, J. Miller, M. Nagarajan, M.-T Schmidt, A. Sheth and K. Sheth. Web Service Semantics - WSDL-S. http://www.w3.org/ Submission/WSDL-S/, W3C Member Submission, 2005.

[16] J. Farrell and H. Farrell. Semantic Annotations for WSDL and XML Schema (SAWSDL). Recommendation, W3C, 2007.

[17] A.A. Farrell, S.A. Oundhakar, A.P. Sheth, and K. Sheth. Meteor-s web service annotation framework," in WWW '04: Proceedings of the 13th international conference on World Wide Web. New York, NY, USA: ACM Press, 2004, pp. 553-562. Available at http://dx.doi.org/10.1145/ 988672.988747.

[18] A. Sheth, S. Kona, L. Simon and T.D. Simon. A universal service-semantics description language. In *Proceedings of the Third European Conference on Web Services*, ECOWS '05, pages 214–, 2005, Washington, DC, USA, IEEE Computer Society.

[19] Tobias, N., F. Maruis, H. Gerald, P. André, and J. Uwe . The ServFace Builder - A WYSIWYG Approach for Building Service-Based Applications. In: *ICWE*, Vol. 6189, Springer, p. 498-501, 2010.

[20] Y. Lin, D. Strasunskas, S. Hakkarainen, J. Krogstie and A. Sølvberg. Semantic annotation framework to manage semantic heterogeneity of process models. In E. Dubois and K. Pohl, editors, *CAiSE*, volume 4001 of *Lecture Notes in Computer Science*, pages 433–446, 2006.

[21] K. Belhajjame, S.M. Embury, N.W S.M., R. Stevens, and C.A, Stevens. Automatic annotation of Web services based on workflow definitions. *ACM Trans. Web* 2, 2, Article 11 (May 2008), 34 pages. DOI=10.1145/1346237.1346239 http://doi.acm.org/10.1145/1346237.1346239.

[22] J. Stevens, T. van Lessen, D. Karastoyanova, and F. Leymann. BPEL for Semantic Web Services (BPEL4SWS). In *On the Move to Meaningful Internet Systems (OTM Workshops), 2007*.

[23] Y. Chabeb and S. Tata. Yet Another Semantic Annotation for WSDL (YASA4WSDL). In *IADIS WWW/Internet 2008 Conference*.

[24] S. Speiser and A.H. Taking. the LIDS off data silos. In *I-SEMANTICS*, 2010.

[25] M. Lanthaler and C. Gütl. A semantic description language for RESTful data services to combat semaphobia, in 5th IEEE International Conference on Digital Ecosystems and Technologies (DEST), May 31 - Jun 03, 2011. Daejeon, South Korea: IEEE.

[26] R. Alarcón and E. Wilde. Linking data from RESTful services, in Proc. of the 3rd Workshop on Linked Data on the Web (LDOWS 2010).

[27] Gomadam K., Ranabahu A. and Sheth A.. SA-REST: Semantic Annotation of Web Resources. Member submission, W3C, April 2010.

[28] S. Balzer, T. Liebig, and M. Wagner. Pitfalls of OWL-S - A Practical Semantic Web Use Case. In ICSOC' 04: Proceedings of the 2nd International Conference on Service Oriented Computing, pages 289-298, New York, NY, USA, November 2004.

Thabet Slimani graduated at the University of Tunis (Tunisia Republic) and defended PhD. thesis with title "New approaches for semantic Association Extraction and Analysis". He has been working as an assistant Professor at the Department of Computer Science, Taif University. He is a member of Larodec Lab (Tunis University). His interests include semantic Web, data mining and web service. He is author of more than 20 scientific publications.

Tab. 1 Comparison between SWS approaches functionalities.

Approach	Year	Conceptual Input	Language	Goal				
				Discovery	Composition	Invocation	Orchestration	Mediation
UPML	1999	PSMs	UPML/LISP	Knowledge-Based Systems development				
DAML-S /OWL-S	2001	Agents	Knowledge-Based Systems development	Semantic annotations of WS				
DIANE	2004	OWL-S	DIANE Elements	√	√	√	**x**	**x**
SWSO	2005	OWL-S	SWSL	√	√	√	√	**x**
USDL	2005	OWL-S	OWL	Language for formally describing the semantics of Web services				
WSDL-S	2005	WSMO, OWL-S, WSDL	XML Shema	Linking semantic annotations to Web services				
WSMO	2005	WSMF, UPML	WSML, RDF	√	√	√	√	√
COWS	2006	DOLCE	OWL	Semantic management of middleware				
GPO	2006	UEMO	OWL	Process modeling				
QuASAR	2006	myGrid	OWL	Integrated platform enabled as Grid and Web services for the storage, dissemination and management of proteomic data				
WSO	2006	OWL-S, WSMO, WSBPEL	OWL	x	√	x	x	x
BPEL4SWS	2007	BPEL4WS, WSMO, SAWSDL	XML Shema	x	x	x	√	x
SAWSDL	2007	WSDL-S	XML Schema	Semantic Annotations for WS WSDL and XML Schema				
FUSION Ontology	2008	SAWSDL, UDDI	OWL-DL	Service registry				
YASA	2008	SAWSDL	XML Schema	Extension of SAWSDL, service discovery				
MicroWSMO/ hRESTS	2008	hRESTS/WSMOLite	HTML with microformat tags	Semantic annotations of RESTful services and Web APIs				
MSM	2008	WSDL, WSMOLite, hRESTS	RDF(S)	√	x	√	x	x
ServONT	2008	OWL	OWL-DL	√	x	x	x	x
WSMO-Lite	2008	SAWSDL, OWL-S, WSMO	RDF(S)	√	√	√	x	x
ServFace	2009	WSDL	XML Schema	For adding of UI-related Annotations to Web service Descriptions (WSDL)				
SSWAP	2009	HTML, Semantic MOBY	OWL	Data and service integration in Biology				
SA-REST	2010	SAWSDL, hRESTS	RDFa	Semantic annotations of RESTful services				
ER Model	2010	ER, BPEL	ER, OWL DL	√	√	x	x	x
LIDS	2010	HTTP, Linked Data	SPARQL	Bridging the gap between data services and Linked Data principles. Lightweight composition				
RELL	2010	REST	RDF / OWL	Description of resource-centered Web APIs in terms of resources				
ROSM	2010	WSMO-Lite, REST	RDFS, SPARQL	Description of resource-centered Web APIs (RESTful services)				
SEREDASj	2011	JSON-LD, REST	JSON, RDF, FOAF	Semantic description of Restful Data Services				

Research on Two Algorithms of Solving Large-scale Tridiagonal Linear Equations

Yu Bencheng[1],Chen Yan[2]

[1]Information and management institute of technology, Xuzhou college of industrial technology ,Xuzhou, China

[2]Information and management institute of technology, Xuzhou college of industrial technology ,Xuzhou, China

Abstract

Based on the analysis of the two kinds of algorithms in solving large-scale tridiagonal linear equations, which are linear interpolation method and the method of double parameters, it is shown that the principle of the linear interpolation method and double parameter method is consistent and it points out that in this principle, the solutions to certain types of tridiagonal equations in the two methods are not stable. But in the case of not so sick, their relative errors of solution are very small, and the situation is very stable.

Key Words: *Tridiagonal Linear Equations, Complexity of algorithm, Stability, Algorithm.*

1. Introduction

There are many applications of tridiagonal equations whose general forms are shown as follows:

$$A_x = f \tag{1}$$

Among which:

$$A = [a_{i-1}, b_i, c_i]_{i-1}^{n} = \begin{pmatrix} b_1 & c_1 & & & \\ a_1 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & a_{n-1} & b_n \end{pmatrix} \tag{2}$$

$f = (f_1, f_2, \cdots, f_n)^T$, $a_i c_i \neq 0 (i = 1, 2, \cdots, n-1)$ .When A is a singular matrix[1], it has a unique solution to equations $A_x = f$.

For such equations, of course, we can use the Gaussian elimination method [2]. But when matrix dimension is very big, the calculation amount by the method of Gaussian elimination will be too large [3]. For its special structure, we can construct some special solutions so as to reduce the amount of calculation. There are a variety of special solutions to tridiagonal equations [4]. This paper mainly introduced two kinds of algorithms and compared in detail the calculation of the two algorithms. Based on the principle of the two methods, it pointed out the similarity between these two algorithms and the instability in the solutions to certain kind of equations.

2. Linear Interpolation Method

Linear interpolation process is described as follows:

The former n-1 of equations of $A_x = f$ is $A_{n-1}{}^{x} = f^{'}$.

f' is made up of the former $n-1$ element of f' .

The latter $n-1$ column vector of A_{n-1} is $\tilde{A}$, among which $\tilde{A}$ is a lower triangular matrix of $c_i (i = 1, \ldots, n-1)$.

For $c_i \neq 0 (i = 1, 2, \cdots, n-1)$, $\tilde{A}$ is a non-singular matrix.

As long as x_1 is given, we can get $\tilde{x} = (x_2, \cdots, x_n)^T$ by

solving $\tilde{A}\tilde{x}$, among which $\tilde{f} = (f_1 - b_1 x_1, f_2 - a_1 x_1, f_3, \cdots, f_{n-1})^T$. Thus we can get $x = (x_1, x_2, \cdots, x_n)^T$, which is a set of solutions to $A_{n-1} x = \tilde{f}$. Then two sets of different solutions are get by taking two different x_1, and the solution of equation $A_x = f$ is obtained through the linear combination of the two sets of solutions [5].

Assuming $x_1^{(0)} = 0$, by the solution of $\tilde{A}\tilde{x}^{(0)} = \tilde{f}$, we can get $x^{(0)} x = (x_1^{(0)}, x_2^{(0)}, \cdots, x_n^{(0)})^T$ which satisfies $A_{n-1} x^{(0)} = f'$. And similarly assuming $x_1^{(1)} = 1$, $x^{(1)} = (x_1^{(1)}, x_2^{(1)}, \cdots, x_n^{(1)})^T$ is get which satisfies $A_{n-1} x^{(1)} = f'$ [6].

For $\forall h \in C$, C is complex number field, $A_{n-1}\left[x^{(0)} + h(x^{(1)} - x^{(0)})\right] = f'$. As long as h is selected to make $x = x^{(0)} + h(x^{(1)} - x^{(0)})$ satisfy the n^{th} equation of $Ax = f$, the solution of $Ax = f$ is obtained and it is easily to get

$$h = \frac{f_n - a_{n-1} x_{n-1}^{(0)} - b_n x_n^{(0)}}{a_{n-1}(x_{n-1}^{(1)} - x_{n-1}^{(0)}) + b_n (x_n^{(1)} - x_n^{(0)})}.$$

3. Double Parameter Method

Double parameter method is slightly different for that x_1 is used as the parameter to determine the relationship between the other variables and x_1, then take a substitution equations $A_x = f$ to determine x_1, thus a solution to $A_x = f$ is obtained [7]. The process is described as follows:

$\tilde{x}$ (Shown as x_1) is calculated by $\tilde{A}\tilde{x} = \tilde{f}$:

$$\begin{cases} x_2 = (f_1 - b_1 x_1)/c_1 \\ x_3 = (f_2 - a_1 x_1 - b_2 x_2)/c_2 \\ \vdots \\ x_n = (f_{n-1} a_{n-2} x_{n-2} - b_{n-1} x_{n-1})/c_{n-1} \end{cases} \quad (3)$$

Take substitutions in a sequence series to change them into a form of containing only constant term and x_1.

Assuming $s_1 = 0, t_1 = 1$, then $x_1 = s_1 + t_1 x_1$, so we can get

$$x_i = s_i + t_i x_1 (i = 2,3,...,n) \quad (4)$$

Among which $s_i, t_i (i = 2,...,n)$ is undetermined parameters. Comparing $x_i = s_i + t_i x_1 (i = 2,3,...,n)$ and

$$\begin{cases} x_2 = (f_1 - b_1 x_1)/c_1 \\ x_3 = (f_2 - a_1 x_1 - b_2 x_2)/c_2 \\ \vdots \\ x_n = (f_{n-1} a_{n-2} x_{n-2} - b_{n-1} x_{n-1})/c_{n-1} \end{cases} \quad (5)$$

We can get recurrence relations of parameter group :

$s_i, t_i, i = 1,....,n.$
$s_1 = 0, s_2 = f_1 / c_1$
$s_i = (f_{i-1} - b_{i-1} s_{i-1} - a_{i-2} s_{i-2})/c_{i-1} (i = 3,4,...,n)$
$t_1 = 1, t_2 = -b_1 / c_1$
$t_i = -(b_{i-1} t_{i-1} + a_{i-2} t_{i-2})/c_{i-1} (i = 3,4,...,n)$

According to the n equation of the equation group $A_x = f$, the way of expression of x_1 can be obtained:

$$x_1 = \frac{f_n - a_{n-1} s_{n-1} - b_n s_n}{a_{n-1} t_{n-1} + b_n t_n} \quad (6)$$

Inserting $x_i = s_i + t_i x_1 (i = 2,3,...,n)$, we can get a solution to $A_x = f$.

4. The relationship between Linear interpolation method and the method of double parameters

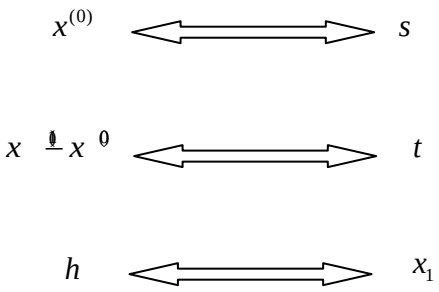
Through the comparison of the two methods, we can find the internal relationship among the parameters [8]. From the expression $x_i = s_i + t_i x_1 (i = 2,3,...,n)$, assuming x_1

is 0 and 1 respectively, we can get

$$x^{(0)} = (s_1, s_2, ..., s_n)^T$$
$$x^{(1)} = (s_1 + t_1, s_2 + t_2, ..., s_n + t_n)^T,$$

in which $x^{(0)}, x^{(1)}$ are shown in linear interpolation method and is shown in double parameter method. The relationship between the two methods is described as below [9]:

Linear Interpolation Double Parameter Method



Take double parameters method for example to go into its computational complexity. 3 n - 5 times multiplying and dividing is needed to calculate s and t, while n-1+5=n+4 times multiplying and dividing is needed to calculate x, so in total 2（3n-5）+n+4=7n-6 times multiplying and dividing is needed. So its computational complexity is: O (7 n).

The above two methods both require the solution to $\tilde{A}\tilde{x} = \tilde{f}$, so the stability of the two methods is closely related to the condition number of $\tilde{A}$. If $\tilde{A}$ is a morbid matrix, a lot of errors may appear by using these two methods to get the numerical solution. These two kinds of methods for tridiagonal equations can be used, but there may be numerical instability for some problems.

5. Numerical Example

Example： A_n is shown as below::

$$A_n = \begin{pmatrix} 3 & 1 & & & \\ 1 & 3 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 3 & 1 \\ & & & 1 & 3 \end{pmatrix}_n \quad (7)$$

Among them, n represents a matrix dimension [10].

The selection of the right end item f is like that, given a n column vector y at random, then left multiply A to get f.

The table below lists the relative error of various methods.

The formula of Relative error is $\in = \|y - x\|_2 / \|y\|_2$,

among which x is the solution by various methods and y is the exact solution of equations set。

Relative error of various methods is shown in table 1.

Table 1. Relative Error

n	chasing	linear interpolation	biparameter	QR decomposition
16	1.9442 E-16	6.3124 E-11	7.6431 E-11	4.3368 E-16
32	2.6459 E-16	2.0342 E-04	3.7341 E-05	4.4902 E-16
64	9.4034 E-17	1.6614 E+10	2.1380 E+09	3.8538 E-16
128	2.0887 E-16	2.3835 E+36	1.5808 E+35	3.0555 E-16
256	2.2368 E-16	*	*	3.3981 E-16
512	2.2412 E-16	*	*	3.4561 E-16
1024	2.2935 E-16	*	*	3.1418 E-16

As seen from the table, although the conditions of A is little (when n=1024, $\|A\|_2 < 5$), the result got from the linear interpolation method and biparameter method may not be so good.

The equations of the lower triangular $\tilde{A}\tilde{x} = \tilde{f}$ should also be required to be solved. Its coefficient matrix $\tilde{A}$ is n - 1 order matrix after getting rid of the first column and the last line of matrix A. So the stability of the two methods has something with the condition numbers of A.

In this example, when n=32,the condition numbers of $\tilde{A}$ are 6.1989E+13，and when n=64，they are up to as high as 1.2721E+17. So it is obvious that it is a sick equation, which leads to the instability of the linear interpolation method. So when the two methods of the linear interpolation method and biparameter method are used, it is required to check if the condition numbers of $\tilde{A}$ are small enough.

6. Conclusion

Through the contrast between the linear interpolation method, biparameter method and chasing method, QR decomposition method, it is apparent that as long as the solution of the problem is not sick, the relative errors of the solutions are very small and stable. In other cases, the calculation speed and algorithm complexity of double parameter method and the linear interpolation method are small, so in actual application process the two methods are very valuable.

7. References

[1]Duan Zhijian, Lv Quanyi, Ma Xinrong, "parallel alternating direction algorithm of linear equations", Computer engineering and application, Vol.45, No. 20, 2009, pp.54-56.

[2]Wu Jianping, Song Junqiang, "Parallel Incomplete Factorization Preconditioning of Block Tridiagonal Linear Systems with 2-D Domain Decomposition", Computational physics, Vol.26, No.2, 2009, pp.191-199.

[3]Cheng Haiying, Xie Jiang,Shao Huagang, "Comparison Between Two Parallel Methods in Solving Tridiagonal Equations, Computer applications and software.Vol.27,No.11,2010,pp.76-78

[4]DUAN Zhijian，YANG Yong, " LV Quanyi,et al. Parallel strategy for solving block-tridiagonal linear systems", Computer Engineering and Applications, Vol.47,No 13,2011,pp.46-49.

[5] LI Tai-quan，XIAO Bo-xun, "Iterated parallel diagonal dominant algorithm for tridiagonal systems", Journal of Computer Applications, Vol.32,No.10,2012,pp.2742- 2744

[6]LIU Yang, "Study of odd-even reduction parallel algorithm for block-tridagonal linear systems", Computer Engineering and Design, Vol. 30, No.13, 2009, pp.3193-3195.

[7]MIAO Sha, ZHENG Xiaowei, "Mult-icore parallel algorithm for cubic spline curve fitting", Journal of Computer Applications, Vol.30, No.12, 2010, pp. 3194-3196.

[8] Chaowei Jiang, Xueshang Feng, “A Unified and Very Fast Way for Computing the Global Potential and Linear Force-Free Fields”, Solar Physics,Vol. 281, No.2,2012,pp. 621-637.

[9]Kyle A. Gallivan, Efstratios Gallopoulos, Ananth Grama,Bernard Philippe,Eric Polizzi,Yousef Saad, Faisal Saied, Danny Sorensen, “Parallel Numerical Computing from Illiac IV to Exascale—The Contributions of Ahmed H. Sameh” ,High-Performance Scientific Computing, 2012, pp.1-44

[10]Ji-teng Jia, Qiong-xiang Kong, Tomohiro Sogabec, “A fast numerical algorithm for solving nearly penta-diagonal linear systems”, International Journal of Computer Mathematics, Vol.89, No.6,2012, pp.851-860.

A Light-weight Relevance Feedback Solution for Large Scale Content-Based Video Retrieval

Zimian Li[1], Ming Zhu[2]

[1] The Key Lab of Network Communication System & Control, The Chinese Academy of Sciences, The Key Lab of Network Communication System & Control, Anhui
Department of Automation, University of Science and Technology of China

[2] The Key Lab of Network Communication System & Control, The Chinese Academy of Sciences, The Key Lab of Network Communication System & Control, Anhui
Department of Automation, University of Science and Technology of China

Abstract

This paper addresses the problem of large scale content-based video retrieval with relevance feedback. We analyze the common methods which leverage local feature detectors to extract feature descriptors from video collections and perform multi-level matching after indexing and retrieval of feature vectors. Instead of learning similarity-preserving codes, we introduce the relevance feedback approach in a light-weight way. A relevance model is proposed to merge semantic similarity with the original distance matching at descriptor level. By learning several weights using canonical correlation analysis (CCA), the resulting candidate list of similar videos changes according to relevance feedback. Finally, we demonstrate the improvement of the proposed method by experiments on a standard real world dataset.

Keywords: *Content-based Video Retrieval, Relevance Feedback, CCA.*

1. Introduction

With the rapid growth of digital video content production on the web, content-based video retrieval (CBVR) has been receiving increasing attention over the last decade. In the computer vision and machine learning community, many approaches focus on multimedia information indexing and retrieval techniques. Compared with individual images, videos have much richer content and therefore need a more complicated structure to describe, index and retrieve.

Recently, different methods have been proposed for video structure analysis, including shot boundary detection, key frame extraction and scene segmentation. The general procedure of existing work can be summarized as three stages. First, using shot detection methods, videos are segmented into clips, which then represented by one or more key frames. Second, a set of high dimensional feature vectors are extracted by feature detector and descriptor. Finally, the similarity between videos is computed from the feature vectors under spatial and/or temporal sequence matching schemes, see [1] for a comprehensive review.

Unlike video copy detection (VCD) or near-duplicate video detection (NDD), content-based video retrieval searches for a more semantic sense of similarity, moreover, compared with content-based image retrieval (CBIR), some additional spatial/temporal information plays an important role in matching stage. So, how to measure the similarity and to perform nearest-neighbor search are the essential problems. In this paper, we leverage the common initial strategies and focus on the semantic retrieval with relevance feedback.

Approximate nearest neighbors (ANN) search methods are used to perform nearest neighbor search in large scale retrieval, especially for high dimensional datasets. One of the most popular techniques is Locality Sensitive Hashing (LSH), which was first introduced in [3]. LSH function families have the property that objects that are close to each other have a higher probability of colliding than objects that are far apart. For different distance measures, different LSH families have been proposed, e.g. LSH for p-norms based on p-stable distribution [4]. However, recent research [5] [6] shows that the Chi2 distance often leads to better results than Euclidean metric for image and video retrieval task, especially when histogram-based descriptors, such as SIFT and SURF, are used to describe the images and video frames. In this paper, we pursue the new LSH scheme fitted to the Chi2 distance, which was introduced by Gorisse [2] for approximate nearest neighbor search in high-dimensional spaces.

Another important question is how to extract and represent semantic information from video data. In the vision community, most recent approaches concern about learning similarity-preserving binary codes. Large scale image/video collections are represented in generally two

major steps: embedding and binarization. Regressing, classification and clustering techniques are merged with indexing and retrieval approaches to generate semantic structures, e.g. Principal component analysis (PCA) based method [7], Spectral Hashing [8], Kernelized LSH (KLSH) [9], Product Quantization (PQ) [10], Linear Discriminant Analysis based method LDAHash [11] and iterative quantization (ITQ) [12]. In all these approaches, compact binary codes are learned by training examples and the performance of similarity-preserving depends on the sample representativeness. What's more, they pay more attention to content-based image retrieval and less to video scenario. Compared with image retrieval, video retrieval has additional spatiotemporal characteristics and it's almost impossible to learn one compact code to represent a video clip totally.

In this paper, we follow the general procedure: leveraging local feature detectors to extract feature descriptors from video collections and perform multi-level matching after indexing and retrieval of feature vectors. Relevance feedback techniques based on canonical correlation analysis (CCA) are introduced to bridge the gap between semantic notions of search relevance and the low-level representation of video content. The main contribution is as follows:

- We analyze the indexing and retrieval stages in the common framework of content-based video retrieval.
- We leverage the state-of-the-art techniques in content representation, similarity measure selection and multi-level matching and merge them to work in an incremental way.
- We introduce a novel light-weight relevance feedback approach to refine the original resulting list.

The rest of this paper is organized as follows. In Section 2, we present the framework of content-based video retrieval with relevance feedback. Section 3 presents the structure to index SURF descriptors using Locality-Sensitive Hashing under Chi2 distance. Section 4 introduces the light-weight relevance feedback solution using Canonical Correlation Analysis (CCA). Section 5 gives the experimental results and performance analysis of our proposed algorithm. Finally we conclude this paper and give some future work in Section 6.

2. Content-Based Video Retrieval Framework

Figure 1 illustrates the framework of content-based video detection with relevance feedback. The processing consists of three parts: Indexing, Retrieval and Relevance Feedback. Indexing videos are processed by shot detection, key-frame and local feature extraction (we use SURF descriptors in this paper) to generate a set of 64-dimensional feature vectors. Then a video database is built using an indexing structure. In the retrieval parts, the same local features extraction is performed. By retrieving in the database a candidate result set is generated and then multi-level matching methods are applied to get the final similar video result list.

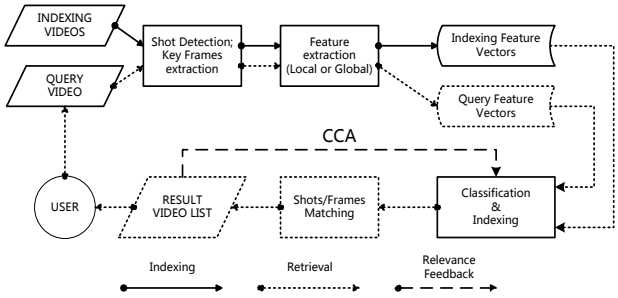


Fig. 1 Framework of CBVR with relevance feedback.

We focus on the indexing structure and relevance feedback techniques in retrieval for large-scale video collections. Different strategies have been proposed for local feature based near-duplicate video detection following the above framework. As to content-based video retrieval, which searches not the copy but the semantic similar ones, after indexing and retrieval in descriptor level, voting-based multi-level matching is not enough. Also, instead of leveraging users' feedback to learn a similarity measure or perform classification/clustering, we "delay" the semantic learning to the shots/frames matching stage.

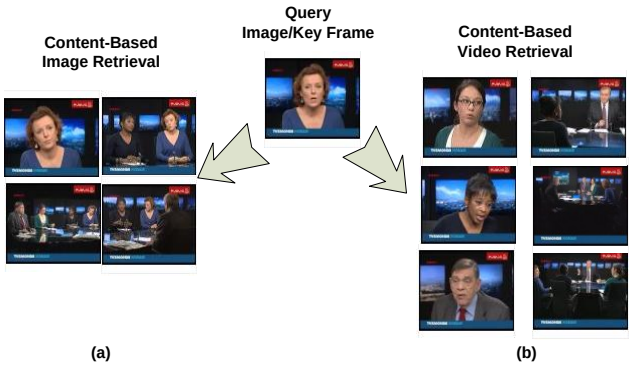


Fig. 2 Difference between CBIR and CBVR

Figure 2 shows the difference of content analysis between image retrieval and video retrieval. In (a) content-based image retrieval, where the query image contains all the semantic information, the results are more intuitive (mainly dealing with some transformations). However, in (b)

content-based video retrieval, where query video consist of many key frames(query images), for one query image, it's necessary to learn/find/create relations with other images extracted from the similar video and make up the semantic lost in representing video with image sequences.

3. Indexing SURF descriptors using Locality-Sensitive Hashing under Chi2 distance

SURF (Speeded-Up Robust Features) detector and descriptor [2] is based on calculating approximate Hessian response for image points and is efficiently implemented on the basis of integral images. SURF is proved to be equal or superior to performance and significantly better computational efficiency in comparison with other local feature methods, such as SIFT, PCA-SIFT. In this paper, we use 64-dimensional SURF descriptors as the feature vectors for key frames extracted from video collections. LSH (Locality Sensitive Hashing) is introduced in [3] for approximate nearest neighbors search in high dimensions. LSH function families have the property that objects that are close to each other have a higher probability of colliding than objects that are far apart. For different distance measures, different LSH families have been proposed. We consider the LSH for chi2 distance [2] since SURF descriptors are designed to be histogram-based descriptors measured by the Euclidean distance. We briefly describe the indexing structure in our scenario and show the characteristics of results after feature vector retrieving. In the basic LSH scheme, a query point is hashed into several buckets in different hash tables to retrieve all points in these buckets, then the distances to each point is computed using the chi2 distance (1):

$$\chi^2(x,y) = \sqrt{\sum_{i=1}^{d} \frac{(x_i - y_i)^2}{x_i + y_i}} \quad (1)$$

For each data point v, k independent hash functions of the form (2) are considered, where a is a d-dimensional vector whose elements are chosen independently from the Normal distribution and b is chosen uniformly from $[0, W]$.

$$h_{a,b}(p) = \frac{\sqrt{\frac{8a \cdot p}{W^2} + 1} - 1}{2} + b \quad (2)$$

Each hash function maps a d-dimensional data point onto the set of integers and the final result is a vector of length k of the form (3).

$$g(v) = \left(h_{a_1 b_1}(v), ..., h_{a_k b_k}(v)\right) \quad (3)$$

Thus, all points in dataset are hashed into buckets labeled by a k-dimensional vector in a hash table. To ensure the accuracy of similarity search, l independent hash tables are generated to construct the LSH indexing structure. Then, in the retrieval step, one can determine near-neighbors by hashing the query point l times to l buckets in l different hash table and retrieving elements stored in buckets containing that point.

Then the candidate feature vector set is used for retrieval on key-frame level and video shot level. Some recent approaches introduced some spatio-temporal matching and sequence matching approaches to further get the final list of similar video results. Note that in the framework based on local descriptor, the effectiveness of retrieving similar key frame (image) is insufficient for retrieving similar video. The improvements on indexing structure can only boost the retrieval efficiency. As mentioned above, similar videos have much more "semantic means" than similar images and the accuracy is defined fuzzy and depends mainly on user's opinions. We leverage the relevance feedback to learn an adaptive similarity score function, which works as a similarity measure in content-based video retrieval.

4. Relevance Feedback using Canonical Correlation Analysis

For each feature vector in each query video, a candidate set is retrieved from the indexing database. Each vector in the candidate set is associated with respective key frame and video shot. As in our scenario, the number of feature descriptors extracted from each key frame is about two hundreds, it's not computationally efficient to perform relevance feedback in descriptor level. We design a similarity function with a correlation matrix to project to semantic space in key frame level. We pick up key frames from the similar videos chosen by users, which are cached during Chi2-LSH retrieval. We represent key frames from the candidate videos as (4) and key frames from feedback video as (5):

$$F_i = \{f_i\}^n \quad (4)$$

$$F_f = \{f_f\}^n \quad (5)$$

where f is calculated by a voting method from below.

$$f_i = \sum_{N_{descriptor}} \sum_{L} w_b \cdot N_{matching} \quad (6)$$

$$f_f = \sum_{N_{descriptor}} \sum_{L} w_b \cdot N_{matching} \quad (7)$$

$N_{descriptor}$ is the number of descriptors extracted from the query frame, $N_{matching}$ is the number of matching descriptors between the two frames in one bucket and w_b is the weight of the corresponding bucket. We use the weight w_b to reduce the impact of large buckets, in which

many descriptors of the same frame match to one query descriptor of the query frame. Then we need a correlation matrix C.

$$C = \{c_{i,j} \mid i = 1,...,m;\ j = 1,...,n\} \quad (8)$$

The matrix C satisfies $F_f = CF_i$. The matrix could be learned by a supervised dimensionality reduction method to capture the result structure in semantic space.

We solve the problem by using the Canonical Correlation Analysis (CCA), which has proven to be an effective tool for extracting a common latent space from two views in a semi-supervised way. The goal of our approach is to find projection directions w_k and u_k for candidate key frame set and relevance key frame set to maximize the correlation between the projected $F_i w_k$ and $F_f u_k$. The problem is represented as:

$$\max C(w_k, u_k) = w_k^T F_i^T F_f u_k$$

subjected to:

$$w_k^T F_i^T F_i w_k = 1, u_k^T F_f^T F_f u_k \quad (9)$$

Solving the above optimal problem can use the generalized eigenvalue solution [15]. Once we get the canonical variables w_k and u_k, the optimal direction to project candidate set to relevance set is determined. Then we obtain the modified retrieval results using the similarity score function as:

$$score_c = \frac{\sum_{N_{frame}} f_m}{N_{frame}} \quad (10)$$

where $f_m \in F_m = CF_i$, N_{frame} is the number of key frames of the query video. An indexing video is considered to be a similar video of the query video if $score_c$ exceed a threshold S_t. The selection of S_t requires a trade-off between recall and precision during retrieval and is data dependent. Finally, the videos with first several highest scores are returned as the results of similar videos with relevance feedback considered.

5. Experiments

To evaluate the semantic effectiveness and robustness of our proposed algorithm, we conducted experiments using the MUSCLE-VCD benchmark [16], which is an evaluation set of the TRECVID 2008. The dataset consists of 101 videos with a combined length of about 100 hours. We divided the indexing videos into about 600 parts and the provided 15 query videos into about 100 parts. Then we performed 60 different similar video searches. Compared with the original process, we manually labeled a certain number of video parts from the candidate set as positive examples for relevance feedback. We modified the open source project E2LSH [4] provided by Alexandr Andoni to process locality sensitive hashing under chi2 distance. Experiments on precision/recall and percentage of relevance are made to illustrate the performance of our proposed method.

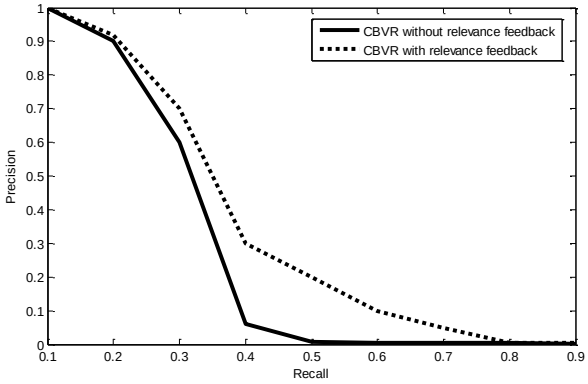


Fig.3 Precision-Recall Curve for CBVR with/without Relevance Feedback

Figure 3 shows the Precision-Recall curves on two different methods. Dot line is the CBVR with relevance feedback and plain line is the original CBVR. We can tell from figure 3 that the novel relevance feedback approach we propose can improve the recall percent for given precision. The recall percent increases 5% when the precision is below 50% and reaches up to 16.7% when the precision is above 50%. We can also see that the CBVR with relevance feedback can achieve higher precision for same recall percent. We can see that with a few similar videos labeled, when the recall percent increases, the precision decreases slower with the relevance feedback. The reason is that the original process ignores some of the similar videos only by distance calculating and locality sensitive hashing. The relevance feedback approach proposed in this paper works as an incremental tool to perform query expansion and boosts the precision in the same recall rate. The relevance feedback improves the efficiency of the retrieval.

Table 1 shows the how the feedback ratio of the candidate set affects the precision and recall percent in our experiments. We can see obviously that the relevance feedback approach could achieve high precision percent while improving recall percent by providing specific ratio of feedback information. The reason is that the artificial semantic information from the feedback of the users complements the semantic similarity loss caused by the

distance calculation fully based on the feature vectors. However, when the feedback ratio exceeds certain threshold such as 30%, the precision percent decreases substantially even the recall percent is still able to maintain increasing. This result shows that too much feedback information from the users reduces the weighting of the distances in multi-level matching and causes the fact that the video retrieval is actually fully depend on the correlation matrix learned from samples provided by the users. We make sure the feedback ratio is below 30% in our experiments in order to maximize the efficiency of relevance feedback approach instead of violating the principle of contend-based video retrieval.

Table 1: Feedback Ratio and Precision-Recall Percent

Feedback Ratio	***Precision***	***Recall***
10%	93%	13.2%
20%	92.7%	15.4%
30%	89.1%	23.4%
40%	55.8%	35.0%
50%	30.9%	40.5%
60%	34.2%	52.3%

6. Conclusion

In this paper, we leverage local feature detectors to extract feature descriptors from video collections and perform multi-level matching after indexing and retrieval of feature vectors using the state-of-the-art techniques in content representation, similarity measure selection. We introduce a novel light-weight relevance feedback approach based on canonical correlation analysis (CCA) to bridge the gap between semantic notions of search relevance and the low-level representation of video content. Experimental results on real world demonstrate the precision gains of our proposed method.

Acknowledgments

We thank Alexandr Andoni for providing his E2LSH binary and the anonymous reviews for their constructive comments and suggestions which greatly improve the quality of this paper. This research was supported by Network Video Communication and Control Project in Sensing China Program of Chinese Academy of Science under Grant XDA06030900.

References

[1] Hu, W. and Xie, N. and Li, L. and Zeng, X. and Maybank, S., "A Survey on Visual Content-Based Video Indexing and Retrieval", Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 41, no. 6, 2011, pp.797-819.

[2] Gorisse, D. and Cord, M. and Precioso, F., "Locality-Sensitive Hashing for Chi2 Distance", Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 34, no. 2, 2012, pp.402-409.

[3] Gionis, A. and Indyk, P. and Motwani, R., "Similarity search in high dimensions via hashing", Proceedings of the International Conference on Very Large Data Bases, 1999, pp.518-529.

[4] Datar, M. and Immorlica, N. and Indyk, P. and Mirrokni, V.S., "Locality-sensitive hashing scheme based on p-stable distributions", Proceedings of the twentieth annual symposium on Computational geometry, ACM, 2004, pp.253-262.

[5] Chapelle, O. and Haffner, P. and Vapnik, V.N., "Support vector machines for histogram-based image classification", Neural Networks, IEEE Transactions on, vol. 10, no. 5, 1999, pp.1055-1064.

[6] Gosselin, P.H. and Cord, M. and Philipp-Foliguet, S., "Combining visual dictionary, kernel-based similarity and learning strategy for image category retrieval", Computer Vision and Image Understanding, Elsevier, vol. 110, no. 3, 2008, pp.403-417.

[7] Gordo, A. and Perronnin, F., "Asymmetric distances for binary embeddings", Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, 2011, pp.729-736.

[8] Weiss, Y. and Torralba, A. and Fergus, R., "Spectral hashing", NIPS, 2008.

[9] Kulis, B. and Grauman, K., "Kernelized locality-sensitive hashing for scalable image search", Computer Vision, 2009 IEEE 12th International Conference on, 2009, pp.2130-2137.

[10] Jégou, H. and Douze, M. and Schmid, C., "Product quantization for nearest neighbor search", Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 33, no. 1, 2011, pp.117-128.

[11] Strecha, C. and Bronstein, A.M. and Bronstein, M.M. and Fua, P., "LDAHash: Improved matching with smaller descriptors", Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 34, no. 1, 2012, pp.66-78.

[12] Gong, Y. and Lazebnik, S. and Gordo, A. and Perronnin, F., "Iterative Quantization: A Procrustean Approach to Learning Binary Codes for Large-scale Image Retrieval", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012.

[13] M. Yeh and K.-T Cheng, "Fast visual retrieval using accelerated sequence matching", Multimedia, IEEE Transactions on, vol. 13, no. 2, 2011, pp. 320-329.

[14] C. Chiu, H. Wang and C. Chen, "Fast min-hashing indexing and robust spatio-temporal matching for detection video copies", ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 6, no. 2, 2010, Article 10.

[15] Foster, D.P. and Kakade, S.M. and Zhang, T., "Multi-view dimensionality reduction via canonical correlation analysis", Technical Report, TR-2008-4, TTI-Chicago, 2008.

[16] Law-To, J. and Joly, A. and Boujemaa, N., "Muscle-VCD-2007: a live benchmark for video copy detection", Available: http://wwwrocq.inria.fr/imedia/civr-bench/, 2007.

[17] P. Haghani, S. Michel, and K. Aberer, "Distributed Similarity Search in High Dimensions Using Locality Sensitive Hashing", In Proceedings of the 12th International Conference on Extending Database Technology (EDBT 09), ACM, 2009, pp. 744-755.
[18] M. Bawa, T. Condie, P. Ganesan, "LSH forest: self-tuning indexes for similarity search", Proceedings of the 14th international conference on World Wide Web (WWW 05), ACM, 2005, pp. 651-660.
[19] Q. Lv, M. Josephson, Z. Wang, M. Charikar, K. Li, "Multi-probe LSH: efficient indexing for high-dimensional similarity search", Proceedings of the 33rd international conference on Very large databases (VLDB 07), ACM, 2007, pp. 950-961.

Zimian Li received the B.S. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2006. He is currently pursuing the Ph.D. degree in the School of Information Science and Technology of USTC. He already published three EI-indexed papers in international conferences and one in Chinese domestic journal. His research interests include self-healing multimedia system, multimedia communication.

Ming Zhu is currently a professor of University of Science and Technology of China. He received B.S., M.S. and Ph.D. degrees in Computer Science from University of Science and Technology of China in 1986, 1989 and 2001, respectively. He became an assistant professor in 1989 and a professor of USTC in 2004. He worked as a visiting scholar in Department of Computing, The Hong Kong Polytechnic University from 1997 to 1998. He is the Director of the Key Lab of Network Communication System & Control, Chinese Academy of Sciences and the Director of the Key Lab of Network Communication System & Control, Anhui. His research interests include intelligent software systems, data mining and network security.

Research on Remote Sensing Image Template Processing Based on Global Subdivision Theory

Xiong Delan[1], Du Genyuan[1]

[1] International School of Education, Xuchang University
Xuchang, Henan, China

Abstract

Aiming at the questions of vast data, complex operation, and time consuming processing for remote sensing image, subdivision template was proposed based on global subdivision theory, which can set up high level of abstraction and generalization for remote sensing image. The paper emphatically discussed the model and structure of subdivision template, and put forward some new ideas for remote sensing image template processing, key technology and quickly applied demonstration. The research has very important significance for improving remote sensing image processing speed, reducing repeated handling of huge amounts of image data, and expanding practical application of remote sensing.

Keywords: Global Subdivision Theory (GST), remote sensing image, subdivision template, template processing

1. Introduction

Remote sensing is a certain science and technology which can measure, analyze and determine a target without contact with the object directly utilizing some sort of sensor devices [1]. The data acquired by remote sensing techniques has some advantages of high real-time, wide range and rich information, which has been widely used in many military and civilian areas such as military reconnaissance, disaster forecasting, environmental monitoring, and resource exploration and so on.

With the development of sensors, remote sensing platforms, and data communication technology, the amount of data obtained by remote sensing expands rapidly, resulting in such a situation that the spatial data production and transmission capacity is far greater than the space data analysis capabilities [2]. At the same time, many application fields have constantly increase requirements about real-time, accuracy and reliability of remote sensing images. The speed has become the bottleneck of the application of remote sensing image.

Aiming at the questions such as vast data, complex operation and time consuming processing for remote sensing image, a new concept of subdivision template was proposed based on global subdivision theory. Subdivision template can set up high level of abstraction and generalization for remote sensing image, and provide convenience for quickly and simply use of remote sensing data resources. This paper emphatically discussed the model and structure of subdivision template, and put forward some new ideas for remote sensing image template processing, key technology and fast demonstrate application. The research would provide cornerstone and feasible solution for template-based change detection and parallel processing. So it has important significance.

2. Global Subdivision Theory

The Global Subdivision Theory (GST) [3] is a lager-scale hierarchical open spatial data management framework. It researches on how to subdivide the Earth (or spherical) into a series of cells with same area and similar shape. It has many advantages such as global scale, continuity, stability, multi-level, uniformity, and so on. So GST may avoid data redundancy effectively and express the levels of the data, which takes advantages over the planar grid system when dealing with the global multi-scale spatial data. With the development of Digital Earth (DE), the expression and management of the global multi-scale spatial information make the limitations of traditional planar grid system become more obviously. To establish a new model for the global multi-scale spatial information is a common concern for most domestic and foreign scholars.

So far, many kinds of subdivision theories were put forward by scholars from various countries. They can be summarized into three kinds: polyhedral subdivision, experience subdivision and wavelet division [4]. The typical subdivision models are Quaternary Triangular Mesh (QTM) by Duttn, Spherical Quaternary Triangle (SQT) by Fekete, Equal Angle Ratio Projection (EARP) by Yuan and so on[5]. The Extended Model Based on the Mapping Division model (EMD) was proposed in paper [6] by Cheng. It made the hierarchical subdivision by longitude and latitude interval, based on traditional mapping division way to achieve the objectives of direct storage and index for existing spatial data. So EMD can

effectively implement management, organization and use of huge spatial data.

GST has been applied in many fields of global spatial information, which can effectively implement storage, extraction and analysis of global scale mass data. It can solve those question that traditional data model limit to manage huge amounts of data on a global scale, multi-scale and hierarchical organization. It can ensure the global spatial data expressing in a global, continuous, hierarchical and dynamic way [7]. But many studies are at the initial stage, most of researches are about theoretical study of subdivision methods, coding model and storage mechanism, but less specific practical application of the results are in shown at present.

3. Subdivision Templates of Remote Sensing Image

3.1 Basic Concepts

GST divide the Earth into cells with regular shape, different coverage area using a certain different division method. The cells have the advantages of hierarchical organization, the uniqueness of the spatial location encoding, storage location and high search efficiency. Subdivision template is structure data set for remote sensing image and matches along with subdivision cell. It contains the data of spatial characteristics set, geographical features set, and control points set of cell. It inherits the advantage of cell, so it is easy to organize and manage to achieve efficient processing applications. Each subdivision template is a comprehensive data set basing on the baseline remote sensing image. It has basic information of image, and combines high-level features and semantic information of remote sensing images in a higher level abstraction and generalization. So it would be more conducive, storage and application for remote sensing images. Here, baseline remote sensing image is a kind of orthogonal projected remote sensing image for some scale.

According to different requirements about spatial location, covering range, resolution of cell in different level, select appropriate orthogonal projected remote sensing image, and convert into a formal, normative image format through a standardized processing step. Then, baseline remote sensing image were gotten. In the procedure of remote sensing image processing, the important task is to set up association with different information and store them in certain way. These information are underlying data and parameters of remote sensing images (such as sensor type, produced time, resolution), contents of covering region (such as regional names, object names), advanced features of the image (such as texture features, color features) and characteristics of cells (such as cell code, division level, cell control points).

3.2 Logical Contents

Through above analysis, we can sure subdivision template is comprehensive information collection of remote sensing image on certain subdivision cell. From a logical point of view, the subdivision template includes the following contents:

Basic Information: contains two aspect information, one is underlying data and relevant parameters of remote sensing images, such as image number, sensor type, resolution, band value, the image source and so on. The other is cell basic information such as cell level, cell code, and cell basic control points data.

Feature Information: including the features and characteristic information of the remote sensing image and cell, such as the image number, image area names, image object names, significant texture features, shape features and cell elevation characteristics, regional characteristics and so on.

Knowledge Comments Information: description and annotation of the subdivision template and cell for some important information, coding style and other related items, such as the template description, regional introduction, image object introduction, and features descriptions.

3.3 Physical Contents

Database management is the mainstream form of remote sensing image data storage and management [8]. High performance cluster-based processing technology based on parallel data processing and grid computing technology based on large-scale distributed processing are the main method for remote sensing image using high-performance processing [9-10]. Therefore, subdivision template can adopt distributed storage management based on relational database. On the whole, the different levels of subdivision template and the remote sensing image are stored in multiple parallel processing units. They can be unified managed using a suitable index structure. In a special storage unit, subdivision template and image are organized, managed and retrieved in a relational database mode, which technology are already quite mature, and many functions can be carried out by structure query language (SQL). According to template contents and image data,

several databases and data tables are created. They can be associated and mapping through primary key, foreign key and other convert operations.

In addition, in order to improve the efficiency of image access, image segment method is adopted, which can divide a huge remote sensing image into a number of smaller physical data block. This way is good to store and manage the huge image. And the data block size will directly affect the system performance, and be considered as an important factor in remote sensing image data storage management. Usually, the size is power of 2. The block size of 256 × 256 pixels or 512 × 512 pixels is often used and was proved better in major environment [11].

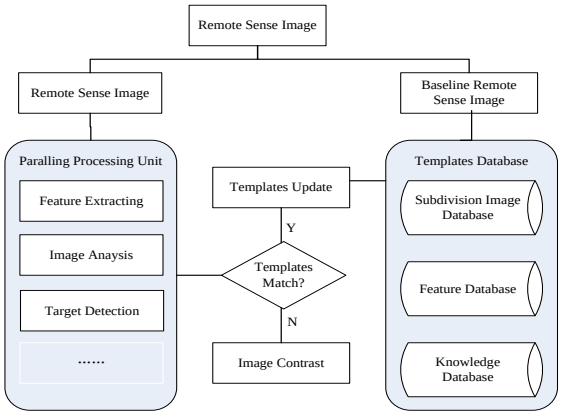


Figure 1. Subdivision template processing framework

4. Subdivision Templates Processing of Remote Sensing Image

4.1 Processing Flow

The processing flow of remote sensing image template based on GST can be described by figure 1.

According to specific subdivision method to determine the size and scope of certain level cells, select a number of continuous levels of cells as research object.

Aiming at different application requirements, initially select the orthogonal projected remote sensing image of typical hot area. And establish subdivision template through standardized processing.

Using some remote sensing image processing software such as Erdas, Envi to extract image features, created knowledge annotation by manual.

In mainstream computer systems developing parallel computing platform for subdivision template, complete image segmentation, feature extraction, manual annotation and other high tasks.

4.2 Key Technology

The key technology of subdivision template processing for remote sensing image are template quickly generation, automatic update, parallel processing and rapidly apply demo.

Templates quickly generation: For cells without establish subdivision template, select appropriate remote sensing images, and develop some automatic processing algorithms for image normalization, feature extraction, feature vector to organize and store, construction association with cell, template automation coding.

Template automatic updates: Automatic update is the key way to ensure uniqueness, independence and timeliness of subdivision template. Templates automatic update can timely detect changes of same region from remote sensing image, and convert new image into templates, and update relate information.

Subdivision templates parallel processing: Research on parallel processing strategy on template-based parallel processing mechanisms, including image preprocessing, image segmentation, feature extraction and similarity measure. Design parallel scheduling algorithm, and analyze its performance.

Rapidly applied demo for subdivision template: utilize subdivision template for image analysis, image region recognition, change detection, target recognition and tracking, object classification, and form some typical application demonstration.

4.3 Preliminary Application

Based on EMD model, we research on shape, feature and coding of cell in certain level, and construct the corresponding subdivision template of remote sensing image. An initial prototype system for subdivision template of remote sensing image processing system was developed. The system was developed in Windows Server 2003, using SQL Server 2005 to manage remote sensing data, and using VC++ to design and develop user interface. Preliminarily, we selected several high-resolution remote sensing images of tourist attractions, and create subdivision template manually. At present, basic functions such as image browsing, template view and feature retrieval according to the specified conditions were completed. In figure 1, if you select some kind of retrieve mode (subject information of image contents), and input some keywords (Henan AND tourism AND spa), the search results will be shown in bottom. You can get main information of image and template, and query detail information by clicking the related buttons.

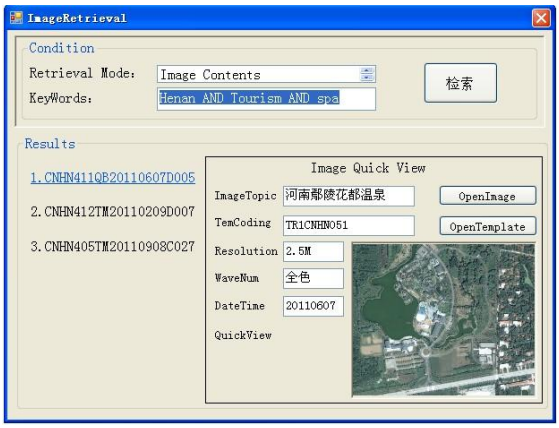


Figure 2. Subdivision template and image retrieve interface.

At present, the system has only completed a test of basic functions of in single computer environment. Preliminary tests showed that the system have highly-targeted, high practicability, high retrieval efficiency and good expandability. More functions would be further developed.

5. Conclusions

Remote sensing images have been widely used in various fields. With the increasing of data amount and expanding of application requirement, the collection, organization, management and sharing of remote sensing data are becoming the most prominent problem to be solved for data producers and users. The global subdivision theory divides the Earth into cells in level with same area and shape, which has advantages of global, multi-resolution, well-proportioned spatial location. It provides a new way to solve the effective organization and management of massive remote sensing data. Combining the features and advantages of subdivision cell, the paper put forward the concept of subdivision template and proposes to set up subdivision template of remote sensing image gradually by level, by batches, and by cell region. Template quickly generation and automatic update would be deeply researched to ensure template uniqueness. With increasing of subdivision template, parallel processing technique would be adopted to improve processing efficiency and sharing service. This study will have a very important strategic significance for extending application fields of remote sensing image, opening practical applications of GST, and enhance the value of spatial data.

Acknowledgments

This work is supported by the Science and Technology Research Project of Henan Province under Grant No. 112102210079.

References

[1] Mather P M.: Computer Processing of Remotely-Sensed Images: An Introduetion (Second Edition).Chichester:John Wiley & Sons,1999.
[2] LI De-ren,"On Generalized and Specialized Spatial Information Grid," Journal of Remote Sensing, vol.9, 2005, pp.513-520.
[3] Goodchild M., Discrete Global Grids for Digital Earth.International Conference on Discrete Global Grids. California: Santa Barbara, 2000.
[4] Gannon, D., Alameda, J., Chipara, O. et. , "Building Grid Portal Applications From a Web Service Component Architecture",in Proceedings of the IEEE, vol.93, 2005,pp.551-563.
[5] Sahr K, White D, Kimmerling A. , " Geodesic discrete global grid systems", Cartography and Geographic Information Science, 30, 2003,pp.121-134.
[6] CHENG Chengqi,GUAN Li, "The Global Subdivision Grid Based on Extended Mapping Division and Its Address Coding", Acta Geodaetica et Cartographica Sinica, vol.39,2010,pp.295-302.
[7] SONG Shu-hua, CHENG Cheng-qi, GUAN Li,et. , "Analysis on Global Geodata Partitioning Models ", Geography and Geo-Information Science,vol.24, 2008, pp. 11-15.
[8]CHENG Qi-min, Remote Sensing Image Retrieve Technolgy. Wuhan:Wuhan University Press, 2011.
[9] Plaza A, Plaza J, Valencia D. , " Impact of Platform Heterogeneity on the Design of Parallel Algorithms for Morphological Processing of High-Dimensional Image Data", Journal of Supercomputing, Vol.40, 2007, pp.87-107
[10]DU Gen-yuan, MIAO Fang, GUO Xi-rong, "A novel network service mode of spatial information and its

prototype system ", Advanced Materials Research, Vol.108, 2010, pp: 319-323.

[11]WANG Hua-bin, TANG Xin-ming, LI Qian-xiang, " Research and implementation of the massive remote sensing image storage and management technology ",Science of Surveying and Mapping, Vol.133, 2008,pp.156-157.

Xiong Delan, Born in 1980, Female, Master Degree achieved at 2006, Instructor, 7 papers publican in Chinese journals, achieved Natural Science Foundation of Henan Province.

Du Genyuan, Born in 1974, Male, Doctor degree achieve at 2011, Associate Professor, more than 10 papers publican in Chinese journals, achieved Scientific and technological Project of Henan Province under Grant.

Study of RBF Nerve Network Tuning PD Control Algoritm of Bilateral Servo System

Guang Wen
School of Machinery and Engineering, Pan-zhihua University
Pan-zhihua, 617000, China

Abstract

In construction tele-robot system. When p-f architecture force feedback was used, the impact of large feedback force result in the strike-like feeling on the operator's hand. If the amplitude is high, it will cause the control unstable. So a improved force feedback control method with the feature of a T-S fuzzy feedback coefficient, which could be modified online nonlinearly and continuously, is developed. A RBF-PID force controller is also designed, and formed a bilateral hydraulic servo control system. The experimental results indicate that the new improved control method reduced the impact of the feedback force, enhanced the compliance and transparency of the tele-operation of construction tele-robot system.

Keywords: Fuzzy feedback coefficient; Force Feedback; Construction Tele-robot

1. Introduction

Master, slave tele-operation robot system works can be inaccessible in the human person harmful to the environment or to complete work. Operators in a safe place that only a true and accurate force tele-presence information, they can control engineering task robot to accurately complete the operation. Pairs of force tele-presence tele-operation robot control system design projects, to ensure system stability, reliability, and tracking performance, So effective control methods, rational design of the controller is the critical to ensure the reliable operation of control systems[1~5]. Authors based on the existing control methods and their application in the field of robotics research, combined with force tele-presence tele-operation robot works bi-directional hydraulic servo control system characteristics, relying on electro-hydraulic proportional valve controlled by the master, from the hydraulic swing motor experimental platform consisting of tele-presence. Force feedback servo-control for bi-directional feedback exists in the impact force the issue, proposed to improve the force feedback control method to increase the smoothness. Experimental results show the effectiveness of the method[6].

This document is set in 10-point Times New Roman. If absolutely necessary, we suggest the use of condensed line spacing rather than smaller point sizes. Some technical formatting software print mathematical formulas in italic type, with subscripts and superscripts in a slightly smaller font size. This is acceptable.

2. Improved bi-directional force feedback servo-control method

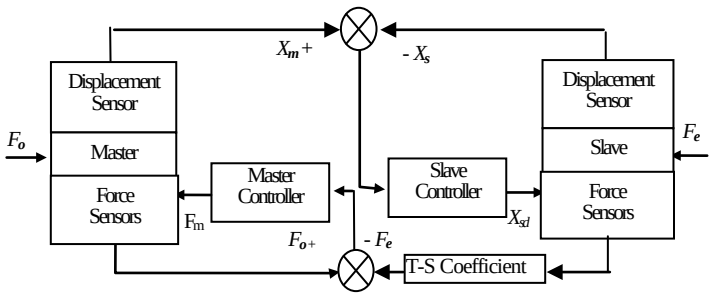


Fig.1 Structure of Improved force feedback tele-robot control system

The choice of the feedback coefficient, the usual practice is based on specific system feedback force range, select an appropriate ratio constant. However, this approach will force when feedback is small, the transparency of the system decreased significantly. The feedback coefficients should be expected of such a nature, force feedback when the feedback coefficient and a small number in order to enhance the sensitivity of force feedback. When the feedback force is large, the feedback factor should be smaller to ensure that force feedback can be put in the scope of the manpower in order to reduce the impact effects. In this paper, TS-type fuzzy model is constructed using non-linear changes in a continuous feedback coefficient to improve the force feedback control method.
The improved control system schematic shown in Fig. 1, the main features of the control is the location of the Slave hand depend on the main and the secondly hand's position deviation between the control, the main force hand

feedback force by the product of F_e, K_{ef} and the operation of force bias control, the control law as follows:

$$F_m=K_f(F_o-K_{ef}*F_e), \quad X_{sd}=K_s(X_m-X_s)$$

Where:
F_m--Drive Master Hand Vector;
K_f--Main hand gain matrix;
F_e--force vector from the secondly hand and the environment;
F_o--The operator control force vector;
X_{sd}--Expect position vector of the Slave hand;
K_s--Displacement gain matrix of the Slave hand;
X_m--Main hand displacement vector;
X_s--Slave hand displacement vector;
K_{ef}--Feedback coefficient.

3.The Structure of System

The tele-operation system with force tele-presence includes following components: the manipulator, electro-hydraulic servo drive system, displacement servo control system, visual tele-presence system, wireless communication system and force sensors, displacement sensors etc, shown in Fig.2.

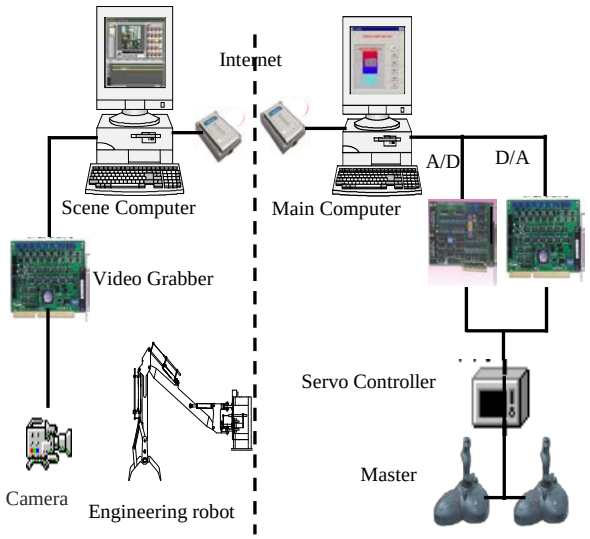


Fig. 2 Master-slave system for remote control

The figure shows that the system is constructed as a master-slave system and that both manipulators for the master and the slave consist of 4-DOF type actuators. Moreover, it is illustrated that a machine tool for grinding is implemented at the end-effecter of the slave manipulator.

In a tele-operated master-slave system as shown in Fig. 1, the master has to play two roles, firstly as a reference input device to the slave, and secondly as a sense of force device. Here, the term “sense of force” means a function that allows the operator to feel a force that is fed back to him from the slave[7~9].

This research deals with a remote-control system applicable to the machining fields, such as grinding, polishing, assembling, and shaping. In machining works that require high speed, high power, and high rigidity in the operation, the attributes of hydraulic actuators make them suitable for these applications. In this study, we deal with a master-slave system composed of serial links by hydraulic cylinders. First, the serial links treated here are assumed to be of 1-DOF, and then of 4-DOF for general use, because we mainly are concerned with developing a new sense of force device.

At the first control stage, the remote control computer handclasps with the worksite one, and then the worksite computer reads the information of the joints’ displacement and velocity through the A/D continuously, and transmits all these information to the remote control computer to initialize the graphic robot and make the operator present the worksite robot’s state. Simultaneously, the worksite computer and the graphic computer receive the control instructions in the manner of event-driven. The graphic computer refreshes the virtual robot motion state on real time. The worksite computer explains the operator’s instructions into the motion angles of every joint by arithmetic, where the sampling and controlling interval is ten mms. The process is shown in Fig.3.

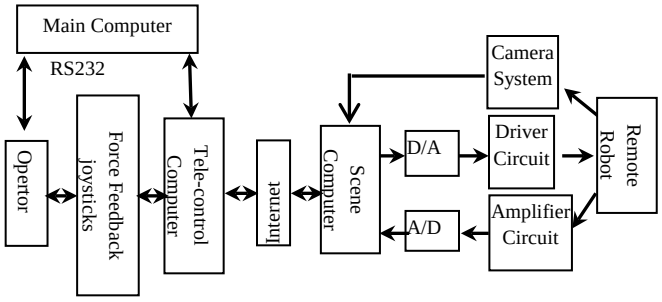


Fig.3 Remote Robot Control System Principle

When the operator operates the remote worksite robot facing to the simulation robot, the video information is needed to be watched on real time because the model errors between the graphic robot and the virtual environment are inevitable [5], and the worksite environment can also not be predicted. All these were completed by the equipments fixed on the remote robot

such as camera, video emitter, video receiver and so on[10,11].

Comparing with the tele-operation which is operated only by the video pictures transmitted from the worksite, the operation with high tele-presence prompt manner may enhance the work efficiency by 30%~50%. Simultaneously, it is not only favor for conquering the influence of time delay, but also can provide friendly graphical user interface. The operator can change video point and video angle of the conceals level forward feeds network [12~13]. It is non-linear from input to the output mapping, but it is linear from the conceals level space to the output space mapping, thus speeds up the study speed greatly and avoids the partial minimum problem.

4.Control algorithm realization

4.1RBF nerve network model

The Radial Basis Function nerve network is proposed by J.Moody and C.Darken in the end of 1980s, it is has three conceals level forward feeds network [14]. It is non-linear from input to the output mapping, but it is linear from the conceals level space to the output space mapping, thus speeds up the study speed greatly and avoids the partial minimum problem. RBF network architecture shows in Fig.4.

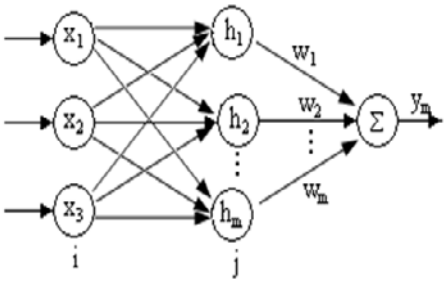


Fig.4 RBF Nerve Network Frame Chart

4.2 RBF nerve network PID control algorithm [15]

PID control to simple structure, robustness, good, able to adapt to the complex system control, etc., in the control engineering has been widely used [16,17]. But the parameters of conventional PID control system by setting a hard-line adjustment of the non-linear and non-deterministic system control result is not very satisfactory. To do this in cooperation with the Ziegler-Nichols method of digital simulation to determine the PID parameters of stable operation of the system K_p, K_i, K_d, after the initial value. Designed to use RBF (Radial Basis Function) neural network model for online identification system, and adjust the PID parameters, the formation of RBF-PID controller to meet the remote operation of robot control system engineering nonlinear and uncertain requirements, its block diagram as follows in Fig.5.

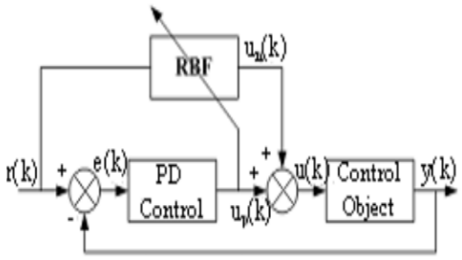


Fig.5 RBF Nerve Network PD control

RBF network is a three-layer feed forward network, from input to output mapping is nonlinear, while the hidden layer space to output space mapping is linear, thus speeding up the learning speed and avoid local minima problems [18~20]. This design network model has six input nodes, eight hidden layer nodes, three output nodes.

According to the system model equation, to be ts as the sampling period, matching all the pole-zero conditions, to obtain the discrete model of the system:

$$y(k) = a_1 y(k-1) + a_2 y(k-2) + a_3 y(k-3) + a_4 u(k-1) + a_5 u(k-2) + a_6 u(k-3) \quad (1)$$

Where:

$y(k)$—the output of the system of k time

$u(k)$—the control input of K time

a_i —known constants of System model

$$X = [x_1, x_2, x_3, x_4, x_5, x_6]^T \quad (2)$$

Then the RBF network input vector：

Where:

$x_i = y(k-i)$ $\quad i=1,2,3$

$x_j = u(k-j+3)$ $\quad j=4,5,6$

Hidden layer nodes to take Gaussian kernel function：

$$h_j = \exp\left(-\frac{\|X - c_j\|^2}{2b_j^2}\right) \qquad j=1,2,\ldots 6 \quad (3)$$

Where: The first j nodes hj center vector $c_j = [c_{j1}, c_{j2}, \ldots c_{j6}]^T$; the base width $b_j = [b_1, b_2 \ldots b_6]^T$; knot vector $H = [h_1, h_2, \ldots h_6]^T$. Take the network weight vector $W = [w_1, w_2, \ldots w_6]^T$, using gradient descent method c_{ji}, b_j, w_j, h_j of the iterative algorithm is as follows:

$$y_m(k) = w_1 h_1 + w_2 h_2 + \cdots + w_6 h_6 \quad (4)$$

$$w_j(k) = w_j(k-1) + \eta\left(\begin{array}{l} y(k) - y_m(k) h_j + \\ \alpha\left(w_j(k-1) - w_j(k-2)\right)\end{array}\right) \quad (5)$$

$$b_j(k) = b_j(k-1) + \alpha\left(b_j(k-1) - b_j(k-2)\right) + \eta\Delta b_j \quad (6)$$

$$c_{ji}(k) = c_{ji}(k-1) + \alpha\left(c_{ji}(k-1) - c_{ji}(k-2)\right) + \eta\Delta c_{ji} \quad (7)$$

Which take learning rate η=0.2, Momentum factor α=0.05。

$$\frac{\partial y(k)}{\partial u(k)} = \frac{\partial y_m(k)}{\partial u(k)} = \sum_{j=1}^{m} w_j h_j \frac{c_{ij} - u(k)}{b_j^2} \quad (8)$$

Incremental PID coefficient adjustment method is as follows:

$$E(k) = R(k) - y(k) \quad (9)$$

Where: *R(k)*—The system reference input.

E(K)—System error.

$$\Delta k_p = -\eta \frac{\partial J}{\partial k_p} = \eta E(k) \frac{\partial y(k)}{\partial u(k)} \left(E(k) - E(k-1)\right) \quad (10)$$

$$\Delta k_i = -\eta \frac{\partial J}{\partial k_i} = \eta E(k) \frac{\partial y(k)}{\partial u(k)} \quad (11)$$

$$\Delta k_d = -\eta \frac{\partial J}{\partial k_d} = \eta E(k) \frac{\partial y(k)}{\partial u(k)} \begin{pmatrix} E(k) - 2E(k-1) \\ +E(k-2) \end{pmatrix} \quad (12)$$

$$u(k) = u(k-1) + K_p\left(E(k) - E(k-1)\right) + K_i E(k) + K_d\left(E(k) - 2E(k-1) + E(k-2)\right) \quad (13)$$

5.Experiment results

To realize a tele-operated manipulation system as shown in Fig. 1, it is necessary to constitute a master-slave system, in which the master and the slave correspond, respectively, to a sense of force and an actuating manipulator. In this section we therefore discuss a master-slave hydraulic system equipped with the new sense of force proposed.

5.1System Constitution

In Fig.6, a schematic diagram of the experimental apparatus for the present study is shown. The total system consists of a master system and a slave system. The operator's force Fop, detected by a force sensor, is sent to the computer in order to actuate a master-side piston. In the slave system, a spring of stiffness k is attached at a frame of apparatus for simulating a load-force of operation. To detect the load-force, a force sensor is set at the inertial load through a plate spring. Two displacements of pistons in master and slave xm, xs, and two forces Fop =Fm, Fs are detected by each sensor and then sent to the computer. Subsequently, two control inputs um and us for actuating the master and the slave are calculated in the computer, according to a bilateral algorithm. For an algorithm of each controller for the master and slave, a proportional control algorithm was adopted. The sampling time was chosen to be 1 ms. In the experiment, two kinds of load, that is tires and hardwood, were tested. With respect to servo-valves and cylinders for constructing two servo-systems in the master and the slave, different types were adopted intentionally between the two systems. Namely, the master-slave system was tested in the experiment for a system with rather different dynamic characteristics between the master and the slave.

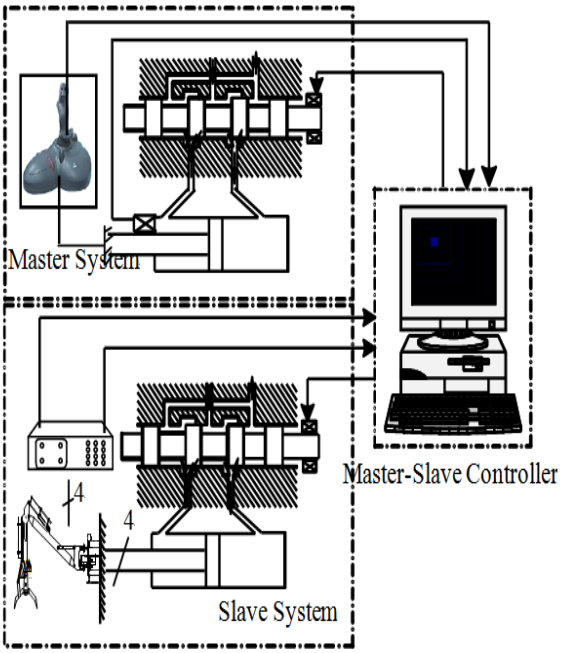


Fig. 6 Diagram of experimental apparatus

It is adopted that a bilateral control methodology for controlling the master-slave system. Concerning system constitutions for bilateral control, the following two types are well known as representative ones: (a) Force reflecting servo type and (b) Parallel control type.

5.2 Experimental Results

In the master-slave system shown in Fig.7, two types of bilateral controls are adopted, that is, a force reflecting servo type and an improved parallel control method. By comparing the force functions between two types of systems, we investigate experimentally the applicability of the proposed system. In the experiment, time responses of two forces Fm and Fs were measured together with those of two displacements X_m and X_s.

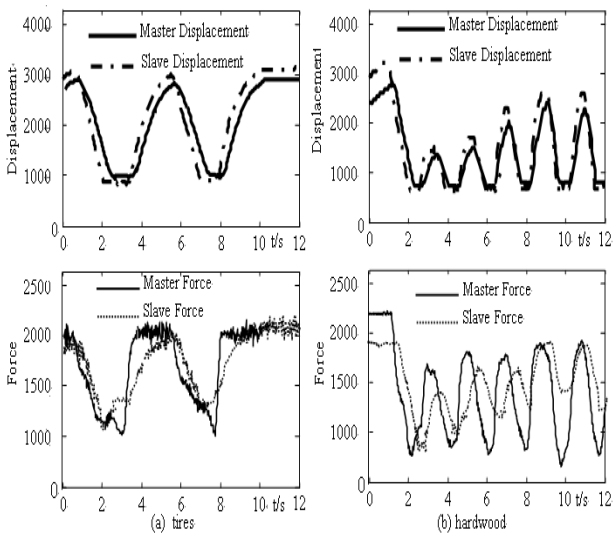


Fig.7 Experimental results of force reflecting servo type

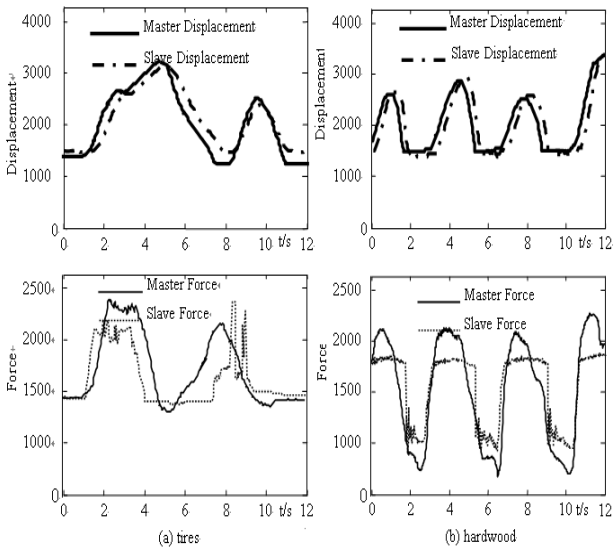


Fig.8 Experimental results of parallel control method

First, response curves for the force reflecting servo system are shown in Figs.7(a) and (b). These figures correspond, respectively, to the results for the tires and the hardwood. Observing Fig. 7(a), it is shown that the slave force Fs is detected almost at the instant that the slave touches the tires. Subsequently, the force Fs is controlled in good agreement with the master force Fm is shown in the figure. In this experiment, the operator was able to feel a softness of tires through the sensing function of the sense of force. On the other hand, Fig.7(b) shows that the response curve of Fs is accompanied by a tendency toward vibration. The vibration appears from the instant that the slave touches at the tires. In addition, the tendency of such a vibration affects the waveforms of displacements Xm and Xs . In this experiment, it was difficult to control the system stably. Secondly, the same kinds of results as seen in Fig. 7 are shown in Figs. 8 (a) and (b) as a result of the parallel control. Through comparing Fig.8(a) and Fig. 7(a), it is shown that both results coincide well with each other. The operator in this experiment was able to feel a softness of tires as in the previous experiment in Fig.7(a). Furthermore, the result for the hardwood from Fig.8(b) is improved distinctly compared with the result in Fig.7(b). The system was kept stable in this experiment under various system conditions. As a result, the operator was able to feel a hardness of wood. Correspondingly, it is observed in Fig.8(b) that the amount of piston dis-placement is smaller than in Fig.8(a), in spite of the fact that a larger force than that in (a) is given to the system.

6. Conclusions

In view of the novel force feedback bilateral servo control system, it was proposed that one kind of RBF nerve network tuning PD on-line from study, adaptive control strategy, which can approach willfully the continuous function characteristic using the RBF nerve network by the free precision, through optimizing two parameters of PD by RBF nerve network, it can improve dynamic characteristic of master-slave control system. Through simulation and experiments, firstly, it can be theoretically proved that this control algorithm of novel force sense bilateral servo system is practical and feasible; next, this algorithm can realize master-slave position tracking and let the operator feel "the force sense" well from feedback, thus improves the human and the environment interaction characteristic and enhances the working efficiency. Moreover, this control arithmetic has taken on control briefness, constringency rate rapidness, real-time well, strong robustness, self-adapted and the rapidity.

References

[1] S.Munir, W.j.Book, "Internet Based Tele-operation using wave variables with prediction", in proceedings of IEEE/ASME International conference on advanced intelligent mechatronics, 2001, Vol. 1, pp. 43-50.

[2] K.Hidetoshi，H.Yamada,T.Muto,"Mater-Slave Control for a Tele-Operation System of Construction Robot", Transactions of the Japan Fluid Power System Society, 2003, Vol. 34,No. 2, pp. 27-33.

[3] M.D.Gong, D.X.Zhao, T.Ni etal.,"Design of an Isomeric Slave Arm for Engineering Robot", Construction Machinery and Equipment, 2003, No. 12, pp. 1-3.

[4] S. Kudomi, H. Yamada, and T. Muto, "Development of a Hydraulic Parallel Link Force Display Improvement of Manipulability Using a Disturbance Observer and its Application to a Master-slave System",Joural of Robotics and Mechatronics,2003,Vol. 15, No. 4, pp. 391-397.
[5] X.X.Tang,H.Yamada,D.X.Zhao,T.Ni,"Haptic Interaction in Teleoperation Control System of Construction Robot Based on Virtual Reality",in Proceedings of the 2009 IEEE International Conference on Mechatronics and Automation ,2009,pp.78-83.
[6] Y.Ye, Y.J.Pan, Y.Gupta,"Time domain passivity control of teleoperation systems with random asymmetric time delays", In Proceedings of the 48th IEEE Conf. on Decision and Control, 2009, pp. 7533–7538.
[7] J. Ware,Y.J. Pan，"Realisation of a bilaterally teleoperated robotic vehicle platform with passivity control",IET Control Theory & Applications, 2011,vol.5, no.8, pp.952-962.
[8] E.Slawinski,V.A.Mut,P.Fiorini,L.R.Salinas, "Quantitative Absolute Transparency for Bilateral Teleoperation of Mobile Robots," Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on , 2012,vol.42, no.2, pp.430-442.
[9] T. M. Lam, H. W. Boschloo, M. Mulder, and M. M. Van Paassen, "Artificial force field for haptic feedback in UAV teleoperation",IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, 2009,Vol. 39, No. 6, pp. 1316–1330.
[10] E. Slawiñski and V. A. Mut,"Control scheme including prediction and augmented reality for tele-operation of mobile robots", Robotica, 2010,Vol. 28, No. 1, pp. 11–22.
[11] Y.Wagatsuma,Y.Toda,N.Kubota,"Formation behavior of multiple robots based on tele-operation,"In Proceedings of 2011 IEEE International Conference on Fuzzy Systems ,2011, pp.713-720, 27-30.
[12] C.Ishii,H.Mikami,T.Nakakuki and T.Hashimoto,"Bilateral Control for Remote Controlled Robotic Forceps System with Time Varying Delay",In Proceedings of 2011 4th International Conference on Human System Interactions,2011, pp.330-335.
[13] C. Ishii, K. Kobayashi, Y. Kamei and Y. Nishitani,"Robotic Forceps Manipulator with a Novel Bending Mechanism",IEEE/ASME Transactions on Mechatronics, 2010,Vol.15, No.5, pp.671-684.
[14] S.C.Cramer,"Brain repair after stroke",New England Journal of Medicine, 2010 .VOL.362,pp.1784–1787.
[15] L.Dipietro, H.I.Krebs, S.E.Fasoli, B.T.Volpe, N.Hogan,"Submovement changes characterize generalization of motor recovery after stroke", Cortex,2009,Vol.45,No.3,pp.318–324.
[16] P.W.Duncan, R.Zorowitz,B.Bates, J.Y.Choi etal.,"Management of adult stroke rehabilitation care: a clinical practice guideline". Stroke,2005,Vol. 36,No.9,pp.100–143.
[17] Y.Hsieh, C.Wu, W.Liao, K.Lin, K Wu etal.,"Effects of treatment intensity in upper limb robot-assisted therapy for chronic stroke: a pilot randomized controlled trial". Neurorehabil Neural Repair,2011, Vol.25,No.6,pp.503–511.
[18] D.E.Nathan ,M.J.Johnson, J.M.McGuire,"Design and validation of a low-cost assistive glove for assessment and therapy of the hand during ADL-focused robotic stroke therapy",J Rehabil Res Dev,2009,Vol.46,No.5,pp.587–602.
[19] W.S.McCombe ,W.Liu,J.Whitall,"emporal and spatial control following bilateral versus unilateral training",Human Movement Science ,2008, Vol.27,No.5,pp.749-758.
[20] A.C.Lo,P.Guarino,L.G.Richards,etal.,"Robot-Assisted The - rapy for Long-Term Upper-Limb Impairment after Stroke",New England Journal of Medicine,2010,Vol.362, pp.1772-1783.

Guang Wen received a B.E.degree from Sichuan University of science and engineering, Zigong, China, in 1997 and a Master Degree in Mechanical and Electronic Engineering from Jilin University, Changchun, China, in 2008.He is now a professor at Panzhihua University. His research interests cover Tele-operation robotics, Numerical Control.

A Novel Block-DCT and PCA Based Image Perceptual Hashing Algorithm

Zeng Jie

College of Information Engineering, Shenzhen University
Shenzhen, Guangdong, P.R.China

Abstract

Image perceptual hashing finds applications in content indexing, large-scale image database management, certification and authentication and digital watermarking. We propose a Block-DCT and PCA based image perceptual hash in this article and explore the algorithm in the application of tamper detection. The main idea of the algorithm is to integrate color histogram and DCT coefficients of image blocks as perceptual feature, then to compress perceptual features as inter-feature with PCA, and to threshold to create a robust hash. The robustness and discrimination properties of the proposed algorithm are evaluated in detail. Experimental results show that the proposed image perceptual hash algorithm can effectively address the tamper detection problem with advantageous robustness and discrimination.

Keywords**:** *image hash; perceptual hash; tamper detection; PCA.*

1. Introduction

Image perceptual hashing, also known as image robust hashing, is defined as mapping images to a short bit string following the human perception [1]. In contrast to classic hash functions (MD5, SHA-1), which is highly sensitive to every bit of input data, image perception hashing is sensitive to image content rather than the integrity of image date. The two principal properties of image perception hashing are robustness and discrimination. Robustness means that the hash algorithm should result in the same out bit string for images with the same underlying content. For example, the raw image, its added noise version, its compressed version, its changed brightness version and its rotation angle version have the same underlying content and should share the same hash value. Discrimination implies that the hash values for any two distinct images should be different and random. That is to say, image perceptual hash functions are statistically independent to different image content.

Image perceptual hash value can be used for content identification and digital signature. The former is mainly used in content indexing and analysis, large-scale image database management. The latter is mainly used in the image certification and authentication, digital watermarking. According to the needs of applications, image perceptual hashing should also meet other two properties — randomness and scale-independence. Randomness means that the hash function should withstand all kinds of forgery attack since the hash values are impossible to be reconstructed by the attacker. Scale-independence implies that the length of hash values should always be an even number, although the input images are in different resolution.

Many image perceptual hash functions have been proposed in the literature. Bian Yang[2] uses the mean of image blocks to obtain a perceptual hash. J.Fridrich[3] extracts perceptual features by projecting image blocks onto key based random patterns and thresholds to create a robust hash with the median. R.Venkatesan[4] and M.K. Mihcak[5] selects the low-frequency sub-band of wavelet coefficients to generate a perceptual hash. F.Lefbvre[6] and J.S.Seo[7] uses radon transform to produce a perceptual hash. Hui Zhang[8] creatively introduces the human visual system to obtain a image perceptual hash.

This article addresses the problem of the tamper detection problem of images with image perceptual hashing. Although there are so many image perceptual hash methods proposed, the tradeoff between robustness and discrimination is relatively few discussed. In this article, we aim at proposing a robust and discriminative image perceptual hash algorithm, and explore the algorithm in the application of tamper detection.

The rest of this paper is organized as follows. Section Ⅱ describes the proposed robust and discriminative image perceptual hash algorithm. The experimental results are detailed in Section Ⅲ. Section Ⅳ contains the properties of the hash value. Conclusion and future work are introduced in Section Ⅴ.

2. Robust and Discriminative Image Perceptual Hashing

The main idea of the algorithm is to integrate color histogram and low-frequency Discrete Cosine Transform (DCT) coefficients of image blocks as perceptual features, then to compress perceptual features as inter-features with

Principal Component Analysis (PCA), and to threshold to create a robust hash. The framework of this algorithm is shown in Figure 1.

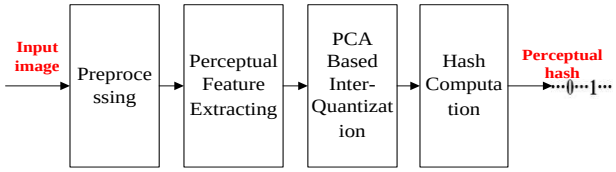


Figure 1. Flow chart of our image perceptual hash method

2.1 Image Preprocessing

The research of cognitive psychology and human visual system show that the sensitivity of eyes to chroma signal is much weaker than to luminance signal, and that brightness is the main features of the image signal[1]. So, only the luminance information is considered in preprocessing. The input image is first converted to a standardized image (64*64) via resampling and interpolation.

Preprocessing not only reduces the computational complexity of follow-up steps (perceptual feature extracting, PCA based inter-quantization, hash computation), but also ensures that the algorithm is independent of scale.

2.2 Perceptual Feature Extracting

During perceptual feature extracting, we adopt the block images strategy, and integrate color histogram and low-frequency Discrete Cosine Transform coefficients of every image block as perceptual features. The process is detailed as follows:

- Divide the standardized image into 64 blocks (block size: 8*8).
- Calculate the color histogram of blocks successively, and the calculation formula is as follows:

$$hist(i) = count(\lfloor ima_{gray} / 32 \rfloor), i = 0, 1, \cdots, 7 \quad (1)$$

- Select the DCT coefficients (DC coefficient and 7 AC coefficients) of blocks successively, and integrate the color histogram as perceptual feature, shown in figure 2.

The energy of the image will be gathered into some DCT coefficients after DCT transformation, DC coefficients contain the main information of the original data matrix, AC coefficients contain the detail information of the data matrix. Meanwhile, they are the most sensitive information of human visual system.

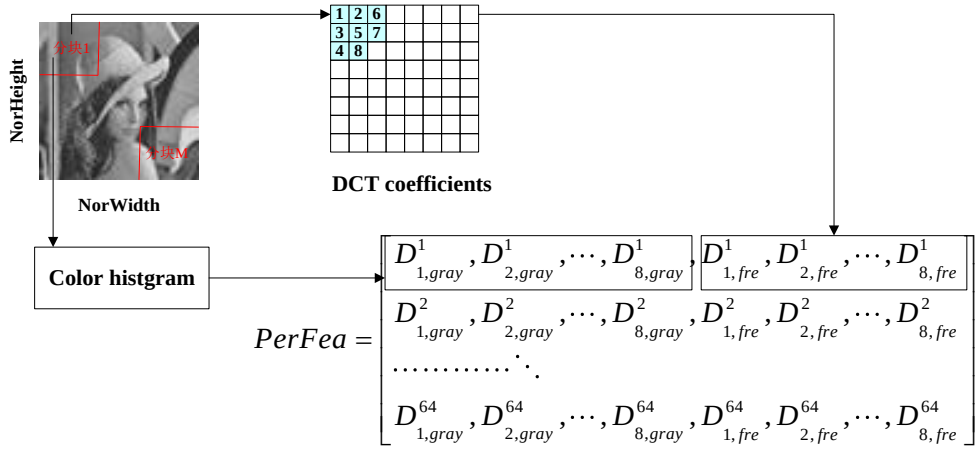


Figure 2. Process of perceptual feature extracting

2.3 PCA Based Inter-Quantization

Each column of perceptual feature (matrix) is an indicator, reflecting appropriate information of the input image. For example, the first column is constituted with each block's occurrences of pixels between 0 and 31, reflecting the distribution of the input image on the pixel interval. However, there is inherent correlation among adjacent blocks. So, each column of perceptual feature matrix contains some redundant information.

Principal Component Analysis is used to reduce, even elimilate, the redundant information in perceptual feature matrix. Then, the perceptual feature is compressed into a inter-feature matrix, with a smaller dimension (10*64) and few redundant information.

2.4 Hash Computation

Once the inter-feature matrix is generated, we can obtain the perceptual hash of input image via binarizing. Each column is binarized using the median of the rank-ordered coefficients.

If the subset of rank-ordered coefficients is denoted as $c(i), i = 0,1,\cdots,K$, then their median is calculated as $\mu = (c(K/2) + c((K+1)/2))/2$. Then, the perceptual hash of input image is obtained by thresholding each column with the median μ as follows:

$$hash_i = \begin{cases} 1, & c(i) \geq \mu \\ 0, & c(i) < \mu \end{cases}, i = 0,1,\cdots,K \quad (2)$$

3. Experimental Results

The proposed algorithm has been implemented with matlab scripts. The evaluation was based on 72 distinct images. These images are all from corel image galley (URL: http://calphotos.berkeley.edu/).

The bit error rate(BER) [9] [10] is denoted as the rate of mismatched bits by comparing two perceptual hashes.

3.1 Robustness to Image Operation

Robustness implies that perceptual hash functions should be robust to all kinds of image operation (contrast increase, median filter, JPEG compression, noise addition, histogram equalisation, laplace sharpen, rotation), since the underlying content is never changed. That's to say, the BER between the perceptual hash of the raw image and the perceptual hash of the operated image should be infinitely close to 0.The experimental results of robustness evaluation is presented in table 1.

TABLE I. RESULTS OF ROBUSTNESS EVALUATION

Image Operation		*Mean of BER*
Contrast Increase	-30%	8.50%
	-20%	3.27%
	20%	4.10%
	30%	9.20%
Median filter		10.06%
JPEG Compression	10%	3.00%
	20%	8.22%
	40%	17.50%
Noise Addition	Gaussian	7.10%
	Peper and Salt	8.50%
Histogram Equalisation		11.27%
Laplace Sharpen		20.24%
Rotation	-5	23.39%
	-3	15.70%
	+3	17.33%
	+5	22.00%

Table 1 shows the mean of BER between the raw image's perceptual hash and the operated image's perceptual hash. Notice that most of the results are lower than 0.1 and stay close to the theoretical value 0, only when the operation (JGEP compression 40%, rotation 5) has changed the underlying content. Meanwhile, the more the underlying content changes, the bigger the mean of BER is. For example, the result is 0.0300 while the JGEP compression is 10%, and 0.1750 while the JGEP compression is 40%.

3.2 Discrimination to Different Images

Discrimination means that the hash functions are statistically independent for different perceptual content, so that any two distinct images result in different and apparently random perceptual hash.

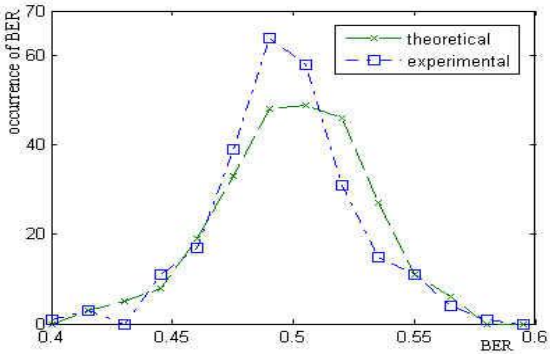


Figure 3. Distribution of BER between distinct images

Assume that two distinct images (i and i') are perceptually different, the theoretical optimal value of their BER $M_{BER}(pHash)$ can be estimated as follows:

$$M_{BER}(pHash) = E[BER(i,i')] \quad (3)$$

where i and i' are taken independently and randomly from a given image set, and $E[]$ denotes mathematical expectation. According to Baris Coskun's and Kevin Hamon's analysis and proof in article [9] and [10], the theoretical optimal value is speculated to be 0.5.

Without loss of generality, we calculate BER between 72 test images, using 255 BER computation overall. Then we count the occurrence of each BER value and obtain the experimental distribution of BER, as shown in Figure 3. Meanwhile, we also plot the theoretical probability density function with $\mu = 0.5$ and $\sigma^2 = 0.0009$ in Figure 3.

The BER between perceptual hash of distinct images has a Gaussian distribution around the mean value of 0.4996, which is close to the theoretical optimal value 0.5. Thus, the perceptual hash of different images can be regarded as statistically independent as expectation.

3.3 Detection to the Tamper—Addition of a Logo

Tamper always brings in malicious changes to the original content of raw image. Typical image tampering operations include adding LOGO, image mosaic and so on. In this

article, we mainly concern the tamper of adding LOGO, since the addition of a LOGO causes minimal change and is widely used with the increasing spread of Internet. The sample of adding LOGO is shown in Figure 4:

Figure 4. Sample of adding a LOGO

The content of tampered image shown in Figure 4 is much similar with the law image, only with malicious changes via adding a logo on the left. In order to detect such tampering operation to image, the BER between raw image and tampered image should be bigger than the BER of robust operation. Meanwhile, in order to distinguish tampered image from distinct image, the BER between raw image and tampered image should also be smaller than the BER of distinct images. We calculate 12 BER between raw image and tampered image, as shown in Figure 5:

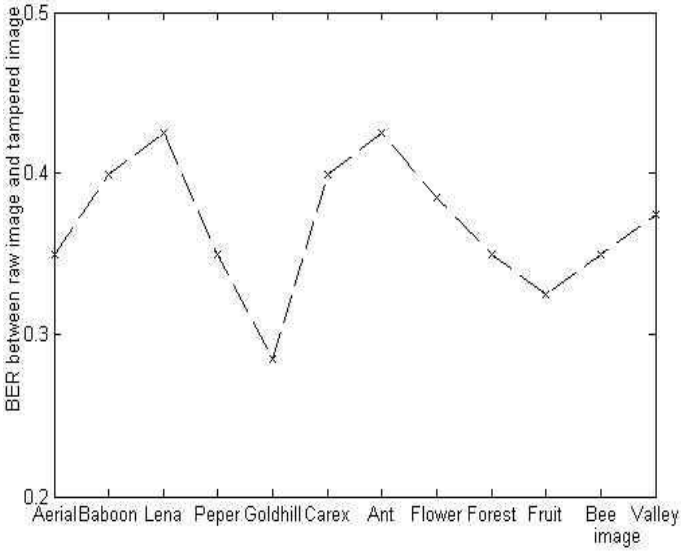


Figure 5. BER between raw image and tampered image

In Figure 5, most of the BER between raw image and tampered image is in the range of [0.3, 0.45]. Note that the most BER of robust operation is lower than 0.1 and the BER of distinct images has a Gaussian distribution around the mean value of 0.4996. Thus, the proposed image perceptual hash algorithm can effectively address such tamper detection problem with advantageous robustness and discrimination.

4 Conclusions

In this article we proposed a robust and discriminative image perceptual hash algorithm in order to address the problem of the tamper detection problem of images. We integrate color histogram and low-frequency DCT coefficients of image blocks as perceptual feature, then compress perceptual feature as inter-feature with PCA, and threshold the inter-feature to create a robust hash. Experimental results show that the proposed algorithm is advantageous at robustness since the most BER of robust operation is lower than 0.1, but it is a pity that the robustness toward rotation is not so perfect as assumed. It's also found that the proposed algorithm has a very discriminative power because the BER of distinct images has a Gaussian distribution around the mean value of 0.4996, which is close to the theoretical optimal value 0.5. With such advantageous robustness and discrimination, the proposed algorithm can effectively detect the tampering operation of adding a LOGO.
Future investigation will address the problem of verification of the tamper detection ability toward other tampering operations. Meanwhile, we will extend this still-image perceptual hash method to video clip to address the problem of video authentication and copyright protection.

Acknowledgments

This work was partially funded by Natural Science Foundation of China through the Free Application Program under contract 61103174, Science and Technology Program of Shenzhen (China) through the Basic Research Program under contract JC201105170647A, and Laboratory and Device Management Research Foundation of Shenzhen University through the Basic Research Program under contract 2011045.

The Authors would thanks to Natural Science Foundation of China and Shenzhen Municipal Science and Technology Trade and Industry and Information Technology Commission (China) for their funding of our projects. The Authors also express gratitude to Laboratory and Device Management Department of Shenzhen University for their support and collaboration.

References

[1] NIU Xia-mu ,JIAO Yu-hua, "An Overview of Perceptual Ha shing," ACTA ELECTRONICA SINICA.China, Beijing, Vol.36, No. 7, 2008, pp. 1405-1411.

[2] Bian Yang, Fan Gu, Xiamu Niu, "Block Mean Value Based Image Perceptual Hashing," International Conference on Intelligent Information Hiding and Multimedia Signal Processing(IIH-MSP '06), 2006, pp.167-172.

[3] J. Fridrich, "Visual hash for oblivious watermarking," Proc .IS&T/SPIE 12th Annu. Symp., Electronic Imaging, Security and Watermarking of Multimedia Content II, San Jose, CA, Jan. 2000, pp.286-294.

[4] R. Venkatesan, S.Koon, M. Jakubowski, and P. Moulin, "Robust image hashing," Proc. IEEE Int. Conf. Image Processing 2000, vol. 3, pp.664-666.

[5] M. K. MIhCak,R Venkatesan, "New Iterative Geometric Methods for Robust Perceptual Image Hashing," Proc of ACM Workshop on Security and Privacy in Digital Rights Management . Philadelphia :LNCS ,2001,pp.13-21.

[6] F.Lefbvre,B.Macq,J.-D.Legat, "RASH: RAdon Soft Hash algorithm," Proc. EUSIPCO,Toulouse,France,2002. pp.54-61.

[7] J.S.Seo, J.Haitsma, T.Kalker,C.D.Yoo,"A robust image fingerprinting system using the radon transform," Signal Process.: Image Commun., vol. 19, no. 4, 2004,pp. 325–339.

[8] Hui Zhang,Haibin Zhang,Qiong Li, Xiamu Niu, "A Multi-Channel Combination Method of Image Perceptual Hashing," Fourth International Conference on Networked Computing and Advanced Information Management,2008. NCM '08.Volume 2, 2008,pp.87-90.

[9] Baris Coskun, Bulent Sankur, Nasir Memon, "Spatio-Temporal Transform Based Video Hashing," IEEE Transactions on Multimedia,VOL. 8,NO. 6, 2006. pp. 1190-1208.

[10] Kevin Hamon, Martin Schmuker, Xuebing Zhou, "Histogram-based perceptual hashing for minimally changing video sequence," The Second IEEE International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution , 2006.

Zeng Jie received his M. Eng. Degree in Signal and Information Processing from the Tianjin University, Tianjin, China in 2001. From 2001 to 2006, He worked as a software engineer in Huawei Technology Ltd. Now he is a lecturer in the College of Information Engineering, Shenzhen University, China. His current research intersts include wireless networks, wireless commmunication, and cooperative wireless networks.

An Improved Interference Cancellation Scheme for Two-User MIMO-MAC

Xinji Tian, Cheng Song

School of Computer Science and Technology, Henan Polytechnic University
Jiaozuo, Henan 454000, China

Abstract

Multi-Input Multi-Output (MIMO) Multiple Access Channels (MAC) for two-user suffers from co-channel interference. For this problem, an interference cancellation scheme based on limited feedback is proposed. Through diagonalization processing for transmitted signals according to feedback information, the co-channel interference is eliminated. Not only the reliability is improved, but also each signal can be Maximum Likelihood (ML) decoded separately. Simulation results show that, compared to the existing interference cancellation scheme, the gain of the proposed scheme is 2dB at the Bit Error Rate (BER) of 10^{-3}.

Keywords: *Multi-input Multi-output, Multiple Access Channels, Co-channel Interference, Maximum Likelihood*

1. Introduction

Multiple-Input Multiple-Output (MIMO), which is one of the mandatory techniques for the next generation wireless communication systems, has the features of spatial multiplexing and spatial diversity [1]. MIMO Multiple Access Channels (MAC) offers substantial capacity improvements and has attracted considerable research attention [2]. There is serious co-channel interference over MIMO-MAC since multiple users send signals simultaneously in the same frequency [3]-[7]. It not only affects the system reliability, but can also increase the decoding complexity.

In [5], a multiuser detection method is presented for the problem of co-channel interference over MIMO-MAC, based on interference suppression scheme in single-user MIMO systems. In [6], an improved interference cancellation method is proposed for two-user MIMO-MAC. However, both schemes only eliminate partial co-channel interference at the receiving terminal, and performance can be improved further. Therefore, transmission schemes over MIMO-MAC with limited feedback information are studied, in which partial interference is eliminated through preprocessing at the transmitters.

In [9], an Alamouti code based transmit scheme with a phase feedback is presented for two-user MIMO-MAC. The transmit power, which is feedback information, is derived with the goal of maximizing the Signal-to-Noise Ratio (SNR) at the receiver. However, only partial interference is canceled. The Maximum Likelihood (ML) decoding by single symbol is impossible, and the decoding complexity can be lowered further.

In order to mitigate co-channel interference and reduced decoding complexity, an interference cancellation scheme based on limited feedback is proposed for two users MIMO-MAC. Through diagonalization processing for transmitted signals according to feedback information, the co-channel interference is eliminated. The reliability is improved, and each signal can be ML decoded separately as well. Theoretical analysis shows that the decoding complexity of the proposed code downgrades to 20% and 33% as required by [9] for modulation order of 4 and 16, respectively. Simulation results show that, the gain of the proposed scheme is at least 2dB at the Bit Error Rate (BER) of 10^{-3} compared to the scheme in [9].

2. System Model

The system model is shown in Fig. 1. There are two users and one receiver each with two antennas. Let $\boldsymbol{H}$ and $\mathbf{G}$ denote the channel matrix from user 1 and user 2 to the receiver respectively, which are given by $\boldsymbol{H}=[\boldsymbol{h}_1,\boldsymbol{h}_2]$ and $\boldsymbol{G}=[\boldsymbol{g}_1,\boldsymbol{g}_2]$, where $\boldsymbol{h}_i=[h_{1i},h_{2i}]^T$, $\boldsymbol{g}_i=[g_{1i},g_{2i}]^T$, $i=1,2$.

The codeword of two users are defined as

$$\boldsymbol{S}=\begin{bmatrix} s_1 & -s_2^* \\ s_2 & s_1^* \end{bmatrix} \quad \boldsymbol{X}=\begin{bmatrix} x_1 & -x_2^* \\ x_2 & x_1^* \end{bmatrix}$$

where s_i and x_i are the modulated signals for user 1 and user 2, respectively, $i=1,2$.

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ denote the precoding matrix for user 1 and user 2 respectively, which are defined as

$$\boldsymbol{A}=\begin{bmatrix} p_1 & 0 \\ 0 & q_1 \end{bmatrix} \quad \boldsymbol{B}=\begin{bmatrix} p_2 & 0 \\ 0 & q_2 \end{bmatrix}$$

where p_1, q_1, p_2 and q_2 are complex numbers, satisfied $|p_1|^2+|q_1|^2=2$ and $|p_2|^2+|q_2|^2=2$. p_1 and p_2 are feedback information, while q_1 and q_2 are calculated at the transmitter according to p_1 and p_2.

The received signal vector $[\boldsymbol{r}_1,\boldsymbol{r}_2]$, with dimension of $N\times 1$, can be expressed as

$$[\boldsymbol{r}_1,\boldsymbol{r}_2]=\boldsymbol{HAS}+\boldsymbol{GBX}+[\boldsymbol{n}_1,\boldsymbol{n}_2] \tag{1}$$

where $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ are $N\times 1$ noise vectors.

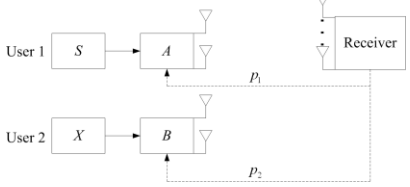


Fig. 1 System model of the proposed scheme

3. Calculation of Feedback Information

The receiver forms a rearranged vector as follows

$$\begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2^* \end{bmatrix}=\boldsymbol{H}'\begin{bmatrix} \boldsymbol{s} \\ \boldsymbol{x} \end{bmatrix}+\begin{bmatrix} \boldsymbol{n}_1 \\ \boldsymbol{n}_2^* \end{bmatrix} \tag{2}$$

where $\boldsymbol{s}=\begin{bmatrix} s_1 \\ s_2 \end{bmatrix}$, $\boldsymbol{x}=\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $\boldsymbol{H}'$ is expressed as

$$\boldsymbol{H}'=\begin{bmatrix} p_1\boldsymbol{h}_1 & q_1\boldsymbol{h}_2 & p_2\boldsymbol{g}_1 & q_2\boldsymbol{g}_2 \\ q_1^*\boldsymbol{h}_2^* & -p_1^*\boldsymbol{h}_1^* & q_2^*\boldsymbol{g}_2^* & -p_2^*\boldsymbol{g}_1^* \end{bmatrix} \tag{3}$$

Multiply both sides of Equation (2) by matrix $(\boldsymbol{H}')^H$ to achieve

$$\boldsymbol{y}=\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}=\begin{bmatrix} \alpha & 0 & \rho & \varepsilon \\ 0 & \alpha & -\varepsilon^* & \rho^* \\ \rho^* & -\varepsilon & \eta & 0 \\ \varepsilon^* & \rho & 0 & \eta \end{bmatrix}\begin{bmatrix} s_1 \\ s_2 \\ x_1 \\ x_2 \end{bmatrix}+\begin{bmatrix} n_1 \\ n_2 \\ n_3 \\ n_4 \end{bmatrix} \tag{4}$$

$$\alpha=|p_1|^2\|\boldsymbol{h}_1\|^2+|q_1|^2\|\boldsymbol{h}_2\|^2$$

$$\eta=|p_2|^2\|\boldsymbol{g}_1\|^2+|q_2|^2\|\boldsymbol{g}_2\|^2$$

$$\rho=p_1^*p_2(\boldsymbol{h}_1^*)^T\boldsymbol{g}_1+q_1q_2^*(\boldsymbol{h}_2)^T\boldsymbol{g}_2^*e^{j\theta-j\beta}$$

$$\varepsilon=p_1^*q_2(\boldsymbol{h}_1^*)^T\boldsymbol{g}_2e^{j\beta}-q_1p_2^*\boldsymbol{h}_2^T\boldsymbol{g}_1^*e^{j\theta}$$

$$[n_1,n_2,n_3,n_4]^T=(\boldsymbol{H}')^H\begin{bmatrix} \boldsymbol{n}_1 \\ \boldsymbol{n}_2^* \end{bmatrix}$$

s_i and $x_i\,(i=1,2)$ keeps orthogonal in their transmission if $\rho=\varepsilon=0$, so that, symbol by symbol ML decoding can be realized. Equations (5) can be obtained with $\rho=\varepsilon=0$.

$$\begin{cases} ap_1^*p_2+bq_1q_2^*=0 \\ cp_1^*q_2-dq_1p_2^*=0 \end{cases} \tag{5}$$

where $a=(\boldsymbol{h}_1^*)^T\boldsymbol{g}_1$, $b=(\boldsymbol{h}_2)^T\boldsymbol{g}_2^*$, $c=(\boldsymbol{h}_1^*)^T\boldsymbol{g}_2$, $d=\boldsymbol{h}_2^T\boldsymbol{g}_1^*$. Through solving equation (5), we get

$$p_1=\sqrt{\frac{2bd}{bd-ac}} \qquad p_2=\sqrt{\frac{2bc}{bc-ad}} \tag{6}$$

Thus, q_1 and q_2 are obtained as

$$q_1=\sqrt{\frac{2ac}{ac-bd}} \quad q_2=\sqrt{\frac{2ad}{ad-bc}} \tag{7}$$

4. Decoding method

From the above analysis, symbol by symbol decoding can be realized when Equation (6) and (7) are satisfied. Specific decoding procedure is as follows.

Step 1, computer $\boldsymbol{H}'$ according to (3);

Step 2, the receiver forms a rearranged vector $\begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2^* \end{bmatrix}$, and then multiply $\begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2^* \end{bmatrix}$ by $(\boldsymbol{H}')^H$ to (3) obtain $\boldsymbol{y}=[y_1,y_2,y_3,y_4]^T$;

Step 3, computer α, and obtain the detected symbol of user 1 by Equation (8).

$$s_i'=\arg\min_{\hat{s}_i\in C}\|y_i-\alpha\hat{s}_i\|^2 \tag{8}$$

where C denotes the constellation point set.

Step 4, computer η, and obtain the detected symbol of user 2 according to $y_k\,(k=1,2)$ and η by

$$x_i'=\arg\min_{\hat{x}_i\in C}\|y_i-\eta\hat{x}_i\|^2 \tag{9}$$

4. Computational Complexity of Decoding

In this section, two schemes are compared in terms of computational complexity of decoding.

$(72N+12M-8)$ flop is required in the process of decoding for the proposed scheme, where M is modulation order.

Minimum Mean Squared Error Successive Interference Cancellation (MMSE-SIC) is adopted as detection in [9]. Since there are several kinds of MMSE-SIC,

We assume [9] uses the low complexity MMSE-SIC presented by [10]. In this condition, $(432N + 244)$ flop is required in the process of decoding for [9].
By calculation, the decoding complexity of the proposed code downgrades to 20% and 33% as required by [9] for modulation order of 4 and 16, respectively.

4. Simulation Results

The BER performance of the proposed system and [9] is investigated. We consider uncoded systems with 4QAM and 16QAM constellations. Fully spatially uncorrelated channels are employed. Assume that the elements of channel and noise are obtained from an independent and normal distribution.
Fig. 2 and Fig. 3 show BER versus SNR at the transmitter for different modulation, with N=2 and N=3, respectively. As observed in these figures, the performance of the proposed scheme significantly outperforms that of [9]. The reason is that, the proposed scheme mitigates all interference rather than partial interference, as in the scheme of [9]. The gain of the proposed scheme is 2dB and 3dB at the BER of 10^{-3} for $N = 2$ and $N = 3$, respectively. Thus, the gain increases with the increment of transmit antennas.

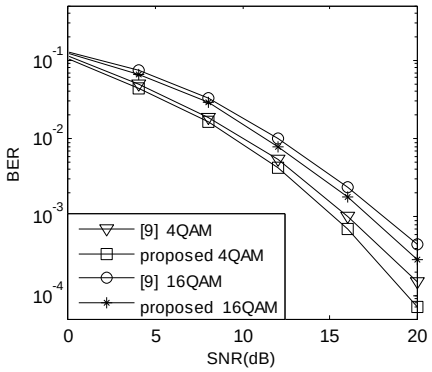


Fig. 2 BER of the two schemes for QPSK with $N = 2$

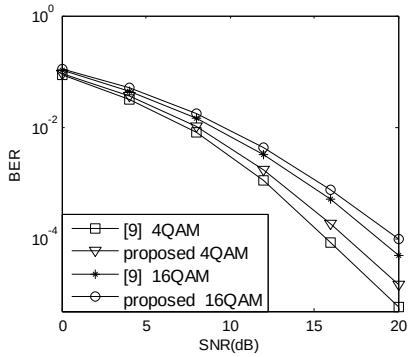


Fig. 3 BER of the two schemes for QPSK with $N = 3$

4. Conclusions

A phase is required to feedback to only one transmitter in [9] while one complex number is required to feedback to each transmitter in the proposed scheme. Since every complex number contains two real numbers, the proposed scheme needs three more real numbers as feedback information, which is the disadvantage. However, the system performance is enhanced, and the decoding complexity is also reduced. In addition, the gain increases with the increment of transmit antennas.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 61202286; the National Natural Science Foundation of China under Grant No.61104079.

References

[1] G. David, S. Mansoor, S. Dashan, S .J. Peter, N Ayman, "From theory to practice: an overview of MIMO space-time coded wireless systems", IEEE J. Select. Areas Commun., vol. 21, no. 3, 2003, pp. 281-302.
[2] K. K. Raj, C. Giuseppe, "Channel state feedback over the MIMO-MAC", IEEE Trans. Inf. Theory, vol. 57, no. 12, 2011, pp. 7787-7797.
[3] L. Feng, J. Hamid, "Interference cancellation and detection for more than two users", IEEE Trans. Commun., vol. 59, no. 3, 2011, pp. 901-910.
[4] K. Javad, C. A. Robert, "Multiuser interference cancellation and detection for users with more than two transmit antennas", IEEE Trans. Commun., vol. 56, no. 4, 2008, pp. 574-583.
[5] K. Javad, C. A. Robert, "Multiuser interference cancellation and detection for users with more than two transmit antennas", IEEE Trans. Commun., vol. 56, no. 4, 2008, pp. 574-583.
[6] M. R. Bhatnagar, A. Hjrungnes, "Improved interference cancellation scheme for two-user detection of Alamouti code", IEEE Trans. on Signal Process., vol. 58, no. 8, 2010, pp. 4459-4465.
[7] L. Feng, J. Hamid, "Multiple-antenna interference cancellation and detection for two users using precoders", IEEE Journal of selected topic in signals processing, vol. 3, no. 6, 2009, pp. 1066-1078.
[8] J. T. Wang, "Joint MMSE equalization and power control for MIMO system under Multi-user interference", IEEE Commun. letters, vol. 16, no. 1, 2012, pp. 54-56.
[9] Y. J. Kim, C. H. Choi, and G. H. Im, "Space-time block coded transmission with phase feedback for two-user MIMO-

MAC," in Proc. IEEE Intl Conf. on Communications (ICC), Kyoto, Japan June 2011.

[10]Tsung H L, Yu L L, "Modified fast recursive algorithm for efficient MMSE-SIC detection of the V-BLAST system", IEEE Trans. on Wireless Communications, vol. 7, no. 10, 2008, pp. 3713-3717.

Xinji Tian received Doctor degree at Beijing University of Posts and telecommunication in 2011. She has been employed at Henan Polytechnic University since the summer of 2011. She has been supported by the National Natural Science Foundation of China. Her research fields are MIMO technology and space time code.

Cheng Song born in 1980, Ph.D. . His research interests include information security, Trusted Computing and Internet of things, etc.

A Sort of Web Service Selection Strategy Based on the Fusion of QoS and Service Reliability

Yucheng Liu[1], Yubin Liu[2]

[1] College of Electrical & Information Engineering, Chongqing University of Science & Technology
Chongqing, 401331, China

[2] School of Continuing Education, Panzhihua Univerisity
Panzhihua, 617000, China

Abstract

Aimed at the Web function being too similar to quickly filter out Web services, the paper proposed a sort of selection strategy in Web service based on the fusion of QoS and reliability. The paper made the analysis on the limitations of the current selection mechanism and evaluated the Web service selection from two aspects in subjectivity and objectivity and constructed the model of Web service selection based on the model of QoS monitoring, target consumption group, service quality estimation and feedback evaluation. It took three cases as example to make the simulation and validated that the proposed strategy not only could adapt dynamic varying environment, but also could ensure the actual service quality and overcame the individual difference of evaluation. The research results demonstrate that the proposed strategy of service selection can filter speedily out the Web service needed by requester from the set of Web service in abundance.

Keywords: *Web Function, Web Service, Fusion, Quality of Service, Service Reliability, Selection Strategy.*

1. Introduction

The selection mechanism of Web service has gone through two stages, namely the QoS based selection and user evaluation feedback based selection. The former is a sort of service selection method based on objective evaluation. Its selective model is shown in Fig.1. The module of ServiceMatchMake in Fig.1 is used for match evaluation in QoS of Web services that meet the functional requirements, but the precondition is that the data in QoS must be truly credible. Practically, there are a large number of false services in the network. The model ignores the dynamic changes in QoS, therefore, the selected service based on QoS may not be the best service. The later is based on user evaluation feedback. Its service selection model is shown in Fig.2. Its accuracy depends on the actual effect of service evaluation from the service consumer. It is a sort of service evaluation method based on subjective evaluation. Therefore, it is difficult to obtain the high quality Web service.

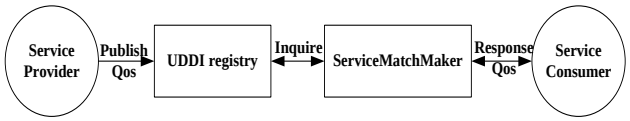


Fig. 1 Web service selection model based on QoS.

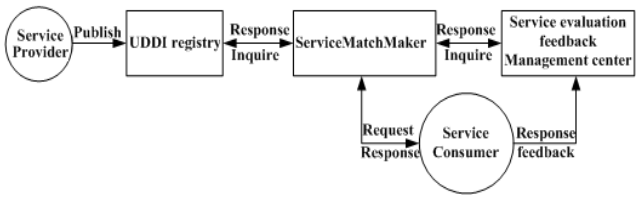


Fig. 2 Web service selection model based on feefback.

In the evaluation methods mentioned above, the user feedback owns one-sidedness, and the subjective evaluation has its limitations. In addition, the selective method suffers the influence of context dependency in user feedback. Aiming at the puzzles mentioned above, lots of scholars have researched deeply into these puzzles. Zhang Wen-bo and Shi Wei-feng researched the dynamic Web services composition based ion BPEL and QoS[1]. Gao Ya-chun and Zhang Wei-qun explored the Web Service Description and Selection Mechanism Based on QoS Ontology[2]. Chen Li-jin and Zhou Ya researched Dynamic Web Service Selection based Multi-QoS constraints[3]. Yang Mo and Wang Li-na researched Web service reliability enhancement method based on trust fault tolerant[4]. The authors think what cause puzzles of selection model based on QoS or based on user evaluation feedback is that the model built can not reflect completely the essence of objects. Therefore, the paper made further study on selection strategy of Web service based on QoS and confidence fusion in order to get a better selection model.

2. Improvement on Selection Strategy

Aimed at the puzzles of one-sidedness and limitation mentioned above, in order to avoid false QoS made by service provider in the attribute value of QoS for objective evaluation, the monitoring mechanism of QoS was introduced. Moreover, the model of object consumer group was introduced to overcome the influence of context dependency in user feedback. And the model can distinguish effective and invalid evaluation so as to put an end to the influence of invalid evaluation for Web service.

2.1 QoS Monitoring Model

The monitoring model was shown in Fig.3. It can make the monitoring and update of QoS attribute value periodically to ensure the confidence and real time effectiveness of QoS attribute value. Suppose the Web service to be as S_i , QoS attribute value as $Q_i = \{q_{i,1}, q_{i,2}, \ldots, q_{i,n}\}$, then Q_i can be updated as shown in Eq.(1).

$$Q_i = w \times Q'_{i,0} + (1-w) \times f(Q'_{i,1}, Q'_{i,2}, ..., Q'_{i,j}) \quad (1)$$

Where, $Q'_i = \{Q'_{i,1}, Q'_{i,2}, ..., Q'_{i,j}\}$ is the QoS data collected by monitor, $Q'_{i,0}$ is the initial QoS data provided by service provider. w is the weight of initial QoS and it expresses by exponential function $1/2^n$, and n is the data number collected by monitor. Therefore the weight value of initial QoS can be adjusted dynamically with increase of collected data amount, namely the specific weight of $Q'_{i,0}$ gets more and more small in the computation. f is a statistical function as shown in Eq.(2).

$$f(Q'_{i,1}, Q'_{i,2}, ..., Q'_{i,j}) = \frac{1}{j}\sum_{m=1}^{j} \lambda^{date(t-t_m)} Q'_{i,m} \quad (2)$$

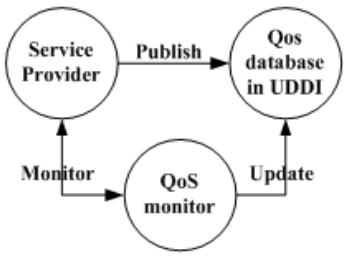


Fig. 3 QoS monitoring model.

2.2 Target Consumption Group Model

The model of target consumption group was shown in Fig.4. For a certain service, the consumers can be divided into m individual group. With regard to Web service S_i, it puts up two announcement parameters, namely the announcement vector QoS of service quality and target service group of the service. For any target group of them, it can be identified by a unique service quality vector, namely $TCGroup_{i,k} = TCGQoS_{i,k}, 0 < k \le m$. For such a target group, it can also be expressed as $TCGroups_i = [TCGQoS_{i,1}, TCGQoS_{i,2}, \cdots, TCGQoS_{i,m}]$.

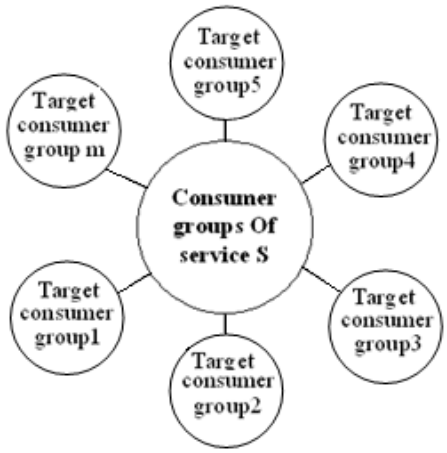


Fig. 4 Target consumer group model.

The flowchart of classification for target consumer was shown in Fig.5. The service provider can locate several consumption groups for service S_i, it can be expressed as $TCGroups_i = [TCGroup_{i,1}, TCGroup_{i,2}, \cdots, TCGroup_{i,m}]$. For service consumer C, the requirement quality is expressed as Q_c, and after computing the similarity between Q_c and $TCGroup_{i,1}$, $TCGroup_{i,2}$, ..., $TCGroup_{i,m}$, it sorts according to rule from big to small, elects target consumption group of biggest similarity, and makes it join in the target consumption group. The measure of similarity difference is adopted by cosine similarity value, because it mainly focuses on the difference in direction of two vectors, and not distance or length. The similarity value is directly mapped into the interval [-1, 1], and the dependency value lies on the between from -1 to 1. In which, "1" shows completely positive correlation, and "-1" represents completely negative correlation.

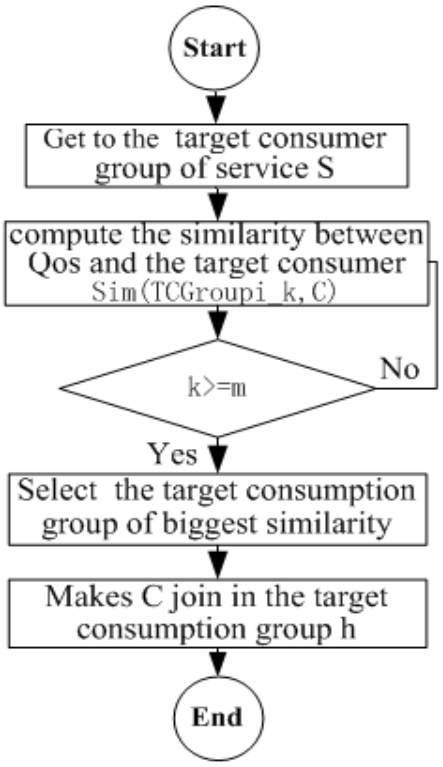


Fig. 5 Flowchart of position target consumer groups for service consume.

Assume the requirement QoS of service consumer C to be as $Q_c = \{q_{c,1}, q_{c,2}, \ldots, q_{c,n}\}$, and the QoS feature of target consumption group $TCGroup_{i,k}$ of service provider is $TCGQoS_{i,h} = \{tq_{i,h,1}, tq_{i,h,2}, \ldots, tq_{i,h,n}\}$, then the similarity between the both can be expressed by Eq.(3). If the value is smaller then it shows that the similarity between the both is bigger.

$$Sim(Q_e, TCGQoS_{i,h}) = \frac{\sum_{j=1}^{n}(q_{c,j} tq_{i,h,j})}{(\sum_{j=1}^{n} q_{c,j}^2 \sum_{j=1}^{n} tq_{i,h,j}^2)^{\frac{1}{2}}} \quad (3)$$

2.3 Service Evaluation Model

The model of service evaluation makes evaluation service from two aspects, namely subjective and objective evaluation. The service selection model based on QoS monitoring and evaluation classification can make QoS own the real effectiveness through introducing time factor to update the QoS dynamically. The confidence of consumer for Web service comes from direct confidence and indirect recommendation confidence, and after respectively computing it can obtain the totality confidence of candidate Web service. It is shown as Eq.(4).

$$T = w_t \times DirectTrust + (1 - w_t) \times IndirectTrust \quad (4)$$

In which, w_t represents the confidence weight.

2.4 Evaluation Feedback Model

The user evaluation feedback management model can be shown in Fig.6. In which, each service has m target consumption groups, and anyone service consumer must be assigned into a target consumer group. The service evaluation of one and the same target consumption group also must be put into the same storage pool. If the feedback information does not belong to the malicious evaluation then it must be encouraged and rewarded, else it must be punished for evaluation of service consumer. The mechanism of evaluation of rewards and punishments can make corresponding reward and punishment according to average evaluation similarity of belonged target consumer group. If it is high for the average evaluation similarity of the service consumer group, then it illustrates that the evaluation confidence is higher, and when it is greater than a certain getting value then it can be considered that the evaluation made by service consumer is impartial, it should be encouraged and rewarded. In opposite it should be punished. The mechanism of rewards and punishments can reduce the malicious evaluation.

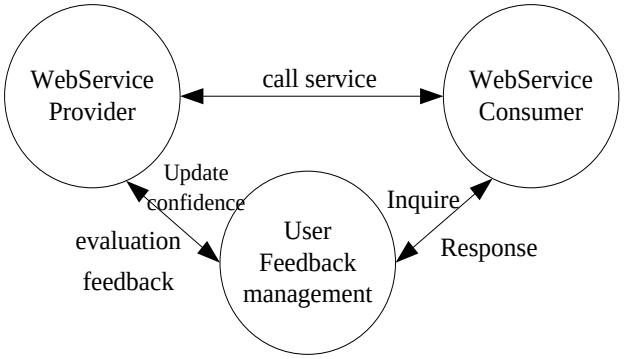


Fig. 6 User feedback management model.

3. Service Selection Strategy Based on Fusion

By means of service confidence evaluation and the third side monitor of service quality, based on the service selection mechanism of service classification and confidence, suitable user service can be selected according to the objective service quality value of comprehensive evaluation and user subjective evaluation value.

3.1 Service Selection Algorithm Flow

The flow of service selection algorithm was shown in Fig.7. The steps of the algorithm: 1)To make semantics

matching and short-cut process constraint processing for announcement vector set of candidate service; 2)To make evaluation for announcement vector of candidate service. In which, the vector formed by correlation attribute value of all QoS of anyone service is called as service announcement vector; 3)To find out the target consumption group in the confidence evaluation, to compute direct confidence value, to find out all evaluation of the target consumption group, and to compute indirect confidence value of the target consumption group for the service consumer; 4)To dispose the above evaluation results synthetically, and return the final evaluation result.

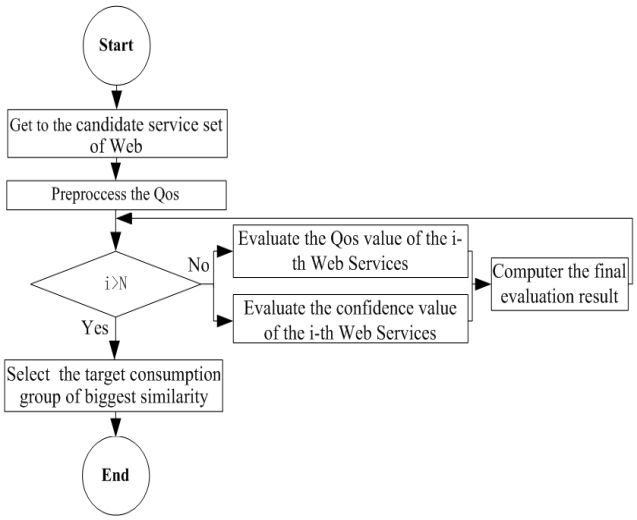


Fig. 7 User feedback management model.

3.2 QoS Constraint Processing

Semantics matching of QoS. It first makes semantics matching for service request QoS and service announcement QoS[5-7], and it can obtain the service announcement vector set Q_v of semantics matching satisfied by the service request vector.

Short-cut process constraint of QoS. In order to unify the computing and comparative analysis, by means of unit transform mode of UnitConversion record in ontology Database, it makes single standard processing for measure mode of QoS service announcement vector. The classification of short-cut process constraint of QoS constraint consists of numerical value type and Boole type and grade type, and the short-cut process constraint makes short-cut process constraint processing for service announcement set of satisfied QoS semantics matching.

3.3 Constraint Processing of QoS

After QoS constraint processing, it can obtain the Web service candidate set $S=\{S_1,S_2,\ldots,S_m\}$, and each service has n pieces of QoS attribute, therefore it can constructs a $m\times n$ matrix Q as shown in Eq.(5). In the matrix Q, each row represents a Web service, and each column represents one and the same QoS attribute.

$$Q=\begin{bmatrix}Q_1\\Q_2\\\vdots\\Q_m\end{bmatrix}=\begin{bmatrix}q_{1,1} & q_{1,2} & \cdots & q_{1,n}\\q_{2,1} & q_{2,2} & \cdots & q_{2,n}\\\vdots & \vdots & & \vdots\\q_{m,1} & q_{m,2} & \cdots & q_{m,n}\end{bmatrix} \quad (5)$$

In the matrix Q, the large the numerical value is, the beneficial for service request. According to the QoS attribute after quantization it can compute the QoS evaluation value of Web service, and it is shown as in Eq.(6).

$$QoS(S_i)=\frac{1}{n}\sum_{p=1}^{n}q_{i,p} \quad (6)$$

3.4 Confidence Evaluation

It has been researched on confidence[8-10], but they ignored all the correlation of context, and it results in different for the same service confidence in different service consumer. Aimed at the shortage mentioned above, the paper firstly seeks that the service consumer belongs to which target consumption group of the service, then after seeking attribute target consumption group it can find direct confidence value through computing. Finally according to Eq.(7), it can find the indirect confidence value.

$$Indirect-Trust(C,S_i,t)$$
$$=(\sum_{r=1}^{z}RP_{c,r}\times valfr_{i,r}\times\lambda^{date(t-t_r)})/z \quad (7)$$

In which, $RP_{c,r}$ gives the confidence of service consumer of $fr_{i,r}$, $valfr_{i,r}$ is the evaluation score of consumer for service S_i.

3.5 Web Service Selection

After completing the evaluation $S=\{S_1,S_2,\cdots,S_n\}$ of all services, it can obtain an evaluation matrix $ER=\{ER_1,ER_2,...,ER_n\}$ of comprehensive considering QoS as well as confidence. Sorting the element in ER from big to small, it can obtain the biggest value ER_i in ER, and the service S_i of the biggest value ER_i must be the only section. The above flow joined the evaluation computing

of user evaluation value of requirement similarity, and the evaluation value joined more actual subjective judgment information, therefore it enhanced the service precision ratio, and it can satisfy the QoS selection requirement of user Web service in a certain grade.

4. Implementation of Service Selection Strategy

4.1 Service selection frame

The total frame of service selection mechanism implementation was shown as in Fig.8.

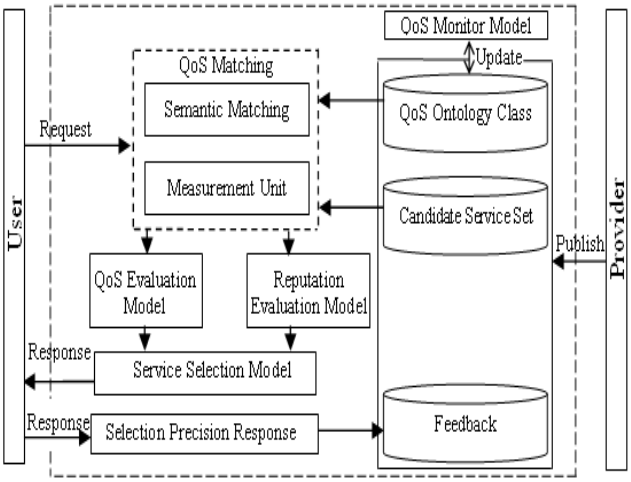


Fig. 8 Service selection frame.

1)The service monitor module QoSMM is in charge that it monitors the QoS vector of service announcement, stores the history data of monitoring, and updates the parameter value of QoS announcement vector periodically.
2) The pretreatment module QMM makes pretreatment for service request QoS vector and service announcement QoS vector, including QoS parameter semantics matching and constraint treatment of QoS parameter short cut process, and it makes all QoS parameter standardization.
3) The evaluation module QEM makes average value computing for result set QoS, and finds out the QoS evaluation value. The input of module QEM is the result set of pretreatment module QMM.
4) For confidence evaluation module TDEM, the input is the module QMM. It produces candidate service set after through QoS matching. It finds out the target consumption group stated by service request in the candidate service according to the service request QoS vector, and then it makes evaluation estimate according to the evaluation value provided by target consumption group.
5) The service selection module SSM makes the sum of evaluation value produced by QEM and TDEM according to the weight value, and then makes sorting for candidate selection, and finally it selects the service in front of M service of candidate set, and return to the service consumer.
6) Module SPRM of user satisfaction investigation is in charge of collecting user satisfaction degree of process execution result in service selection.

4.2 Response Mechanism of Service Request

The flow of response mechanism was shown as in Fig.9. After receiving the QoS service request, through four processing flow it can provides the response result. 1) It can create a new QoS matrix after the candidate service set through the QoSMM QoS pretreatment. 2) QEM module completes the evaluation for QoS. 3)TDEM module finds out the target consumption group according to the requirement QoS of service consumer, and it computes the user evaluation value and completes the confidence evaluation. 4) SSM module is in charge of service selection, and makes its result provide to service consumer.

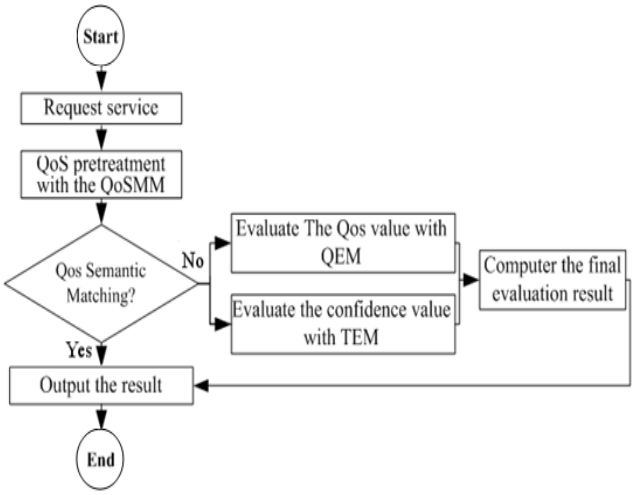


Fig. 9 Response flow of service request.

In the confidence model of the proposed mechanism, it owns the comparability because of the criterion being based on the evaluation of the same target consumption group, and therefore it can finds out the most suitable service of requirement itself for user.

5. Experiment Simulation

For convenience of comparison and description, here the algorithm of Web service selection based on QoS is called as SMQ, and the mechanism of service selection based on

QoS and user feedback is called as WSMQF. Compared WSMQF with SMQ algorithm, it increased the steps of feedback evaluation, and therefore it increased the time cost. Compared WSMQF with SMF algorithm, it increased two steps of evaluating QoS and seeking target consumption group, therefore its time cost is also increased a little. But after locating the belonged target consumption group, because the number of user evaluation is reduced compared with the original, so the time cost is also reduced. The following is the test results of simulation contrast for the time cost of three sorts of algorithm under different conditions.

1) For the service consumer C, under the condition of invalid evaluation being fixed, with the increasing of valid evaluation number, the simulation result of time cost is shown as in Fig.10 under different mechanism. With the increasing of valid evaluation, the time cost of WSMQF is also increased. Related to SMF, under the condition of the same number of invalid and valid evaluation, the time cost of WSMQF is less than SMF. SMQ does not deal with user feedback and only deals with static state QoS evaluation, and therefore its time cost is less, and basically it is not changed steadily.

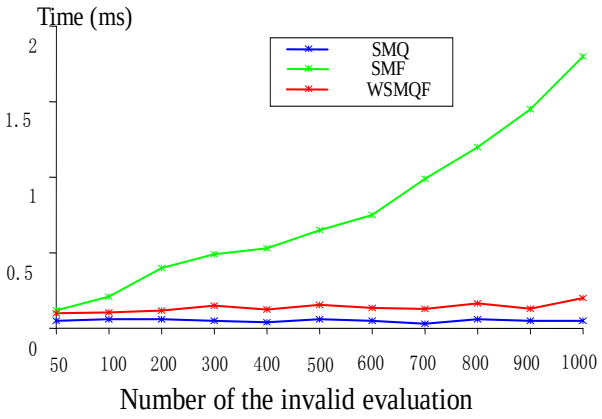


Fig. 10 Time spending under different invalid evaluation.

2)For service consumer C, under the condition of valid evaluation number being fixed, with the increasing of invalid evaluation, the simulation result of time cost of WSMQF is shown as in Fig.11. With the increasing of invalid evaluation, the time cost of WSMQF keeps a stable value basically, and the time cost of SMF is increased with invalid evaluation increasing. And the time cost of SMQ is the same as 1), and it is in a stable status.

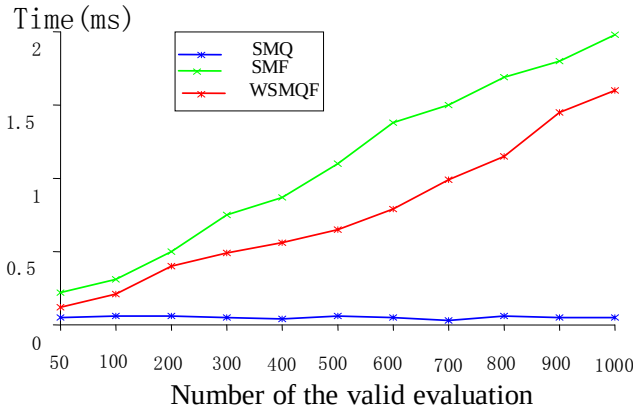


Fig. 11 Time spending under different valid evaluation.

3)Simulation experiment on the precision ratio

The precision ratio shows that after according to the requirement QoS of service consumer selecting the service for consumer, the selected service number of times takes a percentage of ideal service number of times of service request in the service selection system. Aimed at the precision ratio, the statistical result of simulation for proposed algorithm is shown as in Fig.12.

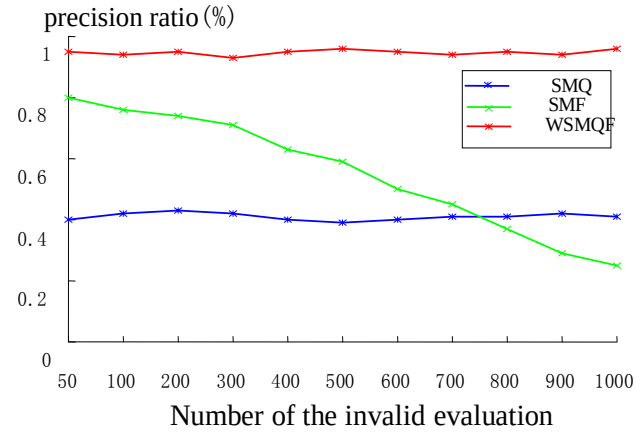


Fig. 12 Selection precision ratio under different invalid evaluation.

Compared with the simulation result it can be seen that the precision ratio of SMQ is lower, and the precision ratio of SMF is reduced with the increasing of invalid evaluation, but the precision ratio of WSMQF is also to keep a higher level, and it can not be reduced with the increasing of invalid evaluation. From the simulation experiment, it can also be seen that the time cost of SMQ is less, but the precision ratio is less. The time cost of SMQ SMF is higher than SMQ does, but related to SMQ it is still high.

Compared with SMF, the time cost of WSMQF is less, but compared with SMQ, the time cost is big a little, and the precision ratio is the highest and the most stable in the three sorts of selection mechanism. From the above mentioned, we can see that under the condition of acceptable time cost, the time cost of WSMQF has obvious improvement in precision ratio.

6. Conclusions

The paper explored in detail the selection strategy of Web service based on the fusion of subjective and objective evaluation for QoS aiming at the puzzle of Web service selection for functional similarity. By means of simulation experiment, the rationality and effectiveness of the service selection strategy had been validated preliminarily. In view of the complexity of service selection mechanism, it is still necessary to make further research to some puzzles, such as how to make reasonable partition for target consumption group according to the actual requirement, how to determine initial confidence value of service consumer so as to restrain the malicious evaluation of service consumer.

Acknowledgments

This work was supported by science & technology project of Chongqing municipal education committee (No. KJ111414) and Chongqing Natural Science Foundation (NO.CSTC, 2010BB2285).

References

[1] W.-B. Zhang, and W.-F. Shi, "Research on Dynamic Web Services Composition Based on BPEL and QoS", Computer Technology and Development, Vol.19, No. 11, 2009, pp.72-75.
[2] Y.-C. Gao, and W.-Q. Zhang, "Web Service Description and Selection Mechanism Based on QoS Ontology", Computer science, Vol.35, No. 12, 2008, pp. 273-276.
[3] L.-J. Chen, and Y. Zhou, "Research on Dynamic Web Service Selection Based Multi-QoS Constraints", Microcomputer Information, Vol.25, No.11-3,2009, pp.209-211.
[4] M. Yang, and L.-N. Wang, "Research of Web Service Reliability Enhancement Method Based on Trust Fault Tolerant", Journal on Communications, Vol.31, No. 9, 2010, pp.131-138.
[5] Y.-D. Liu, and J. Wu, "Research on Web Services Discovery Model Based on Decision-making of the Multiple Attributes of Quality of Service", Aeronautical Computing Technique, Vol. 38, No. 4, 2008, pp.78-83.
[6] K.-F. Liu, H. Wang, and Z.-P. Xu, "A Web Service Selection Mechanism Based on QoS Prediction", Computer Technology and Development, Vol. 17, No. 8, 2007, pp.103-109.
[7] S.-H. Zhao, G.-X. Wu, S.-F. Zhang, Q. Fang, and K. Yu, "Review on SOA of Quality of Service Research", Computer Science, Vol. 36, No. 4, 2009, pp.16-20, 46.
[8] Z.-P. Liu, L. Han, and Z.-T. Liu, "Semantic Web Service Selection Based on Similarity of Qos", Journal of Jiangxi Normal University (Natural Sciences Edition), Vol. 32, No. 2, 2008, pp.189-191,218.
[9] K. Yue, W. Liu, X. Wang, and J. Li, "An Approach for Measuring Quality of Web Services Based on the Superposition of Uncertain Factors", Journal of Computer Research and Development, Vol. 46,No. 5, 2009, pp.841-849.
[10] J.-G. Xu, Y.-L. Luo, and D.-C. Wang, "QoS-aware Web services optimize selection based on reputation model", Journal of Computer Applications, Vol. 28, No. 12, 2008, pp.322-325.

Yucheng Liu achieved the engineering bachelor degree in 1984 and the engineering master degree in 2005. He has been working in Chongqing University of Science and Technology since 1992. He was a member of the program committee of the international conference ICEICE2011, ICEICE2012, ICISE2011 and ICECC2012. He is mainly engaged in the automation professional. He obtained three provincial research projects and published more than thirty papers. His current research interests are computer control technology.

Yubin Liu achieved the science bachelor degree in 1990 and the engineering master degree in 2007. He has been working in Panzhihua Univerisity since 2006. He published more than twenty papers.

User Behavior Prediction based Adaptive Policy Pre-fetching Scheme for Efficient Network Management

Yuanlong Cao[1], Jianfeng Guan[1], Wei Quan[1], Jia Zhao[3], Changqiao Xu[1,2], Hongke Zhang[1,3]

[1] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications Beijing, 100876, China

[2] Institute of Sensing Technology and Business, Beijing University of Posts and Telecommunications Wuxi, Jiangsu 214028, China

[3] National Engineering Laboratory for Next Generation Internet Interconnection Devices, Beijing Jiaotong University Beijing 100044, China

Abstract

In recent years, network management is commonly regarded as an essential and promising function for managing and improving the security of network infrastructures. However, as networks get faster and network centric applications get more complex, there is still significant ongoing work addressing many challenges of the network management. Traditional passive network censoring systems lack of adaptive policy pre-fetching scheme, as a result, preventing malicious behavior (such as hacker, malware etc.) is big challenging. In this paper, we propose a novel user behavior prediction based adaptive policy pre-fetching scheme for efficient network management. A newly Distributed web User Behavior Prediction model (DUBP) is introduced first to cognize and predict user behavior. It extends a distributed DHT network to fix the bottleneck in traditional Client-Server (C/S) architecture occurred by large-scale network service requesting and massive user log analyzing and calculating. Based on user behavior sensing and prediction provided by DUBP, a further Adaptive Policy Pre-fetching and Caching scheme (APPC) is addressed for fine-grained and efficient network management. Our Universal Network (UN) will employ DUBP and APPC scheme to justify its advantages in secure network service.

Keywords: *network management; user behavior analysis, policy pre-fetch; Universal Network*

1. Introduction

Network censoring is getting increasingly important due to the immense growth of Internet users and service providers, such as Internet Content Providers (ICPs), Internet Service Providers (ISPs) and so on. On the one side, data gathered based passive network censoring has been regarded as the common solution for advanced network censoring and security systems that require fine-grained performance measurements, such as Deep Packet Inspection (DPI) [1]. On the other side, adaptive policy pre-fetching scheme is the key feature for fine-grained and efficient network management.

As networks get faster and network centric applications get more complex, sensing malicious behavior then pre-fetching policy gets more difficult. To solve this problem, log records have been proven effective in detecting and combating these harmful behaviors [2]. As mentioned in a report on data breaches investigated by the Verizon Corp Business Risk team reported in 2008 that 66% of organizations investigated had "sufficient evidence available within their logs to discover the breach had they been more diligent in analyzing such resources" [3]. Actually, there are more and more researches focus on network log [4-5]. As addressed above, it is very clear that logs can play a more important role for security event detection, mitigation and prediction.

However, log server employed in current user behavior analysis and network censoring systems usually base on the traditional C/S architecture. For example, work [6] proposed a central log tracker to collect and analyze large scale of users' viewing behavior on Video-on-Demand (VoD) streaming. But with more and more user behavior records arising, the traditional C/S tracker server will inevitably become a bottleneck during communicating, caching and analyzing required to process large-scale network service request.

Since balancing the load of file storage and transfer with fully distributed design, Peer-to-Peer (P2P) networks has been widely applied in distributed applications over internet in recent years, such as P2P file sharing [7], P2P Grid computing [8], P2P SIP transfer [9] and multimedia content delivery [10] As the typical structured P2P networks, Distributed Hash Table (DHT) [11] i.e., Chord [12] becomes a promising solution to avoid the flooding search by tightly coupling data or indices of data hereby mechanism that each node has an M-bit identifier by

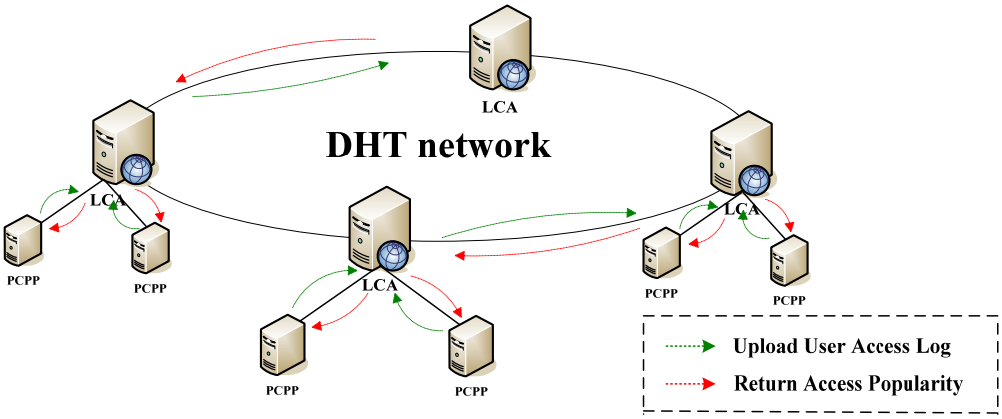


Figure 1. Architecture of the DUBP model

Table I. The description of DUBP model

Component	Description
DHT network	The DHT overlay network is constructed by Chord topology. It consists of LCA. Functions of DHT network include: 1) save user access service logs; 2) compute user behavior popularity; 3) disseminate $\Re$ to PCPP; and 4) self-update once LCA join or leave.
LCA	The LCA's functions involve: 1) receive user access service logs from the PCN; 2) analyze user access service logs and evaluate $\Re$ periodically; 3) share $\Re$ with other LCA in the DHT network; and 4) return $\Re$ to requested PCPP.
PCPP	The PCPP's functions aim to: 1) caputring packet from router or switch etc.; 2) extracting desired log information accordance with specified criteria; 3) uploading log information to its connected LCA. And 4) Request $\Re$ from it's connected LCA then prefetch correspoinding policy for p_i with higher $\Re$ from the PD.

hashing the IP address and other information using a base hash function such as SHA-1 [13].

In this paper, we propose a novel user behavior prediction based adaptive policy pre-fetching scheme for efficient network management. A newly Distributed web User Behavior Prediction model (DUBP) is introduced first to cognize and predict user behavior. It extends a distributed DHT network to fix the bottleneck in traditional Client-Server (C/S) architecture occurred by large-scale network service requesting and massive user log analyzing and calculating. Based on user behavior sensing and prediction provided by DUBP, a further Adaptive Policy Pre-fetching and Caching scheme (APPC) is addressed for fine-grained and efficient network management.

The organization of this paper as follows, Section 2 introduces the overview of DUBP model, as well as its functions design. Section 3 details the proposed adaptive policy pre-fetching and caching scheme for fine-grained and efficient network management. Section 4 introduces the design of test bed, which is based on our universal network architecture. A necessary conclusion and future work will be shown in Section 5.

2. DUBP Model Description

Fig. 1 shows the architecture of the proposed DUBP model. The goal of DUBP is to analyze user access network services log and estimate user behavior popularity by means of distributed manner, per predicting user behavior to improve the speed of censoring policy responding and the performance of overall network censoring systems. In the DUBP model, all Log Collection and Analysis node (LCA) are chosen to construct the Chord to storage, disseminate and analyze users' access network service log. Packet Capturing and Policy Pre-fetch node (PCPP) is in charge of capturing packet from network equipments (such as router, switch etc.), extract desired log information accordance with specified criteria, then upload those log information to its connected LCA, and pre-fetch policy accordance with accessed page popularity provided by the LCA. Comparing to the traditional C/S-based log tracker and analysis model, the proposed DUBP model can support high scalability for large number of users with efficiently and robustly.

The DUBP model consists of two structures. The upper layer is a DHT network which is consisted by LCA. The lower layer is a *C/S* structure that connects LCA with PCPP. Detailed functions of the DUBP's components are described in Table 1.

Assume 1: Each web user had registered his/her basic information (such as age), and got his/her own access identifier (such as IP address, user ID etc.) in *User Information Database* (UID). Besides, current researches omit that people in different classification (such as age, the used core network etc.) have different interesting, so we play attention to which core network user used as well to implement a more fine-grained censoring.

Definition 1: A website S_i consists of a set of webpage which can be represented by (p_1, p_2, … , p_n). And $p_i \in (p_1, p_2, \ldots, p_n)$ has own identification (URL) and popularity $\Re$.

The overall workflow of the DUBP model can be briefed as follows: The PCPP capturing HTTP-based packet from router or switch etc., extracts and use *Source IP* to get ***C**ore **N**etwork ID* (CNID) from its ***IP R**egular **D**atabase* (IPRD), then *PCPP* upload the record consisted of four tuples <CNID, DIP, DPort, URL> (DIP denotes Destination IP, DPort is on behalf of Destination Port, URL is the webpage path user requested) to its connect LCA; Once a record arriving at the LCA, the LCA share the record and inquire desired count URLs of the accessed website which has higher $\Re$ over DHT network, then a *replying* consisted of desired count URLs and source IP will be returned to the PCPP, the PCPP pre-fetches related policies from ***P**olicy **D**atabase* (PD) to support real-time and adaptive user behavior censoring. The LCA also stores and analyzes the $\Re$ periodically.

This section details how DUBP performs the functions of web user log analysis and behavior prediction for network censoring systems. For convenience, we define some useful annotations as described in Table II.

Table II. Annotation description used in DUBP model

Annotation	Description
LCA(*i*)	The *Log Collection and Analysis* node *i*
PCPP(*i*)	The *Packet Capturing and Policy Pre-fetch* node *i*
LCAID(*i*)	The ID of the LCA *i*
SIP(*i*)	The Source IP *i*
DIP(*i*)	The accessing Destination IP *i*, namely the web user *i*
DPort(*i*)	The Port of Destination IP *i*
Hash()	The hash function of DHT network. For example, *Hash*(*CNID*, *DIP*(*i*), *DPort*(*i*)) = LCA(*i*) means that map tuple < *CNID*, *DIP*(*i*), *DPort*(*i*)> to the LCA whose ID is *LCAID*(*i*)

Below subsections address the major stages of DUBP model.

2.1 Web Access Records Uploading

As mentioned in Section II, when a web user initiates web access request, the PCPP will capture access logs then upload logs to its connected LCA, as well as receives $\Re$ information from the LCA. That information will be exchanged between LCAs over DHT network. How the designed process of uploading access logs work is illustrated as follows:

Step 1: Assume that a user *SIP*(*i*) request a website S_i, and PCPP(*p*) captured the *Get* packet from router/switcher, then the PCPP(*p*) extract *Source IP* and use it to get CNID from IPRD (CNID is a integration type such as 1,2,…, n which stands for the core network the user used such as the telecom network, the education network and so on respectively defined in the IPRD). Fig. 2 shows the example how to get CNID.

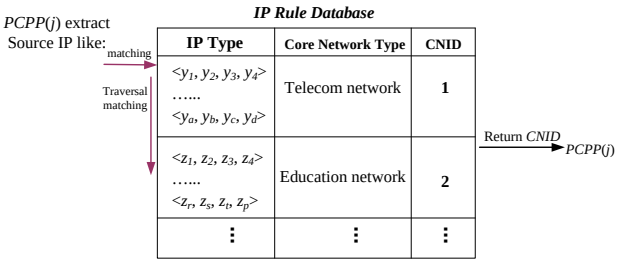


Figure 2. Diagram of an example for getting CNID

Step 2: The PCPP(*p*) use DPI [1] to extract requested page URL, then transmits those access record formed by <*CNID*, *DIP*, *DPort*, *URL*> to its connected LCA periodically. Fig. 3 shows how the PCCP uploads access records to its connected LCA.

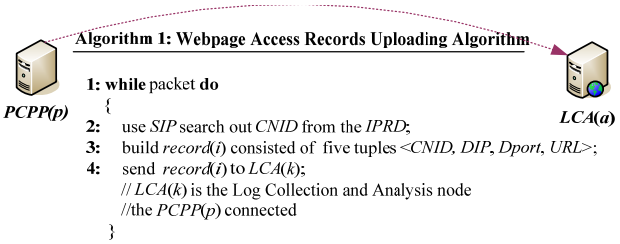


Figure 3. web user access records uploading algorithm

Step 3: When the *LCA* receives access records, it will run *hash* function as Eq.(1) to disseminate the record to others *LCA* over DHT network whose ID equals *Successor*(*key*). *Successor*(*key*) denotes the successor node of key value which has been detailed in [14].

$$key = hash(CNID + DIP + DPort) \quad (1)$$

Fig. 4 illustrates the algorithm that how *LCA*(*a*) disseminates access record uploaded from *PCPP*.

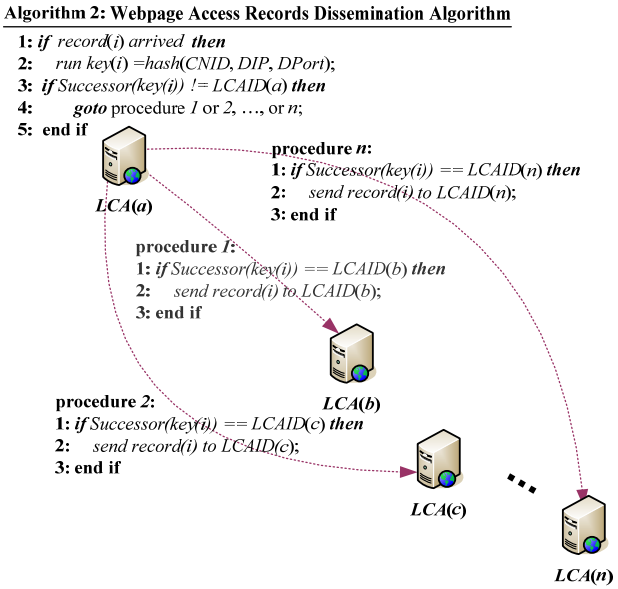


Figure 4. web user access records dissemination algorithm

Step 4: After *record*(*i*) arrived, the LCA(*k*) will cache a new record to its *Webpage Access Records Caching* Table shown in Table III.

Table III. Webpage Access Records Caching Table

CNID	DIP	DPort	URL

After the *PCPP*(*p*) uploading records to its connected *LCA*, then it will send *request*(*SIP*) to the *UID*; then the *UID* returns user information to the *PCPP*(*p*).

2.2 Webpage Access Popularity Analyzing

To reduce LCAs' stress, a period τ is set for webpage access popularity analysis. Webpage access popularity will be stored according with Table IV.

Table IV. Webpage Access Popularity Table

CNID	DIP	DPort	URL	$\Re$	*LastAccessTime*

When τ is coming, the LCAs will analyze webpage popularity as below algorithm detailed as Algorithm 3.

Algorithm 3: Webpage Access Popularity Analysis Algorithm

Definition:

R_i^3 the i_{th} record in Table 3, it consists of $< CNID_i^3 , DIP_i^3 , DPort_i^3 , URL_i^3 >$

$CNID_i^3$ the *CoreNetwork ID* value in R_i^3

DIP_i^3 the *Destination IP* value in R_i^3

$DPort_i^3$ the *Destination Port* value in R_i^3

URL_i^3 the viewed URL record in R_i^3

R_j^4 the j_{th} record in Table 4, it consists of $< CNID_j^4 , DIP_j^4 , DPort_j^4 , URL_j^4 , LAT_j^4 >$

$CNID_j^4$ the *CoreNetwork ID* value in R_j^4

DIP_j^4 the *Destination IP* value in R_j^4

$DPort_j^4$ the *Destination Port* value in R_j^4

URL_j^4 the *URL* record in R_j^4

LAT_j^4 the *Last Access Time* in R_j^4

1: **for** i = 1 to m //m is the total records count in Table 3
2: **for** j = 1 to n //n is the total records count in Table 4
3: **if** ($CNID_i^3 == CNID_j^4$) && ($DIP_i^3 == DIP_j^4$) && ($DPort_i^3 == DPort_j^4$) && (URL_i^3 equal URL_j^4) **then**
4: $\Re$ ++; //URL popularity $\Re$ increases 1
5: set LAT_j^4 as *LCA*(*k*)'s current system time ;
6: delete R_i^3 from Table 3; //cleanup cached record
7: **else**
8: create new record in Table 4;
9: set $\Re$ = 1; //set default $\Re$ for new viewed URL
10: set *LastAccessTime* as LCA(*k*)'s current system time;
11: delete R_i^3 from Table 3; //cleanup cached record
12: **end if**
13: **end for**
14: **end for**

To reduce the load of *LCA*s and improve prediction accuracy, the redundant URLs which had never been viewed for a long time should be removed periodically. Thus, We set a time threshold (denoted as *thresh*) to remove related redundant records once *timerange* is larger than *thresh* . Where *timerange* is defined by

$$timerange = CurrentTime - LastAccessTime \quad (2)$$

When a LCA leaves the DHT network, it should transfer its *Webpage Access Popularity* Table to one of its neighbor peers along the DHT network. When a LCA joins the DHT network, it should get some webpage access popularity information from its neighbor peers as original information. We notice that the LCA may leave the DHT network unexpectedly during popularity analyzing. If so, once the LCA leaves, a task flag will be set to false to ensure the LCA finish its analysis once it comes back to DHT network again.

3. Adaptive Policy Pre-fetching and Caching

Based on user behavior sensing and prediction provided by DUBP, a further Adaptive Policy Pre-fetching and Caching scheme (APPC) is addressed in this section for fine-grained and efficient network management. The major stages of APPC scheme are addressed as below.

Once a record (i.e. *record*(*i*)) arriving at LCA(*k*), it will trigger the LCA(*k*) executing some steps as follow immediately:

1) uses <CNID, DIP, DPort> to obtain the total matched URL counts (denoted as $Count_{Existing}$) from Table 4.
2) inquires the specified URL counts (denoted as $Count_{Specified}$) which is set by administrator in web console.
3) gets κ value via E.q (3). Then return a *replying* consisted of URLs with $\Re$ in top κ and source IP to the PCPP(p) where the record(i) comes from.
4) if there are more than κ URLs has same popularity $\Re$, the URL with less *timerange* will be selected to reach a more fine-grained and accurate predict.

$$\kappa = \begin{cases} Count_{Existing}; & Count_{Existing} < Count_{Specified} \\ Count_{Specified}; & Count_{Existing} \geq Count_{Specified} \end{cases} \quad (3)$$

After *replying* returned, the PCPP(*p*) prefetches corresponding policies with user information and the κ URLs from the PD.

To achieve policy pre-fetching more efficiently, we also employ reinforcement learning to construct a prediction model to predict web user navigating behavior.

a) ***Reinforcement Learning*** (RL). RL is learning what to do-how to map situations to actions, so as to maximize a numerical reward signal. As in Fig. 5, A RL Agent learns knowledge via interacting with the Environment. That is, once the Agent changes its state from one to another, a reward will be returned the environment. From the rewards, the Agent will learn a policy how to gain a more benefit rewards. Previous work [15] detailed the RL problem using Markov decision process (MDPs).

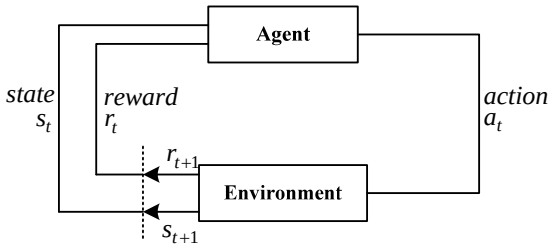


Figure 5. The reinforcement learning illustration

b) *Q*-learning is an off-policy temporal difference control algorithm [15] to learn *state-action* values. E.q (4) shows the on-step Q-learning:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \times [r_{t+1} + \gamma \times \max_{a'} Q(s_{t+1}, a) - Q(s_t, a_t)] \quad (4)$$

Where s_t denotes current state of the *Agent*, a_t is the action adopted by the *Agent*. $Q(s_t, a_t)$ represents how good the a_t in the s_t. α denotes the learning rate. r_{t+1} is the reward returned from the *Environment*. γ is a discount parameter. As mentioned above, an optimal policy can be generated via the $Q(s, \alpha)$ function.

An improved and efficient *Q*-learning algorithm mentioned in [6] is employed in our DUBP model to decide which URLs should be pre-fetched and corresponding policies should be cached.

4. Universal Network based Testbed Design

We have already implemented the basic functions of BPPP mentioned above and confirmed their operations in our *Universal Network* (UN) [16-17]. Our UN is a novel next generation-oriented network architecture which is based on the well-known identifier/locator separation protocol [18]. To help the reader in understand the idea of Universal Network based testbed, we first introduce the context of identifier/locator separation.

4.1 Identifier/locator Separation

The networks with identifier/locator separation commonly consists of two parts, *transit core* and *edge networks*. Fig. 6 shows a basic wireless network topology with identifier/locator separation. We briefly detail how to forward packets in such network topology with assumptions that 1) the two terminal users are in different edge networks; and 2) routing in edge networks is separated from routing in the transit core.

Assuming that the User B with ID_B wants to open a connection to the User A with ID_A, following steps will be complied [19-20].

i. The User A first issue a data packet to its *Ingress Tunnel Router* (ITR). In this case, ID_A and ID_B act

as the source and the destination of the packet, respectively.

ii. Every ITR maintains a table to cache some recently used identifier-to-locator (ID2LT) mappings. When the ITR receives the packet from the User A, it looks up a locator for ID_B in its local identifier-to-locator (ID2LT) mapping table which is used to cache some recently used ID2LT mappings, as shown by ② in Fig. 6. If the cache hits, go to step iv; otherwise, go to step iii.

iii. The ITR resolves the locator(s) for ID_B by querying a mapping server (shown by ③ in Fig. 6). When the ITR receives the resolved locators for ID_B, it caches them into its local ID2LT mapping table.

iv. Denote the locator of ITR and the resolved locator for ID_B by $Locator_1$ and $Locator_2$, respectively. The ITR encapsulates the received packet with an outer header whose destination and source are $Locator_1$ and $Locator_2$, respectively. Then the ITR then sends the encapsulated packet out with destination of *Egress Tunnel Router* (ETR), as shown by ④ in Fig. 6.

v. When the ETR receives the encapsulated packet, it 1) strips the outer header of the encapsulated packet; 2) stores the mapping from ID_A onto $Locator_1$ into its local ID2LT mapping table for possible future usage; and 3) sends the decapsulated packet to corresponding destination, namely User B in the Fig. 6.

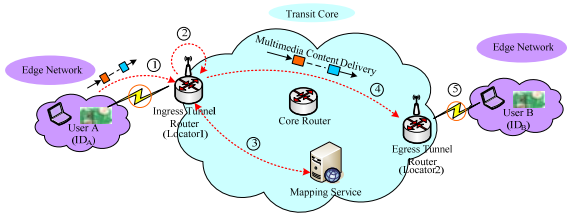


Figure 6. A basic wireless network with identifier/locator separation

4.2 The Framework of UN-based Testbed

The framework of testbed is shown as Fig.7. It consists of two layers, *management layer* and *switch routing layer*, which are described in details next.

a) ***Management Layer***. In order to meet the information security requirements for multimedia communication, computation and service in the identifier/locator separation context, it is necessary to consider some useful management components that refer to multimedia security. For this purpose, the framework of testbed provides management layer to launch security management. The management layer includes *Identifier Mapping Server* (IDMS), *Access-control Policy Database* (ACPD), *User Registrant/Authentication Server* (URAS), *Service Registrant/Authentication Server* (SRAS) and DUBP. Table V describes the components in Management Layer.

b) ***Switch Routing Layer***. As shown in figure 3, Switch Routing Layer consists of Ingress Tunnel Router (ITR), Egress Tunnel Router (ETR), and Core Router (CR) and so on. Next the major functions of these components are detailed. ITR in the testbed still keeps capabilities same as that in original identifier/locator separation context, such as 1) acts as an access point for terminals in Edge Network access to Transit Core; 2) caches and provides ID2LT mappings; and 3) forwards packet. Moreover, for sake of security management, it 4) enables APPC to pre-fetch and cache the policy instances sent from ACPD; 5) requests UTag and STag from URAS and SRAS, respectively; 6) decides whether to grant access in conjunction with the policy instances, UTag and STag. ETR in the testbed still keeps capabilities same as that in original identifier/locator separation context. Previous work [19-20] detailed the functions of ETR. The CR still keeps capabilities same as that in original identifier/locator separation context. That is, it just routes and forwards packet in *transit core*.

5. Conclusions and Future Work

Users' behavior log attracts more and more attentions in current researches. However, Log server employed in current user behavior analysis, network censoring systems usually base on the traditional Client-Server (C/S) architecture. With more and more user behavior records arising, the single C/S tracker server will inevitably become a bottleneck during communicating, caching and analyzing required to process large-scale network service request. In this paper, a novel Distributed web User Behavior Prediction model (DUBP) for network censoring systems is proposed, which extends a distributed DHT network structure. The DUBP makes all nodes' available resources and predicts user behaviors fine-grained by means of users' core network type and historical access records, it can support high scalability for large number of users with efficiently and robustly. Based on user behavior sensing and prediction provided by DUBP, a further Adaptive Policy Pre-fetching and Caching scheme (APPC) is addressed for fine-grained and efficient network management.

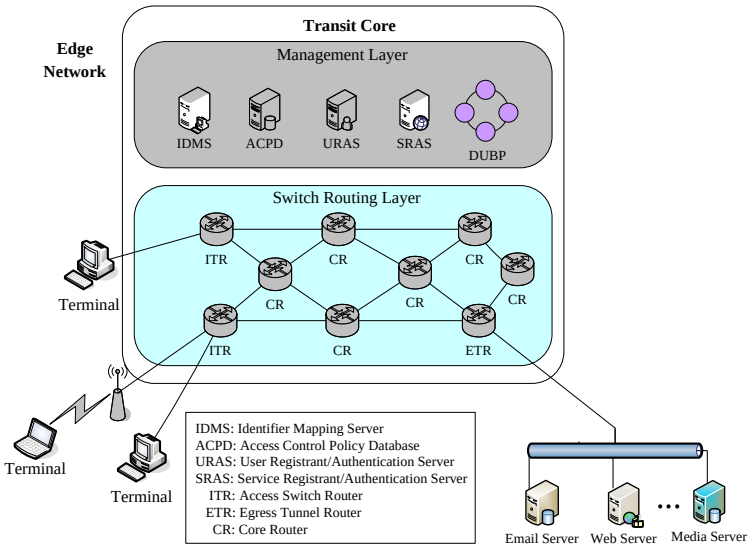


Figure 7. The framework of Universal Network based testbed

Table V. Components in Management Layer

Components	Descriptions
IDMS	IDMS stores identifier-to-locator (ID2LT) mappings for each terminal and subnet which are under its control.
ACPD	ACPD stores the policy instances pre-fetched by APPC and sends them to ITR and ETR periodically. Moreover, a new or updated policy instance will be sent to ITR and ETR instantly by ACPD.
URAS	URAS includes three function modules, which User Registration Module (URM) provides a user interface (i.e. web) for user registration, while User Tag Generator (UTG) creates a User Tag (UTag) for each registered user accordance with 1) user basic information (i.e. age, interest, education and so on); and 2) dynamic information (i.e. malicious behavior) provided by DUBP. And User Authentication Module (UAM) creates a User ID (UID) for each registered user to access Internet. A user without UID will be failed to access Internet by UAM.
SRAS	like URAS, SRAS also includes three function modules, which Service Registration Module (SRM) provides a registration interface (i.e. web) for Internet Service Provider (ISP) or Internet Content Provider (ICP) to register services, while Service Tag Generator (STG) creates a static Service Tag (STag) for each registered service accordance with 1) service basic information (i.e. fee, language, constraint-level and so on) provide by ISP/ICP; and 2) dynamic information (i.e. constraint content). And Service Authentication Module (SAM) creates a Service ID (SID) for each registered media. A service without SID cannot be deployed into Internet.
DUBP	DUBP aims to store and analyze user access behavior (i.e. malicious behavior), and send out analysis results to URAS.

For the experimental validation, we are now extending the UN for optimizing the communication capability with our previous work [21-27]. We will then evaluate and compare the performance of UN with or without the proposed scheme mentioned in this paper during it implements network management.

Acknowledgments

This work was partially supported by the National High-Tech Research and Development Program of China (863) under Grant No. 2011AA010701, in part by the National Basic Research Program of China (973 Program) under Grant 2013CB329102, in part by the National Natural Science Foundation of China (NSFC) under Grant No. 61001122, 61003283, 61232017, in part by the Fundamental Research Funds for the Central Universities under Grant No. 2012RC0603, and in part by the Jiangsu Natural Science Foundation of China under Grant No. BK2011171.

References

[1] M. Grossglauser and J. Rexford, "Passive traffic measurement for IP Operations," in The Internet as a Large-Scale Complex System, 2005, pp. 91–120.

[2] Shenk, Jerry, "SANS Annual 2009 Log Management Survey," Technical Report, SANS, 2009.

[3] Baker, Wade, A. Hutton, C. David Hylender, C. Novak, C. Porter, B. Sartin, P. Tippett, and J. Andrew Valentine, "Data Breach Investigations Report," Technical Report, Verizon Business RISK Team, 2009.

[4] J. Myers, M. Grimaila, R. Mills, "Log-Based Distributed Security Event Detection Using Simple Event Correlator," In Proceedings of the 44th Hawaii International Conference on System Science, Jan. 2011.

[5] E. Hilgenstieler, E. Duarte, G. Mansfield-Keeni, N. Shiratori, "Improving the Precision and Efficiency of Log-based IP Packet Traceback," In Proceedings of IEEE Global Telecommunications Conference (GLOBECOM'07), Nov. 2007.

[6] T. Xu, W. Wang, B. Ye, W. Li, S. Lu, Y. Gao, "Prediction-based prefetching to support VCR-like operations in gossip-based P2P VoD systems," In Proceedings of the 15th International Conference on Parallel and Distributed Systems (ICPADS), Shenzhen, China, Dec. 2009.

[7] J. Pouwelse, P. Garbacki, D. Epema, and H. Sips, "The BitTorrent P2P File-Sharing System: Measurements and Analysis," in Proceedings of Fourth Int'l Workshop Peer-to-Peer Systems (IPTPS), 2005.

[8] I. Foster and A. Iamnichi, "On Death, Taxes, and Convergence of P2P and Grid Computing," In Proceedings of the Second Int'1 Workshop Peer-to-Peer Systems (IPTP3'03), Feb. 2003.

[9] I. Kelenyi, J.K. Nurminen, M. Matuszewski, "DHT Performance for Peer-to-Peer SIP-A Mobile Phone Perspective," In Proceedings of 7th IEEE Consumer Communications and Networking Conference (CCNC'10), 2010.

[10] C. Xu, E. Fallon, Q. Yuansong, Z. Lujie and M. Gabriel-Miro, "Performance Evaluation of Multimedia Content Distribution Over Multi-Homed Wireless Networks," IEEE Transactions on Broadcasting, vol. 57, no. 2, June 2011.

[11] G. Urdaneta, G. Pierre, M. Steen, "A Survey of DHT Security Techniques," ACM Computing Surveys, 2009.

[12] R. Zhou and K. Hwang, "GossipTrust for Fast Reputation Aggregation in Peer-to-Peer Networks," IEEE Transactions on Knowledge and Data Engineering, vol.20, no. 9, pp. 1282-1295, Sept. 2008.

[13] Y. Liu, W. Xue, K. Li, *et al*. "DHTrust: A Robust and Distributed Reputation System for Trusted Peer-to-Peer Networks," In Proceedings of 2010 IEEE Global Telecommunications Conference (GLOBECOM'10), Dec. 2010.

[14] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, "Chord: A scalable peer-to-peer lookup service for Internet applications," In Proceedings of ACM SIGCOMM'01, Aug. 2001.

[15] R. S. sutton and A. G. Barto, "Reinforcement Learning: An Introduction," MIT Press, Cambridge, MA, USA, 1998.

[16] Hongke Zhang, "An Architecture of Universal Network Services," Patent Application, no. 200510134579.1, 2005.

[17] Hongke Zhang, "A method of implementing pervasive service in Universal Network," Patent Application, no.200610169727.8, 2006.

[18] D. Farinacci, V. Fuller, D. Meyer, D. Lewis, "Locator/ID Separation Protocol (LISP)," IETF Internet Draft, draft-ietf-lisp-23.txt (work in progress), May 2012.

[19] H. Luo, H. Zhang, and C. Qiao, "Efficient Mobility Support by Indirect Mapping in Networks with Locator/Identifier Separation," IEEE Transactions on Vehicular Technology, vol.60, no.5, pp.2265-2279, June 2011.

[20] H. Luo, H. Zhang, and M. Zukerman, "Decoupling the design of identifier-to-locator mapping services from identifiers," *Computer Networks*, vol. 55, no. 4, pp. 959–974, March 2011.

[21] C. Xu, T. Liu, J. Guan and H. Zhang, G.-M. Muntean, "CMT-QA: Quality-aware Adaptive Concurrent Multipath Data Transfer in Heterogeneous Wireless Networks," IEEE Transactions on Mobile Computing, vol.PP, no.99, Aug. 2012.

[22] Y. Cao, C. Xu, J. Guan, F. Song, H. Zhang, "Environment-aware CMT for Efficient Video Delivery in Wireless Multimedia Sensor Networks," International Journal of Distributed Sensor Networks, vol.2012, Article ID 381726, 12 pages, 2012.

[23] Y. Cao, C. Xu, J. Guan, H. Zhang, "Background Traffic-based Retransmission Algorithm for Multimedia Streaming Transfer over Concurrent Multipaths," International Journal of Digital Multimedia Broadcasting, vol.2012, Article ID 789579, 10 pages, 2012.

[24] Y. Cao, C. Xu, J. Guan, J. Zhao, H. Zhang, "Cross-layer Cognitive CMT for Efficient Multimedia Distribution over Multi-homed Wireless Networks," In Proceedings of IEEE Wireless Communications and Networking Conference (WCNC'13), accepted.

[25] Y. Cao, C. Xu, J. Guan, H. Zhang, "Cross-layer Retransmission Approach for Efficient VoD Transfer over Multi-homed Wireless Networks," International Journal of Digital Content Technology and its Applications, vol.6, no.23, pp.98-109, Dec. 2012.

[26] Y. Cao, C. Xu, J. Guan, et al., "Relational Analysis Based Concurrent Multipath Transfer Over Heterogeneous Vehicular Networks," International Journal of Computer Science Issues, vol.9, issue 5, no.2, pp.1-10, Sep. 2012.

[27] Y. Cao, C. Xu, J. Guan, "A record-based retransmission policy on SCTP's Concurrent Multipath Transfer," In Proceedings of 2011 International Conference on Advanced Intelligence and Awareness Internet, pp.67-71, Oct. 2011.

Yuanlong Cao received his B.S. degree from Nanchang University of China in 2006, received his M.S degree from Beijing University of Posts and Telecommunications (BUPT) in 2008. During 2007-2009, he worked as an intern in BEA China Telecommunications Technology Center (BEA TTC) and IBM China Development Lab (IBM CDL). During 2009-2010, he worked as a software engineer in DT Research (Beijing). He is currently working toward the Ph.D. degree in the Institute of Network Technology, BUPT. He is broadly interested in computer networks, multimedia communications, wireless networking, network security, and next generation Internet technology.

Jianfeng Guan received his B.S. degree from Northeastern University of China in July 2004, and received the Ph.D. degrees in communications and information system from the Beijing Jiaotong University, Beijing, China, in Jan. 2010. He is a Lecturer in the Institute of Network Technology at Beijing University of Posts and Telecommunications (BUPT), Beijing, China. His main research interests focus around mobile IP, mobile multicast and next generation Internet.

Wei Quan received his B.S. degree in information and computer science from China University of Petroleum (Beijing) in 2009. He is currently working toward the Ph.D. degree in the Institute of Network Technology, Beijing University of Posts and Telecommunications (BUPT). He is broadly interested in computer network technology. In particular, his research interests include wireless sensor network, cognitive wireless network, mobile IP, and next generation Internet technology.

Jia Zhao received the M.S. degree in electrical engineering from Beijing Jiaotong University, China, in 2011. He is pursuing the Ph.D. degree at national engineering laboratory for next generation Internet interconnection devices, Beijing Jiaotong University. His research interests include traffic engineering, overlay routing, game theory, social mobility and mobile ad hoc networks.

Changqiao Xu is an Associate Professor in the Institute of Network Technology and Associate Director of the Next Generation Internet Technology Research Center at Beijing University of Posts and Telecommunications (BUPT), China. He received his PhD degree in Computer Applied Technology from Institute of Software, Chinese Academy of Sciences (ISCAS) in Jan 2009. He was an Assistant Research Fellow in ISCAS from 2002 to 2007, where he held role as a project manager in the research & development area of communication networks. During 2007-2009, he worked as a researcher in Software Research Institute at Athlone Institute of Technology, Ireland. He joined BUPT in Dec 2009 and was a Lecturer from 2009 to 2011. His research interests include computer networks, multimedia communications, wireless networking, network security, and next generation Internet technology.

Hongke Zhang received his M.S. and Ph.D. degrees in Electrical and Communication Systems from the University of Electronic Science and Technology of China in 1988 and 1992, respectively. From Sep. 1992 to June 1994, he was a post-doc research associate at Beijing Jiaotong University. In July 1994, he jointed Beijing Jiaotong University, where he is a professor. He has published more than 100 research papers in the areas of communications, computer networks and information theory. He is the director of the National Engineering Laboratory for Next Generation Internet Interconnection Devices.

The Research on Improving the Order Picking Efficiency in Medical Logistics Area of CPL Based on Serial Partition Relay Picking Model

Xu Wei[1,2], ChongyangShi[1,*], Hantao Song[1]

[1] School of Computer Science, Beijing Institute of Technology
Beijing, 100081, China

[2] China Postal Express & Logistics Co., Ltd
Beijing, 100031, China

Abstract

The medical business of China Post Logistics is developing day by day. China Post Logistics have been using the Parallel partition picking mode for a long time, but too many shortages have appearing in these two years, especially in the order picking cost and efficiency. This paper analyzes the current order picking mode of China Post Logistics and compare with the Parallel partition picking mode. By analysis the advantages and disadvantages of these two modes and combine with the actual situation, we choose the serial partition relay picking model as the picking mode of CPL in medical logistics area. And then it optimizes the order picking route in view of the current deficiency combined with the use of Ant Colony Optimization (ACO). The example simulation result shows that this optimizing is effective and the order picking cost decrease 17.36% and the route decrease 9.80% than that as before. This research not only to Chain Post Logistics but also to other logistics company which runs the medical business has certain reference.

Keywords: *China Post Logistics; Medical Logistics; Serial Partition Relay Picking; Ant Colony Optimization*

1. Introduction

Medical logistics has a big difference with other industry logistics, and it has complex classifications, too much emergency distribution request and so on. It is the second difficulty industry only to the automotive industry logistics [1]. In the warehousing field, it is very complicated due to the multi-vendor, multi-volume, small batches of mixing operations and the special characters of the drugs. In general, a large pharmaceutical warehouse usually divided into several libraries, such as fragmented library, acceptance library, and to be transported library in the base of the business standard. And at the same time, it will be dived into normal temperature, cool library, and easy odor libraries in the base of the GSP standard.

The INFOR system has been the Supply Chain Execution system for China Post Logistics by the tender selections since 2010. The implementation of this system has dealt with most of the problem that China Post Logistics met in its warehousing part of medical business area. Even thought, there are still many disadvantages, such as the high cost of its order picking and lower efficiency of its order picking. These disadvantages reduce the efficiency of China Post Logistics' operations, and it hindering the development of this field in China Post Logistics. It is very urgent for China Post Logistics to optimize the warehousing efficiency. In the base of these, this paper put up relevant remedies (replace the parallel partitions Relay picking mode with serial partition Relay picking mode) after analyzing the current order picking situation and comparing the two modes. At the end of this paper, it planned the order picking route under the serial partition Relay picking mode by the means of Ant Colony System, and draw some relevant conclusions.

2. The outline of parallel partition picking mode

Order sorting is a process which do as quick as possible and as accurately as possible to pick out the goods from the shore, classification in a certain way and waiting for compatibility as the customers' and the distribution's order. Order picking routing is to determine the sorting list of goods from picking orders through the heuristic optimization to reduce the order pickers' walking distance. There are lots of studies show that the order pickers' walking time account for about 50% of the total order time, so it is very important to reduce the order pickers' walking time.

Currently, order picking mode in the field of Medicine of China Post Logistics is parallel partition relay picking mode. Parallel partition relay picking refers to the process in which all the goods were chosen gradually. After finished picking goods from all regions, the pickers

transferred containers (boxes or pallets) that were full of all kinds of items to the nest partition or handed them over to the next picker who was responsible for his partition. This picker was responsible for sorting all the goods in this partition. The whole order sorting job wasn't completed until the last picker finished his job. Just like the assembly line in manufacturing factory, the following process can start only if the previous process is finished. The product was completed when the last process was finished [2, 3]. The specific operational processes of the picking mode is shown in Fig 1.

In Fig 1 the solid box represents every storage area in the warehouse, usually storage area is closed. Bar graph in solid box indicates storage cargo space of drugs In this mode, a picker is not responsible for the completion of picking a full order, and all items in one order are selected from different regions. Therefore, these items are selected by different pickers, and then were sent to the final tally by belt [2].

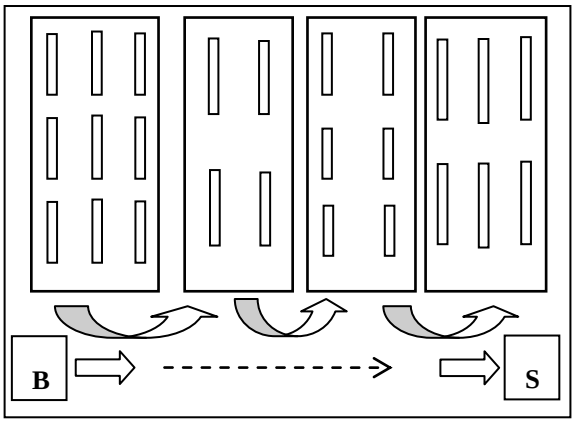


Fig.1 Schematic diagram of parallel partition relay picking mode

In this mode, each picker was assigned to the named shop or zone, and they will be very familiar with the items and its locations in their sorting regions, which can improve the efficiency and accuracy of sorting jobs. At the same time, there is no more than one picker in one corridor, which can avoid the congestion to the maximum extent. In this mode, Pickers have little or no chance to chat with other staff because of the regional segmentation which is helpful for warehouse managers to control and supervise the staff.

However, in the chosen mode, the previous partition pickers did not complete the order picking when the following partition picker was back to the junction points. This picker cannot help the previous pickier to sort the items by crossing the junction point(will not take the initiative to help the former pickers) At this point, the picker will choose to wait which will result in a waste of time .This order picking efficiency cannot be achieved the highest [3-4]. In this mode, the system needs to analysis the orders or arranges the staff, which in some extent increases the warehousing cost of China Post logistics. And, when the order volume is not up to a certain quantity, each partition requires at least one picker, which tremendous waste warehouse labor costs particularly in high labor costs society.

3. Serial partition relay picking mode

3.1 Overview of serial partition relay picking mode

Serial partition relay picking model is an improved model of parallel partition relay picking model. The specific operation of this mode is shown in Fig 2.

In this mode, the order is not spitted, so the pickers sort the items from one picking area to one picking area. After finishing the sorting job in this picking area, the pickers transferred the containers (boxes or pallets) to the next picking area to continue sorting. The whole order sorting job wasn't completed until the last item in this order was sorted. In this mode, because each picker is responsible for one or several orders and orders are not divided. This picking model can greatly improve the response speed of the order, and also can eliminate cost to classify the orders [5], which is particularly important for responding emergency orders in medical logistics field in the China Post Logistics

Compared to with parallel partition relay picking mode, serial partition relay picking mode has the following four advantages:

1) Fast reaction speed of orders. This is the biggest advantage of serial partition relay picking mode, especially for frequent emergency orders ordered by customers in pharmaceutical warehousing field in China Post Logistics. In the serial partition relay picking mode, each picker is responsible for a complete order, pickers can quickly deal with the orders. It is easier to get overstock in parallel partition relay picking mode, which reduces the speed of response to the orders. In China Post medicine field, we often receive a large number of emergency orders, thus the reaction speed to order of is particularly important.

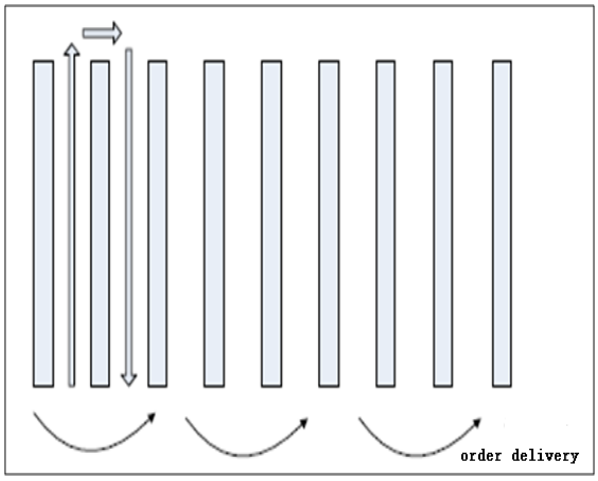


Fig.2 Schematic diagram of Serial partition relay picking mode

2) Reduced cost of labor. In the serial partition relay picking mode each storage area doesn't require one pickers. One picker sorts all the items in one order. It is goods for pharmaceutical warehouse of China Post who has a large number of different sorting areas especially in the case that the human cost is very expensive.
3) Reduced cost of order classification. In parallel partitions relay picking mode, it requires to split each order firstly, which requires a single system or machine to split the orders. This is actually a waste of behavior. However, it does not need to split orders, which save the cost of order classification.
4) High accuracy of selection. In this mode, each picker is responsible for the entire order, which makes pickers clear of the location of cargo in order not to cause the omission. Goods from different orders in one picking area are required to be reallocated according to the order after sorting which makes the low accuracy of order picking in the serial partition relay picking mode
Of course, the serial partition relay picking mode has some shortcomings, specifically the following 2 points:
1) Congestion in warehouse channel. This is more important for warehouse with higher degree of mechanization operation, but less important for warehouse with generally degree of mechanization operation. The degree of mechanization operation in the field of China Post Logistics is not very high, so it will not cause congestion
2) Meaningless exchanges between employees. In the serial partition relay picking mode, because there are a large number of the intersection between different pickers , it makes an increase in unnecessary exchanges between the pickers , which wastes the sorting time and reduces the efficiency of the order picking.
There are a large number of emergency orders in medical logistics warehousing chosen field in China Post Logistics. These orders need short time to reaction, and there is not expensive sorting equipment in the field of postal logistics warehouse which basically needs human to sort. Based the two above aspects as well as the advantages and disadvantages of the above serial partition relay picking mode with parallel partition relay picking model, this paper selects the serial partition relay picking mode to enhance the efficiency of warehouse order picking in China Post Logistics in the field in order to be able to improve the efficiency of the supply chain execution system in the areas.

3.2 The key to optimize the serial partition relay picking model

The biggest advantage of China Post Logistics' medical warehousing field is the time saving. Therefore, save as much time as possible is very important, especially for China Post Logistics which have a lot of urgent orders. The customers in nowadays consider the order speed more and more, how to response the order as quick as possible becomes more and more important.
There is a research found that the time savings from two main parts in this mode: saving the identify time and the walking time, and the walking time accounted as much as more than 70% of the entire order picking time. So the key to reduce the walking time when picking an order is to reduce the pickers' unnecessary walking.
Reduce the order pickers' unnecessary walking can perform in two ways: Firstly, strict control the picker' order picking operation; secondly, design a rational picking route in order to assure the order picker walk the shortest way. Our main concern here is to design the optimal path of picking makes the picker walk the shortest distance to make the order picking efficiency maximum. The purpose of the serial partition Relay picking mode is to reduce picking time, improve the response time of the orders, the key is to design a rational picking route to make the order picker walk the shortest way. We can improve the reaction speed of the order and to reduce the response time by this way, so that it can improve the Supply Chain Execution System efficiency of China Post Logistics' medical warehousing field.

4. Ant Colony Optimization design based on serial partition relay picking model

Symmetric TSP problem is the most basic lining problem, this problem can be seen as the a single traveler who stars from one city, travelling to the other cities, and minimum the walking distance, it is a typical NP-hard problem and difficult to find the accurate solution, especially in solving large scare problems [6-7]. Since the TSP problem was put

forward, many scholars get a deep research in this issue and improved the solving method. For the general method of solution is to use intelligent algorithms, such as ant colony optimization [8]. The paper below shows the ant colony optimization applied to serial partition relay pick mode path optimization research and hope that it can draw a conclusion.

The algorithm above is similar to the traditional TSP problem [9, 10]. And the mathematical model of this problem can be constructed as below.

$$\min \quad z = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} x_{ij}$$

$$s.t. \quad \sum_{j=1}^{n} x_{ij} = 1 \qquad i = 0,1,2\cdots n$$

$$\sum_{i=1}^{n} x_{ij} = 1 \qquad j = 0,1,2\cdots \mathrm{n}$$

$$x_{i_1 i_2} + x_{i_2 i_3} + \cdots + x_{i_k i_1} \le k-1 \qquad k = 2,3,\cdots, \mathrm{n}-1$$

$$x_{ij} = 1 \ or \ 0$$

We set w_{ij} Means the distance between the goods i and goods j, x_{ij} =1 means the cargo move from position i to position j the moment, x_{ij} =0 means the picking truck don't cross the route i and route j.

In the formula listed above

$$\min \quad z = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} x_{ij}$$

Means the shortest distance of the order pickers' walking distance.

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad i = 0,1,2\cdots n$$

$$\sum_{i=1}^{n} x_{ij} = 1 \qquad j = 0,1,2\cdots \mathrm{n}$$

These two formulas mean that the loop can only go through each vertex once and the only once.

$$x_{i_1 i_2} + x_{i_2 i_3} + \cdots + x_{i_k i_1} \le k-1 \qquad k = 2,3,\cdots, \mathrm{n}-1$$

This formula means the model will not include the loop with k vertexes.

When we use the ant colony optimization model to solve this problem, we suppose that there are m ants, and each ant has the following features: the ants select the next visiting city by the probability of the pheromone and the distance between the current city and the next one ($\tau_{ij}(t)$ refers to the probability of line $e(i, j)$ at t time).

We set the ants goes the legal route and unless the ant visited all the cities, there are not allowed to visited the city which they have visited before. In order to prevent this situation to happen, we set a tabe table (we set that $tabu_k$ is the table of ant k and the $tabu_k(s)$ is the sth element). When the ant completed a circle, the ant left the pheromone on every route it walked.

The amount of pheromone on each route is the same at the initial time, and we set $\tau_{ij}(0)=C$ (C for constant). When the ants are moving, the ant choose the next city by the amount of pheromone on the linking route, and $p_{ij}^{k}(t)$ means the probability of ant k moving from city I to city J at time t.

$$p_{ij}^{k} = \frac{\tau_{ij}(t) \bullet \eta_{ij}^{\beta}(t)}{\sum_{s \in allowed_k} \tau_{is}^{\alpha}(t) \bullet \eta_{is}^{\beta}(t)} \qquad if \quad j \in allowed_k$$

$$p_{ij}^{k} = 0 \qquad otherwise$$

$allowed_k = (0,1\cdots \mathrm{n}-1) - tabu_k$ refers to the next city allowed to select for ant k, and we use $tabu_k(k = 1,2,\cdots \mathrm{m})$ to record the cities the ant visited, it can be dynamic adjustment with the evolution process. η_{ij} stands for the inspiration function, and it refers to the arc of visibility, it can be find out by some heuristic algorithms, in general, $\eta_{ij} = 1/d_{ij}$,and d_{ij} means the distance between city i and city j. a refer to the important degree of the path, β refers to the importance of the visibility. After n periods of time, the ants complete a cycle, and the amount of information need to be adjusted according to the following formula:

$$\tau_{ij}(t+n) = (1-\rho) * \tau_{ij}(t) + \Delta\tau_{ij}$$

$$\Delta\tau_{ij} = \sum_{k=1}^{m} \Delta\tau_{ij}^{k}$$

In the formula above, ρ means the volatile coefficient of the pheromone, $\rho \in [0,1]$, refers to the amount of pheromone left by the ant k on route ij, and $\Delta\tau_{ij}$ refers to the increment of pheromone on the route ij in the cycle.

The steps to simulate this mathematic model can be describes as below:

1th step: Initialize all parameters: $\alpha, \beta, \rho, C, Q$, posing the number of ants is m, the number of nodes is n, place the ants to node (that is picking containers that goods need to be sorted) as each ant's initial position;

2nd step: Calculate the distance between each node (Euclidean distance), and calculate the visibility of t

moment, the visibility is the reciprocal of the distance between the nodes;
3rd step: Start the iteration of the algorithm;
4th step: ants move from node i to the next node. The moving probability is decided by probability function above. At the same time update the table tabu. If the table is not full so continue searching until a path search is complete and then return to point of origin;
5th step: Record the shortest path, and release the tabu lists, if the number of the iterations is less than the specified number of iterations, then restart the interation from the 3rd step;
6th step: Find the shortest path as an end result from all optimal paths and results;
7th step: Output the best path.
The steps of this simulation described above and the ant colony algorithm's iterative process is shown in Fig 3 below.

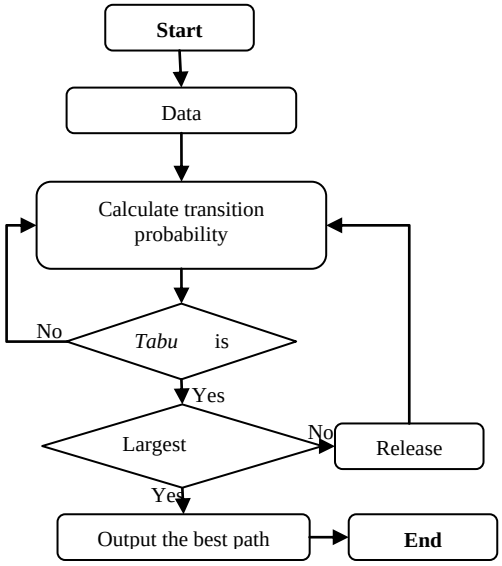


Fig. 3 Ant colony Optimization's iterative process

5. Numerical example and effectiveness

We use MATLAB R2010a to program this algorithm to completed the following examples. The examples assume that a digital picking machine positioning the place by the lights on the shelves to pick up the goods, we assume that the tally at coordinates (0, 0) and the other various coordinates are（36，15）、（1，22）、（37，13）、（34，35）、（32，16）、（23，22）、（41，4）、（31，9）、（36，8）、（37，27）、（6，36）、（27，49）、（23，16）、（33，9）、（15，16）、（18，21）、（31，23）、（37，22）、（36，28）、（29，28）、（43，21）、（39，18）、（37，23）、（34，23）、（34，31）、（23，24）、（30，20）、（25，27）、（28，26）、（20，25）,and these 30 kinds of medicine are in 6 libraries. In the analysis of this example, we select the algorithm to interaction 7 times and observing the final results, these results show in Table 1 below:

Table.1 Iteration results

Num. of Calculations	1	2	3	4	5	6	7	Mean
Num. of Iterations	30	50	100	150	200	250	300	
Routings	282	278	276	282	282	280	284	280.6

The average optimal solution in table 1 is 280.60, and the optimal solution is 276. The results in table 1 shows: it is almost the same when the integration times is more than 100, and the best interaction time is 100 times. With the increasing of iterations, the computation time required is increasing, the compute efficiency is decreasing, and too short interactions lead the result not too accurate. Therefore, we interact 100 times to generate as the optimal solution. Fig 5 and Fig6 shows the specific picking route and the average total route when interact 100 times.

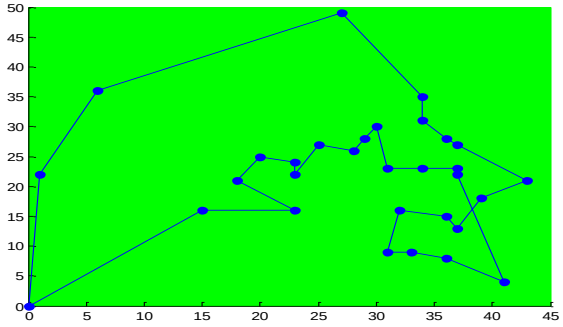


Fig.4 The origin order picking route under parallel partition picking mode

Before the optimization of the system, the order picking mode is parallel partition relay picking model, in the mode, 6 order pickers responsible for an order, each of the take responsible for good picking in their own library district. Such picking strategy enormous waste of the human resources, and when a picker finish his own work, he sent the order to the next order in the neighbor picker, it is a big waste for China Post Logistics' warehouse. In a cost analysis, we find that in the original picking system, every order need 1.21 order pickers and the average picking distance is as much as 310.42 units (show in the Fig4). In

the new order picking mode, each order only need one order picker and the walking distance is only on the average of 280.60. Not only to this, this mode don't need to splitting the orders.

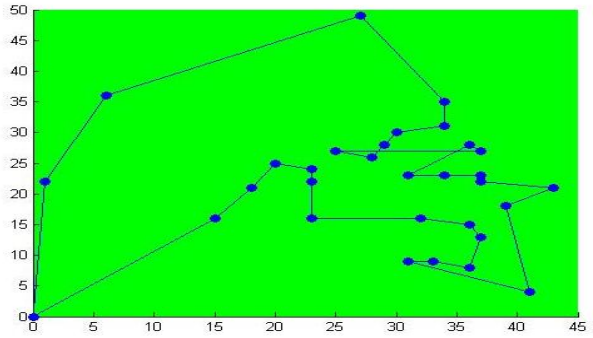


Fig.5 The specific optimal order picking route under Serial partition relay picking model

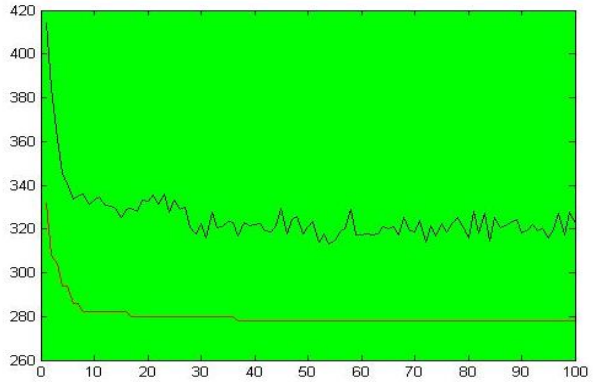


Fig.6 The average total optimal order picking route under Serial partition relay picking model

Consider from the human resources cost consumption, the cost of order picking in the warehouse reduce 17.36%. Consider from the distance of the order picking path, more than 9.80% decrease than that as before. The results show the order picking mode and the route planning is effective.

6. Conclusions

This paper design a new order picking mode in China Post Logistics' medical logistics business area by analyzing the existing problem of this field, and then design the order picking route by using the Ant Colony Optimization. At last, it made numerical examples through MATLAB simulations. The example simulation result shows that this optimizing is effective and the order picking cost decrease 17.36% and the route decrease 9.80% than that as before. This research not only to Chain Post Logistics but also to other logistics company which runs the medical business has certain reference.

Acknowledgments

The work is funded by the National Natural Science Foundation of China (No.61003065, and 61272169).

References

[1] Mazzo S, Loiseau L. Ant colony algorithm for the capacitated vehicle routing [J]. Electronic Notes in Discrete Mathematics,2004(18):181-186.
[2] Eyan A, Rosenblatt M J. Establishing Zone in Single command Class-based Rectangular AS/RS [J]. IEEE Transactions. 1994:38-46.
[3] Jane C C, Storage Location Assignment in a Distribution Center[J].International Journal of Physical and Logistics Management.2000(1):55-57.
[4] Jane C C, Laih Y W. A Clustering Algorithm for the Item Assigement in a Synchronized Zone order Picking System[J]. European Journal of Operational Reaseach,2008,116:489-496.
[5] S Henn, G Wascher. Tabu search heuristics for the order batching problem in manual order picking systems [J]. European Journal of Operational Research, 2012.
[6] CH. V. Raghavendran, G. Naga Satish, P. Suresh Varma, Intelligent Routing Techniques for Mobile Ad hoc Networks using Swarm Intelligence[J], International Journal of Intelligent Systems and Applications, IJISA Vol. 5, No. 1, December 2012.
[7] Wu Ying-ying, Wu Yao-hua, Wang Yan-yan. Modeling and Emutational Analysis of Sorting Systems with Random Orders [J]. Journal of System Simulation, 2011(1).
[8] Xiaochun Li, Analysis on Efficiency of Zone Order Picking in Warehousing[J], Packing engineering. Vol.29, No.8, 2008.
[9] P Luo, H Fu, L Hou, Y Zhang, W Fan. Parallel Ant Colony based inter-domain routing algorithm in WSON[J]. Technology,2011.
[10] M Dorigo, T Stutzle. Ant colony optimization: Overview and recent advances[M], Handbook of metaheuristics, 2010.

Xu Wei is a Phd student School of Computer Science of Beijing Institute of Technology. He received his master degree of computer science from Beijing University of Posts and Telecommunications in 1999. Now He is a deputy director of China Postal Express & Logistics Co., Ltd. And his researches include Data Picking System and E-Commerce.

Chongyang Shi received his Phd degree from Beijing Institute of Technology in 2011.Now he is a lecturer in Beijing Institute of Technology. And his researches include web mining and knowledge organizing in computer science.

Hantao Song is a professor in School of Computer Science of Beijing Institute of Technology. He received his bachelor degree from Tsinghua University in 1965. And his researches include web mining, distributed database and multimedia.

Effect of Fuel Types on the Performance of Gas Turbines

Naeim Farouk[1] , Liu Sheng [2]

[1] College of Power and Energy Engineering, Harbin Engineering University
Harbin, 150001, China

[2] College of Automation, Harbin Engineering University
Harbin, 150001, China

Abstract

In this paper we investigate the effect of the different types of fuels used during the same period in order to see how the efficiencies were affected. There are two types of fuels which can be used in the power plant station under consideration, one of them is liquefied petroleum gas (LPG) and the other is the light diesel oil (LDO). The efficiency of the plant is high while using liquefied petroleum gas (LPG) when compared to light diesel oil (LDO). This due to two reasons , the first is the high low heat value (LHV) of LPG , The second is the mixture of LPG and the air is more homogenous than LDO mixture during the combustion process.

Keywords: *Gas turbine, combined cycle, configuration system, Calorific value.*

1. Introduction

Gas turbines are increasingly used in combination with steam cycle, either to generate electricity alone, as in combined cycles, or to cogeneration both electrical power and heat for industrial processes [1], a wide variety of fuels, solid, liquid and gases can be used. A combined cycle featuring one or several gas turbines and a steam cycle is a power plant option commonly used for power production that offers high efficiency [2]. For any gas turbine-manufacturer, the fuels that will be used will have a profound effect upon both the machine design and the materials of construction [3]. When using natural gas, the combined cycle with unfired heat recovery steam generator can achieve the highest net plant efficiency (about 60%) of all fossil-fueled power plants used mid to upper output range, since the fuel heat is only supplied at a high temperature level to the working fluid in the combustion chamber of the gas turbine [4].

2. System configuration

The plant consists of two gas turbines with type of PG6581B and rated capacity of 38 MW, one unit of steam turbine with rated capacity of 36 MW and heat recovery steam generator (HRSG) is made by Harbin Boiler Works (China). Heat Recovery Steam Generator (HRSG) is the important component of combined cycle power plant used to recover waste heat from the high temperature of the exhaust of the gas turbines and generate steam .High efficiency; low energy losses and long expected life are the important factors which make combine cycle power plants unique in compression with other type of plants. The steam turbine type L36-6.70 is also the product of Nanjing Turbine & Electrical Machinery Group Co .Ltd Other main ancillary systems consist of air compressor system, firefighting system, potable water generation plant, waste water treatment plant, heating ventilation and air condition (HVAC) SYSTEM. DC system, uninterruptible power supplies system (UPS), etc. A schematic diagram of the plant is shown in Fig 1.

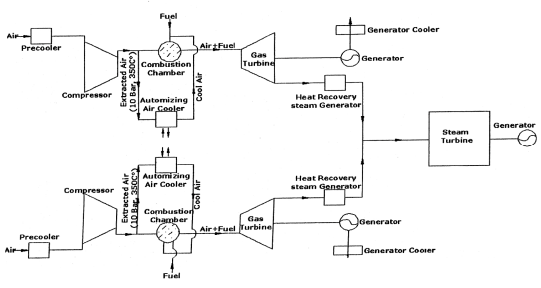


Fig. 1 Schematic Diagram of the Combined Cycle Power Plant

3. Parameters of Main Equipments

3.1 Gas turbine generator unit

The gas turbine generator unit was manufactured by Nanjing Turbine Group Company limited. The power output is 32.551MW under the following design condition:

Ambient temperature	40°C
Atmosphere pressure	0.966 bar
Ambient humidity	38%
Inlet air pressure drop	100 mm H_2O
Exhaust pressure drop (under combined cycle)	350 mm H_2O
Fuel	Light diesel oil (LDO)
Power factor	0.80
Rated frequency	50 Hz

3.2 Heat recovery steam generator

The HRSG was supplied by Harbin marine boiler & turbine research institute. The HRSG output parameters are:

Maximum continues output	63.78 t/h
Output steam temperature	6.9 M Pa
Output steam temperature	468°C
Exhaust gas temperature	<154°C
Feed water temperature	104 °C

3.3 steam turbine generator unit

Steam turbine also was manufactured by Nanjing Turbine Company limited. It is a single case, condensing type turbine. The main parameters are as followings:

Main steam pressure	6.7 M Pa
Main steam temperature	456°C
Rated process steam flow	6 t/h
Process steam pressure	0.9 M Pa
Process steam temperature	244.3°C
Rated main steam flow	127.56 t/h
Exhaust steam pressure	0.0099 M Pa
Generator power factor	0.80
Frequency	50 Hz

4. The Effect of Ambient Temperature on Efficiency

The data used for the analysis is obtained from the manufacturer data sheet of power plant [5, 6]. The output heat (Q_{out}), input heat (Q_{in}) and thermal efficiency η_{Th} is calculated by equation (1-3).

$$Q_{out}(KJ) = Q_{out}(MWh) \times 10^3 \times 3600 \quad (1)$$

$$Q_{in}(KJ) = M_{LPG} \times Low(CV)_{LPG} + M_{LDO} \times (CV)_{LDO} \quad (2)$$

$$Low(CV)_{LPG} = 45125 Kj/Kg$$

$$Low(CV)_{LDO} = 42679.2 Kj/Kg$$

$$\eta_{Th} = \frac{Q_{out}}{Q_{in}} = \frac{Q_{out}}{M_{LPG} \times Low(CV)_{LPG} + M_{LDO} \times Low(CV)_{LDO}} \quad (3)$$

Where $(M)_{LPG}$ is the mass of liquid petroleum gas, $(CV)_{LPG}$ calorific value of liquefied petroleum gas, M_{LDO} mass of light diesel oil, and $(CV)_{LDO}$ calorific value of light diesel oil .

Table1: Effect of Fuel Types on Efficiency, Year (2007)

day	Mf(Tons)LDO	Mf(Tons)LPG	Heat input (KJ)	Heat output (KJ)	Efficiency %
15/2	103.66	-	4.426 E+10	1.188 E+10	0.26832
1/3	-	145.74	6.576 E+10	1.976 E+10	0.300526
25/3	-	202.03	9.117 E+10	2.859 E+10	0.313609
15/4	172.38	-	7.360 E+10	2.029 E+10	0.27565
15/10	190.15	-	8.119 E+10	2.199 E+10	0.27088
3/11	-	167.7	7.567 E+10	2.147 E+10	0.2837

From table 1 : the efficiency of the plant is quite high when using liquid petroleum gas (LPG) And the difference vary from 5 to 8%, which indicated the type of fuel being used has a significant effect on gas turbine thermal efficiency and consequently power output.

In comparison of using LPG and LDO as fuels for operating the plant it was found that the efficiency using small amount of LPG than LDO is higher, but due to shortage in the production of LPG locally it cannot be used continuously.

LPG is more pure than LDO which result in less failure of the turbine hot parts.

References

[1] F. Haglind, A review on the use of gas and steam turbine combined cycles as prime movers for large ships, Part III: Fuels and emissions, Energy Converts. Manage. 49 (12) (2008) 3476-3482.

[2] Y. Boissnin, Combined cycle power plants: a practical guide to the right choice, ASME Cogen-Turbo, 1989, pp. 333-345.
[3] D.M. Todd, GE combined cycle experience. 33rd GD Turbine State-of-the-Art Tech Seminar, Paper No. GER-3585A, 1989.
[4] E. Wittchow, Advanced Materials and Technologies, Volume: 3A-DOI: 10.1007/b71804, ISBN: 978-3-540-42943-2 , Publisher: Springer-Verlag · Copyright: 2002.
[5] Power plant daily reports, 2007.
[6] Wafaa E, effect of air temperature on the efficiency of gas turbines in Gerri power plant, 2010.

Sheng LIU, dean of the automation college in HEU, his interests are stochastic process control, the theory and application of robust control system, electromagnetic compatibility, digital signal process, optimal estimation and control of stochastic system.

Naeim Farouk Ph.D. degree in control theory and control engineering from HUE, his interests are Fuzzy control, diesel engines analysis and control, power machinery.

Effect of Ambient Temperature on the Performance of Gas Turbines Power Plant

Naeim Farouk[1], Liu Sheng [2], Qaisar Hayat[3]

[1] College of Power and Energy Engineering, Harbin Engineering University
Harbin, 150001, China

[2] College of Automation, Harbin Engineering University
Harbin, 150001, China

[3] College of Power and Energy Engineering, Harbin Engineering University
Harbin, 150001, China

Abstract

Efficiency and electric-power output of gas turbines vary according to the ambient conditions. The amount of these variations greatly affects electricity production, fuel consumption and plant incomes. The purpose of the present study is to investigate the effect of the ambient temperature on the performance of gas turbines. We observed that the power decreases due to reduction in air mass flow rate (the density of the air declines as temperature increases) and the efficiency decreases because the compressor requires more power to compress air of higher temperature.

Keywords: *Gas Turbine, Combined cycle, configuration System, Efficiency*

1. Introduction

Several gas turbines are being widely used for power generation in several countries all over the world. Obviously, many of these countries have a wide range of climatic conditions, which impact the performance of gas turbines [1]. Gas turbines are increasingly used in combination with steam cycle, either to generate electricity alone, as in combined cycles, or to cogeneration both electrical power and heat for industrial processes [2]. A combined cycle featuring one or several gas turbines and a steam cycle is a power plant option commonly used for power production that offers high efficiency.

Kakaras [3] reported that the gas turbine output and efficiency is a strong function of the ambient air temperature. Depending on the gas turbine type, power output is reduced by a percentage between 5 to 10 percent of the ISO-rated power output (15^{o}C) for every 10 K increase in ambient air temperature. At the same time the specific heat consumption increases by a percentage between 1.5 and 4 percent. Lamfon [4] investigated the performance of a 23.7 MW gas turbine plant operated at ambient temperature of 30 to 45^{o}C. The net power output is improved by 11 percent when the gas turbine engine is supplied with cold air at the inlet. At the ambient temperature of 30^{o}C the net power output increases by 11 percent at ISO-rated condition, accompanied by a 2 percent rise in thermal efficiency and a drop in specific fuel consumption of 2 percent.

Mohanty [5] presented that by increasing the inlet air temperature from the ISO-rated condition to a temperature of 30^{o}C, would result in a 10 percent decrease in the net power output. For gas turbine of smaller capacities, this decreased in power output can be even greater. He also indicated that a rise in the ambient temperature by 1^{o}C resulted in 1 percent drop of the gas turbine rated capacity

Ameri [6] reported that in a 16.6 MW gas turbine when the ambient temperature decrease from 34.2^{o}C to ISO-rated condition, the average power output can be increased by as much as 11.3 percent. He also indicated for each 1^{o}C increase in ambient air temperature, the power output will decrease by 0.74 percent.

Dawoud [7] presented the results from the study of gas turbine plant in two locations in Oman. The results showed that fogging cooling is accompanied with 11.4 percent more electrical energy in comparison with evaporative cooling in both locations. On the other hand, absorption cooling offers 40 percent and 55 percent more energy than fogging cooling.

Aihazmy [8] reported that an average power output increment of 0.57 percent for each 1^{o}C drop in inlet temperature. The power output is increased by 10 percent during cold humid conditions and by 18 percent during hot humid condition.

Boonnasa [9] presented the results from the study of combined cycle power plant operated in Bangkok. The results showed that decreasing temperature from 35^{o}C to ISO-rated condition increase the power output of a gas turbine by 10.6 percent and the combined cycle power

plant by 6.24 percent annually. The gas turbine was rated at 110.76 MW.

2. System configuration

The plant consists of two gas turbines with type of PG6581B and rated capacity of 38 MW, one unit of steam turbine with rated capacity of 36 MW and heat recovery steam generator (HRSG) is made by Harbin Boiler Works (China). Heat Recovery Steam Generator (HRSG) is the important component of combined cycle power plant used to recover waste heat from the high temperature of the exhaust of the gas turbines and generate steam .High efficiency; low energy losses and long expected life are the important factors which make combine cycle power plants unique in compression with other type of plants. The steam turbine type L36-6.70 is also the product of Nanjing Turbine & Electrical Machinery Group Co .Ltd. Other main ancillary systems consist of air compressor system, firefighting system, potable water generation plant, waste water treatment plant, heating ventilation and air condition (HVAC) SYSTEM. DC system, uninterruptible power supplies system (UPS), etc. A schematic diagram of the plant is shown in Fig 1.

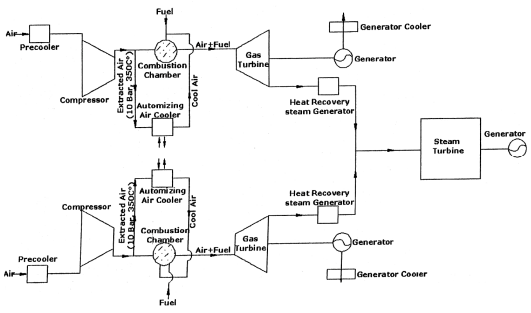


Fig. 1 Diagram for gas turbines in power plant

4. Parameters of Main Equipments

4.1 Gas turbine generator unit

The gas turbine generator unit was manufactured by Nanjing Turbine Group Company limited. The power output is 32.551MW under the following design condition:

Ambient temperature	40°C
Atmosphere pressure	0.966 bar
Ambient humidity	38%
Inlet air pressure drop	100 mm H_2O
Exhaust pressure drop (under combined cycle)	350 mm H_2O
Fuel	Light diesel oil (LDO)
Power factor	0.80
Rated frequency	50 Hz

4.2 Heat recovery steam generator

The HRSG was supplied by Harbin marine boiler & turbine research institute. The HRSG output parameters are:

Maximum continues output	63.78 t/h
Output steam temperature	6.9 M Pa
Output steam temperature	468°C
Exhaust gas temperature	<154°C
Feed water temperature	104 °C

4.3 steam turbine generator unit

Steam turbine also was manufactured by Nanjing Turbine Company limited. It is a single case, condensing type turbine. The main parameters are as followings:

Main steam pressure	6.7 M Pa
Main steam temperature	456°C
Rated process steam flow	6 t/h
Process steam pressure	0.9 M Pa
Process steam temperature	244.3°C
Rated main steam flow	127.56 t/h
Exhaust steam pressure	0.0099 M Pa
Generator power factor	0.80
Frequency	50 Hz

4.4 The Effect of Ambient Temperature on Efficiency

The data used for the analysis is obtained from the manufacturer data sheet of power plant [10, 11].
All the finding obtained from plant was analyzed. The analysis showed results (Table I), which have been plotted in graphs, Figure (2 and 3). The graphs provide and depict result of the calculated power output and efficiency. The calculations took into account the average reading of nine days for each month.

Table 1: Effect of ambient conditions on performance, years (2006-2007)

	Month	Ta°C Average	Heat input (KJ)- Average	Heat output (KJ)- Average	Efficiency %
2006	3	28.5	4.68E+10	1.54E+10	36.48
	4	31.5	4.91E+10	1.80E+10	36.77
	5	34	5.01E+10	1.69E+10	34.5
	6	34	4.30E+10	1.55E+10	36.25

	7	32	4.97E+10	1.78E+10	35.79
	8	31	5.00E+10	1.77E+10	35.32
	9	32.5	4.74E+10	1.72E+10	36.35
	10	32	4.33E+10	1.60E+10	37.17
	11	27.5	2.55E+10	1.01E+10	36.75
	12	24.5	4.74E+10	1.69E+10	37
2007	2	22	3.91E+10	1.48E+10	38.02
	3	25	7.72E+10	1.73E+10	37.87
	4	31	5.27E+10	1.83E+10	35.86
	5	35	5.54E+10	2.22E+10	36.84
	6	33	5.13E+10	1.80E+10	40.11
	7	27	5.87E+10	2.45E+10	42.89
	8	28	5.80E+10	2.18E+10	41.92
	9	31	5.89E+10	2.63E+10	38.01
	10	31	4.86E+10	1.87E+10	38.42
	11	30	4.44E+10	1.71E+10	40.94
	12	26	4.95E+10	2.05E+10	41.09

The output heat (Q_{out}), input heat (Q_{in}) and thermal efficiency (η_{Th}) are calculated by equation (1-3).

$$Q_{out}(KJ) = Q_{out}(MWh) \times 10^3 \times 3600 \quad (1)$$

$$Q_{in}(KJ) = M_{LPG} \times Low(CV)_{LPG} + M_{LDO} \times (CV)_{LDO} \quad (2)$$

$$Low(CV)_{LPG} = 45125 Kj/Kg$$

$$Low(CV)_{LDO} = 42679.2 Kj/Kg$$

$$\eta_{Th} = \frac{Q_{out}}{Q_{in}} = \frac{Q_{out}}{M_{LPG} \times Low(CV)_{LPG} + M_{LDO} \times Low(CV)_{LDO}} \quad (3)$$

Where M_{LPG} is the mass of liquid petroleum gas, $(CV)_{LPG}$ calorific value of liquefied petroleum gas, M_{LDO} mass of light diesel oil, and $(CV)_{LDO}$ calorific value of light diesel oil .

5. Results and discussions

From figure 2: illustrates the variation of temperature and efficiency during the whole year. In March when the temperature is 28.5°C the corresponding efficiency is 36.48%, for April the efficiency is nearly same as March, but when the temperature increases the efficiency decreases as shown in May and June (summer season). For the remaining months of the years; September, October, November and December the efficiency is observed to as the temperatures drops in those months.

From figure 3: the efficiency decreases gradually as the average temperature increase from February to April. In July the efficiency reach maximum value when the ambient temperature is 27°C due to the rainy season. The efficiency then decrease again in September and October due to increasing in the temperature. The efficiency again rises as the temperature drops in December (winter season).

In general the ambient conditions under which a gas turbine operates have a noticeable effect on both the power output and efficiency.

It is clear from the above that the efficiency is greatly affected by the ambient temperature of the air entering the compressor.

There is variation in power and efficiency for a gas turbine as a function of ambient temperature compared to the reference international organization for standards (ISO) condition at sea level and 32.78 °C.

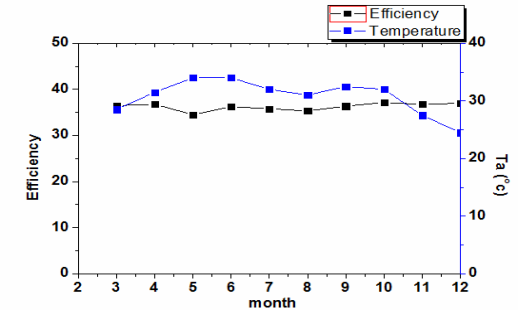


Fig. 2 Thermal efficiency and ambient temperature during the year (2006)

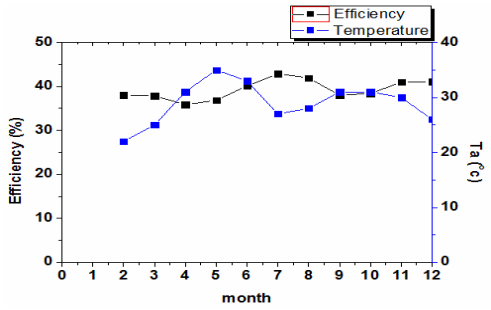


Fig. 3 Thermal efficiency and ambient temperature during the year (2007)

References

[1] Amir Abbas Zadpoor, Ali Hamedani Golshan, Performance improvement of a gas turbine cycle by using a desiccant-based evaporative cooling system, Energy 31 (2006) 2652–2664
[2] F. Haglind, A review on the use of gas and steam turbine combined cycles as prime movers for large ships, Part III: Fuels and emissions, Energy Convers. Manage. 49 (12) (2008) 3476-3482.
[3] Kakaras, E. (2006) , Inlet Air Cooling Methods for Gas Turbine Based Power Plant, ASME vol.128, pp. 312-317.
[4] Lamfon, J.N .(1998) , Modeling and Simulation of Combined Gas Turbine Engine and heat Pipe System for Waste Heat Recovery and Utilization, ENERGY CONVERS vol.39, pp. 81-86.
[5] Mohanty, B. (1995) , Enhancing Gas Turbine Performance By Intake Air Cooling Using an Absorption Chiller, HEAT RECOVERY SYSTEMS & CHP vol.15, pp. 41-50.
[6] Ameri, M. (2004) , The Study of Capacity Enhancement of The Chabahar Gas Turbine Installation Using an Absorption Chiller, APPLIED THERMAL ENGINEERING vol.24, pp. 59-68.
[7] Dawaud, B. (2005) , Thermodynamic Assessment of Power Requirements and Iimpact of Different Gas-Turbine Inlet Air Cooling Techniques at TwoLocations in Oman, APPLIED THERMAL ENGINEERING vol.25, pp. 1579-1598.
[8] Alhazmy, M.M. (2004), Augmentation of Gas Turbine Performance Using Air Coolers, APPLIED THERMAL ENGINEERING vol.24, PP .415-429.
[9] Boonnasa, S. (2006) , Performance Improvement of The Combined Cycle Power Plant By Intake Air Cooling Using an Absorption Chiller, ENERGY vol.31, pp. 2036-2046.
[10] Power plant daily reports, 2007.
[11] Wafaa E, effect of air temperature on the efficiency of gas turbines in Gerri power plant, 2010.

Sheng LIU, dean of the automation college in HEU, his interests are stochastic process control, the theory and application of robust control system, electromagnetic compatibility, digital signal process, optimal estimation and control of stochastic system.

Naeim Farouk Ph.D. degree in control theory and control engineering from HEU, his interests are Fuzzy control, diesel engines analysis and control, power machinery.

Qaisar Hayat was born in Pakistan. He received his bachelor's degree in electrical engineering from University of Engineering and Technology, Taxila (Pakistan) in 2001 and master's degree from Lahore University of Management Sciences Lahore (Pakistan) in 2006. Then joined to HEU for PhD in Power and Energy Engineering

Knowledge representation with SOA

Daniela Gotseva[1] and Ioannis Dimakopoulos[2]

**[1] Computer Systems Department, Technical University of Sofia
Sofia, Bulgaria**

**[2] Computer Systems Department, Technical University of Sofia
Sofia, Bulgaria**

Abstract

This paper addresses the problem of supporting the software development process through the artificial intelligence. The expert systems could advise the Domain Engineer in programming without the detailed experience in programming languages. He will use and integrate, with the help of deductive database and domain knowledge, the previously developed software components to new complex functionalities.
The objective of this document is to provide the knowledge representation about atomic Web Services which will be registered as the facts in the deductive database. The author proposes to use the decision rules in decision tables for representing the service model which consists of semantic specification, interface description, service quality (QoS), non-functional properties. Also the use of Domain Specific Languages (DSL) for modeling Domain Engineer's re-quests to the expert system will be considered within this document. As the illustrative use case for described knowledge representation the author proposes the domain of SOA-based geographic information systems (GIS) which represent a new branch of information and communication technologies.

Keywords: *domain engineering, Services Oriented Architecture, deductive database, expert system, Domain Specific Languages, service model, complex service.*

1. Introduction

The aim of this document is to propose a new approach of soft-ware development supported by the artificial intelligence. The Services Oriented Architecture (SOA), especially the Web Services go towards the need of developing software families through Domain Engineer which has no detailed experience in computer programming, but has strong expert knowledge. This process could be supported by expert systems.

The background of the consideration is the Domain Engineering approach [8] which relies on developing software families from reusable components which are parts of common domain system. In the future, the software can be named service-ware, where all resources are services in a Service Oriented Architecture. The main idea of this approach is that business processes engineer operates on atomic services, not on the software or hardware that implements the service [9].

The method proposed within this paper could be used in large companies enabled on SOA for realizing business processes management (BPM) applications. Web Services are considered as a promising technology for Business-to-Business (B2B) integration. A set of services from different providers can be composed together to provide new complex functionalities.

2. Concept

Fig. 1 presents the overview of the approach considered within this document. Expert system plays the role of decision supporting system. Its task is to provide the proposition of complex service (workflow of atomic Web Services) basing on the Domain Engineer's request explained by means of Domain Specific Language (DSL). The facts in the deductive database are delivered by Soft-ware Developer which implements new functionalities fashioned as the Web Services compliant with enterprise SOA infrastructure. Software Developer registers the atomic service model into facts database and also the service instance in SOA registrars.

The author of this paper proposed in previous work [3] the proof of concept prototype based on the Java framework for intelligent discovery and matchmaking atomic Web Services within integrated workflow called complex service. Thus, the problem of knowledge representation in Services Oriented Architecture will be considered in next sections.

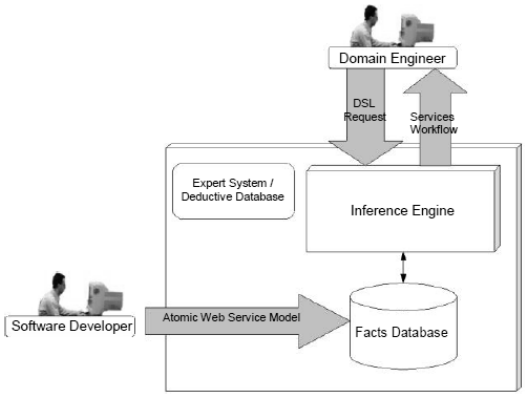


Fig. 1 Overview of the approach.

3. Problem statement and challenges

The solution issue of writing computer program through other computer program is very idealistic challenge, so it seems to be realistic when some assumptions have been fulfilled. The Services Oriented Architecture based on a collection of Web Services that communicate with one another within the distributed systems, which are self-contained and do not depend on the context or state of the other services, allows for discovery of new program functionalities by expert system. The next assumption is that all actors of Fig. 1 should use common domain namespace (domain objects) expressed through domain ontologies (for instance Web Service Modeling Ontology [20]).

The aim of research work described within this document is to provide the sufficient knowledge representation about Web Services which consists of service models, that involve interface de-scription and semantic specification as well as information about service quality (QoS) and non-functional properties.

The properly defined models of atomic Web Services registered as the facts in expert system will enable inferring knowledge about enterprise software resources by Domain Engineer and matchmaking them as the new applications.

4. Related work

The author of [1] describes the semantic service specification, which is the basis for the composition of services to application service processes. Semantic-specified services are a precondition for the development of complex functionality within application service processes. If the user wants to use a service with a desired functionality he sends the semantically specified request and checks which existing services can fulfill this request. The semantic service specification specifies the characteristics of a service. It means, semantic service specification defines what the service does, not how the service doest it. The characteristics of a service contain for example the input parameter, the results, the effects (changing of the world) and the conditions for a successful execution of the service. The first requirement of the semantic service specification is an existing domain ontology, which describes the domain specific concepts and associations and attributes of these concepts. A further requirement for the description of the semantic service specification is a unified description language. The F-Logic language [17] and its extension called Flora-2 [19] have been used. F-Logic is a deductive, object oriented database language which combines the declarative semantics and expressive-ness of deductive database languages with the rich data modeling capabilities supported by the object oriented data model [1].

The authors of this paper propose other approach to explain the service models using Java language expressions. The main objective for this solution is to combine in one programming language: knowledge about services, expert system/rule engine compliant with JSR-94 specification (implementation of the Java Rule En-gine API known as JSR94, which allows for support of multiple rule engines from a single API [16]) as well as J2EE [18] middle-ware and software patterns which is the powerful development platform for Services Oriented Architecture [2]. In the previous paper author proposed the architecture for complex services prototyping and proven the feasibility of this approach on the Java plat-form using the developed prototype [3].

A proper service description answers three questions about a service: what the service does (including its non-functional description), where it is located, and how it should be executed [4]. The Fig. 2 presents the atomic service model proposed by authors of this paper which answers these questions.

Web Services are software applications with public interfaces described in XML. According to the established standards, Web Service interfaces are defined in Web Service Description Language (WSDL) [5]. Published in Universal Description, Discovery and Integration (UDDI) registrars [10] could be discovered and invoked by other software components. These systems interact with Web Services using XML-based message in Simple Object Access Protocol (SOAP).

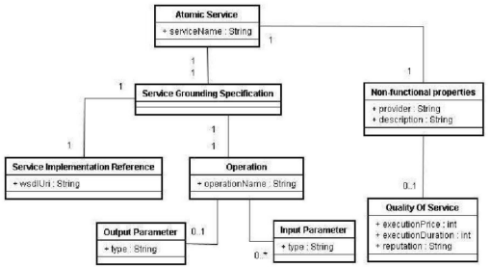


Fig. 2 Atomic Service Model.

Service Grounding Specification (see Fig. 2) refers to the WSDL description. WSDL consists of a hierarchy of objects (proposed within [3] to use domain ontology to define these objects), from the most basic data type, through message, operation, port type, binding and port to service itself [5]. Its wsdlUri attribute is the Unified Resource Identifier (URI) and refers to the service WSDL file. WSDL does not provide methods to describe non-functional service properties.

Quality of Service (QoS) in service oriented platforms is a crucial attribute in assessing proper operation of services. Loosely coupled distributed systems in service discovery, composition and execution have emerged as a new paradigm in building virtual organizations. In order to support rapid and dynamic composition of services it should be possible to locate services that meet user's functional requirements. Moreover, it should be possible to select the best service based on their QoS. It is important to stress the difference between non-functional (NF) and QoS parameters. QoS parameters are a subset of NF parameters. NF parameters may include some information that is not directly computable, for example textual service description, phone numbers to service developers (providers), date of service preparation etc. As a result of that, when using either NF or QoS concepts, one should distinguish that NF relates to a whole set of non-functional parameters, and QoS refers to those NF parameters that may be computationally processed, compared and verified with greater ease [4].

In service arena it is suggested that the term QoS should refer not only to such basic, originating from networking parameters as bandwidth, latency, error rate or availability (the probability that the service is available), reliability (stability of a service function-ality, i.e. ability of a service to perform its functions under stated conditions). Therefore, additional aspects come into consideration, such as speed of operation, robustness, accuracy of operation, de-pendability, capacity (a limit of concurrent requests for guaranteed performance), throughput (the number of requests served in a given time period), response time (the time taken by a service to process its sequence of activities), execution cost (the amount of money for a single service execution). Even parameters such as operating system and storage capacity of the executing system may by considered QoS parameters, as they affect end-to-end op-eration of a service [4] [7].

Currently, most approaches that deal with quality of services address only some generic parameters such as execution price, execution duration, service availability and reliability [6]. These parameters may be defined as follows [4]:

- Execution price – the amount of money that a service requestor has to pay to the service provider for using the Web Service.
- Execution duration (also called latency time) – mea sures the expected delay in seconds between the moment when a request is sent and the moment when the service is rendered.
- Execution duration is a sum of the processing time and the transmission time.
- Reputation (also called Service quality reputation) is a measure of service trustworthiness. It depends mainly on end user's experience of using the service. Different users may have different opinions on the same service.

5. Implementation

All service instances available in particular domain are treated as the knowledge representation system and can be explained as the decision table which contains production rules. Decision tables specify what decisions should be made when some conditions are fulfilled [11]. This document considers the knowledge reasoning problem employing decision tables' formalism

$$K = (U, A) \quad (1)$$

Where K is the knowledge representation system, U is a nonempty, finite set, called universe, and A is a nonempty set of primitive attributes.

The knowledge representation system which distinguishes the condition and decision attributes can be called decision table T:

$$T = (U, A, C, D) \quad (2)$$

Where C and D called condition and decision attributes are two subsets of attributes.

Any implication

$$\Phi \rightarrow \Psi \quad (3)$$

Is considered as the decision rule and Φ, Ψ are called predecessor and successor respectively.

If Eq. (3) is decision rule and P contains all attributes occurring in Φ (condition attributes) and Q contains all attributes occurring in Ψ (decision attributes) then this decision rule can be called PQ-rule.

Let's consider the real decision table (see Table 1), which represents the knowledge system from geographic information systems domain in Services Oriented Architecture and the facts are explained as the PQ-rules. The use case scenario and the services landscape were described within [3].

Table 1: Real Decision Table.

Operation Name	*Input Parameters*	*Output Parameter*	*Provider*	*Execution Price*	*Execution Duration*	*Reputation*	*Service Name*
P1	P2	P3	P4	P5	P6	P7	Q1
getMap	{Coordinates}	Map	TeleAtlas	5$	12ms	high	GisMap
provideMap	{Coordinates}	Map	GISAtlas	0$	24ms	medium	PrintMap
drawPoint	{Coordinates, Map}	PointMarket	GIS Company	0$	2ms	high	DrawPoint
drawSegment	{Coordinates, Coordinates, Map}	SegmentLine	GIS Company	2$	5ms	high	DrawSegment
computeDistance	{Coordinates, Coordinates}	Distance	ITS	0$	1ms	medium	ComputeSegment Distance

The columns P1-P7 represent the condition attributes and column Q1 represents the decision attribute of the PQ-rule. These PQ-rules are stored as the facts in expert system database.

The Eq. (4) formalizes a possible representation of PQ-rule from Table 1 in accordance to the Eq. (3).

P1=getMap and P2={Coordinates} and P3=Map → Q1=GISMap (4)

The authors of this paper prepared the facts database in terms of production rules regarding Eq. 4 and Table 1 as the Java class which is loaded into the Working Memory of expert system (see code listing 1).

Code listing 1: FactsDatabase Class.

```
public class FactsDatabase { WorkingMemory
rulesEngineMemory;

public FactsDatabase(WorkingMemory rulesEngineMemory) {
this.rulesEngineMemory = rulesEngineMemory;
}

public void activateFacts() { AtomicService as;
Collection inputParameters; QoS qos;
// PQ rule
// P-attributes

as = new AtomicService();
as.setOperationName("getMap");
inputParameters = new ArrayList();
inputParameters.add(new
Coordinates().getClass().getName());
as.setInputParameters(inputParameters);
as.setOutputParameter(new Map().getClass().getName());
as.setProvider("TeleAtlas");
qos = new QoS(); qos.setExecutionPrice(5);
qos.setExecutionDuration(12); qos.setReputation("high");
as.setQos(qos); //Q-attributes
as.setServiceName("GisMap");
as.setServiceDescription("Service creates a map
according to provided longitude and latitude.");
rulesEngineMemory.insert(as);
}
```

As the expert system the JBoss DROOLS [12] rule engine based on the RETE algorithm [13] has been used. Drools imple-ments and extends the Rete algorithm which is called ReteOO, what signifying that Drools has an enhanced and optimized im-plementation of the Rete algorithm for Object Oriented systems [14].

The Domain Engineer models the request to the deductive database as the production rules presented in Eq. (3) manner, also to infer conclusions which results in actions ***" When <conditions> then <actions>"***. The advantage of using rules engine is the declarative programming. Rules are much easier to read than source code. Also the ability of creation of executable domain knowledge repository plays the important role. Domain experts are often a wealth of knowledge about business rules and processes. They typically are non-technical, but can be very logical. Rules can allow them to express the logic in their own terms [12].

The production rule example (code listing 2) shows the strength of proposed approach. The Domain Engineer models the request to the deductive database as the one rule instead of a lot of source code lines and nested loops in structural programming languages or SQL statements. But, the production rules modeling could be much easier through usage of Domain Specific Language (DSL). It is the way of extending the rule to problem domain. Simple DSL can be implemented by lexical processing. In addition, DSL can be used to create front-ends to existing systems or

to express complicated data structures. A DSL is a programming language tailored especially to an application domain: rather than being for a general purpose, it captures precisely the domain's semantics [15]. DSL can act as "patterns" of conditions or actions that are used in rules, only with parameters changing each time [12]. Rules expressed in Domain Specific Language have human-readable form and match the expression used by domain experts [15].

Code listing 2: Production Rule Example.

```
rule "serviceProposition1" when
#conditions
as : AtomicService( outputParameter == "soa-
rules.ontology.Map", qos.executionDuration < 20 , ser-
viceName : serviceName, serviceDescription : serviceDe-
scription )
then #actions
System.out.println( "Proposed service1: " + serviceName
+ " - " + serviceDescription);
End
```

Code listing 3 shows how the rule can be transformed to "patterns" of DSL.

Code listing 3: DSL patterns.

```
 [conditions]

DSL Language expression:
There is an Atomic Service where Rule mapping:
AtomicService(serviceName : serviceName, ser-
viceDescription : serviceDescription)
DSL Language expression:
- output parameter equals "{value}" Rule mapping:
outputParameter == "{value}"
DSL Language expression:
- executionDuration is less than "{value}" msec Rule
mapping:
qos.executionDuration < "{value}"

[actions]

DSL Language expression:
Print service name and service description Rule mapping:
System.out.println( "Proposed service1: " + serviceName
+ " - " + serviceDescription);
```

The usage of "patterns" of Domain Specific Language allows the Domain Engineer to model the request to the expert system and find the desired Web Service in friendly manner as shown on code listing 4.

Code listing 4: Usage of DSL patterns.

```
rule "serviceProposition1"

when

        #conditions
        There is an Atomic Service where
        - output parameter equals "soa-
        rules.ontology.Map"
        - executionDuration is less than "20" msec

then

        #actions
        Print service name and service description

End
```

6. Conclusion

The presented approach allows supporting the Domain Engineer in developing applications from business processes management area. The Domain Engineer has no detailed experience in computer programming, but has strong expert knowledge. He can model the requests to the deductive database as the production rules in human-readable format with usage of Domain Specific Languages instead of several lines and nested loops of Java or SQL code The author discussed within this paper the knowledge representation in SOA explained as the decision tables with atomic service models which involve semantic specification, interface description (WSDL), non-functional properties and quality of services (QoS).

The further research will be focused on refinement of reasoning process with usage of other techniques of the artificial intelligence, development of domain specific languages for GIS domain, storage of the facts before loading to production memory (the traditional solution as the text files is not enough convenient to hold on objects) as well as discovery and matchmaking workflows of complex services.

References

[1] Donath Steffi, Automatic Creation of Service Specifications, 6th Annual International Conference on Object-Oriented and Internet-Based Tech-nologies, Concepts, and Applications for Networked World, Net.ObjectDays Proceedings, pp.79-89, September 19-22, 2005

[2] Hansen Mark, SOA Using Java Web Services, Person Education Inc., Prentice Hall, 2007

[3] Grobelny Piotr, Rapid Prototyping of Complex Services in SOA Architecture, IX International PhD Workshop OWD'2007, Conference Archives PTETiS, vol. 23(1), pp.71-76, 2007

[4] Kowalkiewicz Marek, Current challenges in non-functional service description – state of the art and discussion on rese arch results, Net.ObjectDays Proceedings, Spetember 19-22, 2005 pp.91-96

[5] Christensen, E., F.Curbera, et al. Web Services Description Language (WSDL) 1.1, World Wide Web Consortium (W3C), 2001

[6] Zeng, L., B. Benatallah, et al., Quality driven Web Services Composition. Proceedings of the 12th international conference on World Wide Web (WWW) Budapest, Hungary, ACM Press 2003

[7] Kokash Natallia, D'Andrea, Vinzenzo, Evaluating Quality of Web Ser-vices: A Risk-Driven Approach, Business Information Systems, Witold Abramowicz (Ed.) 10th International Conference BIS 2007 proceedings, LNCS 4439, pp.180-194, Springer, 2007

[8] Czarnecki Krzysztof, Eisenecker Ulrich: Generative Programming – Methods, Tools and Applications, Addison Wesley, Boston, MA, 2000

[9] Ekelhart Andreas et al.: Security Issues for the Use of Semantic Web in E-Commerce, Business Information Systems, Witold Abramowicz (Ed.) 10th International Conference BIS 2007 proceedings, LNCS 4439 pp.1-13, Springer, 2007
[10] UDDI Specifications, http://www.oasis-open.org/committees/uddi-spec/doc/tcspecs.htm, accessed January 2007
[11] Pawlak Zdzisław: ROUGH SETS Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, 1991
[12] Proctor Mark et al.: Drools Documentation, http://downloads.jboss.com/drools/docs/4.0.4.17825.GA/html_single/ in-dex.html, accessed January 2008
[13] ForgyC., RETE: A Fast Algorithm for the Many Pattern Many Object Pattern Match Problem, Artificial Intelligence, 19(1), pp.17-37 Sept. 1982
[14] Doorenbos Robert B., Production Matching for Large Learning Systems (Rete/UL), PhD thesis, Carnegie Mellon University, January 31, 1995
[15] Spinellis Diomidis, Notable design patterns for domain-specific languages, The Journal of Systems and Software 56 (2001) pp. 91-99, El-sevier 2001
[16] Toussaint Alex, Java Rule Engine API™ JSR-94, Java Community Proc-ess, http://jcp.org/en/jsr/detail?id=94, BEA Systems, September 2003
[17] Kifer Michael et al.: Logical Foundations of Object-Oriented and Frame-Based Languages, Journal of the Association for Computing Machinery, May 1995
[18] Sun Microsystems, Simplified Guide to the Java 2 Enterprise Edition, http://java.sun.com/j2ee/reference/whitepapers/j2ee_guide.pdf, Accessed January 2008
[19] Yang Guizhen et al.: FLORA-2: A Rule-Based Knowledge Representation and Inference Infrastructure for the Semantic Web, Second International Conference on Ontologies, Databases and Applications of Semantics (ODBASE), Catania, Sicily, Italy, November 2003
[20] Dumitru Roman et al.: Web Service Modeling Ontology, Applied Ontology, 1(1), pp. 77-106, 2005

Daniela Gotseva is associate professor, PhD and Vice Dean of Faculty of Computer Systems and Control, Technical University of Sofia, from 2008 with primary research interest of programming languages, system programming, and fuzzy logics. She is a member of the IEEE and the IEEE Computer Society.

Ioannis Dimakopoulos is PhD student at Faculty of Computer Systems and Control, Technical University of Sofia, from 2010, with primary research interest of system programming and service oriented architecture (SOA).

Detection of false alarm in handling of selfish nodes in MANET with congestion control

Ms. I.Shanthi[1] and Mrs. D. Sorna Shanthi[2]

**[1]Department of computer science and engineering,
Rajalakshmi engineering college,
Chennai, Tamilnadu, India-602105**

**[2]Department of computer science and engineering,
Rajalakshmi engineering college,
Chennai, Tamilnadu, India-602105**

Abstract

In a mobile ad hoc network, the mobile nodes will have the characteristics of mobility and constraints in resources. Since, the mobility is high, the nodes may move randomly and fastly, which lead to network partitioning. The resource constraints leads to a big problem as decrease in performance and the network partitioning leads to poor data accessibility. To improve the data accessibility, we have proposed several data replication techniques. Most of the users at different places assume that mobile nodes co-operate fully in terms of sharing their memory space. But In reality, some nodes may decide as not to co-operate with others or partially co-operate with other nodes. The behavior of these selfish nodes leads to decrease in over all data accessibility of the network. We have explored the impression of selfish nodes in a MANET from the perspective of replica allocation and developed selfish node detection algorithm that considers the partial selfish node and fully selfish node as selfish replica allocation. The replica will be allocated using specific SCF tree concept. An alarm will be raised based on the selfish behavior of overall nodes called overall selfishness alarm. But the alarm will also be initiated because of network disconnections too but it seems and treated as overall selfishness alarm, it will affect the overall performance of the network. The concept of the paper deals with detection of false alarm as differentiated from overall selfishness alarm and to inform the other nodes at route as exactly where the disconnections occur to select the next best alternative path and also to increase the performance with increased congestion control. Detection of attacker node in the network and should be informed to all others in the network.

Keywords: *mobile ad hoc network, selfish nodes, selfish replica allocation, crcn*

1. Introduction

MANET (Mobile ad hoc network) are dynamic networks populated by mobile stations. Stations in MANETs are usually laptops or mobile phones. These devices feature Bluetooth or Wi-Fi network interfaces and communicate in a decentralized manner. Mobile ad hoc networks are composed of a set of communicating devices able to spontaneously interconnect without any pre-existing infrastructure for it. Devices in specific range can communicate in a point-to-point fashion. More and more people are interested in mobile ad hoc networks.

Mobile networking is one of the most important technologies supporting pervasive computing. Mobility is a vital feature of MANET. Because of the high cost and lack of flexibility of such networks, experimentation is generally achievable through simulation.

During the last years, advances in both hardware and software techniques have resulted in mobile hosts and wireless networking common and diverse. Generally there are two different approaches for enabling wireless mobile units to communicate with each other:

1) Infrastructure - Based network: Wireless mobile networks usually been based on the cellular concept and depend on good infrastructure support, in which mobile devices communicate with access points like base stations connected to the stable network infrastructure.

2) Infrastructure less network: In infrastructure less approach there is no central administration for the entire network. The mobile wireless network is infrastructure less in manner commonly known as a mobile ad hoc network (MANET). A MANET is a collection of wireless nodes that can dynamically form a network to exchange information without using any pre-existing stationary network infrastructure.

MANETs have attracted a lot of attention due to the admiration of mobile devices and the advances in wireless communication technologies. Each node in a MANET must act as a router, and it should communicate with each other [1]. Network partitions can occur frequently because usually the nodes will move freely in MANET. But it cause

some data to be often inaccessible to some of the nodes. Hence, data accessibility is often a significant performance metric in a MANET.

Due to its great features of mobility and flexibility, MANET attracts different real world application areas whereas topology changes very quickly. MANET is more vulnerable than wired network due to node's mobility, dangers from compromised nodes inside the network, limited security, dynamic topology, scalability and lack of centralized management [2]. Because of these vulnerabilities criteria, MANET is more susceptible to malicious attacks.

Devices in MANET should be able to detect the presence of other devices around and it should perform the necessary set up to facilitate communication and sharing of data and service. Nodes that lie within each other's communication range can communicate directly are responsible for dynamically discovering each other. In order to enable communication between nodes that are not directly within each other's communication range, intermediate nodes acts as routers that relay packets generated by other nodes to their respective destination. The nodes at MANET is often energy-constrained such as battery, memory space. In our point of view we have taken the memory space as the constraint to find out the behavior of the node.

In MANET, breaking of communication link is very frequent, as nodes are free to move anywhere. The density of nodes and number of nodes are depends on the applications in which we are using MANET. The dynamic topology of MANET results in route changes and frequent network partitions and possibly packet losses. [3]

Data are usually replicated at nodes, other than the unique owners, to increase data accessibility to handle with frequent network partitions. A large amount of research has recently been proposed for replica allocation in a MANET. In general, replication can simultaneously improve data accessibility and reduce query response time if node have space to hold both all the replicas and the original data.

However, there is often a trade-off between data accessibility and query delay, because the most of the nodes in a MANET have only limited memory space [1]. For example, a node may hold a part of the frequently accessed data items locally to reduce its own query delay to get good performance. However, if there is only limited memory space and many of the nodes hold the same replica in their local memory space. Some data items would be replaced and missing by the replication process. Thus, the overall data accessibility would be decreased. A node should not hold the same replica that is also held by many other nodes. But however because of this replication process, there will be an increase in its own query delay [1].

1.1 MANET Features

MANET has the following features:

1).Autonomous terminal: In MANET, each mobile node is an independent node, which could function as both a host and a router. The ability and functions as a host, the mobile nodes can also perform switching functions as a router. So generally endpoints and switches are indistinguishable in MANET.

2).Distributed operation: Since there is no background network for the central control of the network operations, the control of the network is distributed among the nodes. The nodes involved in a MANET should cooperate with each other and communicate among themselves and each node acts as a relay as needed, to implement specific functions such as routing and security.

3).Multi hop routing: Basic types of ad hoc routing algorithms can be single-hop and multi hop. Based on the diverse link layer attributes and routing protocols. When delivering data packets from a source to its destination out of the wireless broadcast range, the packets would be forwarded through one or more intermediate nodes. When a node tries to send information to other nodes which is out of its communication range, the packet should be forwarded via one or more intermediate nodes.

4).Dynamic network topology: Since the nodes are movable in nature, the network topology may change quickly and randomly and the connectivity among the terminals may differ with time. The nodes in the MANET dynamically establish routing among themselves as they travel around, establishing their own network on the fly.

5).Light-weight terminals: In maximum cases, the nodes at MANET are mobile with less CPU capability, low power storage and small memory size.

1.2 MANET Applications

The set of applications for MANETs are ranging from large-scale to small scale environment, highly mobile and small and dynamic networks that are constrained by power sources. Some of the typical applications include:

1).Military battlefield: Military equipment now consistently contains some sort of computer equipment which will be useful for the security of the country. The mobile Ad hoc networking would allow the military to take advantage of common place network technology to maintain an information network in the military area. The communication deals between vehicles and soldiers.

2).Commercial sector: Ad hoc can be used in emergency/rescue operations for disaster relief efforts, e.g. in fire, flood. Emergency rescue processes would takes place where non-existing or damaged communications infrastructure and rapid deployment of a communication network is needed

3).Local level: MANET can autonomously link an instant and temporary multimedia network using notebook computers or palmtop computers to spread and share information among participants at a place e.g. conference or classroom.

4).Personal Area Network (PAN): Short-range MANET can simplify the intercommunication between various mobile devices (such as a PDA, a laptop, cellular phone). Tedious wired cables are replaced with wireless connections.

5).Sensor network: This technology is a network composed of very large number of small sized sensors. These types of sensors can be used to detect any number of properties of an area. For Example, temp, pressure. The capabilities of each sensor are very limited and each must rely on others in order to forward data to a central server.

Automotive applications: Automotive networks are widely at research. Cars should enabled to talk to the road, to traffic lights and to each other. The network will provide the information about road conditions, congestions to the driver to optimize the traffic flow.

2. Selfish Node Behavior in MANET

Several nodes will be participated in the MANET for data forwarding and data packets transmission between source and destination. All the nodes of MANET will perform the routing function as mandatory. They must forward the traffic which other nodes sent to it. Among all the nodes some nodes will behave selfishly, these nodes are called selfish nodes.

MANET are Dynamic Topologies Bandwidth-constrained, variable capacity links Power-constrained operations Limited physical security.

A).Dynamic topologies Nodes are free to move arbitrarily; thus the topology of the network, may change randomly and rapidly at unpredictable times in network. Modification of transmission and reception parameters such as power may also impact the topology.

B).Bandwidth constrained: variable capacity links Wireless links will continue to have significantly lower capacity than their hard-wired counter parts. The relatively low to moderate link capacities will leads to the congestion rather than the exception.

C).Power-constrained operations: Some or all the nodes in a MANET rely on batteries for their energy. Thus, for these nodes, the most vital design problem may be that of power conservation.

Any node in MANET may act selfishly, which means, using its limited resource only for its own profit, since each node in a network has resource constraints, such as storage and battery limitations. A node would like to enjoy the profits provided by the resources of other nodes in the network, but however it should not make its own resource accessible to help others. Existing exploration on selfish behaviors in a MANET mainly focus on network concerns. For network problems at MANET may be as some selfish nodes may not transmit data to others to conserve their own battery constraints. Even though network disputes at MANET are important, replica allocation is also critical, ever since the vital goal of using a MANET is to provide data services to users [1].

We address the problem of selfishness in the context of replica allocation in a MANET. The problem because of replica allocation refers as if a selfish node may not share its own memory space to store replica for the benefit of other nodes. Selfish replica allocation refers to a node's

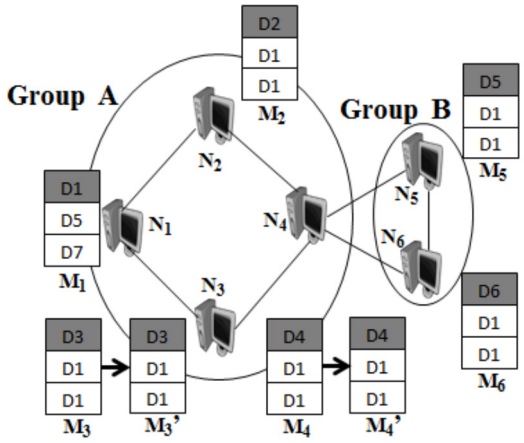


Fig 1.1 Example of selfish replica allocation

Non cooperative action, such that the node refuses to cooperate fully in sharing its memory space with other nodes. According to the figure, where nodes N1, N2... N6 maintain their memory space of nodes as M1, M2... M6, respectively, with the specified access frequency information in Table. As shown in Figure DCG seeks to minimize the duplication of data items in a group to achieve high data accessibility

Table 1.1 Access frequency table

Data	Nodes					
	N_1	N_2	N_3	N_4	N_5	N_6
D_1	0.65	0.25	0.17	0.22	0.31	0.24
D_2	0.44	0.62	0.41	0.40	0.42	0.46
D_3	0.35	0.44	0.50	0.25	0.45	0.37
D_4	0.31	0.15	0.10	0.60	0.09	0.10
D_5	0.51	0.41	0.43	0.38	0.71	0.20
D_6	0.08	0.07	0.05	0.15	0.20	0.62
D_7	0.38	0.32	0.37	0.33	0.40	0.32
D_8	0.22	0.33	0.21	0.23	0.24	0.17
D_9	0.18	0.16	0.19	0.17	0.24	0.21
D_{10}	0.09	0.08	0.06	0.11	0.12	0.09

2.1 Behavioral States of Selfish Nodes

We can define the behavioral states of nodes as three types in MANET from the viewpoint of certain constraints at memory space.

Type-1 node: The nodes are non-selfish nodes. These nodes can hold replicas allocated by other nodes within the limits of their memory space

Type-2 node: The nodes are fully-selfish nodes. These are the nodes which do not hold replicas allocated by other nodes, but it will allocate replicas to other nodes for their own accessibility.

Type-3 node: The nodes are partially- selfish nodes. These partially selfish nodes would use their own memory space partially for allocated replicas by other nodes. Their memory space may be separated logically into two parts: one is selfish area and another one is public area. These partially selfish nodes allocate replicas to other nodes for their accessibility.

The detection of the type-3 nodes is always complex, since they are not always selfish. In some situation, a type-3 node may be considered as non-selfish, since the node shares part of its memory space. But in our paper,, we have considered it as selfish node only, since these nodes also leads to the selfish replica allocation problem. Note that selfish and non-selfish nodes perform the same procedure when they receive a data access request, even though they behave differently in consuming their memory space.

2.2 Actions of Each Nodes at Specific Period

Each node detects the selfish nodes based on credit risk scores. Each node makes its own topology graph and builds its own SCF-tree by excluding selfish nodes. The topology graph may be of partial according to the particular node. Based on the concept of SCF-tree, each node in the network allocates replica in a fully distributed manner [1].
The CR score is updated during the query processing phase. With the degree of selfishness which we have measured, we intend a tree that represents relationships among nodes in a MANET, for replica allocation, termed the SCF-tree [1]. The SCF-tree models human friendship management in the real world.

When a node Ni makes an access request to a data item typically sending a query, the particular node will checks its own memory space first. If the requested item is present in its own local memory space then the request got successful. If it does not hold the original or replica, the request will be broadcasted to other nearby nodes which is connected to the node Ni. The request is also successful when Ni receives any reply from at least one node connected to Ni with one hop or multiple hops of nodes, which holds the original or replica of the targeted data item. Otherwise, the request fails.

Whenever a node Ni receives a data access request, it either 1) serves the request by sending its original or replica if it holds the target data item, or 2) forward the request to its neighbors if it does not hold the target data item.

3. Proposed Strategy

This chapter focuses on the mobile ad hoc network having selfish nodes and the way of replica allocation in the system helps to solve the data accessibility problem and about the simulation of system.

3.1 Detecting Selfish Nodes

In our strategy, each node calculates a CR score for all the nodes to which it is connected as its neighborhood. Each node at the network shall estimate the “degree of selfishness” for all of its connected nodes based on the CR score. We describe selfish features that may lead to the selfish replica allocation problem to determine both expected value and expected risk. Credit risk will

be calculated as the ratio of expected risk to the expected value.

Selfish features are classified into two groups: n query processing-specific and node-specific feature. In the query processing-specific feature, we develop the ratio of selfishness alarm which is the ratio of Node N1's data request being not served by the expected node Nk due to Nk's selfishness in its memory space. The query processing-specific feature can represent the expected risk of a node [1]. To effectively identify the expected node, Node N1 should know the (expected) status of other nodes' memory space. The SCF-tree-based replica allocation techniques support this assumption.

Node-specific features can be explained by considering the following case: A selfish node may share part of its own memory space, or a small number of data items, like partially selfish node. In this occasion, the size of shared memory space or the number of shared data items can be used to represent the degree of selfishness. The node-specific features can be used to represent the expected value of a node.

3.2 Building SCF-TREE

The main objective of our novel replica allocation techniques is to attain high data accessibility while reduce in traffic overhead. High Data accessibility is the prominent concern in all networks. If the replica allocation techniques allocate replica of the specified data item without any other node's concern, the traffic overhead will decrease.

Since the SCF-tree consists of only non-selfish nodes, we need to measure the degree of selfishness to apply in real-world friendship management to replica allocation in a MANET. We use the value of credit risk for building the tree. Before constructing or updating the SCF tree, node Ni eliminates selfish nodes from the base group INi.

The key strength of the SCF-tree-based replica allocation techniques is that it can minimize the communication cost, even though achieving high data accessibility. The high data accessibility is possible because each node detects selfishness and makes replica allocation at its own pleasure, without forming any group in the network.

Each node has a parameter d, the depth of SCF-tree. When N1 builds its own SCF-tree, N1 first appends the nodes that are connected to N1 by one hop to N1's child nodes. Then, N1 checks recursively the child nodes of the combined nodes, until the depth of the tree is equal to d. we assume that all nodes are non-selfish nodes for simplicity [1].

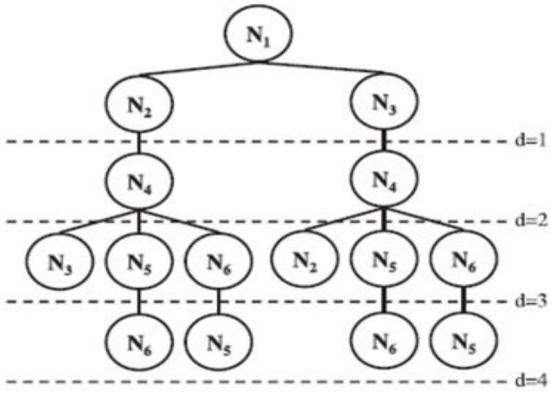


Fig 1.2 SCF tree for N1

As seen in Fig, the SCF-tree may have multiple routes for some nodes from the root node and that confer high stability. At every specific relocation period, each node in the network updates its own SCF-tree based on the network topology of that moment.

3.3 Replica Allocation

Each node allocates replica at its discretion based on Table and Fig mentioned above. When each node receives a request for replica allocation from Nk during a specific relocation period, the specific node solely determines whether to accept or reject the request. If the request is accepted by other node, the specific node will maintains its Mp based on the nCRki given by credit risk Table. If the highest nCRhi among the nodes which allocated replica to Ni, is greater than nCRki, Ni replaces replica allocated by node with replica requested by Nk.

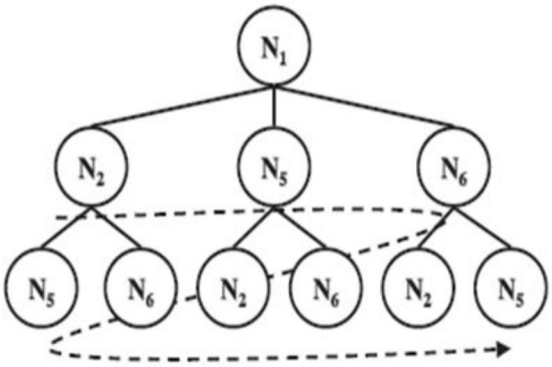
Table 1.2 Credit Risk table

N_i	N_k					
	N_1	N_2	N_3	N_4	N_5	N_6
N_1	·	0.30	0.85	0.80	0.45	0.22
N_2	0.40	·	0.80	0.90	0.30	0.50
N_3	0.25	0.35	·	0.75	0.65	0.75
N_4	0.45	0.44	0.51	·	0.23	0.37
N_5	0.30	0.60	0.85	0.40	·	0.21
N_6	0.40	0.50	0.90	0.52	0.30	·

Each node Ni executes this algorithm at every relocation period after building its own SCF-tree.

The node determines the priority for allocating replica items. The allocating replica priority is based on Breadth First Search (BFS) order of the SCF tree [1]. The dotted arrow in Fig.1.3 represents the priority for allocating replica. For example, in Fig.1.3, N1 selects N2 as the first target of the allocation in the tree. After allocating a replica to the last target node (i.e., N5 in Fig), the first node of the tree, N2 will be the next target in a round-robin manner. In our strategy, the target node could be the expected node

Fig 1.3 SCF based replica allocation



3.4 Detection of Attacker Nodes

A wired network under a single administrative domain allows for discovery, repair, response, and forensics of misbehaving nodes. A MANET is normally not under a single administrative area, making it challenging to perform any kind of centralized management or control. Hence here malicious nodes may enter and leave the immediate radio transmission range at random intervals, may collude with other malicious nodes to disrupt network activity and avoid detection [5].

After detecting some nodes as selfish node at the network, we would select an alternative path for the data packet transmission for a while. After a specific period of time every node will again start detecting the selfish nodes by measuring the degree of selfishness of all nodes in the network. But sometimes the attacker node, drops the packet without forwarding which will be very dangerous to our network also considered as selfish nodes. The attacker node in a network will be very prone to all sorts of attacks. Until the node misbehaves or alters any data in the network we cannot able to find it is an attacker. We will not take any actions against that node since us assuming that node as selfish node. In fact the node behaving as normal changed to selfish node only because of lack of the energy constraints and memory power. So a node cannot be selfish forever, whenever it get the constraints back it come back as normal node by indulging itself in normal data forwarding and sharing the space for other node's data.

If any attacker intrude inside a particular network it will leads to reduce in security of the data in the network. So we must identify the attacker node along with the selfish node. If a node acts as selfish for more than the predefined threshold value time then it will be considered as malicious or attacker. Each node at the network employs the mechanism that utilizes the neighborhood information to detect the misbehaving character of its neighbors.

At a specific interval the nodes will calculate the degree of selfishness of other nodes. After some 10 specific intervals, a particular node remains as selfish for all the time without any change then the node will be taken in consideration for the analysis of finding out whether it is malicious node or not. The neighborhood node of the particular node uses the detection mechanism to detect the misbehavior of specific node [5]. The mechanism is defined as whenever a particular node behaves abnormally the other node sent request will increase the malcount of the particular node. If the malcount of a particular node exceeds the predefined threshold value, then all the nodes of the network will be informed about the particular node.

After receiving that information all the other nodes at the network will be checks their local malcount for the broadcasted malicious node by analyzing the history and add the result to the initiator's response. All the nodes of the network will be constantly monitors the behavior of its neighbors and analyses it to detect if the neighbor has been compromised.

If neighborhood node detects the specific node as malicious, it propagates that information to throughout the network and waits for their responses. If two or more nodes has been reported about the particular node as same means then the malicious node will be isolated by other nodes. All the nodes which are using the malicious node as a route for their transmission will be in the process of discovering new routes. The detection of attacker node will be useful for avoiding future attacks.

3.5 False Alarm Detection

The false alarm will be differentiated from the overall selfishness alarm. If any alarm generated means we should verify the reason of the alarm. We should calculate the degree of selfishness again and to confirm the behavior of selfish nodes at the network. If the number of selfish nodes exceeds the threshold value means it will get confirm as overall selfishness alarm else the alarm has been raised because of the network disconnections. We should

diagnose the network disconnections by use of false detection algorithm. If it became true we should neglect the alarm with of less concern. The detection of this false alarm leads to better performance in the overall network.

The system using DSR protocol for the data transmission. The key distinguishing feature of DSR is the use of source routing. In the protocol, the sender knows the wide-ranging hop - by-hop route to the endpoint and these routes are stored in a route cache. The source route is carried by data packet in the packet header. When any node in the network tries to send a data packet to a destination for which it does not already know the route, then it follows the process of route discovery to discover the new path with dynamism.

Generally thuds protocol is composed of two processes that work together to allow the discovery and maintenance of source routes in the ad hoc network:

Route Discovery is the mechanism by which a node S wishing to send a packet to a destination node D obtains a source route to Destination. Usually the route discovery mechanism is used only when Source node attempts to send a packet to Destination and does not already know a route to Destination.

Route Maintenance is the mechanism by which Source node is trying to explore, while using a source route to the Destination, if the network topology has changed such that it can no longer use its route to the Destination, because a link along the route no longer works.

There may be any network disconnections can happen over the route. When Route Maintenance specifies and informs that the source route is broken or destroyed, the Source node which needs to send data can attempt to use any alternate route it chances to know the Destination node, or can invoke Route Discovery again to find a new route. Route Discovery and Route Maintenance each operate entirely on request. In particular, DSR does not require periodic packets of any kind at any level within the network.

Cognitive radio network is a new emerging exploration area. Cognitive radio network enhances the existing software-defined radio, whose physical layer behavior is mostly defined in software. Usually Cognitive radio has the following characteristics. Initially, it is aware of its environment and its capabilities. Next, it is able to independently alter its physical layer behavior based on its previous experience and its current atmosphere. Lastly, it is capable of carrying out the complex adaptation strategies according to the cognitive cycle shown in.With these capabilities, when spectrum environment changes around cognitive user, it is proficient of recognizing these changes and independently changing its physical layer settings.

Though, there is no existing simulator that is suitable for the demand of cognitive radio simulations. Several researchers implemented their algorithms for cognitive radios on existing network simulator such as NS-2, OPNET, and QUALNET. There is a demand to extend existing simulators to support cognitive radio simulators.NS2 is the most popular simulator to implement the concepts related to wireless networks We make use of existing NS-2 to lengthen it to support cognitive radio network simulation.

In a group of nodes Source will discover a route for the destination to reach the data packets and keep sending the data through that route. CRCN always checking the transmission range around those networks. It will find the network disconnections when the first selected paths node got move from out of transmission range. Whenever the mobility range of any node at the specific route gets higher and it leads to network disconnections. The CRCN will detect the disconnections by recognizing when the transmission range of the nodes exceeds from the specific limit of range. After disconnections occur, false alarm will be raised. All the nodes will be intimated as this alarm is a false alarm and to ignore this alarm. But the nodes at the specific route alone should be intimated as where the disconnection occurred and should search an alternative path to reach the destination.

Already a lot of disconnections and link failure will be in the network. So we have to send the false alarm through the path which having no disconnections. But anyway we will have some half incomplete messages in the way of routing nodes. After disconnections we have to inform those node to delete those incomplete messages to avoid the waste of memory space.

4. Congestion Control

In mobile ad hoc network (MANET), congestion is one of the most important restrictions that deteriorate the performance of the whole network. It is essential to adjust the data rate used by each sender in order not to overload the network, where multiple senders compete for link bandwidth. The Packets at the network may be dropped when they reach the router and cannot be forwarded. Many packets are dropped while excessive amount of packets arrive at a network bottleneck. The packets dropped would've traveled long way and in

addition the lost packets often trigger retransmissions. This intimates that even more packets are sent into the network. And so, network throughput is still more worsened by the phenomenon called network congestion. There is a high probability of congestion collapse where almost no data is delivered successfully if no appropriate congestion control is performed. [6]
Using CRCN in MANET we can reduce the congestion rate at the network. The cognitive radio senses the data flow of the network and provides the increased data rate to the network. The use of cognitive radio cognitive networks provides increase in the packet delivery ratio of the network.
The simulation shows that the performance has been increased in packet delivery ratio and reduced communication cost with the use of cognitive radio cognitive network.

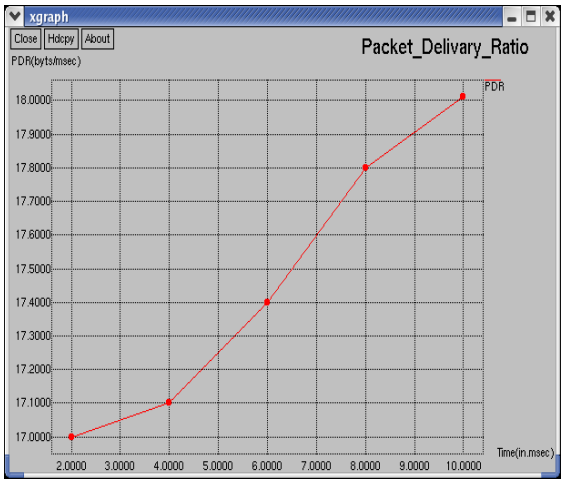


Fig 1.4 Packet Delivery Ratio

In the system we have implemented a selfish node detection method and novel replica allocation techniques to handle the selfish replica allocation. The proposed schemes are inspired by the real-world observations in economics in terms of credit risk and in human friendship management in terms of choosing one's friends completely at one's own decision. We have applied the concept of credit risk from economics to detect selfish nodes. Each and Every node in a specific network calculates credit risk information on other connected nodes individually to measure the degree of selfishness.

The system performance are extensively significant in the detection of attacker and to provide congestion control at MANET. Extensive simulation shows that the proposed strategies outperform the cooperative replica allocation techniques in terms of data accessibility, communication cost, and query delay. The False alarm at selfishness will decrease the data flow of the network. By using our technique we will pass the information as it is not by selfishness. So no significant change will occur except choosing for alternative routes. As a part future, we plan to consider all the replication strategies and network disconnections suited for various consistency level and with increase in security against various attacks. Our next goal will be to conduct an analytical study of the impact of node mobility on network performance with misbehaving nodes. We plan then to design and evaluate a collaborative security scheme that solves the selfishness problem, analyzing the effects of such mechanism on network throughput and communication delay.

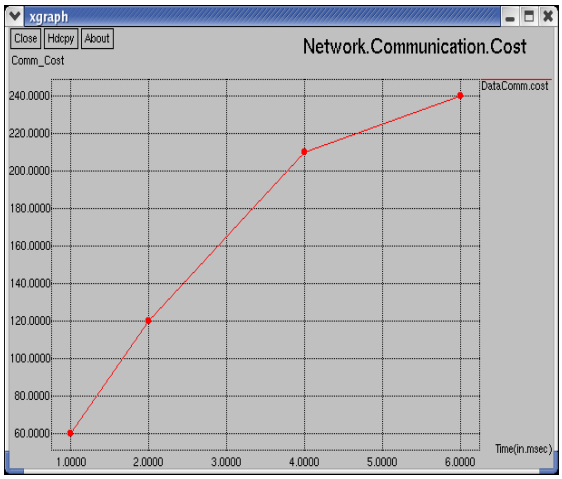


Fig 1.5 Network Communication Cost

5 Bibliography

[1]Jae-Ho Choi, Kyu-Sun Shim, SangKeun Lee, and Kun-Lung Wu"Handling Selfishness in Replica Allocation over a Mobile Ad Hoc Network" IEEE Transactions on mobile computing, vol. 11, no. 2, February 2012.
[2]Priyanka goyal,Rahul rishi, vinti parmar "MANET: Vulnerabilities, Challenges, Attacks, Application" IJCEM International Journal of Computational Engineering & Management, Vol. 11, January 2011
[3]Mohit Kumar and Rashmi Mishra "An Overview of MANET: History,Challenges and Applications" Indian Journal of Computer Science and Engineering (IJCSE).
[4]Sonali Bhargava and Dharma P. Agrawal " Security Enhancements in AODV protocol forWireless Ad Hoc Networks"2004
[5]C. Lochert, B. Scheuermann, M. Mauve, "A Survey on Congestion Control for Mobile Ad-Hoc Networks", Wiley Wireless Communications and Mobile Computing 7 (5), pp. 655-676, June 2007
[6]T. Hara, "Effective Replica Allocation in Ad Hoc Networks for Improving Data Accessibility," Proc. IEEE INFOCOM, pp. 1568- 1576, 2001.
[7]L. Anderegg and S. Eidenbenz, "Ad Hoc-VCG: A Truthful and Cost-Efficient Routing Protocol for Mobile Ad Hoc Networks with Selfish Agents," Proc. ACM MobiCom, pp. 245-259, 2003.

[8]K. Balakrishnan, J. Deng, and P.K. Varshney, "TWOACK: Preventing Selfishness in Mobile Ad Hoc Networks," Proc. IEEE Wireless Comm. and Networking, pp. 2137-2142, 2005.
[9]H. Li and M. Singhal, "Trust Management in DistributedSystems," Computer, vol. 40, no. 2, pp. 45-53, Feb. 2007.12. M. Li, W.-C. Lee, and A. Sivasubramaniam, "Efficient Peer-to-PeerInformation Sharing over Mobile Ad Hoc Networks," Proc. WorldWide Web (WWW) Workshop Emerging Applications for Wireless andMobile Access, pp. 2-6, 2004
[10]S. Marti, T. Giuli, K. Lai, and M. Baker, "Mitigating Routing Misbehavior in Mobile Ad hoc Networks," Proc. ACM MobiCom,pp. 255-265, 2000.
[11]K. Paul and D. Westhoff, "Context Aware Detection of Selfish Nodes in DSR Based Ad-Hoc Networks," Proc. IEEE Global Telecomm. Conf., pp. 178-182, 2002.
[12]J. Zhai, Q. Li, and X. Li, "Data Caching in Selfish Manets," Proc.Int'l Conf. Computer Network and Mobile Computing, pp. 208-217,2005.
[13]T. Hara and S.K. Madria, "Consistency Management Strategies for Data Replication in Mobile Ad Hoc Networks," IEEE Trans. Mobile Computing, vol. 8, no. 7, pp. 950-967, July 2009.
[14]S.-Y. Wu and Y.-T. Chang, "A User-Centered Approach to Active Replica Management in Mobile Environments," IEEE Trans. Mobile Computing, vol. 5, no. 11, pp. 1606-1619, Nov. 2006.
[15]N. Laoutaris, G. Smaragdakis, A. Bestavros, I. Matta, and I. Stavrakakis, "Distributed Selfish Caching," IEEE Trans. Parallel and Distributed Systems, vol. 18, no. 10, pp. 1361-1376, Oct. 2007.
[16]P. Michiardi and R. Molva, "Simulation-Based Analysis of Security Exposures in Mobile Ad Hoc Networks," Proc. European Wireless Conf., pp. 1-6, 2002.

I.Shanthi acquired her B.Tech degree in Information Technology under Anna University at Arunai engineering college in 2011 and currently pursuing M.E Degree in computer science and engineering under Anna University Chennai at Rajalakshmi engineering college, Chennai. Her area of interest includes networks and image processing and analysis.

Mrs.D.Sorna Shanthi has a teaching experience spanning over 9 years in the field of computer science and engineering. She acquired her B.E degree in Kamaraj engineering college at virudhunagar and completed her M.Tech degree in Sathyabama University in specialization of computer science. She is presently working at Rajalakshmi engineering college as Assistant Professor. She was formerly in valliamai engineering college. Her area of interest includes networks.

Modeling of Multipath Transport

Chang Liu[1], Fei Song[2], Huan Yan[3], Sidong Zhang[4]

National Engineering Laboratory for Next Generation Internet Interconnection Devices
Beijing Jiaotong University, Beijing, China

Abstract

In this paper, we propose a model for evaluating the transmission performance of multipath transport. Previous researches focused exclusively on single pair users in simple scenarios. The distinct perspective in this paper is to build models for analyzing the performance when multipath transport is used in the entire network scope. We illustrate the influences on the transmission performance caused by the variation of network topologies, the services' arrival rate, the services' size and other parameters. We demonstrate through simulation that multipath transport could conditionally increase the throughput than single-path transport. And it has the capability to support higher services' arrival rate in various network topologies. And higher multi-parent probability will be beneficial for multipath transport to take its advantages.

Keywords: *multipath transport, service model, load balance, throughput gain*

1. Introduction

Multi-interface (3G, WLAN, WMAN, etc.) terminals which enable users to not only have access to services anywhere anytime from any network, but also through several interfaces, are increasing in user numbers, and can be expected to be the most common type of Internet device in the near future. Such trend has induced the emergence of the idea of 'multipath transport' which refers to the method of sending data over multiple available paths simultaneously. Multipath transport enables network resources to be used concurrently, and improves user experience. There are two key benefits of multipath transport. One is that the resilience of connectivity is strong with multiple paths, and the other is that it increases the efficiency of resource usage, and thus increases the network capacity available to end hosts [1]. As to the transport protocols at the transport layer, multipath functions are not new. Concurrent Multipath Transfer (CMT) [3-5] is a kind of multipath transport supported by the Stream Control Transmission Protocol (SCTP) [2]; Multipath TCP (MPTCP) [1,6] is a modified version of Transmission Control Protocol (TCP), whose original goal was to support multipath transport. Many subsequent researches on path selection [7-10], load sharing [11-14], retransmission judgment [15-16], throughput estimation [17-20], receiver buffer size [5,21], have been conducted to improve transmission performance of these two protocols.

However, the majority of the current works on multipath transport focus exclusively on the performance of single pair users, leaving out the consideration for the entire network topology and the possibility of multiple pairs of users. The simulations in these works were often set up based on environments con two terminals, and the two terminals accessed network though multipath. They paid more attentions on the performances between the single pair users, which is not close to the actual scenarios.

The motivation of this paper is to discuss the performance of multipath transport when it is used in the entire network by multiple pairs of users instead of just single pair. An analytic model was proposed in order to achieve this object. And we also analyze the multipath transport performances in different kinds of network topology based on analytic model.

In this paper, we pay more attentions to the performances of the entire network. Different from other researches, we analyze what will happen in the network if all of the endpoints use multipath transport, and make a comparison with all of them use single-path transport. First of all, we have made a topological model to construct the network topology, and made a services model to simulate the arrival services. With these two models, we have demonstrated that multipath transport could obtain higher throughput than single-path transport, especially when multipath transport is used in the entire network. And we have also demonstrated that multipath transport could

support higher services' arrival rate than single-path transport. Secondly, with changing the network topology of simulations, we have demonstrated that higher multi-parent probability will be beneficial for multipath transport to take its advantages.

2. Multipath Transport and Single-path Transport

2.1 Single-Path Transport

At transport layer, TCP is a common protocol which provides reliable, ordered delivery. Due to the basic design principle, the original TCP does not support multihoming terminals, so it only supports single-path transport. Major Internet applications such as the World Wide Web, email, remote administration and file transfer rely on TCP, and single-path transport is widely used in current Internet. It can be said that the current Internet is based on this kind of single-path transport applications.

SCTP is another transport protocol which provides reliable, ordered delivery just as TCP. One key different between SCTP and TCP is SCTP could support multihoming terminals. This feature makes SCTP suits for modern network better than TCP. However, although SCTP supports multihoming terminals, the original SCTP still support single-path transport only. It could build more than one path among two terminals, but it only allows using one path to transmit data, and the other paths are backup paths. So in essence, the original SCTP is still using single-path transport.

2.2 Multipath Transport

Recently, the user requirements keep increasing rapidly, especially reflected in bandwidth. Single-path transport is hard to satisfy these situations, and it needs several single paths to combine together to provide better services. So multipath transport is presented.

Multipath TCP (MPTCP) [1] is a modified version of TCP. It adopts the idea of subflows to realize multipath transport, different subflows are sent to different paths. The mechanism of each subflow is just likes original TCP's. This research direction has been accepted by IETF [1].

Concurrent Multipath Transfer (CMT) [3-5] is based on SCTP. With the feature of supporting multihoming in SCTP, CMT builds multipath among single pair, and uses these path to transmit data concurrently, thus it realizes multipath transport.

Both of them are hot topics. There are many researches related with multipath transport (MPTCP and CMT), such as the problem of out of order packets, lost packet and retransmission judgments, receiver buffer size, load balance among multipath, etc.. These researches have already demonstrated that multipath transport can provide better end to end services than single-path transport could.

However, multipath transport means it will seize more network resources. Although multipath transport can improve end to end services, if each user uses multipath transport, it is difficult to predict the transmission effect. So before multipath transport is used widely in the Internet, it needs to do many analyses of the entire network with multiple pair multipath transport users.

2.3 Load Sharing in Multipath Transport

The load sharing is one of the main questions in multipath transport. In multipath transport, there are several available paths we could use, how to use them, how many data will assign to them is a real question. So many researches are about this question, but they have different emphases, such as path selection [7-9] (it is about how each packet selects its sending path), data distribution [10-11], load balance [12-14], etc.. Most of those researches have a consensus, the wider path should send more data, and this also coincidences the common understanding.

Therefore, in our paper we assume that multipath transport assigns data to each path based on the ratio of real-time bandwidths.

3. Construction of Network Topology

3.1 Node Classification

Here, we propose a classification method for the nodes in the current Internet structure. We divide the nodes into different levels based on each's switching performance. For instance, a 10Gb/s node can be classified into core level, and a 10Mb/s node can be classified into leaf level. Nodes at different levels also have different topological properties. Higher level nodes often have higher Connectivity Degree, which means that they have more connections.

For convenience of demonstration, we create a three-level network structure. Note that in reality, the network structure created using our method is not limited to three levels. In the structure, the nodes at the lowest level are leaf nodes, while the rest are core nodes.

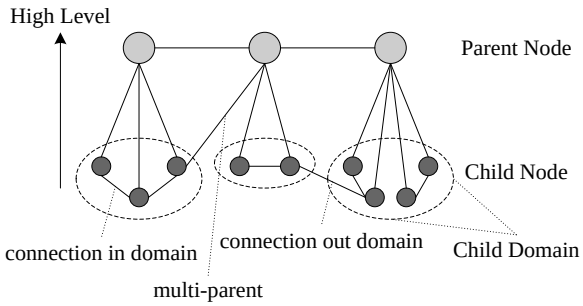


Fig.1 Different level nodes in networks

Nodes are also classified as parent/child in our structure. The definition and characteristics of the parent/child nodes are as follows:

- A parent node connects to its child nodes, and operates at the level atop them. A child node that is not at the bottom level of the network structure can also be the parent of its own children.
- The number of child nodes connected to a parent node is random, and we assume this number obeys to the uniform distribution. μ_l^{NC} is used to express the mean value of the number at the level l. This parameter is majorly determined by the scale of the network.

The relationship between parent nodes and child nodes is shown in Fig.1.

3.2 Domain

A domain is defined as the set of all the child nodes connected to a parent node. One example of a domain on the Internet would be a subnet. The set of the terminals that are using the same network interface can also be considered as a domain. For instance, the computers using wired access mode are in a different domain from the ones using wireless access mode, although they might be close in distance. In this paper, the probability of a connection between any two nodes depends on the domains they are in. The features of a domain are as following:

- The value of the probability of a connection between any two nodes within the same domain is based on their network level. P_l^{in} is used to express this probability at the level l.
- Similarly, P_l^{out} is used to express the probability of a connection between any two nodes that are at the same level l, but in different domains.
- In our network structure, the more distance there is between any two nodes in different domains, the smaller chance there is that they will directly connect. We propose a threshold value D_{thrsh}, so that the probability of a connection between them P_l^{out} will become invalid when the distance between any two nodes falls below this value.
- In general P_l^{in} of any node is greater than its P_l^{out}. These two parameters are directly influenced by the topology of the network. For instance, a mesh network will see a higher probability of a connection between the nodes than a star network or a tree network.

3.3 Multi-parent

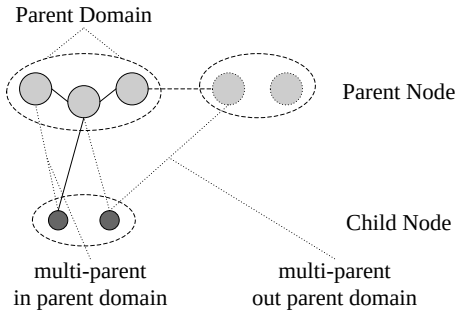


Fig.2 Multi-parent probability

It is also possible that a node connects with multi-parent nodes, especially when it is at high level. In the past, a terminal usually had only one network interface, or was only able to access the network through one interface at a time, and thus the number of multi-parent nodes was small. However, as multi-interface terminals have become widely used in the past decade, multi-parent nodes are now expected to become the major network nodes in the near future.

Here, we use P_l^{MP} to express the overall probability of a node at the level l connecting to multiple parent nodes at the level $l-1$.

Still, it is difficult for a node to connect with multiple parents that are too distant from each other at the same time. Therefore, we will only consider two types of multi-parent connection, as shown in Fig.2:

- For a child node, its first parent node is defined as its original parent node, and it is set during the creation of the node. Then we define the domain of its original parent node to be its parent domain. Any node within this domain will have a probability P_l^{MPin} to be its parent node.
- A child node can only connect with a new parent outside of its parent domain, when the outside parent to be is directly connected with the original parent node. When this condition is fulfilled, this probability is valid and is expressed as P_l^{MPout}.

a) Make random number of Level 1nodes, average number = μ_1^{NC};

b) for each Level 1 node:
  b.1) make random number of Level 2 nodes and connect with the parent node, average number = μ_2^{NC} ;
  b.2) for each Level 2 node:
    b.2.1) has a probability of P_2^{MPin} to connect with another Level 1 parent node;
  b.3) in the domain, two Level 2 nodes have a probability of P_2^{in} to be connected;

c) outside the domain, two Level 2 nodes have a probability of P_2^{out} to be connected;

d) for each Level 2 node:
  d.1) make random number of Level 3 nodes and connect with the parent node,, average number = μ_3^{NC} ;
  d.2) for each Level 3 node:
    d.2.1) has a probability of P_3^{MPin} to connect with another Level 2 parent node which is in its parental domain;
    d.2.2) has a probability of P_3^{MPout} to connect with another Level 2 node which is outside its parental domain;
  d.3) in the domain, two Level 3 nodes have a probability of P_3^{in} to be connected;

e) outside the domain, two Level 3 nodes have a probability of P_3^{out} to be connected.

Fig.3 Pseudocodes for network topology construction

3.4 Parameters of the Network Structure

Summing up the above, there are five major parameters in our network structure: μ_l^{NC} , P_l^{in} , P_l^{out} , P_l^{MPin} and P_l^{MPout} . By varying the parameters, we could obtain and run simulations for different network topologies.

3.5 Process of Constructing Network Topology

Basing on the above analyses, we can now construct network topologies using the steps shown as pseudocodes in Fig.3.

4. Modeling for Service Transmission

Internet network services occupy network resources and create data flows from one terminal to another through a series of routers. In this paper, we built a model for such data flows in order to analyze their transmission process.

We assume that a service is provided by a leaf node, transmitted through a series of nodes, and in the end received by another leaf node. This assumption is based on the fact that in our model, higher level nodes are the abstractions of the real-world routers, and the leaf nodes represent real-world terminals. We refer to the routers (higher level nodes) a particular data flow has travelled through as a path. According to the common routing rules, each node has a specific transmission capability, which reflects the node's maximum forwarding rate. It's not hard to infer that a path's bandwidth is the minimum of all the nodes' bandwidth along this path. All the data flow passing through a node will occupy its transmission capability averagely. We define a term 'services' size', which stands for the total size of the data packets a service creates. If a service has finished transmitting the data, it will release the network resources. Also, a random number of new services will arrive at a certain rate.

Based on the above model, simulations could be carried out both for single-path transport and multipath transport scenarios to examine their transmission performance and for further comparison.

4.1 Network Services

In our model, a number of new services will emerge on a fixed time basis. According to the classical queue model, we assume that the services' arrival rate obeys Poisson distribution P(λ), as shown in Formula (1). Adjustments to the services' arrival rate can be achieved by varying the parameter λ.

$$\text{Number of new services} \sim P(\lambda) \quad (1)$$

4.2 Service Paths

In our model, we will adapt existing routing algorithms, such as OSPF, RIP, etc which will select a path with the minimum hop count, for both single-path and multipath transport.

In Fig.4, the arrows indicate the direction of the data flow. And the single path is the lines with the single arrow, which indicates the minimum hop count from the source to the destination. For the purpose of demonstration, we adopted a two-path transmission model. Since multipath transport is made possible by terminals accessing the network through different interfaces, the first and the last hops in our model should be separate. However, it is possible that the two paths cross each other at a certain node inside the network.

In our model, multipath transport finds its two paths by following steps:

- Find the path with the minimum hop count from the source to the destination, and mark it as Path 1;
- Mask the first and the last node in Path 1 temporarily;
- In the new topology, repeat Step 1, then mark the resulting path as Path 2, which will have different first and last hop from Path 1;
- If Path 2 cannot be found, multipath transport will not be carried out;
- Unmask the first and the last node in Path 1.

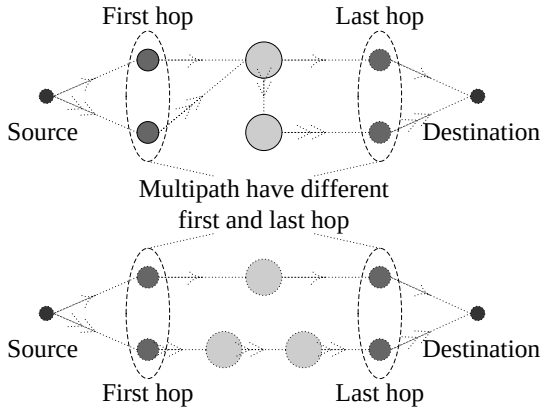


Fig.4 Two examples of multipath transport

Fig.4 shows the resulting two paths. One is marked by a series of single arrows, and the other by double arrows. Note that these two paths have separate first hops, and separate last hops, too. In each path, there will be an independent flow for the specified service.

4.3 Transmission Capability of The Nodes

In actual networks, routers receive packets and forward them based on its routing algorithms, and their forwarding rate can suffice as an abstracted representation of a router in a mathematical model because it is usually considered to be the most fundamental property. In our particular model, this abstracted property will be called transmission capability of a node, and nodes at a higher level will have greater transmission capability due to our node classification method. Also, transmission capability will be the same for the node that are at the same level.

According to 4.2, there will be multiple data flows coming from different services passing through one node. We assume all services are equal users of a node's transmission capability splitting its resources evenly. For this reason, we introduce the concept of a node's bandwidth, which stands for the value of transmission capability a service occupies at a certain node. Its mathematical manifestation is shown in Formula (2).

$$b_i^{node} = \frac{C_i^{node}}{Num_i^{node\_flow}} \quad (2)$$

In formula (2), i is the serial number of a certain node; b_i^{node} is the node's bandwidth of Node i; C_i^{node} is the transmission capability of Node i, and $Num_i^{node\_flow}$ is the number of data flows passing through Node i.

4.4 Transmission Throughput of A Service

The transmission throughput of a service's data will be equal to the lowest node bandwidth along the transmission path, as Formula (3) shows. b_p^{path} is the bandwidth of Path p, and b_i^{node} is the bandwidth of Node i.

$$b_p^{path} = \underset{i \in Path_p}{Min}(b_i^{node}) \quad (3)$$

It is not difficult to see that, for single-path transport, the transmission throughput of a service's data is equal to the value of its path's bandwidth.

For the multipath transport in our model, we will discard the interaction between the two paths because they can be reduced using many available methods, such as using optimized transport protocol, enhancing the receiver buffer's capacity, etc. Therefore, in our model, the transmission throughput of a service's data with multiples paths can be expressed as the sum of all paths' bandwidth, as Formula (4) shows.

$$TP = \sum_p b_p^{path} \quad (4)$$

4.5 Size of A Service

The size of a service here refers to the overall size of all the data the service needs to transmit. In our model, we assume that this property obeys Uniform Distribution, as Formula (5) shows. $S_k^{t_0}$ is the initial size of a certain service k; μ_S is the mean value of all the data sizes a service could produce .

$$S_k^{t_0} \sim U(0.5\mu_S, 1.5\mu_S) \quad (5)$$

We have also considered alternatives of data size distribution pattern, e.g. Exponential Distribution and Constant Value Distribution when running the simulations, but the results were very similar. Therefore, we only demonstrate through the uniform distribution in this paper.

When all the data of a certain service has reached its destination, this service will terminate itself and release the path.

4.6 Final Model for Service Transmission

In each time unit, a service will transmit a bulk of data, whose size is equal to the value of the transmission throughput of this service from the origin to the destination, and the size of the residual data will decrease correspondently until it has reached 0, meaning this service is done and is closing itself.

We will use S_k to express the size of the residual data of Service k, and thus the iterative formula could be expressed as Formula (6). S_k^{t+1} and S_k^t are the sizes of the residual data at the time $t+1$ and the time t; TP_k^t is the transmission throughput at the time t.

$$S_k^{t+1} = S_k^t - TP_k^t \quad (6)$$

Finally, with the formulas from this chapter, we were able to build a mathematical model for network service transmission.

$$\begin{cases} S_k^{t_0} \sim \mathrm{U}(0.5\mu_S, 1.5\mu_S) \\ S_k^{t+1} = S_k^t - \sum_j \underset{m \in \mathrm{Path}_j}{\mathrm{Min}} \left(\frac{C_m^{t\,\mathrm{node}}}{\mathrm{node\_flow}} \right) \\ \qquad\qquad\qquad\qquad Num_m^t \end{cases} \quad (7)$$

4.7 Service Modeling Processes

We summarized the processes for building this model, as Fig.5 shows.

For each unit time:

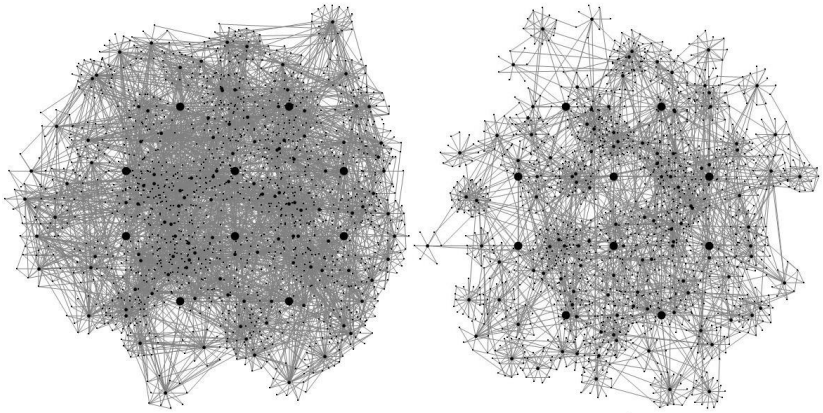
a) create random number of new services which obeys P(λ);

b) for each new service:
  b.1) source and destination are choosing from Level 3 nodes randomly;
  b.2) find 1 or 2 shortest path for the service; (for single-path transport or multipath transport)
  b.3) add the number of flows for each node which is in the shortest path;

c) for each service:
  c.1) calculate the bandwidth of its path;
  c.2) calculate the throughput of service at this unit time;
  c.3) decrease service size due to its throughput;
  c.4) if service size ≤ 0, delete the service and the flows in its paths;

d) end one unit time.

Fig.5 Pseudocodes for Service Transmission Modeling



(a) **(b)**

Fig.6 Network Topologies for Simulations

Table 1 Parameters of the Network Topologies for Simulations

Level of the Nodes	Number of Child Nodes	Probability of A Connection		Probability of Multi-parent Scenarios	
	μ_l^{NC}	P_l^{in}	P_l^{out}	P_l^{MPin}	P_l^{MPout}
Level 1	10	100%	-	-	-
Level 2	10	20%	5% (range 30)	5%	-
Level 3	10	10%	2% (range 30)	10%	5%

5. Simulations and Results

This chapter discusses the network simulations we ran and the corresponding results we obtained. 5.1 talks about the construction of network topologies in the simulation environment according to the node classification methods in Chapter 3. 5.2 compares the single-path transport and the multipath transport in many aspects under the topologies we constructed. 5.3 discusses the effects of varying network topology on the service's models.

5.1 Constructing the Network Topology

According to Chapter 3, we constructed the network topologies for simulation network with 3 levels of nodes, where Level 1 is the highest and consists only of the core nodes, and Level 3 is the lowest and consists only of the leaf nodes. The parameters for constructing these network topologies were set according to Table 1.

Fig.6(b) demonstrates a topology with a low multi-parent probability, which was obtained by multiplying the multi-parent probability of all the nodes by a factor of 0.2. This topology is very similar to the actual network topologies of today, because, though having multiple network interfaces, current terminal devices usually solely use one interface to access the networks. On the other hand, Fig.6(a), which was created using larger multi-parent probability, will better model a topology where terminals accessing the network through multiple interfaces simultaneously.

5.2 Simulation for Single-path Transport and Multipath Transport

In the simulations, the unit of time was set as 1 time unit, and the unit of services' size as 1 size unit, and therefore, the transmission throughput is 1 size unit per time unit. All simulation time length is 1000 time units, and all results are average values of the properties tested in these 1000 unit times. Using Mathematic 8.0, we ran simulations of SP (Single-path Transport) and MP (Multipath Transport) for the topologies we built.

A) Service Transmission Throughput

In our simulations, all services were set to have certain throughputs (Part 4.4), and we could get the average value of service throughputs for each unit time, which is determined by the services' arrival rate λ (Part 4.1).

Fig.7 shows the service throughput at each time step, under λ=300 and λ=340, respectively. We concluded from the results that:

- Under λ=300 (Fig.7(a)), the network is steady. The service's throughput varies but is bounded. This phenomenon distinguishes from that under λ=340 (Fig.7(b)), where the SP transport failed to keep the network state steady, which is demonstrated through a graduate decline in the services' throughput. The mechanism for the decline is that a λ as high as 340 is 'unbearable' to this network topology with SP transport. The services' arrival rate becomes higher than the services' finishing rate, resulting in a traffic situation. However, with MP transport, even under a λ as high as 340, the throughput managed to stay steady.
- It can be observed that the services' throughput keeps decreasing until reaching the steady range. This is because the network is set to start with no services at all, and it takes a certain period of time for it to go from its initial state to the steady state. This transient state takes about 20~50 time units, and depends on a number of factors such as the service's size. Further analyses of the factors will be made later in this paper.
- The diagrams show that in this particular network topology, MP transport yields higher service throughput than SP transport. Detailed analyses will be made later in this paper.
- It can also be inferred that, with the same topology and the same λ, MP transport tend to accommodate higher services' arrival rate than SP transport. Detailed analyses will be made in later chapters.

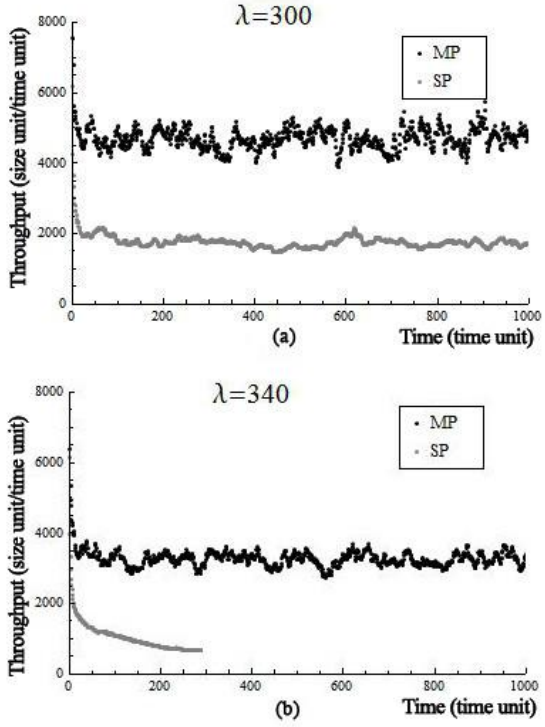


Fig.7 Real-time services' throughputs

In Fig.8, we demonstrate the relationship between the services' throughput and the services' arrival rate. It can be observed that the MP throughput is almost twice as large as the SP throughput at the same arrival rate in our specific topology. We believe that this can further support our inference that in the same network topology and with the same λ, using MP transport tends to result in higher service throughput than using SP transport.

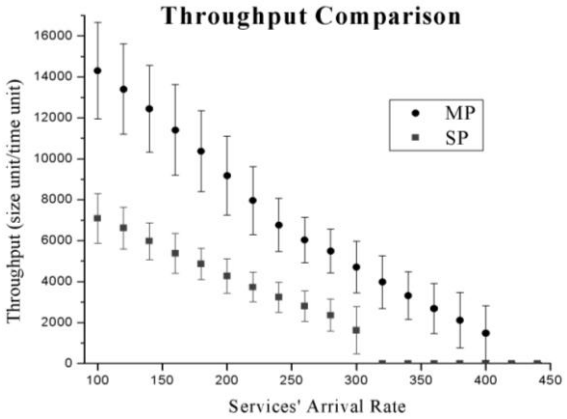


Fig.8 The Services' Throughput With the Services' Arrival Rates

B) Services' Size

Fig.9 demonstrates the influence of the services' size μ_S on the services' arrival rate in our model. It can be inferred from Fig.9 that MP tends to support higher services' arrival rate than SP. Further analysis reveals that the value of the product $\lambda_{max} \bullet \mu_S$ remains constant. For MP, $\lambda_{max} \bullet \mu_S \approx 4150$, whereas for SP, $\lambda_{max} \bullet \mu_S \approx 3050$. This constant in fact reflects the overall transmission capacity of a network topology under a certain type of transport method. In our network topology, MP increases this capacity by about 35% compared to that of the SP. Note that in later chapters of this paper, the simulations were done with the services' size 10.

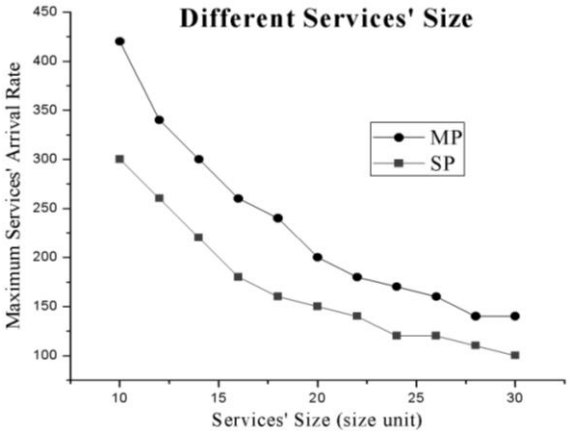


Fig.9 Different Services' Size

C) MP Transport In Single Pair and Entire Network

As Chapter 1 stated, most previous researches focused solely on testing their transport strategies through simulations that were set in a network with only a single pair users, which is not representative of actual networks. Such researches were often conducted in a multivariate analysis fashion, where the network topologies and the service models were held constant, with the transport strategy as the only variable, in order to demonstrate the resulting superior network performance. Here, we adapted a similar fashion in which we experimented with our MP transport strategy, but what is different is that here, it is both the network topology and the transport strategy that are the variables. We applied both the MP and the SP transport strategies respectively to a simple single pair ends, and later to the entire network. The simulation results we obtained are shown in Fig.10.

Fig.10 shows the Throughput Gain (TG), which refers to the ratio of MP throughput to the SP throughput when applied to the same network topology, with respect to the services' arrival rate. This parameter measures the improvements the MP brought compared to the SP. It can be observed that, When SP and MP was applied to a single pair users, the TG reached as high as 1.8, but still, it fails to grow over 2.0. This is due to the fact that our MP strategy, which introduces only one additional path, can increase the overall throughput by 100% at maximum. The fact that the TG was always above 1.0 is enough to demonstrate that the MP is superior to SP in a simple two-end network. This result stays true for the more complex network as well. However, When SP and MP was applied to the entire network, it can be observed that as the services' arrival rate grows, the TG grows over 2.0, and at one point reached as high as 3.0. This is because when MP is applied, the data flow within the entire network can be balanced, enhancing the bandwidth along each path.

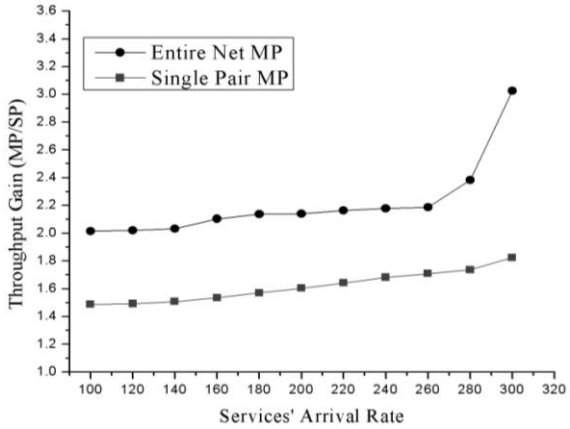


Fig.10 TG in Single Pair Users and Entire Network

5.3 Effect of Variations in the Parameters of the Network Topologies

A) Effects of Variation in the Multi-parent Probabilities

Table 2 Multi-parent Probabilities Variation Range

Level	P_l^{MPin}	P_l^{MPout}
Level 2	4%~20%	-
Level 3	4%~20% (P_2^{MPin})	2%~10% ($P_2^{MPin}/2$)

Multi-parent probabilities P_l^{MPin} and P_l^{MPout} were introduced in 3.3. A low multi-parent probability indicates that an endpoint is less likely to use multiple interfaces to access the network, so it will be hard to find the second path between the source and destination. The aim of this part is to measure the impact of multi-parent probabilities. The variation range we set for the multi-parent probabilities is shown in Table 2.

Fig.11 shows the effects of multi-parent probability P_l^{MPin}. From our observation, it can be inferred that the higher the multi-parent probabilities grow, the higher the TG gets, because when the value of P_l^{MPin} is small, the superiority of MP are seriously compromised as the endpoints find it hard to connect through multiple paths. However, the TG reached a plateau (about 3.276) when P_2^{MPin} hits 10%.

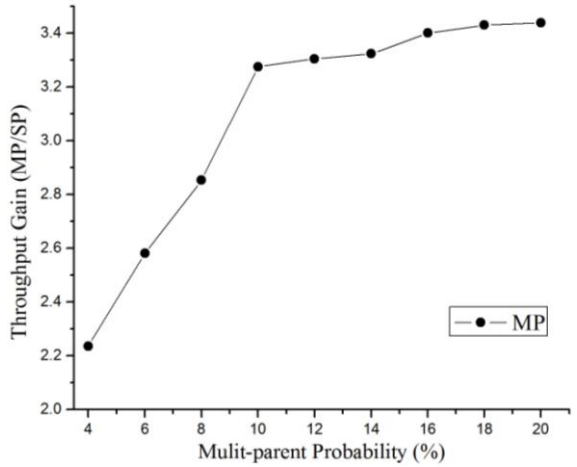


Fig.11 TG With Respect To P_2^{MPin}

B) Effects of Variation in the Number of the Child Nodes

The quantity of the child nodes μ_l^{NC} directly influences the scale of the network and the loading pressure of higher level nodes, as 3.1 stated. Here, we aimed to examine the effects of μ_l^{NC}. The variation range of μ_l^{NC} is set to be 6~16 (under the basic network topology, with μ_l^{NC} =10).

It can be observed in Fig.12 that the larger the child quantity gets, the higher the services' arrival rate for both MP and SP transport there is, bringing a higher transmission capability. The MP, however, is still superior to SP at all times. The relationship between the child quantity and the maximum services' arrival rate is close to being linear.

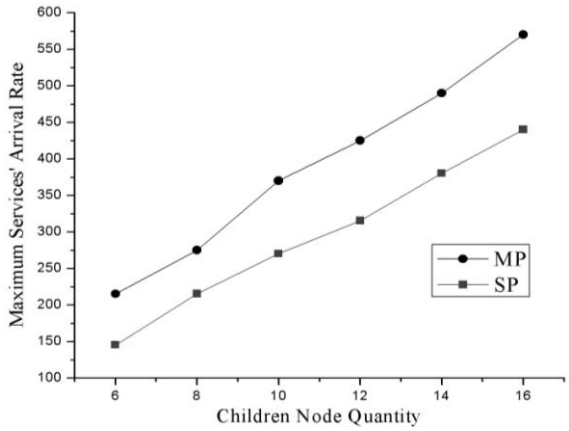


Fig.12 Maximum Services' Arrival Rate With Respect to the Child Node Quantity

6. Conclusions

In this paper, we proposed a topological model to construct the network topology, and created a services model to simulate arrival services. The simulation results demonstrate that, 1) MP could obtain higher throughput than SP. Moreover, if MP is used in the entire network, the improvement of throughput will be more remarkable, for the throughput gain of entire net MP is about 2.1, but the gain of single-pair MP is only about 1.5. 2) MP could also support higher services' arrival rate than SP, and the improvement is about 30%. 3) Services' size will affect the maximum services' arrival rate, but $\lambda_{max} \bullet \mu_S$ will remain a constant value which reflects the overall transmission capacity of a network topology under a certain type of transport method, and MP could increase this capacity by about 35% compared to SP. 4) When multi-parent probability is lower than 10%, there is a linear relationship between multi-parent probability and throughput gain, so higher multi-parent probability will be beneficial for MP to take its advantages. 5) Larger network size will support higher services' arrival rate, and there is also a linear relationship between network size and maximum services' arrival rate, for both MP and SP.

Acknowledgments

This paper is supported in part by the National High-Tech Research and Development Program of China (863) under Contract No. 2011AA01A101, in part by the National Natural Science Foundation of China (NSFC) under Contract No. 61232017, No. 61102049 and No. 61100217, in part by Ph.D. Programs Foundation of Ministry of Education of China under Contract No. 20120009120005.

References

[1] A. Ford, C. Raiciu, M. Handley and et al. "Architectural Guidelines for Multipath TCP Development". RFC6182. 2011.
[2] R. Stewart, Q. Xie, K. Morneault. "Stream Control Transmission Protocol". RFC4960. 2007.
[3] J.R. Iyengar, P.D. Amer, R. Stewart. "Concurrent Multipath Transfer using SCTP Multihoming over Independent End-to-End Paths". Networking. 2006, pp.951-964.
[4] F. Perotto, C. Casetti, G. Galante. "SCTP-based Transport Protocols for Concurrent Multipath Transfer". Wireless Communications and Networking Conference. 2007. pp.2971-2976.
[5] J.R. Iyengar, P.D. Amer, R. Stewart. "Performance Implications of a Bounded Receive Buffer in Concurrent Multipath Transfer". Computer Communications. 2007. pp.818-829.
[6] S. Barré, C. Paasch, O. Bonaventure, "MultiPath TCP: From Theory to Practice", Lecture Notes in Computer Science, 2011, pp. 444-457.
[7] H. Han, S. Shakkottai, C.V. Hollot, R. Srikant, D. Towsley. "Multi-path TCP: a Joint Congestion Control and Routing Scheme to Exploit Path Diversity in the Internet". Networking, 2006. pp.1260-1271.
[8] C. Liu, S.D. Zhang, H. Yan, H.K. Zhang. "Self-Adaptive Path Selection Scheme in Concurrent Multipath Transfer". International Conference on Broadband Network & Multimedia Technology. 2010. pp.8-13.
[9] J.X. Liao, J.Y. Wang, T.H.g Li, X.M. Zhu, "Introducing Multipath Selection for Concurrent Multipath Transfer in the Future Internet", Computer Networks, 2011, 55(4), pp. 1024-1035.
[10] S. Shailendra, R. Bhattacharjee, S.K. Bose. "Optimized Flow Division Modeling for Multi-Path Transport". India Conference. 2010. pp.1-4.
[11] Y. Hasegawa, I. Yamaguchi, T. Hama, H. Shimonishi, T. Murase. "Improved Data Distribution for Multipath TCP Communication". GLOBECOM. 2005. pp.271-275.
[12] C. Casetti, G. Galante, R. Greco. "Load Balancing over Multipaths using Bandwidth-Aware Source Scheduling", Wireless Personal Multimedia Communications. 2004. pp.1551–1556.
[13] M. Fiore, C. Casetti. "An Adaptive Transport Protocol for Balanced Multihoming of Real-Time Traffic". GLOBECOM. 2005. pp.523-530.
[14] S.C. Nguyen, X.F. Zhang, T.M.T. Nguyen, G. Pujolle, "Evaluation of Throughput Optimization and Load Sharing of Multipath TCP in Heterogeneous Networks", Wireless and Optical Communications Networks (WOCN), 2011, pp. 1-5.
[15] J.R. Iyengar, P.D. Amer, R. Stewart. "Retransmission Policies for Concurrent Multipath Transfer using SCTP Multihoming". Networks. 2004. pp.713-719.
[16] P. Natarajan, J. Iyengar, P. Amer, R. Stewart. "Concurrent Multipath Transfer using Transport Layer Multihoming: Performance under Network Failures". Military Communications Conference. 2006. pp.1-7.
[17] J. Iyengar, P. Amer, R. Stewart. "Concurrent Multipath Transfer using Transport Layer Multihoming: Performance under Varying Bandwidth Proportions". Military Communications Conference. 2004. pp.238-244.
[18] K.K. Lam, J.M. Chapin, V.W.S. Chan. "Performance Analysis and Optimization of Multipath TCP". Wireless Communications and Networking Conference. 2011. pp.695-700.
[19] S.C. Nguyen, X.F. Zhang, T.M.T. Nguyen, G. Pujolle. "Evaluation of Throughput Optimization and Load Sharing of Multipath TCP in Heterogeneous Networks". Wireless and Optical Communications Networks. 2011. pp.1-5.
[20] S. Shailendra, R. Bhattacharjee, S.K. Bose, "Improving Congestion Control for Concurrent Multipath Transfer Through Bandwidth Estimation based Resource Pooling", Information, Communications and Signal Processing (ICICS), 2011, pp. 1-5.
[21] J.R. Iyengar, P.D. Amer, R. Stewart. "Receive Buffer Blocking in Concurrent Multipath Transfer". GLOBECOM. 2005. pp.121-126.
[22] M. Handley, S. Floyd, J. Padhye and et al. "TCP Friendly Rate Control (TFRC): Protocol Specification". RFC3448. 2003.
[23] F. Song, H.K. Zhang, S.D. Zhang, F.M.V. Ramos, J. Crowcroft. "Relative Delay Estimator for SCTP-Based Concurrent Multipath Transfer". GLOBECOM. 2010. pp.1-6.

Chang Liu is a full-time Ph.D. candidate in School of Electronic and Information Engineering, National Engineering Laboratory for Next Generation Internet Interconnection Devices, Beijing Jiaotong University. His current research interests are Next Generation Internet Architecture, Transport Protocol and Network modeling.

Fei Song, received his Ph.D. degree from Beijing Jiaotong University in 2010. He is now a Lecturer in School of Electronic and Information Engineering, National Engineering Laboratory for Next Generation Internet Interconnection Devices, Beijing Jiaotong University. His current research interests are Next Generation Internet Architecture, Wireless Communications, Cloud computing.

Zhang Sidong, is now a professor in Beijing Jiaotong University. He has published more than 100 research papers in the areas of wireless communications, computer networks, Ad-hoc networks, sensor networks and information theory. Professor Zhang is also a member of the electronics and information science steering committee of the Ministry of Education, a member of the expert committee of the national Natural Science Foundation of China (NSFC).

Multi–period Optimal Portfolio Decision with Transaction Costs and HARA Utility Function

Zhen Wang[1], Shuling Gao[2]

[1] Institute of Information and System Computation Science, Beifang University of Nationalities
Yinchuan, Ningxia 750021, China

[2] Department of Applied Mathematics, Zhoukou Normal University
Zhoukou, Henan 466001, China

Abstract

Portfolio selection problem is one of the core research fields in modern financial management. While considering the transaction costs in the long term investment makes the portfolio selection problems more complex than there are no transaction costs. In this paper, the general multi-period investment problems with HARA utility function and proportional transaction costs are investigated. By using the dynamic programming method, the indirect utility function is defined for solving the portfolio selection problem. The optimal strategies and the boundary of the no-transaction region are obtained in the explicit form. And the procedure for solving the original portfolio selection problem is given. Numerical example shows the feasibility and effectiveness of the method provided in this paper.

Keywords: *Optimal portfolio, Dynamic programming, Transaction costs, HARA utility function.*

1. Introduction

The portfolio selection problem is one of the most important problems faced to the investors, who need to allocate his or her wealth among different assets or assets classes properly. Determining the optimal portfolio is a rather complex problem which depends on the objective of the investor. In the single period setting, the problem is well understood and can be easily solved by using the mean-variance model [1] or other static models (see [2]). In the multi-period setting, the problem is more complex than the single period one. The multi-period portfolio problem was proposed by [3] and [4]. Explicit solutions for these problems are only available under some assumptions: investment opportunities are constant; there are no transaction costs; the short sale is allowed and the market is complete.

It is well known that an investor who ignores the transaction costs would end up bankrupt. Several authors have made important contributions to the effect of the transaction costs in the multi-period setting (see, for example, [5-18]). Kamin [5] introduced the transaction cost into the dynamic portfolio selection model, and found that the investor's behavior is systematically different from the one without transaction costs. Constantinides [6] extended Kamin's model to the HARA utility function. Magill and Constantinides [7] developed a method to determine the impact of trading costs on capital market equilibrium. Constantinides [8] showed that in the case of proportional transaction costs and power utility, the no-transaction region is of great importance for all practical applications, and believed that these boundaries cannot be obtained analytically. Then, he developed approximate solutions for the case of the investor with a power utility (see [9]). In [10], they studied the optimal consumption and investment decision with the transaction costs for an investor and gave an algorithm for solving the free boundary problem.

These solutions usually deal with the case in an infinite time horizon. But it is more realistic to analyze a finite terminal time. In this case, Gennotte and Jung [11] developed a numerical approximate value of the boundaries. Akian et al. [12] considered the n risky assets situation, and gave the viscosity solution. Boyle and Lin [13] extended Gennotte and Jung's work, and illustrated the solution procedure in which the returns on the risky asset follow a multiplicative binomial process. Framstad et al. [14] showed that the solution in a jump diffusion market has the same form as in the pure diffusion case. Jang [15] investigated an optimal portfolio selection problem with transaction costs when an illiquid asset pays cash dividends and there are constraints on the illiquid asset holding, and provided the closed form solutions for the problem.

Motivated by the above results, we extend research by Boyle and Lin to include the case where the investor has the HARA utility functions. We provide an explicit closed form solution to the finite horizon problem when there are proportional transaction costs and the investor has the HARA utility function. A procedure to derive the boundaries of the no-transaction region is also given.

2. THE MODEL

2.1 HARA utility function

The definition of the general class of HARA utility function is introduced in this subsection. This kind of utility function is very general indeed since it contains the most used utility functions.

Definition 1. *A utility function U is said to have harmonic absolute risk aversion (HARA) if the inverse of its absolute risk aversion is linear in wealth.*

Remark 1. According to Definition 1, the HARA utility function can be written as:

$$U(x) = a \cdot \left(b + \frac{x}{c} \right)^{1-c} \qquad (1)$$

with the domain $b + \frac{x}{c} > 0$. The constant parameters a, b, and c satisfy the condition: $a(1-c) > 0$.

Usually four subclasses are distinguished. When $c = -1$, the utility function is the quadratic utility function. As $c \to \infty$, the utility function takes the form

$$U(x) = ab \cdot \exp\left(-\frac{x}{b} \right)^{1-c},$$

which is often called the CARA (Constant Absolute Risk Aversion) or exponential utility function. If $b = 0$, $c \neq 1$ the utility function is the CRRA (Constant Relative Risk Aversion) or power utility function, which formed

$$U(x) = a \cdot \frac{x^{(1-c)}}{1-c}.$$

Because of $\lim_{c \to 0} (x^c - 1) \cdot \frac{1}{c} = \ln x$, $U(x) = \ln x$ can be considered as another special case of HARA utility function.

2.2 Dynamic programming with utility functions

Consider a financial market where an investor can make decision for his sequential investment at T trading times, indexed as $t = 1, 2, \cdots, T$, over a finite planning period. There are two securities: one riskless asset and one risky asset at each time. Denote P_0^t the price of the riskless asset and P_t the prices of the risky assets at time t. For $t = 1, 2, \cdots, T-1$, $r^0 = \frac{P_{t+1}^0}{P_t^0}$ is the total return on riskless asset and $r_t = \frac{P_{t+1}}{P_t}$ is the total return on the risky assets respectively. Thus, r^0 is a constant and r_t is a random variable.

Assume that an investor holds a portfolio with $x_1^0 \geq 0$ dollars of the riskless asset and $x_1 \geq 0$ dollars of the risky asset at the initial time $t = 1$. At each trading time $t = 1, 2, \cdots, T-1$, the investor may make his investment decision to maximize his expected utility of terminal wealth. Let x_t^0 be the dollar amounts of the riskless asset and x_t be dollar amounts of the risky asset in the portfolio at time t before trading. It is assumed that there is a transaction cost proportional to the amount of

each risky asset traded. Let θ be the unit transaction cost for buying or selling the risky asset. We use u_t to denote investment decision at time t. u_t is the amount of the risky asset traded, $u_t \geq 0$ for buying and $u_t \leq 0$ for selling. Thus, the total transaction costs could be $\theta \cdot |u_t|$. Then, the following relationships can be built:

$$y_t^0 = x_t^0 - u_t - \theta \cdot |u_t|, \tag{2}$$

$$y_t = x_t + u_t, \tag{3}$$

here y_t^0 is the dollar amounts of the riskless asset, y_t is the dollar amounts of the risky asset at time t after trading. We also assume that y_t^0 and y_t are non-negative.

Thus, at time $t+1$ the portfolio amounts before trading can be written as:

$$x_{t+1}^0 = y_t^0 r^0 = (x_t^0 - u_t - \theta \cdot |u_t|) r^0, \tag{4}$$

$$x_{t+1} = y_t r_t = (x_t + u_t) r_t. \tag{5}$$

Equation (4) and (5) describe feasible investment decisions. The objection is to find an optimal sequential investment strategy that maximizes the expected utility of terminal wealth, namely:

$$\max_{u_t, t=1,2,\cdots,T-1} \quad E[U(x_T^0, x_T)], \tag{6}$$

for the given initial portfolio (x_1^0, x_1). Here $U(\cdot)$ represents the HARA utility function.

From above preparation, the model for the investment problem can be presented as:

$$\begin{aligned} \max_{u_t, t=1,2,\cdots,T-1} \quad & E[U(x_T^0, x_T)] \\ subject \quad to \quad & x_{t+1}^0 = (x_t^0 - u_t - \theta \cdot |u_t|) r^0, \\ & x_{t+1} = (x_t + u_t) r_t, \\ & t = 1, 2, \cdots, T-1. \end{aligned} \tag{7}$$

Problem (7) can be solved by a dynamic programming technique. In [13], it was assumed that the terminal utility function U must be a concave, homogeneous differentiable function with some degree, say α. As we assumed above, the terminal utility function U is the HARA utility function which taken the form of (1) is a concave and differentiable function to the terminal total wealth. But it is not homogeneous. Thus we need to do the following transformations. Let

$$\begin{aligned} & \bar{x}_T^0 = x_T^0 + bc, \\ & \bar{x}_t^0 = x_t^0 + bc \cdot \left(\frac{1}{r^0}\right)^{T-t}, \qquad t = 1, 2, \cdots, T-1, \end{aligned} \tag{8}$$

Thus,

$$U(x_T^0, x_T) = \tilde{U}(\bar{x}_T^0, x_T) = a\left(\frac{\bar{x}_T^0 + x_T}{c}\right)^{1-c}. \tag{9}$$

Then, problem (7) is equivalent to problem (10):

$$\begin{aligned} \max_{u_t, t=1,2,\cdots,T-1} \quad & E[\tilde{U}(\bar{x}_T^0, x_T)] \\ subject \quad to \quad & \bar{x}_{t+1}^0 = (\bar{x}_t^0 - u_t - \theta \cdot |u_t|) r^0, \\ & x_{t+1} = (x_t + u_t) r_t, \\ & t = 1, 2, \cdots, T-1. \end{aligned} \tag{10}$$

To apply dynamic programming, we define the indirect utility function V_t, $t = 1, 2, \cdots, T$, as follows:

$$V_t(\bar{x}_T^0, x_T) = \begin{cases} \tilde{U}(\bar{x}_T^0, x_T), & t = T, \\ \max\limits_{u_t} E_t V_{t+1}(\bar{x}_{t+1}^0, x_{t+1}), & t = 1, 2, \cdots, T-1. \end{cases} \tag{11}$$

Here, E_t denotes the expectation over r_t conditional on x_t^0 and x_t. According to the Bellman principle of optimality, the variable u_t, which maximizes $E_t V_{t+1}(\bar{x}_{t+1}^0, x_{t+1})$, $t = 1, 2, \cdots, T-1$, forms the optimal trading strategy of the problem (7).

3. OPTIMAL STRATEGY

In this section, we develop the procedures for solving problem (7) and (10). The derivation is based on the main theorem of [13].

Let

$$\begin{aligned} g_t(u_t, \bar{x}_t^0, x_t) &= E_t V_{t+1}(\bar{x}_{t+1}^0, x_{t+1}) \\ &= E_t V_{t+1}((\bar{x}_t^0 - u_t - \theta \cdot |u_t|) r^0, (x_t + u_t) r). \end{aligned} \tag{12}$$

Then, we give the definition of the no-transaction region. For any portfolio in this region, the expected value will not be increased by buying or selling the risky asset.

Definition 2. When the set of portfolio Φ_t satisfies

$$\Phi_t = \left\{ \left(\bar{x}_t^0, x_t \right) \mid g_t\left(u_t, \bar{x}_t^0, x_t\right) \le g_t\left(0, \bar{x}_t^0, x_t\right), \text{ for all } u_t \right\} \quad (13)$$

Φ_t is called the no-transaction region at time t.

Let $\partial^+ g_t / \partial u_t$, $\partial^- g_t / \partial u_t$ denote the right and left derivatives of g_t respectively. According to the main theorem of [13], if $0 < a_t \le b_t < \infty$, the no-transaction region can be written as:

$$\Phi_t = \left\{ \left(\bar{x}_t^0, x_t \right) \mid a_t \le x_t / \bar{x}_t^0 \le b_t \right\},$$

where

$$a_t = \min\left\{ x_t \mid \frac{\partial^+ g_t(0,1,x_t)}{\partial u_t} = 0, x_t \ge 0 \right\},$$

$$b_t = \max\left\{ x_t \mid \frac{\partial^- g_t(0,1,x_t)}{\partial u_t} = 0, x_t \ge 0 \right\}.$$

The optimal transaction strategies for problem (10) are given in the bellowing theorem.

Theorem 1. *If $a_t \le x_t / \bar{x}_t^0 \le b_t$, then there will be no buying or selling the risky asset.*

If $a_t > x_t / \bar{x}_t^0$, then

$$\max_{u_t} g_t(u_t, \bar{x}_t^0, x_t) = g_t(u_t^+, \bar{x}_t^0, x_t) = g_t(0, y_t^{0+}, y_t^+), \quad (14)$$

where

$$\begin{aligned} u_t^+ &= \frac{\bar{x}_t^0 a_t - x_t}{1 + (1+\theta) a_t}, \\ y_t^{0+} &= \bar{x}_t^0 - (1+\theta) u_t^+, \\ y_t^+ &= x_t + u_t^+, \end{aligned} \quad (15)$$

with $(y_t^{0+}, y_t^+) \in \Phi_t$ and $y_t^+ / y_t^{0+} = a_t$.

If $b_t < x_t / \bar{x}_t^0$, then

$$\max_{u_t} g_t(u_t, \bar{x}_t^0, x_t) = g_t(u_t^-, \bar{x}_t^0, x_t) = g_t(0, y_t^{0-}, y_t^-), \quad (16)$$

where

$$u_t^- = \frac{\bar{x}_t^0 b_t - x_t}{1 + (1-\theta) b_t},$$

$$\begin{aligned} y_t^{0-} &= \bar{x}_t^0 - (1-\theta) u_t^-, \\ y_t^- &= x_t + u_t^-, \end{aligned} \quad (17)$$

with $(y_t^{0-}, y_t^-) \in \Phi_t$ and $y_t^- / y_t^{0-} = b_t$.

Proof. Case 1, $a_t < \infty$. since $u_t^+ > 0$,

$$\begin{aligned} &\frac{\partial g_t(u_t^+, \bar{x}_t^0, x_t)}{\partial u_t} \\ &= -r^0 (1+\theta) E_t \frac{\partial V_{t+1}\left((\bar{x}_t^0 - (1+\theta)\cdot u_t^+) r^0, (x_t + u_t^+) r_t \right)}{\partial \bar{x}_t^0} \\ &\quad + E_t r_t \frac{\partial V_{t+1}\left((\bar{x}_t^0 - (1+\theta)\cdot u_t^+) r^0, (x_t + u_t^+) r_t \right)}{\partial x_{t+1}} \\ &= (\bar{x}_t^0 - (1+\theta)\cdot u_t^+)^{-c} \cdot \left\{ -r^0 (1+\theta) E_t \frac{\partial V_{t+1}\left(r^0, a_t r_t \right)}{\partial \bar{x}_t^0} \right. \\ &\quad \left. + E_t r_t \frac{\partial V_{t+1}\left(r^0, a_t r_t \right)}{\partial x_{t+1}} \right\} \\ &= (\bar{x}_t^0 - (1+\theta)\cdot u_t^+)^{-c} \cdot \frac{\partial g_t^+(0,1,a_t)}{\partial u_t}. \end{aligned}$$

As we mention above that $\frac{\partial^+ g_t(0,1,a_t)}{\partial u_t} = 0$, thus

$$\frac{\partial^+ g_t(u_t^+, \bar{x}_t^0, x_t)}{\partial u_t} = 0. \quad (18)$$

It is shown that u_t^+ is a maximum point.

Case 2, $a_t = \infty$. As g_t is non-decreasing, u_t^+ is the right end point of its domain, thus u_t^+ is the maximum point.

Similarly, u_t^- is the maximum point when $b_t < x_t / \bar{x}_t^0$. □

Thus, Theorem 2 gives the optimal strategies of the original problem (7) below.

Theorem 2. *If $a_t \le \dfrac{x_t}{x_t^0 + bc(1/r^0)^{T-t}} \le b_t$, then $\bar{u}_t = 0$, i.e. there will be no buying or selling the risky asset.*

If $a_t > \dfrac{x_t}{x_t^0 + bc(1/r^0)^{T-t}}$, *then*

$$\begin{aligned} \tilde{u}_t^+ &= \frac{x_t^0 a_t - x_t + bc(1/r^0)^{T-t} \cdot a_t}{1+(1+\theta)a_t}, \\ \tilde{y}_t^{0+} &= x_t^0 + bc(1/r^0)^{T-t} - (1+\theta)\tilde{u}_t^+, \\ y_t^+ &= x_t + \tilde{u}_t^+, \end{aligned} \tag{19}$$

with $(\tilde{y}_t^{0+}, \tilde{y}_t^+) \in \Phi_t$ *and* $\tilde{y}_t^+ / \tilde{y}_t^{0+} = a_t$.

If $b_t < \dfrac{x_t}{x_t^0 + bc(1/r^0)^{T-t}}$, *then*

$$\begin{aligned} \tilde{u}_t^- &= \frac{x_t^0 b_t - x_t + bc(1/r^0)^{T-t} \cdot b_t}{1+(1-\theta)b_t}, \\ \tilde{y}_t^{0-} &= x_t^0 + bc(1/r^0)^{T-t} - (1-\theta)\tilde{u}_t^-, \\ y_t^- &= x_t + \tilde{u}_t^-, \end{aligned} \tag{20}$$

with $(\tilde{y}_t^{0-}, \tilde{y}_t^-) \in \Phi_t$ *and* $\tilde{y}_t^- / \tilde{y}_t^{0-} = b_t$.

Proof. We know that $\bar{x}_t^0 = x_t^0 + bc(1/r^0)^{T-t}$, $t = 1, 2, \cdots, T$. Then, substituting $\bar{x}_t^0 = x_t^0 + bc(1/r^0)^{T-t}$ into (15) and (17), we will get the optimal strategies of the original problem (7). □

Now, how to calculate a_t, b_t and the indirect utility function are presented.

Definition 3. *V is a piece-wise linear utility function with respect to the function U, if there is a sequence of increasing numbers q_j, $j = 1, 2, \cdots, s$, and non-negative constants α_{ij} and β_{ij} with respect to the underlying probability space $\{\omega_i; i = 1, 2, \cdots, I\}$ such that*

$$V(x^0, x) = \sum_{i=1}^{I} U(\alpha_{ij} x^0, \beta_{ij} x) \Pr(\omega_i), \tag{21}$$

$$\text{for } q_j \le x / x_0 \le q_{j+1}. \tag{22}$$

Assuming that V_t is a piecewise linear utility function with respect to U. Let $\{r_t^k; k = 1, 2, \cdots, K\}$ be all possible outcomes for r_t, and $\{\omega_i\}$ be all possible outcomes from $(r_{t+1}, r_{t+2}, \cdots, r_T)$, where $\{\omega_i\}$ represents the set of all future paths of the underlying tree structure starting at a node at time $t+1$. Then starting at time t all the paths of the underlying tree structure can be written as $\{(r_t^k, \omega_i)\}$.

Now, calculate V_t recursively starting at $t = T$. At $t = T$,

$$V_T(\bar{x}_T^0, x_T) = U(\bar{x}_T^0, x_T) = a\left(\frac{\bar{x}_T^0 + x_T}{c}\right)^{1-c}. \tag{23}$$

Suppose that

$$\begin{aligned} V_{t+1}(\bar{x}_{t+1}^0, x_{t+1}) &= \sum_{i=1}^{I} U(\alpha_{ij}\bar{x}_{t+1}^0, \beta_{ij} x_{t+1}) \Pr\{\omega_i\} \\ &= \frac{a}{c^{1-c}} \sum_{i=1}^{I} (\alpha_{ij}\bar{x}_{t+1}^0 + \beta_{ij} x_{t+1})^{1-c} \Pr\{\omega_i\}, \end{aligned} \tag{24}$$

$$q_j \le \frac{x_{t+1}}{\bar{x}_{t+1}^0} \le q_{j+1}, j = 0, 1, \cdots, s; q_0 = 0, q_{s+1} = \infty.$$

Then,

$$\begin{aligned} g_t(0, \bar{x}_t^0, x_t) &= E_t V_t(\bar{x}_t^0 r^0, x_t r_t) \\ &= \sum_{k=1}^{K}\sum_{i=1}^{I} U(\alpha_{ij} r^0 \bar{x}_t^0, \beta_{ij} r_t^k x_t) \Pr\{(r_t^k, \omega_i)\} \\ &= \frac{a}{c^{(1-c)}} \sum_{k=1}^{K}\sum_{i=1}^{I} (\alpha_{ij} r^0 \bar{x}_t^0 + \beta_{ij} r_t^k x_t)^{1-c} \Pr\{(r_t^k, \omega_i)\}. \end{aligned} \tag{25}$$

Let $\tilde{\alpha}_{ij} = \alpha_{ij} r^0$, $\tilde{\beta}_{ij} = \beta_{ij} r_t^k$, thus, when $u_t \ge 0$,

$$\begin{aligned} & g_t(u_t, \bar{x}_t^0, x_t) \\ &= \frac{a}{c^{(1-c)}} \sum_{k=1}^{K}\sum_{i=1}^{I} \{\tilde{\alpha}_{ij}[\bar{x}_t^0 - (1+\theta)u_t] + \tilde{\beta}_{ij}[x_t + u_t]\}^{1-c} \Pr\{(r_t^k, \omega_i)\}; \end{aligned} \tag{26}$$

and when $u_t \le 0$,

$$\begin{aligned} & g_t(u_t, \bar{x}_t^0, x_t) \\ &= \frac{a}{c^{(1-c)}} \sum_{k=1}^{K}\sum_{i=1}^{I} \{\tilde{\alpha}_{ij}[\bar{x}_t^0 - (1-\theta)u_t] + \tilde{\beta}_{ij}[x_t + u_t]\}^{1-c} \Pr\{(r_t^k, \omega_i)\}. \end{aligned} \tag{27}$$

Therefore, a_t is a solution of one of the equations

$$\sum_{k=1}^{K}\sum_{i=1}^{I}(\tilde{\alpha}_{ij}+\tilde{\beta}_{ij}a_t)^{-c}[\tilde{\beta}_{ij}-(1+\theta)]\tilde{\alpha}_{ij}\Pr\{(r_t^k,\omega_i)\}=0, \quad \tilde{q}_j\le a_t<\tilde{q}_{j+1}, j=0,1,2\cdots \tag{28}$$

and b_t is a solution of one of the equations

$$\sum_{k=1}^{K}\sum_{i=1}^{I}(\tilde{\alpha}_{ij}+\tilde{\beta}_{ij}a_t)^{-c}[\tilde{\beta}_{ij}-(1-\theta)]\tilde{\alpha}_{ij}\Pr\{(r_t^k,\omega_i)\}=0, \quad \tilde{q}_j\le a_t<\tilde{q}_{j+1}, j=0,1,2\cdots \tag{29}$$

The indirect utility function can be calculated as follows:

Rearrange all $\left(r^0/r_t^k\right)q_j$, $k=1,2\cdots,K$, $j=0,1,2\cdots$, from smallest to largest and relabeled them as $\tilde{q}_h$ in order of magnitude. Thus, for $l=1,2,\cdots,I$, $\tilde{\alpha}_{lh}=\alpha_{lj}r^0$, $\tilde{\beta}_{lh}=\beta_{lj}r_t^1$,where $\left(r^0/r_t^1\right)q_j\le\tilde{q}_h\le\tilde{q}_{h+1}\le\left(r^0/r_t^1\right)q_{j+1}$; for $l=I+1,I+2,\cdots,2I$, $\tilde{\alpha}_{lh}=\alpha_{l-I,j}r^0$, $\tilde{\beta}_{lh}=\beta_{l-Ij}r_t^2$, where $\left(r^0/r_t^2\right)q_j\le\tilde{q}_h\le\tilde{q}_{h+1}\le\left(r^0/r_t^2\right)q_{j+1}$; … ; for $l=(K-1)I+1,(K-1)I+2,\cdots,KI$, $\tilde{\alpha}_{lh}=\alpha_{l-(K-1)I,j}r^0$, $\tilde{\beta}_{lh}=\beta_{l-(K-1)I,j}r_t^K$, where $\left(r^0/r_t^K\right)q_j\le\tilde{q}_h\le\tilde{q}_{h+1}\le\left(r^0/r_t^K\right)q_{j+1}$.

Change $\tilde{\alpha}_{lh}=\alpha_{l-I,j}r^0$, $\tilde{\beta}_{lh}=\beta_{l-Ij}r_t^2$, where l and h back to i and j to avoid too much notation,

$$g_t(0,\bar{x}_t^0,x_t)=\frac{a}{c^{(1-c)}}\sum_{k=1}^{K}\sum_{i=1}^{I}(\tilde{\alpha}_{ij}\bar{x}_t^0+\tilde{\beta}_{ij}x_t)^{1-c}\Pr\{(r_t^k,\omega_i)\}, \quad \tilde{q}_j\le\frac{x_t}{\bar{x}_t^0}\le\tilde{q}_{j+1}. \tag{30}$$

From Theorem 1, we obtain

$$V_t(\bar{x}_t^0,x_t)=\begin{cases} g_t(0,y_t^{0+},y_t^+), & \dfrac{x_t}{\bar{x}_t^0}<a_t, \\ g_t(0,\bar{x}_t^0,x_t), & a_t\le\dfrac{x_t}{\bar{x}_t^0}\le b_t, \\ g_t(0,y_t^{0-},y_t^-), & \dfrac{x_t}{\bar{x}_t^0}>b_t. \end{cases} \tag{31}$$

Assume that $\tilde{q}_{j1}<a_t\le\cdots<\tilde{q}_{j2}\le b_t<\cdots$. Define

$$\bar{q}_0=0,\bar{q}_1=a_t,\bar{q}_2=\bar{q}_{j1},\cdots,\bar{q}_{j2+j1+2}=b_t,\bar{q}_{j2+j1+3}=\infty,$$

and

$$\bar{\alpha}_{i0}=\frac{\tilde{\alpha}_{i,j1}+a_t\tilde{\beta}_{i,j1}}{1+(1+\theta)a_t},\quad \bar{\beta}_{i0}=(1+\theta)\bar{\alpha}_{i0};$$

$$\bar{\alpha}_{ij}=\tilde{\alpha}_{i,j1+j-1},\quad \bar{\beta}_{ij}=\tilde{\beta}_{i,j1+j-1},\quad j=1,2,\cdots,j2-j1+1;$$

$$\bar{\alpha}_{i,j2-j1+2}=\frac{\tilde{\alpha}_{i,j2}+a_t\tilde{\beta}_{i,j2}}{1+(1-\theta)a_t},\quad \bar{\beta}_{i,j2-j1+2}=(1-\theta)\bar{\alpha}_{i,j2-j1+2}.$$

Hence,

$$V_t(\bar{x}_t^0,x_t)=\frac{a}{c^{(1-c)}}\sum_{k=1}^{K}\sum_{i=1}^{I}(\bar{\alpha}_{ij}\bar{x}_t^0+\bar{\beta}_{ij}x_t)^{(1-c)}\Pr\{(r_t^k,\omega_i)\}, \quad \bar{q}_j\le\frac{x_t}{\bar{x}_t^0}\le\bar{q}_{j+1}, \quad j=0,1,\cdots. \tag{32}$$

Based on the above discussion, the problem (7) can be solved by the following procedures:

First, choose the proper distribution of the r_t and construct the scenario tree to determine the scenario's paths.

Secondly, use (28) and (29) to calculate the boundaries of the no-transaction region.

Finally, determine whether the portfolio lies in the no-transaction region. If it is, the portfolio is the optimal strategy; if not, take the optimal strategies as (19) and (20) shown.

4. NUMERICAL EXAMPLE

The model and the solution procedure presented in section 3 will be illustrated in this section by the numerical examples. We assume $T=5$ and consider the case in which the rate of return for the risky asset in each period is dependent of t and has only two states u and d. The Boyle and Lin [13] parameterization for u and d is used in this section, via, $u=e^{\sigma\sqrt{h}}$, $d=e^{-\sigma\sqrt{h}}$, $r^0=e^{\delta h}$ that $\sigma=1$, $h=0.25$, $\delta=0.05$.

Thus $r^0 = 1.0126$ and the scenario data for the risky asset's return is shown in table 1.

Table 1: The risky asset's return

i th Scenario	r_1	r_2	r_3	r_4
1	1.0205	1.6825	2.7739	4.5733
2	1.0205	1.6825	2.7739	1.6824
3	1.0205	1.6825	1.0204	1.6823
4	1.0205	1.6825	1.0204	0.6189
5	1.0205	0.6189	1.0204	1.6823
6	1.0205	0.6189	1.0204	0.6189
7	1.0205	0.6189	0.3754	0.6189
8	1.0205	0.6189	0.3754	0.2277

Suppose that $a = 1$, $b = 5$, $c = 2$, we can calculate the boundaries of no-transaction region recursively backwards from the last period by using the procedures in the last section. Table 2 shows the values, when $\theta = 0.001$ and $\theta = 0.01$, respectively.

Table 2: The no-transaction bounds

	$\theta = 0.001$	$\theta = 0.01$
$a1$	0.559342	0.405368
$b1$	0.795427	1.07235
$a2$	0.587316	0.35647
$b2$	0.774512	1.5798
$a3$	0.526971	0.158935
$b3$	0.817125	2.547631
$a4$	0.501839	0.083563
$b4$	0.835976	2.963258

As shown in table 2, the transaction costs have a dramatic impact on the no transaction region. When the transaction costs increase, the no-transaction region has become wider. If we set the initial proportion of the risky asset is 0.05 and the initial proportion of the riskless asset is 0.95, then the investor should buy the risky asset to reach the boundary $a_1 = 0.559342$. Allowing for transaction costs, the amount of the risky asset to be purchased at time 1 is 0.5276, which is larger than the one using the power utility function as the terminal utility function in [13]. It shows that the power utility function is more risk aversion than the HARA utility function, when the two type of function have the same parameter c.

To interpret the role of the boundaries of no-transaction region, we assume that the initial wealth is 1000 dollars, of which initial proportion of risky asset is 10% and that of riskless asset is 90%. Optimal investment decisions for problem (7) are shown in Table 3 and Table 4. For each entry in these tables, the first number γ_t represents $\frac{x_t}{x_t^0 + bc(1/r^0)^{T-t}}$. According to Theorem 2, it should be compared with a_t and b_t in Table 2, then we can determine how to calculate u_t. The second number stand for the amount of risky asset one should buy or sell, while the third and forth numbers give the amount of riskless asset and risky asset one should hold, respectively. Comparing the results in table 3 with the one in table 4, we can see that the investors facing the higher transaction costs will behave more risk aversion.

Table 5 gives the optimal investment strategies in each period with different transaction costs. It shows that the investors will change his strategies according to his forecasting of the rate of the risky asset's return. Take the first scenario as an example, the scenario shows the up-up tendency while the proportion of the risky asset increases. Anyway, the above results show the efficiency of the method we proposed in this paper.

5. CONCLUSION

Multi-period investment problems with HARR utility function and proportional transaction costs are investigated in this paper. We have developed the analytical expressions for the indirect utility function and the boundary of the no-transaction region and given the optimal strategy of the investment problem. From the analysis, we can see that the results are independent with the time spacing, thus our researches are also valid for unequal time spacing. The numerical example indicates the efficiency of the method.

REFERENCES

[1] H. Markowitz, "Mean-variance analysis in portfolio choice and capital markets". Blackwell, Oxford, UK, 1992.

Table 3: Numerical results for $\theta = 0.001$

		1	*2*	*3*	*4*	*5*
1	γ_t	0.1099	0.5554	0.9615	1.4272	-
	u_t	262.0647	13.3531	-53.2438	-283.9864	-
	x_t^0	900	655.4606	660.0601	871.5386	1443
	x_t	100	369.487	644.1284	1258.1	3270.6
2	γ_t	0.1099	0.5554	0.9615	1.4272	-
	u_t	262.0647	13.3531	-53.2438	-283.9864	-
	x_t^0	900	655.4606	660.0601	871.5386	1443
	x_t	100	369.487	644.1284	1258.1	1203.2
3	γ_t	0.1099	0.5554	0.9615	0.525	-
	u_t	262.0647	13.3531	-53.2438	0	-
	x_t^0	900	655.4606	660.0601	871.5386	892.646
	x_t	100	369.487	644.1284	426.8134	778.591
4	γ_t	0.1099	0.5554	0.9615	0.525	-
	u_t	262.0647	13.3531	-53.2438	0	-
	x_t^0	900	655.4606	660.0601	871.5386	892.646
	x_t	100	369.487	644.1284	426.8134	286.4352
5	γ_t	0.1099	0.5554	0.3537	0.5223	-
	u_t	262.0647	13.3531	76.0047	0	-
	x_t^0	900	655.4606	660.0601	601.3376	619.0405
	x_t	100	369.487	2393977	319.3284	537.2062
6	γ_t	0.1099	0.5554	0.3537	0.5223	-
	u_t	262.0647	13.3531	76.0047	0	-
	x_t^0	900	655.4606	660.0601	601.3376	619.0405
	x_t	100	369.487	239397	319.3284	197.6323
7	γ_t	0.1099	0.5554	0.3537	0.1922	-
	u_t	262.0647	13.3531	76.0047	126.0125	-
	x_t^0	900	655.4606	660.0601	601.3376	491.3126
	x_t	100	369.487	239397	117.4793	150.6971
8	γ_t	0.1099	0.5554	0.3537	0.1922	-
	u_t	262.0647	13.3531	76.0047	126.0125	-
	x_t^0	900	655.4606	660.0601	601.3376	491.3126
	x_t	100	369.487	239397	117.4793	55.4431

[2] H. Konno and H. Yamazaki, "Mean-absolute deviation portfolio optimization model and its applications to Tokyo stock market", Management Science, 37(5), 1991, pp.519-531.

[3] P. Samuelson, "Lifetime portfolio selection by dynamic stochastic programming", Review of Economics and Statistics, 51(3), 1969, pp.239-246.

[4] R. Merton, "Lifetime portfolio selection under uncertainty: the continuous time case", Review of Economics and Statistics, 51(3), 1969, pp.247–257.

[5] J. Kamin, "Optimal portfolio revision with a proportional transaction cost", Management Science, 21(11), 1975, pp.1263-1271.

[6] G. Constantinides, "Optimal portfolio revision with proportional transaction costs: extension to hara utility function and exogenous deterministic income", Management Science, 22(8), 1976, pp.921-923.

[7] M. Magill and G. Constantinides, "Portfolio selection with transaction costs", Journal of Economic Theory, 13(2), 1976, pp.245-263.

[8] G. Constantinides, "Multi-period consumption and investment behavior with convex transaction costs", Management Science, 25(11), 1979, pp.1127-1137.

[9] G. Constantinides, "Capital market equibilibrium with transaction costs", Journal of Political Economy, 94(4), 1986, pp.842-862.

[10] M. Davis and A. Norman, "Portfolio selection with transaction costs", Mathematics of Operations Research, 15(4), 1990, pp.676-713.

Table 4: Numerical results for $\theta = 0.01$

		1	*2*	*3*	*4*	*5*
1	γ_t	0.1099	0.4031	0.661	1.7875	-
	u_t	190.6707	0	0	0	-
	x_t^0	900	726.0888	745.1131	764.5015	784.2602
	x_t	100	296.6295	499.0791	1384.4	6331.3
2	γ_t	0.1099	0.4031	0.661	1.7875	-
	u_t	190.6707	0	0	0	-
	x_t^0	900	726.0888	745.1131	764.5015	784.2602
	x_t	100	296.6295	499.0791	1384.4	2329.1
3	γ_t	0.1099	0.4031	0.661	1.7875	-
	u_t	190.6707	0	0	0	-
	x_t^0	900	726.0888	745.1131	764.5015	784.2602
	x_t	100	296.6295	499.0791	509.2603	856.7286
4	γ_t	0.1099	0.4031	0.661	0.6575	-
	u_t	190.6707	0	0	0	-
	x_t^0	900	726.0888	745.1131	764.5015	784.2602
	x_t	100	296.6295	499.0791	509.2603	315.1812
5	γ_t	0.1099	0.4031	0.2432	0.2419	-
	u_t	190.6707	0	0	0	-
	x_t^0	900	726.0888	745.1131	764.5015	784.2602
	x_t	100	296.6295	183.584	187.3291	315.1437
6	γ_t	0.1099	0.4031	0.2432	0.2419	-
	u_t	190.6707	0	0	0	-
	x_t^0	900	726.0888	745.1131	764.5015	784.2602
	x_t	100	296.6295	183.584	187.3291	115.938
7	γ_t	0.1099	0.4031	0.2432	0.089	-
	u_t	190.6707	0	0	0	-
	x_t^0	900	726.0888	745.1131	764.5015	784.2602
	x_t	100	296.6295	183.584	68.9174	42.653
8	γ_t	0.1099	0.4031	0.2432	0.089	-
	u_t	190.6707	0	0	0	-
	x_t^0	900	726.0888	745.1131	764.5015	784.2602
	x_t	100	296.6295	183.584	68.9174	15.6925

[11] G. Gennotte and A. Jung, "Investment strategies under transaction costs: the finite horizon case", Management Science, 40(3), 1994, pp.385-404.

[12] M. Akian, J. Menaldi and A. Sulem, "On an investment-consumption model with transaction costs", Journal of Control and Optimization, 34(1), 1996, pp.329-364.

[13] P. Boyle and X. Lin, "Portfolio selection with transaction costs", North American Actuarial Journal, 1(2), 1997, pp.27-39.

[14] N. Framstad, B. Øksendal and A. Sulem, "Optimal consumption and portfolio in a jump diffusion market with proportional transaction costs", Journal of Mathematical Economics, 35(2), 2001, pp.233-257.

[15] B. Jang, "Optimal portfolio selection with transaction costs when an illiquid asset pays cash dividends", Journal of the Korean Mathematical Society, 44(1), 2007, pp.139-150.

[16] H. Shah and S. Bhingarkar, "Data-acquisition, data analysis and prediction model for share market", International Journal of Computer Science Issue, 8(3), 2011, pp.530-534.

[17] M. Neshat, A. Baghi, A. Pourahmad, G. Sepidnam, M. Sargolzaei and A. Masoumi, "Fuzzy hybrid expert system for marketing mix model", International Journal of Computer Science Issue, 8(6), 2011, pp.126-134.

[18] C. Li and Z. Li, "Multi-period portfolio optimization for asset – liability management with bankrupt control", Applied Mathematics and Computation, 218(22), 2012, pp.11196-11208.

Table 5: Proportion of riskless asset and risky asset

Scenario	*Time*	$\theta = 0.001$		$\theta = 0.01$	
		x_t^0	x_t	x_t^0	x_t
1	1	0.9000	0.1000	0.9000	0.1000
	2	0.6395	0.3605	0.7100	0.2900
	3	0.5061	0.4939	0.5989	0.4011
	4	0.4092	0.5908	0.3558	0.6442
	5	0.3061	0.6939	0.1102	0.8898
2	1	0.9000	0.1000	0.9000	0.1000
	2	0.6395	0.3605	0.7100	0.2900
	3	0.5061	0.4939	0.5989	0.4011
	4	0.4092	0.5908	0.3558	0.6442
	5	0.5453	0.4547	0.2519	0.7481
3	1	0.9000	0.1000	0.9000	0.1000
	2	0.6395	0.3605	0.7100	0.2900
	3	0.5061	0.4939	0.5989	0.4011
	4	0.6713	0.3287	0.6002	0.3998
	5	0.5341	0.4659	0.4779	0.5221
4	1	0.9000	0.1000	0.9000	0.1000
	2	0.6395	0.3605	0.7100	0.2900
	3	0.5061	0.4939	0.5989	0.4011
	4	0.6712	0.3287	0.6002	0.3998
	5	0.7571	0.2429	0.7133	0.2867
5	1	0.9000	0.1000	0.9000	0.1000
	2	0.6395	0.3605	0.7100	0.2900
	3	0.0027	0.9973	0.8023	0.1977
	4	0.6532	0.3468	0.8032	0.1968
	5	0.5354	0.4646	0.7134	0.2866
6	1	0.9000	0.1000	0.9000	0.1000
	2	0.6395	0.3605	0.7100	0.29
	3	0.0027	0.9973	0.8023	0.1977
	4	0.6532	0.3468	0.8032	0.1968
	5	0.7580	0.2420	0.8712	0.1288
7	1	0.9000	0.1000	0.9000	0.1000
	2	0.6395	0.3605	0.7100	0.2900
	3	0.0027	0.9973	0.8023	0.1977
	4	0.8366	0.1634	0.9173	0.0827
	5	0.7653	0.2347	0.9484	0.0516
8	1	0.9000	0.1000	0.9000	0.1000
	2	0.6395	0.3605	0.7100	0.2900
	3	0.0027	0.9973	0.8023	0.1977
	4	0.8366	0.1634	0.91731	0.0827
	5	0.8986	0.1014	0.9804	0.0196

Zhen Wang lecturer received her B.A. in Applied Mathematics in 2005, the M.S. degree in Econometrics in 2008, and the Ph.D. degree in Applied Mathematics from Xidian University in 2012. She is currently with Institute of Information and System Computation Science, Beifang University of Nationalities, China. Her main research areas include Mathematical Finance, Portfolio Optimization, and Intelligent Optimization.

Shuling Gao received her M.S. degree in Applied Mathematicsfrom Shaanxi Normal University, China, in 2008. Her main interests include Intelligent Optimization, Theory and Methods of Optimization.

Exploring Verbalization and Collaboration during Usability Evaluation with Children in Context

Mohammadi Akheela Khanum[1] and Munesh C. Trivedi [2]

**[1] Faculty of Engineering, Department of Computer Science, PAHER University
Udaipur, Rajasthan, India**

**[2] Department of Computer Science, DIT
Greater Noida, U.P, India**

Abstract

In this paper, we investigate the effect of context on usability evaluation. The focus is on how children behave and perform when they are tested in different settings. Two most commonly applied usability evaluation methods: the think-aloud and constructive interactions are applied to the children in different physical contexts. We present an experimental design involving 54 children participating in two different configurations of constructive interaction and a traditional think-aloud. The behavior and performance of the children in two different physical contexts is measured by evaluating the results of application of think-aloud and constructive interaction. Finally, we outline lessons on the impact of context on involving children in usability testing.

Keywords: *usability evaluation, children, physical context, think-aloud, constructive interaction.*

1. Introduction

Now a days when a user buy any gadget, be it a mobile phone, laptop, or an ipad, he first check how easy and understandable the gadget functionality is [1]. This indicates that the users nowadays are more particular about the usability of the gadgets. Usability is most often defined as the ease of use and acceptability of a system for a particular class of users carrying out specific tasks in a specific environment [2]. Ease of use affects the user's performance and their satisfaction, while acceptability affects whether the product is used [2].

With the rapid emergence of new technologies in everyday activities, it is common for all age groups to use new devices. Children cannot be left behind when the use of technologies is discussed. Many children nowadays are found to spend hours with the devices such as laptop computers, game consoles, cell phones, digital cameras, or audio players. All these technologies are becoming an essential part of daily lives. "While many adults struggle with comprehending and manipulating digital interfaces, today's young children enthusiastically approach these interfaces with little or no effort, although they may not completely understand how to use it, or what their implications are" [3].

Children are not miniature adults but they have their own set of preferences, perception, style, likes, and dislikes [4]. When designing technology for children their preferences should be taken into account. To do so, usability evaluation* is performed with the children as the testers of technology. During the early design phases of children technology, usability engineers performs usability testing to uncover usability problems that might creep into the product when set to be used in the real context.

Context is a term defined differently by different people. For example, Brown et al. [5] define context as *"location, identities of the people around the user, the time of the day, season, and temperature"*. Ryan et al [6] define context as the *"user's location, environment, identity and time"*. Hull et al [7] included the entire environment by defining context to be *"aspects of the current situation"*. Schilit et al [8] claim that the important aspects of the context are: where you are, who you are with, and what resources are nearby. Dey et al [9] define context to be the *"user's physical, social, emotional or informational state"*.

When evaluating the usability of any system, the behavior of the user is very important. The factors which may affect the user behavior needs to carefully considered because the result of usability evaluations may vary in different settings where the user may exhibit varying behaviors. Product usability doesn't take place in a vacuum; rather, it happens in context [10]. The characteristics of the context (the users, tasks, and environment) may be as important in determining usability as the characteristics of the product itself. Changing any relevant aspect of the context of use may change the usability of the product [11].

Therefore, in this paper we try to find the answer to the following question (i) how does physical context affect verbalizations of perceptions, thoughts, and understandings concerning the interaction in usability evaluations? We address the above stated question by looking at how children perform and behave in lab and in field testing when constructive interaction and think-aloud, methods are applied.

First, we present the literature review on the effect of context during usability testing. Secondly, an experimental design involving 54 children participating in two different configurations of constructive interaction and a traditional think-aloud is presented. Thirdly, we present results from the evaluations by illustrating how the children behaved and perceived the different context when we applied the constructive interaction and think-aloud protocol. Finally, we outline lessons on impact of context on children in usability testing.

2. Related Work

The importance of physical context in usability evaluations have been researched for a long. Out of the many factors that can effect usability evaluations, physical context is considered to directly influence the behaviour of the people involved in the usability evaluations. The physical context may include the location, the temperature, the time, the light etc.

Tsiaousis & Giaglis [12] examined the effects of environmental distractions on mobile website usability. They proposed a model hypothesizing on the effects of environmental distractions on the usability of mobile sites. They categorized the environmental distractions into auditory, visual and social. A preliminary test on 20 users was conducted to investigate the effect of environmental distractions on mobile website usability. Results confirmed that environmental distractions have direct effect on mobile website usability.

Hummel et al. [13] developed a mobile context-framework based on a small wireless sensor network, to monitor environmental conditions such as light, acceleration, sound, temperature, and humidity during the usability experiments. User experiments have been conducted in a laboratory with seven test persons where the environmental conditions were changed. Under varying environmental conditions the performance of the users on the average was decreased in terms of higher error rates and delays.

Kaikkonen et al. [14] carried out usability testing of mobile consumer application in two environments: in a laboratory and in a field with a total of 40 test users. Results indicate that conducting a time-consuming field test may not be worthwhile when searching user interface flaws to improve user interaction. They found that field testing is worthwhile when combining usability tests with a field pilot or contextual study where user behavior is investigated in a natural context.

Razak et al. [15] conducted usability testing with children in both laboratory and field. Drawing applications were tested in their preschool and an educational game was tested in the usability laboratory. The results indicate that field study is more suitable for understanding children experience with technology than it is with testing for usability problems and laboratory study is more suitable for evaluating user interfaces and interaction with the application than it is with understanding children's experience.

Andrrzejczak & Liu [16] examined the effect of location on the user's stress level during usability evaluation. User stress levels were assessed by Spielberger's State-Trait Anxiety Inventory; using the paper survey's baseline and experimental stress scores. In addition, user performance data was recorded through task times and subjective user assessments. The data suggested no significant differences exist between participant data in both baseline and experimental anxiety scores. This implies that remote testing as a cost-efficient way to conduct user testing, may be a viable alternative to traditional lab testing without altering the test's effectiveness.

Madathil [17] performed a synchronous remote usability test using a three-dimensional virtual world, and empirically compared it with WebEx, a web-based two-dimensional screen sharing and conferencing tool, and the traditional lab method. The results suggest that virtual lab method is as effective as the traditional lab and WebEx based methods in terms of the time taken by the test participants to complete the tasks and the number of higher severity defects identified. Test participants and facilitators alike experienced lower overall workload in the traditional lab environment than in either of the remote testing environments.

Baillie & Schatz [18] evaluated a multimodal mobile application through a combination of laboratory and field studies. The users were given a set of four action scenarios to be performed. The results were surprising; only one action scenario was completed in the time frame whereas three out of four action scenarios were completed in lesser time. Error rates were higher in lab than in the field. The reason for such performances by the users could be that the users feel more relaxed in the field.

Donker & Markopoulos [19] studies a comparative assessment of three UEMs namely the Concurrent Think Aloud (CTA), interview and questionnaire. Each of these UEMs requires a different level of verbalization for the children that are performing the evaluation. In order to tests these three evaluation methods, 45 children aged 8-14 years were recruited as the test users. The result indicates that children who think aloud during testing uncover more problems than the children who answer specific questions.

However, to elicit verbal comments the children have to be prompted, which can be an indication that children find it difficult to think aloud. Prompting may cause children feel obliged to mention problems to please the experimenter. This could lead to non problems being reported. The result also suggests that girls thinking out loud report more usability problems than boys.

Baauw and Markopoulos [20] conducted a study to compare UEMs. The study involved twenty four children in the age group of 9-11 year, in the usability testing of the computer game- BioMania. The usability evaluation was carried out to test two UEMs namely the TA and post task interview. The results indicate that there was no significance difference between the problems reported by the two genders. The post task interview allows observation data and verbalization data to be obtained on fly without analyzing tapes. Thus, post task interviews can offer practical benefit at the cost of slightly longer sessions. The number of usability problems identified through the two methods was not significant.

Markopoulos and Bekker [21] presented a framework for characterizing comparative studies of usability testing methods with respect to their appropriateness for children. They found that the ability to verbalize problems in interactions depends on: the ability of translating experiences into verbal statements, on their knowledge of the language and on prior experiences in speaking up to adults. They found that compound tasks and abstract tasks formulations could pose problems to children, as their abstract and logical thinking abilities are not yet fully developed and they are not skilled in keeping multiple concepts simultaneously in mind. The results also indicate that think aloud helps generate more problems reports than questionnaires and interviews.

Vermeeren et al., [22] conducted a study on the use of post task interviewing evaluation technique with 6-8 years old children. The results show that children overall were fairly good at answering the questions. The negative side effects of applying the technique on the outcome of the usability test are minor. Further, the study suggests applying such technique to uncover extra data about possible causes for interaction difficulties. Also to limit the questions by only asking detailed questions about those parts of the design that needs extra attention.

3. Method

3.1 Participants

54 children (24 girls and 30 boys) at the age ranging from 10 years to 13 years old (Mean M=11.63; Standard Deviation SD=0.88) participated as test subjects in the experiment. All the children were 6th and 7th grade pupils from two different English medium schools in the Lucknow city of India. The children did not receive compensation for their involvement in the experiment. The children were assigned as test subjects to one of the four test setups: as individual testers in the lab and in the field for think-aloud sessions, as pairs in lab and field for constructive interaction sessions. Each individual setup had 9 testers (4 girls and 5 boys), and each paired setup had 9 pairs (4 pairs of girls and 5 pairs of boys), Children were randomly assigned to each of the four test setups. Children in pairs were familiar with each other. Table 1 shows the assignment of children to different setups.

Table 1: 54 children assigned as individual testers in think-aloud and as pairs in constructive inetraction

	Constructive Interaction		***Think-aloud***	
	Lab	***Field***	***Lab***	***Field***
Boys	5x2	5x2	5	5
Girls	4x2	4x2	4	4
Total	9x2	9x2	9	9

3.2 Settings

The sessions were held at the school's campus itself, because the school authorities did not permit us to take the children to the place where the usability laboratory was set. We created two labs, one for field testing sessions, and one for laboratory testing sessions. For the field testing, we used the school's computer lab which the students were familiar with and we tried to keep it as it was used by the children. No restrictions were imposed on the people to move in the lab during the test session. This created a perfect field environment for the children. For testing in lab environment, we setup a usability laboratory in one part of the school. The lab environment was different as compared with the field. Lab was located in a quiet place where people not related with the test sessions were not allowed. The lab was only occupied by the test monitors and the test participants at any given time during the test sessions. Fig 1 depicts the usability test session.

Fig. 1 Snapshot of usability test session

3.3 System

The selected system for our experiment was International Children's Digital Library (ICDL). This particular website was selected because digital libraries are becoming a common place for children and many researches are now focusing on how the children are using these new learning tools. During the children's demographic data collection, we also found that none of the children had ever used ICDL. Fig. 2 is the screenshot of ICDL home page.

International Children's Digital Library is a collection of books that features various books for children in different age groups. ICDL has four search tools for accessing the current collection books: Simple, Advanced, Location, and Keyword. Using the simple search, users can search for books using colorful buttons representing the most popular search categories. The advanced search interface allows users to search for books in a compact, text-link-based interface that contains the entire library category hierarchy. By selecting the location based search, users can search for books by spinning a globe to select a continent. Finally, with the Keyword search, users search for books by typing in a keyword.

Fig. 2 Screenshot of ICDL

3.4 Procedure

The first step towards starting the test was to take consent from the school authorities. After clearing the first step we proceeded with taking the consent from the children's parents or guardians. To do so, we handed over the consent forms to the children to get it signed by their parents or guardians. The consent form provided information about the type of test their wards will be involved in and that the choice of allowing their children to take the test was purely voluntary. After receiving consent from 54 children, we scheduled the usability evaluation sessions. At the beginning of the test sessions children were introduced to the experiment by the participating researchers. The researchers explained the children's roles in the experiment and how their participation would contribute to our research.

Hanna et al. [23] guidelines for usability testing with children were followed. We greeted and children and introduced ourselves. Particularly, we focused on stressing the importance of the participation, and stressing that they were not the object of the test. The purpose of the usability test was explained to the children in detail. The children received questionnaires on which they had to provide answers to such as age, name, school, computer/internet experience, number of hours spend each week on computer/internet, and online reading experience. The usability test sessions were conducted in two labs, one a specialized usability laboratory setup in the school and the other was the school's computer lab. During the test sessions, all the screen activities and children's interaction with ICDL were recorded using CamStudio for later analyses. CamStudio is an open source desktop screen recorder

The children were asked to solve five tasks. The tasks involved the use of different search options in ICDL. This included searching books by country, searching books by title, searching books by language, searching award winning books in English and reading a specified book in the language of their preference. We did not specify any time limits for the tasks, but required the participants to try to solve all tasks.

All children were able to solve all specified tasks. On an average, the children spent 11:11 minutes (SD=2:87) in the lab and 9:33 minutes (SD=2:28) in the field on the all the tasks. The individual testers were asked to think-aloud while solving the tasks.

Think-aloud was explained to the individual testers in terms of the descriptions in [24]. The pairs were asked to collaborate with each other while solving the tasks. Constructive interaction was explained to the pairs as described in [24].

After the usability sessions, the children were asked to complete the subjective workload test (NASA-TLX) [25]. The children filled in the test form individually even though they participated in pairs. NASA-TLX is applied to evaluate the workload as experienced by the children in order to compare their behavior in different settings.

4. Data Analysis

36 sessions were completed and then analyzed in detail. The sessions were analyzed based on how well children verbalized (in think-aloud sessions) and collaborated (in

constructive interaction sessions). The different aspects of our analysis were (i) Degree of verbalization and collaboration, (ii) Quality of verbalization and collaboration, (iii) impact of test monitor on solving the tasks, (iv) communication between the test monitor and the user and (v) prompting by the test monitor. The quantitative values were assigned to each of these parameters on a scale of 1 to 5. A score of 1 means the lowest and 5 means the highest. For instance, a score of 5 assigned to verbalization/collaboration means that the children verbalized their thoughts to the maximum during think-aloud sessions and collaborated highest during constructive interaction sessions.

5. Results

The 54 children in the 36 usability test sessions solved all the assigned tasks. The task completion time in the field (M=9.78, SD=2.28) was lesser compared to the time taken in the laboratory (M=10.67, SD=2.87). But no significance difference was found for the task completion times.

5.1 Assessment of verbalization and collaboration in different settings

To assess the four setups we applied six different aspects of verbalization and collaboration in usability tests. These six aspects are illustrated in table 2. The setting whose mean score (M) marked with a plus sign indicates that it has a significant difference with the setting whose M is marked with a minus sign. SD is the standard deviation. Verbalization refers to the verbal comments during think-aloud sessions which would facilitate identification of what the tester is feeling about the interface under test. Collaboration refers to verbalization during constructive interaction sessions.

Interestingly, we found that the quality of verbalization was considerably higher for the constructive interaction sessions compared to the think-aloud sessions. The score in the lab (M=4.0, SD= 0.5) and in the field (M=3.8, SD=0.4) did not differ much amongst the pairs. However, the score was higher in the field (M=2.67, SD=0.67) as compared to lab (M=1.89, SD=0.74) for the individual testers.

The analysis of variance shows significant differences between the four settings on degree of verbalization F (3, 32) = 22.55, p= 4.93811E-08. Since the value of p indicated a significant difference between the settings, we performed a post-hoc test.

The post-hoc analysis showed significant difference at the 1% and 5% level between the pairs and individual testers in the lab and the field during both the constructive interaction and think-aloud sessions, however the difference was not significant amongst the pairs and amongst individual testers in the four settings.

Further, we analyzed the quality of verbalization and collaboration in the test sessions. The quality of the collaboration was higher for both the constructive interaction sessions than the quality of verbalization for think-aloud sessions. Field settings provoked more verbalization and collaboration for the testers. The analysis of variance shows significant difference between all the setups on the quality of verbalization/collaboration F (3, 32) =11.76, p=2.35463E-05. The post hoc analysis showed a significant difference at 1% level between the constructive interaction lab setting and think-aloud lab setting, between constructive interaction field and think-aloud lab setting. At 5% level between constructive interaction lab setting and think-aloud lab setting, between constructive interaction field setting and think-aloud lab setting and also between constructive interaction field and think-aloud field setting.

Table 2: Assessment of verbalization and collaboration in four settings for all testers

Testing parameters	***Constructive Interaction***		***Think-aloud***	
	Lab	***Field***	***Lab***	***Field***
Degree of verbalization/collaboration	M=4.0+ SD=0.5	M=3.8+ SD=0.4	M=1.89- SD=0.74	M=2.67- SD=0.67
Quality of verbalization/collaboration	M=3.2+ SD=0.8	M=3.4+ SD=0.5	M=1.67- SD=0.67	M=2.44- SD=0.68
Impact of test monitor on solving the tasks	M=2.22 SD=0.67	M=2.33 SD=0.71	M=2.56 SD=0.88	M=2.56 SD=0.53
Communication between test monitor and user	M=2.33 SD=0.50	M=2.11 SD=0.60	M=2.44 SD=0.88	M=2.56 SD=0.53
Prompting by the test monitor	M=2.22+ SD=0.67	M=2.22+ SD=0.67	M=3.11- SD=0.33	M=3.00- SD=0.71
Time taken to complete the tasks	M=10.67 SD=3.67	M=8.89 SD=2.24	M=11.56 SD=1.88	M=9.78 SD=2.17

The test monitor plays an important role during usability evaluation. Test monitor is a person who closely monitors the usability test activities and notes the tester's behavior, verbalization, and other such things which may of interest for the usability test under consideration. We analyzed the impact of test monitor on solving the usability tasks. Constructive interaction provides potentially natural thinking-aloud as test subjects collaborate in pairs to solve tasks and therefore, one could expect less influence and interaction with a test monitor. We found that the test monitor has slightly more interaction with the think-aloud subjects compared the constructive interaction subjects, but the difference is not significant $F(3, 32) = 0.5$, $p = 0.684$.

Another factor of our analysis was to assess the level of communication between the test monitor and testers. Test monitor have a slightly higher level of interaction with the testers during think-aloud sessions. However, this difference was not significant $F(3, 32) = 0.78$, $p = 0.515$. We also assessed the level of prompting that was required to make the testers verbalize their actions during the test sessions. Think-aloud required higher level of prompting than the constructive interaction. Also, field testing using think-aloud required lesser prompting compared to lab testing. However, for constructive interaction, prompting in field and lab was not significantly different. The analysis of variance shows significant difference between the setups on the amount of prompting by the test monitor $F(3, 32) = 5.60$, $p = 0.003$. The post hoc analysis showed a significant difference at 5% level between the constructive interaction lab setting and think-aloud lab setting, and between constructive interaction field and think-aloud lab setting.

Finally, we assessed the amount of time spent on solving all the tasks during each test session. Not surprisingly, we found that the testers in think-aloud sessions spent more time on solving the tasks. Field sessions took lesser time compared to their lab counterparts. But this difference is not significant $F(3, 32) = 1.71$, $p = 0.183$.

6. Discussion

In this section, we discuss the qualitative results from the study. We have identified a number of interesting outcomes related to usability testing in context with children.

Outcome 1: *usability testing in field provides natural environment for children to freely verbalize their thoughts.*
The children freely verbalize their actions and thoughts in field during constructive interaction and also during think-aloud sessions. Field testing also resulted in better quality of verbalization during both constructive interaction and think-aloud sessions compared to their lab counterparts. Lesser interaction between the test monitor and testers was found in field for both constructive interaction and think-aloud sessions. Time taken to complete all tasks was lesser in field.

Outcome 2: *constructive interaction provides better degree and quality of verbalization compared to think-aloud*
During the constructive interaction sessions the children were more relaxed but during think-aloud sessions they were nervous. Individual testing made the children feel that it was they who were tested and not the interface. One of the individual testers was so nervous that he gave up the test. Higher prompting was required for individual testers. Verbalizing thoughts while solving tasks made the children uneasy. In one case when the monitor asked the tester to verbalize his thoughts, he stopped working and began to think. Working in pairs made the children more comfortable. They discussed much before taking a move while solving the tasks. However, in some cases of constructive interaction the dominating tester ignored the other partner. Lesser intervention by the test monitor was noticed for constructive interaction sessions.

7. Conclusion

In this paper, we investigate how children perform and behave in different physical settings during usability testing. Our particular focus is on how the children behave and perceive a testing situation when involved in lab and field testing session with traditional think-aloud and constructive interaction. Our results show that field testing with children resulted in better level and quality of verbalization. Field testing can be a feasible option for testing with children. Even though we did not impose any time constraints on the children, our results show that field testing took lesser time to complete the tasks.

Our results also show that the pairing of children had impact on how the children verbalized and collaborated in pairs during the testing sessions. We found that constructive interaction facilitate natural think-aloud as the pairs tended to collaborate well while solving the tasks. The quality of verbalization was fair enough to get them closer to the solution.

We further experienced that the individual testers applying think-aloud tended to be more verbose in the field than in the lab. This could be an indication that it is not only the method that is affecting the usability tests but also the context in which the test is performed.

Our future goal is to further investigate the impact of context by applying other quantitative measures.

References

[1] M.C. Trivedi, and M.A.Khanum, "Role of Context in Usability Evaluations", ACIJ, Vol. 3, No. 2, 2012, pp. 69-78.

[2] N.Bevan, "Measuring usability as quality of use", Software Quality Journal, Vol 4, 1995, pp.115-130.

[3] A. M. Marhana, M. I. Miclea, C. Popaa, G. Preda, "A review of mental models research in child-computer interaction", Procedia - Social and Behavioral Sciences, Vol.33,2011, pp.368-372.

[4] A. Druin, "The Role of Children in the Design of New Technology", HCIL Technical Report No. 99-23, University of Maryland, USA, 1999.

[5] P. J. Brown, J.D. Bovey, and X. Chen, "Context Aware Applications: From the Laboratory to the Marketplace", IEEE Personal Communications, Vol.4, No.5, 1997, pp.58-64. IEEE Press.

[6] N. Ryan, J. Pascoe, and D. Morse, "Enhanced Reality Fieldwork: the Context Aware Archaeological Assistant", In Gaffney, V., van Leusen, M. & Exxon, S. (Eds.) Computer Applications in Archaeology, 1997.

[7] R. Hull, P. Neaves, J.Bedford-Roberts, "Towards Situated Computing", 1st International Symposium on Wearable Computers, 1997, pp. 146-153.

[8] B. Schilit, and M. Theimer, "Disseminating Active Map Information to Mobile Hosts", IEEE Network, Vol.8,No.5, 1994, pp.22-32. IEEE Press.

[9] A.K. Dey, and G.D Abowd, "Towards a Better Understanding of Context and Context-Awareness", In CHI2000 Workshop on What, Who, Where, When and How of Context - Awareness, New York, 2000, ACM Press.

[10] E. Olmsted-Hawala, E.Murphy, S. Hawala, and K. Ashenfelter, "Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability," In Proceedings of the 28th international Conference on Human Factors in Computing Systems, 2010, pp. 2381-2390.

[11] N.Bevan, and M.Macleod, (1994) "Usability Measurement in Context", Behaviour and Information Technology (BIT), Vol.13 (1-2), 1994, pp.132 - 145.

[12] A.S. Tsiaousis, and G.M. Giaglis "Evaluating the Effects of the Environmental Context-of-Use on Mobile Website Usability," Mobile Business, 2008. ICMB '08. 7th International Conference on, vol., no., pp.314-322.

[13] K. A.Hummel, A. Hess, and T. Grill, "Environmental context sensing for usability evaluation in mobile hci by means of small wireless sensor networks", In MoMM '08:Proceedings of the 6th International Conference on Advances in Mobile Computing and Multimedia, 2008,pp. 302–306.

[14] A. Kaikkonen, T. Kallio, A. Kekäläinen, A. Kankainen, A. Cankar, "Usability Testing of Mobile Applications: A Comparison between Laboratory and Field Testing," Journal of Usability studies , Vol.1, No.1 2005, pp 4-16.

[15] F.H.A. Razak, H. Hafit, N. Sedi, N.A. Zubaidi, and H.Haron "Usability testing with children: Laboratory vs field studies", User Science and Engineering (i-USEr), International Conference on , vol., no., 2010, pp.104-109.

[16] C.Andrrzejczak, D. Liu (2010) "The effect of testing location on usability testing performance, participant stress levels, and subjective testing experience", Journal of Systems and Software Vol. 83, No. 7, 2010.

[17] K.C. Madathil, "Synchronous remote usability testing: a new approach facilitated by virtual worlds", Proceedings of the 2011 annual conference on Human factors in computing systems CHI'11.

[18] L.Baillie, R.Schatz, "Exploring Multimodality in the Laboratory and in the Field", Proc. 7th Intern. Conference on Multimodal Interfaces, 2005, pp. 100-107.

[19] A. Donker, P. Markopoulos , "A comparison of think- aloud, questionnaires and interviews for testing usability with children," Proc. HCI 2002, Springer, pp. 305-316.

[20] E.Baauw, P.Markopoulos, "A comparison of think-aloud and post-task interview for usability testing with children," Proc. Interaction Design and Children,2004, pp.115–116.

[21] P. Markopoulos, and M.Bekker, "How to compare usability testing methods with children participants," In Bekker, M., Markopoulos, P., and Kersten-Tsikalkina, M. (Eds.): "Interaction Design and Children". Shaker Publisher, 2002, ISBN 90-423-0200-3, pp. 153-159.

[22] A. Vermeeren, M.M. Bekker, I.E.H. van Kesteren, H. de Ridder, "Experiences with structured interviewing of children during usability tests," In L.J. Ball et al. (eds.) Proc. CI 2007, The 21st British HCI Group Annual Conference, Swindon, UK: BCS.pp.139-146.

[23] L.Hanna, K. Risden, K. Alexander," Guidelines for Usability Testing with Children", Interactions Vol. 4 No. 5,pp. 9-14.

[24] J. Nielsen, (1993) Usability Engineering. Academic Press

[25] R.C.Miller and S.G. Hart, (1984) "Assessing the Subjective Workload of Directional Orientation Tasks",In Proceedings of 20th Annual Conference on Manual Control, NASA Conference Publication, 1984, pp. 85 – 95

The Intensity and the Factors Affecting the Use of Social Network Sites Among the Students of Jordanian Universities

Andraws Swidan[1], Hasan Al-Shalabi[2] , Mustafa Jwaifell[3] , Arafat Awajan[4] and Adnan Alrabea[5]

[1] Computer engineering department, Faculty of Engineering and Technology, University of Jordan, Jordan

[2] Faculty of Engineering, Al-Hussein Bin Talal University, Jordan

3Department of Curriculum and Teaching, Al-Hussein Bin Talal University, Jordan

[4]Computer Science Department, Princess Sumaya University for Technology, Jordan

[5]Faculty of Information Technology, Al Balqa Applied University, Jordan

Abstract

The present paper examines social network sites (SNSs) usage among university students of four Jordanian universities distributed in different regions of the Kingdom. Seven hundred and twenty seven students were sampled and they completed a questionnaire based on the technology acceptance model. In addition, 16 participants, four from each university, were interviewed. The variance in the extent of SNSs usage in relation to university, faculty, gender, age, study level and socioeconomic background was investigated. This study employed a mixed-method model as interviews and questionnaires were employed. The data were qualitatively and quantitatively collected, sorted, analyzed and reported. The results of the qualitative analyses and the quantitative descriptive results suggested that the extent of SNS usage is high among the university students in Jordan. Chi-Square tests used to determine whether the equality use of social networks among Jordanian university according to various parameters for the top four social networks were done. The researchers' recommendations are to make better use of those Social Networks by integrating them in universities' learning management systems.

Keywords: *Social Networks Sites (SNS), Higher Education, University students, Jordanian Universities*.

1. Introduction

Jordanian young adults, as their peers in the world, are using intensively social networks sites. The aim and intensity of such usage is the subject of several research papers Ahmad, Suleiman Alhaji. (2011), Kwon Ohbyung and Wen Yixing. (2009), Leng Goh Say, Likoh Jonathan, Japang Minah, Andrias Ryan Macdonell., and Amboala Tamrin. (2010). It is almost impossible to underestimate the role and effect of using SNS in our daily life. The impact of SNS on the Arab spring movements is still to be studied and investigated. Social, economic and academic effects of SNS were studied by several researchers Boyd Dana and Ellison Nicole. (2007, October), Ellison Nicole B., Steinfield Charles., and Lampe Cliff. (2007), Hewitt Anne and Forte Andrea. (2006), Pempek Tiffany A., Yermolayeva Yevdokiya A., and Clavert, Sandra L. (2009), Valenzuela Sebastian., Park Namsu., and Kee Kerk F. (2008). College and university students are the main SNS users. Certain SNS are more frequently used than others. What are the most commonly used SNS among Jordanian students and what are the factors influencing these phenomena is the subject of this study. Social, economic factors, type of study, gender of the students and level of study were investigated. The two technology acceptance measures: Perceived usefulness (PU) and Perceived ease-of-use (PEOU) were estimated for the most commonly used SNS. The Technology Acceptance Model (TAM) is an information systems theory that models how users come to accept and use a technology. The model suggests that when users are presented with a new technology, a number of factors influence their decision about how and when they will use it. The reverse of this statement is assumed true. If new technologies are adopted then the intensity of their use indicates their PU and PEOU.

.

2. Background of the study

SNSs can be described as online community that gathers people with in same interests. Kwon and Wen (2009) defined SNSs Sites as an individual web page which

enables online, human-relationship building by collecting useful information and sharing it with specific or unspecific people. While Boyd and Ellison (2007) defined SNSs Sites as web-based services that allow individuals to construct a public or semi-public within a bounded system. Facebook can be considered one of the most SNSs that influenced online communications between people, even this relationship shifted to a specific enrollment of relationship . Hewitt and Forte (2006) described the results from ongoing investigation of student/faculty relationships in the online community Facebook to understand how contact on Facebook was influencing student perceptions of faculty, where the result of this survey point to one third of the students they surveyed did not believe that faculty should be present on the Facebook at all. Those finding are very interesting while the Arab universities students are insists to have faculty member e-mail to interact with him as we noticed all the time in our experience.

The students' experiences and uses of SNSs can be differing according to their needs, which may differ from country to another. Pempek, Yermolayeva, and Calvert (2009) investigated experiences of 92 undergraduates students reporting daily time use and responding to an activities checklist to assess their use of the popular Social Network site Facebook, where they concluded that students spend approximately 30 minutes throughout the day as a part of their daily routine, beside this result, the use of Facebook in a style of one-to-many communication tool.

The uses of SNSs can be one-to-one or one-to-many as a part of group uses. This study is trying to survey Jordanian Universities' students to gain a full picture about the intensity of use of different SNSs and how this use depends upon the gender, type of study, social economic situation, level of study and type of faculty. The result of this research can open the doors for Academics and Policy Makers to take advantages of the most common SNSs among the students of Jordanian universities.

Valenzuela, Park, and Kee (2008) found that SNSs and in specific Facebook effect college students. They found positive relationships between intensity of Facebook use and students' life satisfaction, social trust, civic participants and political engagement.

With regard to intensity, Ellison, Steinfield and Lampe (2007) investigated the benefits of Facebook "Friends:" social capital and college students' use of online SNSs; they examined the relationship between use of Facebook and the formation and maintenance of social capital. The study surveyed 286 undergraduate students. The findings of their study showed a strong association between use of Facebook and the three types of social capital, with the strongest relationship being to bridging social capital.

SNSs are playing a great roll in the lives of university students as Leng, Likoh, Japang, Andrias, and Amoala (2010) pointed out. They reach this point of view through a descriptive study conducted to investigate SNS usage among university students in Labuan. The study concluded that the mass adoption of SNS points to evolution in human social interaction regardless of age, culture background, occupations and general demographic profile. Thus it was obvious to them that university students will eventually use the SNS as a main medium of communication to maintain their relationships with friends and family members as well as expanding their niche community.

It is obvious that SNSs became as a demand of interaction between academics and their students whereas SNS are part of university student daily life all over the world. Ahmad (2011) studied the SNSs' usage and students' attitudes towards social behaviors and academic adjustment in Northern Nigerian Universities. His findings revealed that the extent SNS usage depends upon students ethnicity and religion.

3. Problem statement of the study

The population of Jordan consists almost of 65% of youths and most of them are enrolled in universities. Academics and policy makers are committed to make use of Social Networks for the benefits of teaching and learning and integrating SNSs within Learning management systems or even giving educators attention to exploit relative advantages of SNSs academically and implement new innovations of methodologies such as Mobile learning or interact with students through Internets' technologies. This study aims at providing the decision makers and educators to maximize the benefit of use of SNS

4. Questions of the study

To maximize the benefit of SNS the study explores the intensity of SNSs use among university students in Jordan analyzing that use in relation to geographical location which indirectly reflects the socioeconomic atatus, the faculties type, the gender type and the study level iof students. The research questions were:

Q1: what are the most popular SNSs Sites among the students of Jordanian Universities?

Q2: what is the intensity of SNSs use among the students of Jordanian Universities?

5. Method and Measures

The study was conducted as a part of a project of investigating the SNSs uses among students of Jordanian

Universities, while the items of the questionnaire is part of the project conducted by the authors.

To answer the research questions, the researchers set a questionnaire consisted of 9 questions related to the study variables. The questionnaire was distributed to four universities: the University of Jordan in the capital, Princess Sumaya University for Technology, Jordan in the capital denoted as Sumaya, Al-Hussein Bin Talal University Jordan in the south denoted as Hussein and Al Balqa Applied University, Jordan denoted as Balqa. Data collected and analyzed in a descriptive quantitative research.

6. Sample of the study

The sample of the study consisted of (727) students drawn randomly out of four above mentioned Jordanian universities. The study variables are: 1) university, 2) faculty, 3) gender, 4) year level and 5) social level.

7. Results and discussion

Table 1 shows the cross tabulation profile of the participants.

Table 1 Cross tabulation profile of the participants

		Frequency	Percent
University	University of Jordan	280	38.5
	Hussein	155	21.3
	Balqa	135	18.6
	Sumaya	157	21.6
Total		727	100
Faculty	Science	531	73
	Arts	196	27
Total		727	100
Gender	Male	319	43.9
	Female	408	56.1
Total		727	100
Level	First Year	163	22.4
	Second Year	154	21.2
	Third Year	177	24.3
	Fourth Year	233	32.0
Total		727	100

Analysis of the participants profile is shown in figures 1,2,3,4,5,6.

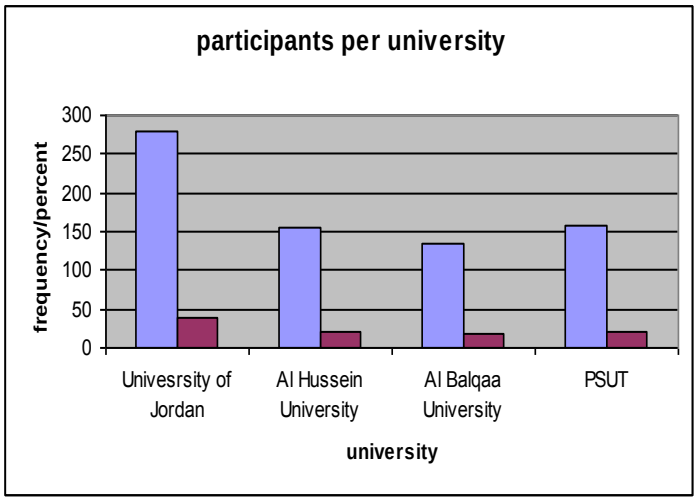


Figure 1. the number of students and their percentage among participants per university

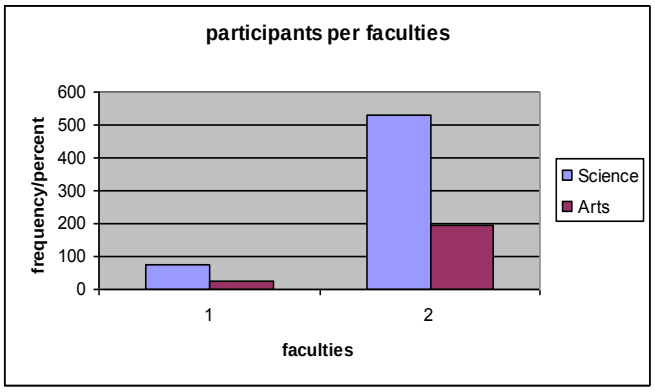


Figure 2. the number of students and their percentage among participants per faculties

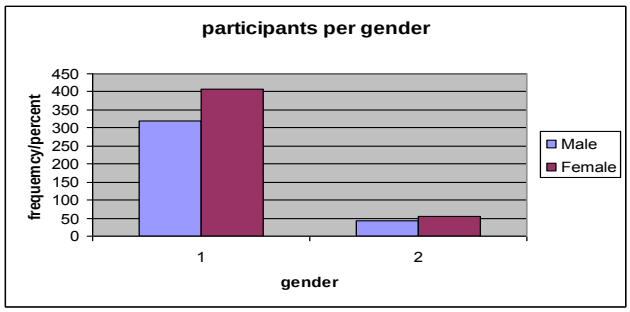


Figure 3. the number of students and their percentage among participants per gender

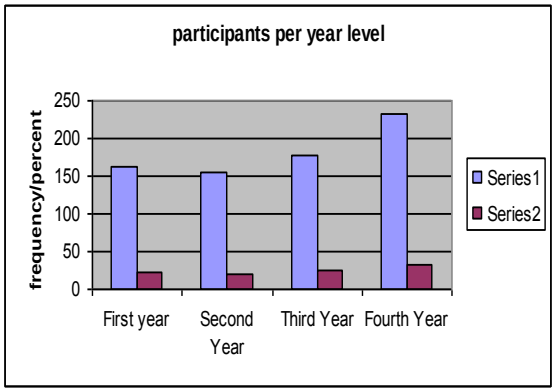


Figure 4. the number of students and their percentage among participants per year level

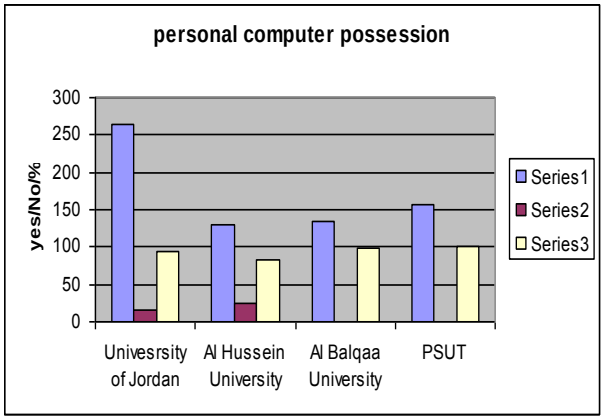


Figure 5. the number of students with personal computers and their percentage among participants

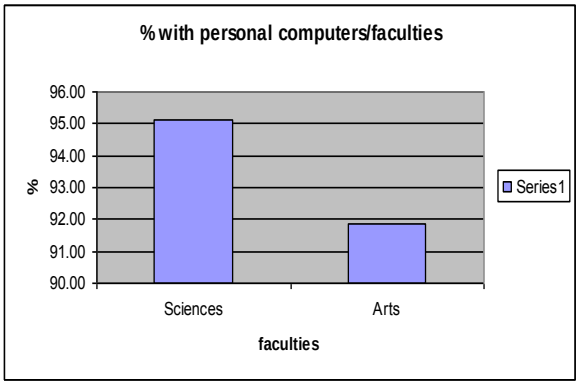


Figure 6. the % of students with personal computers per faculties

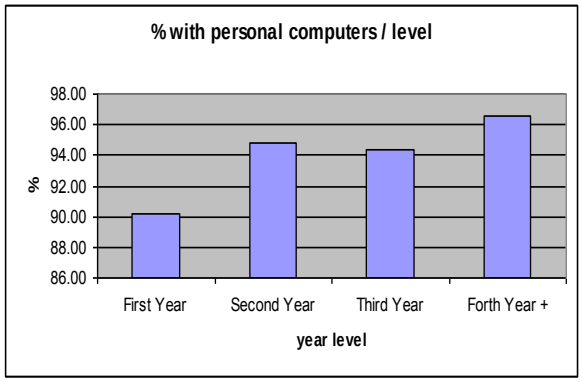


Figure 7. the number of students with personal computers per year level

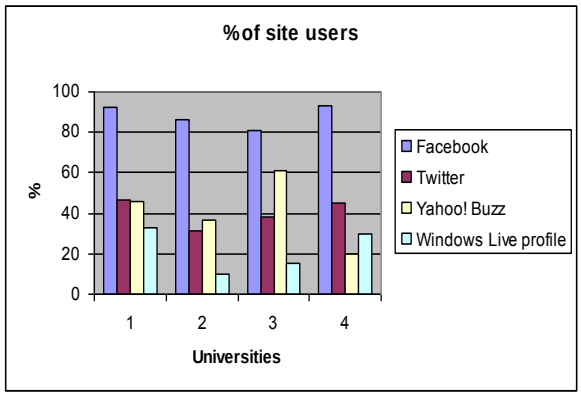


Figure 8. the number of students using most commonly used SNS per university

Table 2 shows the results of the students answer on the availability of personal computers

Table 2 personal computer availability

Variables		Computer		Total
		Yes	No	
University	Univ. of Jordan	264	16	280
	Hussein	130	25	155
	Balqa	134	1	135
	Sumaya	157	0	157
Faculty	Sciences	505	26	531
	Arts	180	16	196
Gender	Male	298	21	319
	Female	387	21	408
Level	First Year	147	16	163
	Second Year	146	8	154
	Third Year	167	10	177
	Forth Year +	225	8	233

A summary of the results of students answers on the type of Social Networks According to Universities (who answered Yes) can be provided as follows:

The top four common social networks used among Jordanian Universities Students are:
1. FaceBook
2. Twitter
3. Yahoo! Buzz
4. Windows Live Profiles

The least four common social networks used among Jordanian Universities Students are:
1. Gaiga Online
2. SodaHead.com
3. BlackPlanet.com
4. Plaxo

Chi-Square tests used to determine whether the equality use of social networks among Jordanian university for the top four social networks are shown in table 3.

Table 3 Chi-Square tests for the top four SNS among universities

University	Social Network	Yes	No	Expected Value	Chi-Square	Sig
Univ. of Jordan	FaceBook	258	22	14.86	15.812	0.001
Hussein		134	21			
Balqa		109	26			
Sumaya		146	11			
Univ. of Jordan	Twitter	131	149	56.08	11.691	0.009
Hussein		48	107			
Balqa		52	83			
Sumaya		71	86			
Univ. of Jordan	Yahoo! Buzz	128	152	55.34	54.79	0.000
Hussein		57	98			
Balqa		82	53			
Sumaya		31	126			
Univ. of Jordan	Windows Live profile	93	187	32.87	36.832	0.000
Hussein		16	139			
Balqa		21	114			
Sumaya		47	110			

All are significant

Chi-Square tests used to determine whether the equality use of social networks among Jordanian university according to faculty for the top four social networks are shown in table 4.

Table 4 Chi-Square tests for the top four SNS among faculties

Faculty	Social Netwotk	Yes	No	Expected Value	Chi-Square	Sig
Sciences	FaceBook	492	39	21	26.933	0.000
Arts		155	41			
Sciences	Twitter	246	285	81	18.586	0.000
Arts		56	140			
Sciences	Yahoo! Buzz	197	334	80	12.325	0.000
Arts		101	95			
Sciences	Windows Live profile	159	372	47	33.496	0.000
Arts		18	178			

All are significant

Chi-Square tests used to determine whether the equality use of social networks among Jordanian university according to gender, for the top four social networks are shown in table 5.

Table 5 Chi-Square tests for the top four SNS among genders

Faculty	Social Netwotk	Yes	No	Expected Value	Chi-Square	Sig
Male	FaceBook	286	33	35	.252	0.615
Female		361	47			
Male	Twitter	139	180	132.51	.968	0.325
Female		163	245			
Male	Yahoo! Buzz	120	199	130.76	2.673	0.102
Female		178	230			
Male	Windows Live profile	69	250	77	2.277	0.131
Female		108	300			

All are not significant. The males and females having equal responses, though, there is no significant differences in using social networks according to gender.

Chi-Square tests used to determine whether the equality use of social networks among Jordanian university according to their level (year of study) for the top four social networks are shown in table 6.
.

Table 7 Chi-Square tests for the top four SNS among year of study level

Level	Social Netwotk	Yes	No	Expected Value	Chi-Square	Sig
Year1	FaceBook	139	14	16.95	3.467	0.325
Year2		141	13			
Year3		158	19			
Year4+		209	24			
Year1	Twitter	57	106	63.97	6.093	0.107
Year2		62	92			
Year3		73	104			
Year4+		110	123			
Year1	Yahoo! Buzz	55	108	63.13	5.279	0.152
Year2		63	91			
Year3		75	102			
Year4+		105	128			
Year1	Windows Live profile	29	134	37.49	10.777	0.013
Year2		37	117			
Year3		38	139			
Year4+		73	160			

All are not significant at $\alpha \geq 0.05$ except the social network Windows Live Profile, where Widows Live Profile is less common among both Year1 and Year2 students.

8. Conclusion

The study revealed that students of Jordanian Universities intensively use SNSs. which can be as an academic tool for communication and interacting with/between educators and students alike. Analysis show that the most commonly used SNSs are FaceBook, Twitter, Yahoo! Buzz and Windows Live Profiles. Chi-Square tests used to determine whether the equality use of social networks among Jordanian university for the top four social networks are significant. Chi-Square tests used to determine whether the equality use of social networks among Jordanian university according to faculty for the top four social networks are significant as well. Chi-Square tests used to determine whether the equality use of social networks among Jordanian university according to gender are not significant. Academics and policy makers can take advantages of SNSs and integrating them into learning management systems.

References:

Ahmad, Suleiman Alhaji. (2011). Social networking sites' usage and students' attitudes towards social behaviors and academic adjustment in Northern Nigerian Universities. Unpublished thesis, UUM College and Sciences. University Utara Malaysia.

Boyd Dana and Ellison Nicole. (2007, October). "Social Network Sites: Definition, History, and Scholarship." *Journal of Computer-Mediated Communication*, 13 (1).

Ellison Nicole B., Steinfield Charles., and Lampe Cliff. (2007). The benefits of Facebook "Friends:" social capital and college students' use of online social network sites. Journal of Computer-Mediated Communication 12 (2007) 1143-1168

Hewitt Anne and Forte Andrea. (2006). Crossing boundaries: Identity management and student/faculty relationships on the Facebook. CSCW'06, November 4-8, Alberta, Canada.

Kwon Ohbyung and Wen Yixing. (2009). An empirical study of the factors affecting social network service use. Computer in Human Behaviour, 26 (2010) 254-263.

Leng Goh Say, Likoh Jonathan, Japang Minah, Andrias Ryan Macdonell., and Amboala Tamrin. (2010). Descriptive study of SNS usage among university students in Labuan, Labuan e-Journal of Muamalat and Society, Vol. 4, 2010, pp. 54-64. Available at:

Pempek Tiffany A., Yermolayeva Yevdokiya A., and Clavert, Sandra L. (2009). College students'

social networking experiences on Facebook. Journal of Applied Developmental Psychology, 30 (2009) 227-238.

Valenzuela Sebastian., Park Namsu., and Kee Kerk F. (2008). Lessons from Facebook: The effect of social network sites on college students' social capital. Submited to the 9th International Symposium on Online Journalism. Austin, Texas, April 4-5, 2008.

Calculation in Parallel Sensitivity Function Using Vector Presentation Algorithm (VPA)

Hamed Al Rjoub
Umm Al-Qura University
Computer Science Department, Makkah-Saudi Arabia

Abstract

This paper presents a new algorithm to solve in parallel linear equations which represent a mathematical model for a large dimension control system and calculates in parallel sensitivity function using n-1 processors where n is a number of linear equation that can be represented as TX=W, where T is a matrix of size $n_r \times n_c$, $X=T^{-1}W$,is a vector of unknowns, and $\partial X/\partial h=-T^{-1}(\ (\partial T/\partial h)X-(\partial W/\partial h))$ is a sensitivity function with respect to variation of system components h. The algorithm (VPA) divides the mathematical input model into two partitions and uses only (n-1) processors to find out the vector of unknowns for original system $x=(x_1,x_2,\ldots,x_n)^T$ and in parallel using (n-1) processors to find the vector of unknowns for similar system $(x^{|})^t=-d^tT^{-1}=(x^{|}_1,x^{|}_2,\ldots,x^{|}_n)^T$ where d is a constant vector .Finally, the sensitivity function (with respect to variation of any component $\partial X/\partial h_i = (x_i \times x^{|}_i)$ can be calculated in parallel by multiplication unknowns $x_i \times x^{|}_i$ respectively, where i=0,1,…n-1 .The running time t is reduced by O(t/2) and the efficiency of (VPA) is increased by 50-60% .

***Key words**: Parallel processing, Vector Presentation, Sensitivity Function, Matrix, Variation, Running Time, Mathematical Model.*

1. Introduction

The ability to develop mathematical models in Biology, Physics, Geology and other applied areas has pulled and has been pushed by the advances in High Performance Computing. Moreover, the use of iterative methods has increased substantially in many application areas in the last years [9, 5]. One reason for that is the advent of parallel Computing and its impact in the overall performance of various algorithms on numerical analysis[1].The use of clusters plays an important role in such scenario as one of the most effective manner to improve the computational power without increasing costs to prohibitive values. However, in some cases, the solution of numerical problems frequently presents accuracy issues increasing the need for computational power. Verified computing provides an interval result that surely contains the correct result [6]. Numerical applications providing automatic result verification may be useful in many fields like simulation and modeling. Finding the verified result often increases dramatically the execution time [2]. However, in some numerical problems, the accuracy is mandatory. The requirements for achieving this goal are: interval arithmetic, high accuracy combined with well suitable algorithms. The interval arithmetic defines the operations for interval numbers, such that the result is a new interval that contains the set of all possible solutions. The high accuracy arithmetic ensures that the operation is performed without rounding errors, and rounded only once in the end of the computation. The requirements for this arithmetic are: the four basic operations with high accuracy, optimal scalar product and direct rounding. These arithmetics should be used in appropriate algorithms to ensure that those properties will be held. There is a multitude of tools that provides verified computing, among them an attractive option is C-XSC (C for extended Scientific Computing) [3]. CXSC is a free

and portable programming environment for C and C++ programming Languages, offering high accuracy and automatic verified results. This programming Tool allows the solution of several standard problems, including many reliable numerical parallel algorithms. The need to solve systems of linear algebraic equations arises frequently in scientific and engineering applications, with the solution being useful either by itself or as an intermediate step in solving a larger problem. In practical problems, the order, n, may in many cases be large (100 – 1000) or very large (many tens or hundreds of thousands). The cost of a numerical procedure is clearly an important consideration — so too is the accuracy of the method. Let us consider a system of linear algebraic equations:

$$Ax = b, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Where $A = \{aij\}_{i,j=1}^{n}$ is a given matrix, and $b = (b1, \ldots, bn)^t$ is a given vector. It is well known (see, for example, [4, 5]) that the solution, x, $x \in R^n$, when it exists, can be found using – direct methods, such as Gaussian elimination, and LU and Cholesky decomposition, taking $O(n^3)$ time; – stationary iterative methods, such as the Jacobi, Gauss- Seidel, and various relaxation techniques, which reduce the system to the form:

$$x = Lx + f, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

and then apply iterations as follows

$$x^{(0)} = f, \; x^{(k)} = Lx^{(k-1)} + f, \; k = 1, 2, \ldots \quad (3)$$

until desired accuracy is achieved this takes $O(n^2)$ time per iteration. – Monte Carlo methods (MC) use independent random walks to give an Approximation to the truncated sum (3)

$$x^{(l)} = \sum_{k=0}^{l} L^k f \quad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

taking time O(n) (to find n components of the solution) per random step. Keeping in mind that the convergence rate of MC is $O(N^{-1/2})$, where N is the number of random walks, millions of random steps are typically needed to achieve acceptable accuracy. The description of the MC method used for linear systems can be found in [6], [7], [8]. Different improvements have been proposed, for example, including sequential MC techniques [5], resolve-based MC methods [1], etc., and have been successfully implemented to reduce the number of random steps. In this paper we study the quasi-Monte Carlo (QMC) approach to solve linear systems with an emphasis on the parallel implementation of the corresponding algorithm. The use of quasirandom sequences improves the accuracy of the method and preserves its traditionally good parallel efficiency. The paper is organized as follows: gives the background - MC for linear systems and a brief description of the quasirandom sequences we use, describes parallel strategies, presents some numerical results and presents conclusions and ideas for parallel processing.

2. RELATED WORK

Solution of large (dense or sparse) linear systems is considered an important Part of numerical analysis, and often requires a large amount of scientific computations [9, 10]. More specifically, the most time consuming operations in iterative methods for solving linear equations are inner products, vector successively updates, matrix-vector products and also iterative refinements [11, 12]. Tests pointed out that the Newton-like iterative method, presents a iterative refinement step and uses a inverse matrix obtained through the backward/forward substitution (after LU decomposition), which are the most time consuming operations. The parallel solutions for linear solvers found in the literature explore many aspects and constraints related to the adaptation of the numerical methods to high performance environments [3]. However, the proposed solutions are not often realistic, and mostly deal with unsuitable models for high performance environments of distributed memory as clusters of workstations. In many theoretical models (such as the PRAM family) the transmission cost to data exchange is not considered, but in

distributed memory architectures this issue is crucial to gain performance. Nevertheless, the difficulty in parallelizing some numerical methods, mainly iterative schemes, in an environment of distributed memory, is the interdependency among data (e.g. the LU decomposition) and the consequent overhead needed to perform inter process Communication (IPC) [3]. Due to this, in a first approach some modifications were done in the backward/ forward substitution procedure [7] to allow less Communications and independent computations over the matrix. Another possible optimization when implementing for such parallel environments is to reduce communication cost through the use of load balance techniques, as we can see in some recent parallel solutions for linear systems solvers [8]. Anyway, their focus was toward the issues related to MPI implementation through a theoretical performance analysis. Few works were found related to numerical analysis of parallel implementations of iterative solvers, mainly using MPI. Moreover, some interesting papers found present algorithm which allow the use of different parallel environments [9]. However, those papers (like others) do not deal with verified computation. We also found some works which focus on verified computing [5] and both verified computing and parallel implementations, but this thesis implement other numerical problems or use a different Parallel approach. Another concern is the implementation of self verified numerical solvers which allow high accuracy operations. The researches already made, show that the execution time of the algorithms using this kind of routines is much larger than the execution time of the algorithms which do not use it [11, 13]. The C-XSC library was developed to provide functionality and portability, but early researches indicate that more optimizations may be done to provide more efficiency, due to additional computational cost in sequential, and consequently for other environments as Itanium clusters. Some experiments were conducted over Intel clusters to parallelize self-verified numerical solvers that use Newton-based techniques but there are more tests that may be done [2,14].

Sensitivity analysis defines the relative sensitivity function for time independent parameters as:

$$S_{i,j} = \partial X_i / \partial h_j \ , \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Where X_i represents the i-th state variable, h_j is the element of the parameter vector. Hence the sensitivity is given by the so-called sensitivity matrix S, containing the sensitivity coefficient Si,j ,equation 5 .The direct approach of numerically differentiating by means of numerical field calculation software will lead to diverse difficulties [1,3]. Therefore, some ideas to overcome those problems aim at performing differentiations necessary for sensitivity analysis prior to any numerical treatment. Further calculations are then carried out with a commercially available field calculation program. Such approach has already been practical successfully [7]. As it considered that the linear system (1) where A is a tridiagonal matrix of order n of the form shown in (6), $x=(x_0,x_1,\ldots..,x_{n-1})^T$ is the vector of unknowns , and $d=(d_0,d_1,\ldots,d_{n-1})^T$ is a vector of dimension n.

$$A= \begin{pmatrix} b_0 & c_0 & & & \\ a_1 & b_1 & c_1 & & \\ & a_2 & b_2 & c_2 & \\ & & .. & .. & \\ & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & a_{n-1} & b_{n-1} \end{pmatrix} \quad \ldots (6)$$

In the LU factorization A, is decomposed into a product of two bidiagonal matrices L and U as A=LU, where

$$L=\begin{pmatrix} 1 & & & & \\ h_1 & 1 & & & \\ & .. & 1 & & \\ & & .. & .. & \\ & & h_{n-2} & 1 & \\ & & & h_{n-1} & 1 \end{pmatrix}$$

$$U=\begin{pmatrix} u_0 & c_0 & & & \\ & u_1 & c_1 & & \\ & & . & . & \\ & & & . & . \\ & & & u_{n-2} & c_{n-2} \\ & & & & u_{n-1} \end{pmatrix}$$

The LU algorithm to solve the linear system (1) then proceeds to solve for y from Ly=d and then finds vector X in parallel:

Step 1. Compute the decomposition of A given by

$u_0 = b_0$,
$h_i = a_i / u_i\text{-}1$, $1 <= i <= n\text{-}1$,
$u_i = b_i - h_i * c_{i-1}$, $1 <= i <= n\text{-}1$,

Step 2. Solve for y from Ly =d using

$y_0 = d_0$,
$y_i = d_i - h_i * y_{i-1}$, $1 <= i <= n\text{-}1$.

Step 3. Compute X by solving ux = y using

$x_{n-1} = y_{n-1} / u_{n-1}$,
$x_i = (y_i - c_i * x_{i+1})/u_i$, $0 <= i <= n\text{-}2$.

First we consider the parallelization of the LU decomposition part of the LU algorithm to solve (1), i.e. Step 1 above. Once the diagonal entries $u_0,u_1,\ldots,u_n$ of U have been calculated, $h_1,h_2,\ldots,h_{n-1}$can subsequently be computed in a single parallel step with n-1 processors. Thus it concentrates on the computation of the u_i's

3. Parallel Algorithm to Calculate Sensitivity Function Using VP Algorithm

The application of high performance programming techniques for solution of Electric Power Systems problems has been increasing. Particularly, parallel processing present's very remising perspectives when heavy amputation is required. It may consist in a feasible alternative for solution of several large-scale problems, which are not well conditioned for a sequential approach. Despite its potentiality in engineering software development, parallel algorithm philosophy is quite different from that adopted by sequential programs. This work presents investigations regarding the application of parallel processing to calculate sensitivity function for a large dimension control system which we can write its mathematical model as a system of linear equations.

3.1 VPA Description

The main goal of **VP** algorithm is resolving in parallel linear equations which represents as AX=W ,and calculate sensitivity function of electric power systems to obtain the result with respect to variation any component of output function F with respect to any component of electric power systems h ($\partial f/\partial h$) . VP algorithm contains the next stages: distribution data(rows matrix A and components vector W) to the p processors where p= n (n is the number of rows ,m is the number of columns) which represents the mathematical model of electric system, and calculate in parallel unknown vector for origin system $X=(x_1,x_2,\ldots,x_n)^T$. Distribution data (at the same time) to p processors, and calculate unknown vector for similar system $(x^|)^t=-d^tT^{-1}=(x^|_1,x^|_2,\ldots,x^|_n)^T$. Multiplication operation for unknown $x_i \times x^|_i$ respectively using n-1 processors to find in parallel sensitivity function for a large dimension system.

3.2 Distribution data stage

In this stage, we defined vectors $v_1^0, v_2^0 \ldots v_n^0$. Figure 1, illustrates this stage.

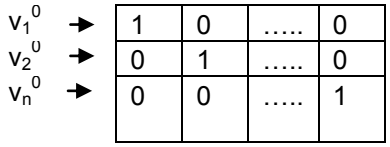


Fig. 1: defined vectors $v_1^0, v_2^0 \ldots v_n^0$.

Given A_1 first row matrix A, A_2 second row matrix A, and A_n last row matrix A with unknown vector w. Figure. 2, illustrates the mentioned above.

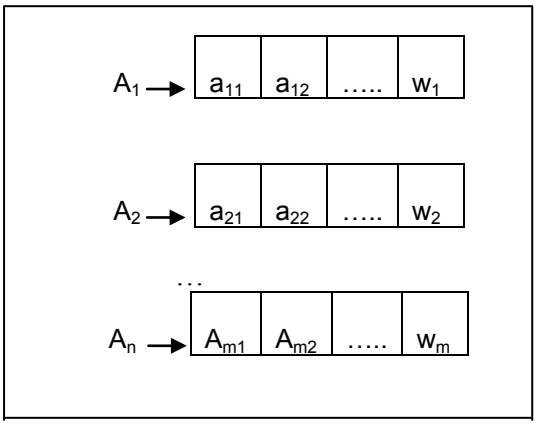


Fig. 2: Distribution rows stage matrix A with unknown vector w .

3.3 Multiplication and division Stages

In this stage we find unknown:
$C_2^1 = (A^1 * v_2^0) / (A^1 * v_1^0)$ and in parallel we calculate variable C until C_m :

$$C^{n-1}_m = A_{n-1} * V^1_m / A_{n-1} * V^1_{m-1}$$

And in parallel we find vector $V_2^1 \ldots V_m^1$:

$V_2^1 = V_2^0 - C_2^1 * V_1^0$

…….

$V_m^1 = v_m^0 - v_m^1 * v_1^0$

Finally we calculate the equations:

$C_m^{n-1} = A_{n-1} * V_m^1 / A_{n-1} * V_{n-1}^1$,

and find unknown $x_1, x_2 .. x_n$ for original system.
$V_m^{n-1} = V_m^1 - C_m^{n-1} * V_{n-1}^1$.

$V_m^{n-1} = (x_1, x_2, \ldots, x_n)^T$.

3.4 Distribution data for similar system

Distribute data to p= n-1 processors, and calculate unknown vector for similar system $(x^l)^t = -d^t T^{-1} = (x^l_1, x^l_2, \ldots, x^l_n)$, (we do that at the same time when we calculated unknown vector for original system $X = (x_1, x_2, \ldots, x_n)^T$ as mentioned above).

3.5 Calculate in parallel sensitivity function algorithm

Step 1.Compute unknown vector for similar system $X^l = (x^l_1, x^l_2, \ldots, x^l_n)$ using next equation :

$$(x^l)^t = -d^t T^{-1} \ldots\ldots\ldots\ldots\ldots\ldots (7)$$

Step 2.multiplicate equation (6) from the right side by matrix T and transpose left and right side to obtain a system with respect to X^l:

$$T^t x^l = d \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

Step 3 . Calculate:

$$\partial X/\partial h = -T^{-1} (\partial T/\partial h) X - (\partial W/\partial h) .. (9)$$

Step 4. Find sensitivity Function f with respect to h:

$$\partial f/\partial h = -d^t T^{-1} (\partial T/\partial h \, X - \partial W/\partial h) .. (10)$$

Step 5. Put the expression (7) in (10) then:

$$\partial f/\partial h = (x^l)^t \partial T/\partial h X - (x^l)^t W/\partial h \ldots\ldots (11)$$

To implement the expression (11) we just need to resolve in parallel the tow linear systems (1) and (8) by using VP algorithm.

4. A numerical Example

Figure. 3, illustrates the electric circuit, in which we want to calculate in parallel the sensitivity function of the output potential v_{out} with respect to resistance g_2, condensers c_1, and c_3, respectively, the mathematical model for this circuit is :

$$\begin{bmatrix} G_1+G_2+{}_sC_1+ {}_sC_2 & G_2 - {}_sC_2 \\ G_2 - {}_sC_2 & G_2+G_3+{}_sC_2+ {}_sC_3 \end{bmatrix} \times \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Using VP algorithm in parallel, we find unknowns vector X for original system:

$$x = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} (3 - j) /5 \\ (2+j) /5 \end{bmatrix}$$

At the same time we find unknowns vector $X^{|}$ for similar system :

$$x^{|} = \begin{bmatrix} v^{|}_1 \\ v^{|}_2 \end{bmatrix} = \begin{bmatrix} -(2+j)/5 \\ (-3+j)/5 \end{bmatrix}$$

Finally we just do the multiplication operation to find the sensitivity function as follows:

$\partial v_{out} / \partial c_1 = {}_s v^{|}_1 v_1 = 1\text{-}j7/25,$

$\partial v_{out} / \partial G_2 = (v^{|}_1 - v^{|}_2)(v_1 - v_2) = -3\text{-}j4/ 25,$

$\partial v_{out} / \partial c_3 = {}_s v^{|}_2 v_2 = 1\text{-}j7/25.$

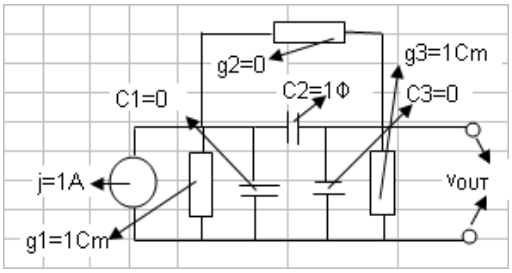


Fig. 3: electric circuit to calculate sensitivity function for v_{out} with respect to variation parameters (C1, G_2, C3).

5. RESULTS

To calculate the accurate time and performance we repeat the process m times then we divide the measured time on m for both single and multi-thread versions, for single thread we start basic multiplication division and subtraction inside the Matrix until we get the upper of that matrix, for multi-threading we use R-1 threads where R is the count of desired matrix rows, we measured the longest thread which is the last one in this present case, then every thread take a part of the matrix basic operations and we do that in parallel for origin and similar systems.
Table 1 shows the time results done on Pentium Due 1.8 GHZ processor with 1 GB Ram and shows the time when used one processor (single thread) and the time when used a multi processors in parallel (multi thread) to calculate the unknown vector. From the table 1, figure 4 and 5 show that time and performance is increased with respect to the size of matrix, which represents the linear system.

Table 1: Comparison between single and multithread

performance	Multi thread, MS	Single thread, MS	Matrix Dimension
1.25	0.000002	0.000005	1x2
6.933	0.00001	0.000105	2x3
13.741	0.000025	0.000325	3x4
29.83	0.000033	0.000809	4x5
53.267	0.000027	0.001718	5x6

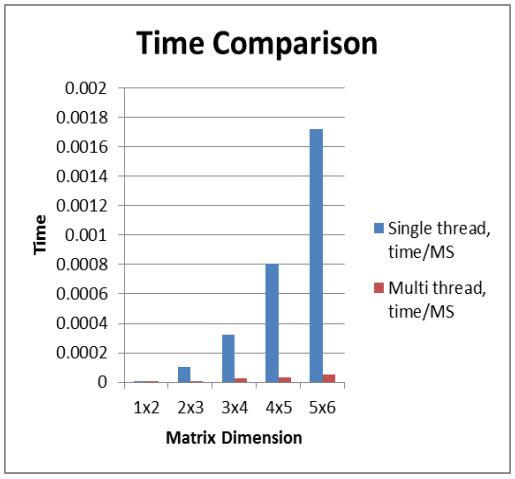


Fig 4: Time comparison between single and parallel to calculate unknown's vector

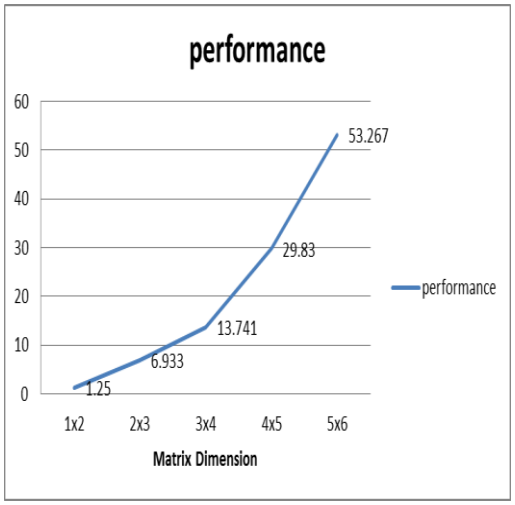


Fig 5: System performance with respect to matrix dimension.

6. CONCLUSION

One of the most important fields in sciences and techniques is to calculate a sensitivity function for large dimension control systems, this research demonstrates new parallel algorithm to do that, the idea of this algorithm is to distribute the coefficient mathematical model studies system to a number of processors and use the parallelism to reduce the running time for solving linear equations of a big size by using original and similar systems which working in parallel, and that is confirmed in the description of VP algorithm , The running time is reduced by O(t/2) and, The efficiency of VP algorithm is increased by 50-60%.

REFERENCES

[1] Duff, I.S. and H.A. van de Vorst, 1999.Developments and Trends in parallel Solution of Linear Systems. Technical Report RAL TR-1999-027.

[2] Ogita, T, S. M. Rump, and S. Oishi, 2005.Accurate Sum and Dot Product.SIAM Journal on Scientific Computing, 26(6):1955-1988.

[3] Klatte, R., U.Kulisch, and A. Wiethoff.1993.C-XSC-A C++ Class Library for Extended Scientific Computing.Spriger-Verlag.

[4] Duff, I.S., 1999.The Impact of High Performance Computing in the Solution of Linear Systems: Trend and Proplems.Technical Report RAL TR-1999-072.

[5] Facius, A., 2000.Iterative solution of linear systems with improved arithmetic and result verification .PhD thesis, University of Karlsruhe.

[6] Bohlender, G., 1990.What Do Need Beyond IEEE Arithmetic? Computer Arithmetic and Self-validation Numerical Methods.Academic Press Professional,Inc.,San Diego,CA.

[7]Eisentat,S.C.,M.T.Heath,1988.Modified cyclic algorithm for solving triangular system on distributed-memory multiprocessor.SIAM J.Stat.Comput.,9(3):589-600.

[8]Holbig,c.a.,P.S.Morandi,2004.Self verifying Solvers for Linear Systems of Equations in C-XSC.In Proceeding of Parallel and Distributed Programming (PPAM),volume 3019,pages 292-297.

[9] Cunha, R.D. and T.Hopkins, 1991.The parallel solution of triangular systems of linear equations. Technical Report 86*, University of kent, Cantenbary, UK.

[10] Saad, Y., 1995.Iterative Methods for Sparse Linear Systems.Boston:PWS Publishing Company.

[11] Feng, T., 2002.A Message-Passing Distributed-Memory Newton-GMRES Parallel Power Flow Algorithm. Volume 3, pages 1477-1482.

[12] Heath, J.W., 1993.Parallel Numerical Linear Algebra.Technical Report UCB CSD-92-703.
[13] Hedayat, G.A., 1993.Numerical Linear Algebra and Computer Architecture. Technical Report UMCS-93-1-5.
[14] Lo, C.G. and D.W.Cheunge, 1997.Efficient Parallel Algorithm for Dense Matrix LU Decomposition with Pivoting on Hypercubes.Computer & Mathematics with Applications, 33(8):39-50.

Hamed Khaled Alrjoub
Alrjoub is an assistant professor in computer science department, deanship of preparatory year, Um Alqura University/ Saudi Arabia. He was a head of computer science department, Irbid national university / Jordan. He got a PhD from international civil aviation university, Kiev / Ukraine, his area of interest is Parallel processing, E-commerce, e-government, system analysis, database, data mining, machine learning.

Multiple Servers - Queue Model for Agent Based Technology in Cache Consistence Maintenance of Mobile Environment

G.Shanmugarathinam[1], Dr.K.Vivekanandan[2]

[1] Research Scholar(external mode) in Bharathiar university Coimbatore, Tamil Nadu,India

[2] Professor, Bharathiar university Coimbatore, Tamil Nadu,India

Abstract

Caching is one of the important techniques in mobile computing. In caching, frequently accessed data is stored in mobile clients to avoid network traffic and improve the performance in mobile computing. In a mobile computing environment, the number of mobile users increases and requests the server for any updation, but most of the time the server is busy and the client has to wait for a long time. The cache consistency maintenance is difficult for both client and the server. This paper is proposes a technique using a queuing system consisting of one or more servers that provide services of some sort to arrive mobile hosts using agent based technology. This services mechanism of a queuing system is specified by the number of servers each server having its own queue, Agent based technology will maintain the cache consistency between the client and the server .This model saves wireless bandwidth, reduces network traffic and reduces the workload on the server. The simulation result was analyzed with previous technique and the proposed model shows significantly better performance than the earlier approach.

Keywords: *mobile database, wireless networks, database cache, Queuing model.*

1. Introduction

Mobile Computing uses portable computing devices such as laptops, PDAs and wearable computers. Example applications used in mobile computing environment include sales force automation, order entry, e-mail, calendar management, financial and news services, insurance companies, emergency services (police, medicals), traffic control, taxi dispatch, etc.

A mobile computing environment is a distributed system, thus when data at the server changes, the client hosts must be made aware of this fact in order to invalidate their cache, otherwise the host would continue to answer queries with the cached values returning incorrect data.

Mobile computing has stringent constraints in network resources, such as bandwidth and connectivity. As such, data in mobile applications are often cached at clients to increase performance, data availability and reliability. Most fault-tolerant schemes for wireless sensor networks focus on power failures or crash faults. Little attention has been paid to the data inconsistency failures

Recent advances in wireless and mobile networks have led to the exponential growth of mobile applications although a number of studies have been made in this subject, few researchers focused on mobile data access. In this design a node as middle server (MS). It is between the server and client. Whenever server data was updated immediately synchronization starts with Middle server and the client. Some of the clients wake up from sleep mode immediately request the Middle server for the updated data and need not request the server. So it reduces the work load in the main server database.

2. Related work

2.1 Updated Invalidation Report (UIR) based cache techniques.

In this approach [1] Invalidation Report based cache management is an attractive approach for mobile environments. In this approach the server periodically broadcasts an IR in which the changed data item are indicated. Since IR arrive periodically, client can go to sleep most of time and only wake up when the IR comes. It has some drawbacks such as long query latency and low hit ratio. There is a long latency problem with a UIR (Updated Invalidation Report) based approach, where a small fraction of the essential information related to cache invalidation is replicated several times within an IR interval, and hence the client can answer a query without

waiting until the next IR. If there is a cache miss the client still needs to wait for the data to be delivered.
It has some drawbacks such as long query latency and low hit ratio. There is a long latency problem with a UIR (Updated Invalidation Report) based approach.

2.2 Adaptive Energy Efficient Cache Invalidation Scheme

In this approach [2] to reduce the bandwidth requirement, the server transmits in one of three modes slow, fast and super-fast .The mode is selected based on thresholds specified for time and the number of clients requesting updated objects.
The mode is selected based on thresholds specified for time and the number of clients requesting updated objects. The updating is less in server if the mode changes to slow, so the client has to wait for long time to utilize cache data during invalidation report

2.3 Smart Server Update Mechanism (SSUM)

In this approach [3] SSUM the server send data updates to the cache node (CN). Request Mobile host that desires a data items sends its request to its nearest query director (QD). If this QD finds the query in its cache, it forwards the request to the CN caching the items, which in turn sends the item to the requesting mobile host. Otherwise it forward it to its nearest QD, If the request traverses all QDs without being found, a miss occurs and it gets forwarded to the server which sends the data item to the request mobile host
In this case, latency is more as well as bandwidth is wasted because of this the client should wait idle for the server to reply.

3. Queuing model for Agent based technology

In this paper proposed Queuing model for Agent based technique for cache consistency for wireless network .In our design does not required to produce an Invalidation report, In mobile computing the mobile user increase and request the server for any updating but most of the time server is busy and client to wait for long time .so we design the node as middle server using between server and client, thread agent maintain a log and thread synchronization in client and server, maintain the cache consistency .
The following subsection describe the proposed algorithm in detail

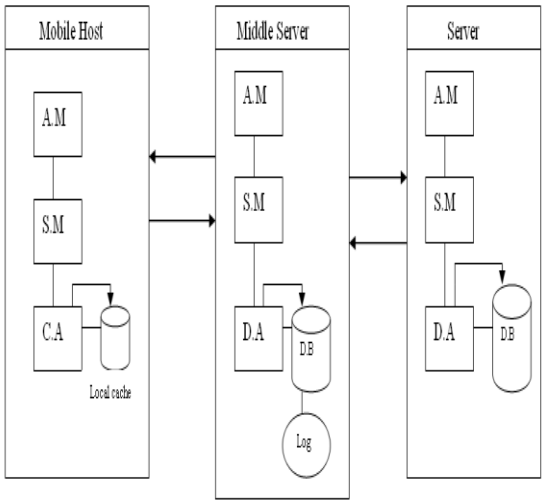


Fig 1: System Architecture

AM –Agent manager

S.M- Security module

C.A-Cache Agent

D.A-Database agent

DB-Database

3.1 Middle server with Agent based technique

Some of client data cache is placed in Middle server. To find the middle server it is based on the network bandwidth, speed of CPU, lower latency , cache hit
Ratio The middle server are near to the client, so the communication cost, energy consumption are very less, easy to update the cache data and easily maintain Consistency. If Data requested is not available in local cache, the client send the broadcast request to the middle server . Middle server receive the packet , search in the cache and send the acknowledge data to the client. The Middle servers satisfy the nearest client request. Advantage is low cost for communication and reduces the network traffic in mobile networks. The Middle server log and Agent synchronization Model maintains consistency between the Server data and Middle server. For each cached data object uses log to maintain consistency between Server, Middle server and Mobile client. When a data $d_{x\ is}$ retrieved by a mobile client log is created to indicate data is valid or not.
If and when the Server receives an updated data object d_x it broadcasts and synchronizes with Agent Manager of client to make cache data object reliable. During this process a log maintained in server is compared with recent log of Middle server If so there is a need of Updation, it processes to perform update function(s). In mobile environments a Mobile Cache is one of two states. (i) Awake or (ii) Sleep. If a Mobile Client is awake an

internal request is shared between Agent Manager at Middle server and Agent Manager at client to ensure that data object is updated. If there is an Updation the SynchM of server synchronizes with SynchM of Middle server and client in order to make as valid data object.
The data objects of a Mobile Client in the sleep state are unaffected until it wakes up. When a mobile client wakes up a new Agent upon is created which holds last accessed log, this log passes to the Middle server, On receiving upon the log it compares with previous log maintained by it. If it is invalid data cache the Agent Manager of middle sever cache starts the synchronizes with client for the updating data.

3.2 Queuing system

A queuing system consist of one or more server that provide services of some sort to arriving .Every day examples can be described as queuing systems ,such Computer system , manufacturing systems ,maintenance systems, communication system. This services mechanism of a queuing system is specified by the number of server each server having its own queue or common queue. The essence of queuing theory is that it takes in to account the randomness of the arrival process and the randomness of the service process. The word randomness refers to the probability distribution used in the arrival as well as service process.
One of the most important queuing models is the Evlang loss models, it assumes that arrivals follows a passion process and that the blocked Mobile host (those who find all server busy) are cleared (that is they are denied entry in to the system so the blocked Mobile Host are lost)

The probability of blocking is Evlang B formula given by

$$B(s,a) = \frac{\frac{a^s}{s!}}{\sum_{k=0}^{s} \frac{a^k}{k!}} \qquad (1)$$

$$a = \lambda t$$

Where "S" is the numbers of server
"a" is equal to $\lambda\tau$ where
"λ" is arrival rate and
τ is the average services time

B(1,0.8)=0.444
S0 44.44% of the arrivals will be blocked
Similarly increase the number of server s the percentage of blocking can be computed from Evlang B formula . but when s and a are large it is hard to calculate directly so interactive scheme or formula is designed as follow :

$$B(n,a) = \frac{a\ B\ (n-1,a)}{n + a\ B\ (n-1,a)} \qquad (2)$$

$$B(10,8) = \frac{B(9,8)}{10 + 8B(9,8)}$$

$$B(10,8) = \frac{8^{10}}{10!} \div \sum_{k=0}^{10} \frac{8^k}{k!}$$

$$B(10,8) = \frac{8^{10}}{10!} \div e^8$$

$$B(10,8) = 0.1217$$

$$B(10,8) = 0.1217$$

This means that when 8 Evlangs of passion traffic is offered to 10 server then about 12 % of the arrivals will be blocked . since the blocked mobile host are cleared from the system .

Table 1 : Shows the Blocking percentage of Mobile Host

No of Server	Blocking Percentage
1	44.4
2	15.09
3	3.86
4	0.76
5	0.12
6	0.016
7	0.0018
8	1.8696652894159525E-4
9	1.6619244255037438E-5
10	1.3295395227262418E-6

4. Performance Evaluation and Results Discussion

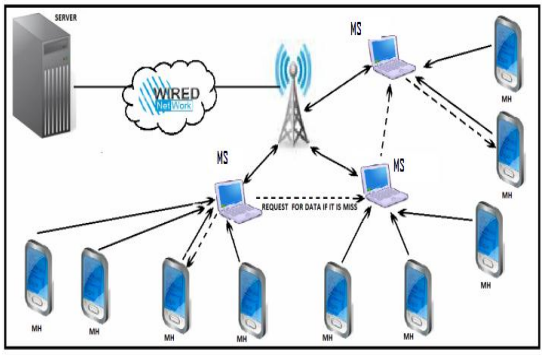


Fig 2: Simulation for Agent Based Model

In simulation created a Mobile Sales, online distributor of clients for interact with a sales system. The workload consists of a set of interactions: number of clients can be added , create new orders, change payment types, check on the status of previous orders, view new products the distributor might have recently added, look at detailed descriptions of products, and make changes to the product log. Each client's test run is defined by a time period that specifies the length of execution. The proposed algorithm compare with SSUM algorithm and the result shows the reduce the average access latency (seconds) .

Fig 3: simulation of mobile Host

Fig 4: simulation of mobile Host

ID	Symbol	ISIN	CPrice
1	BHEL	INE257A01026	789.23
2	AXISBANK	INE238A01026	1070.70
3	HCLTECH	INE860A01027	570.55
4	TATASTEEL	INE081A01012	404.54
5	TCS	INE467B01029	950.56

Fig 5 : Middle server Database

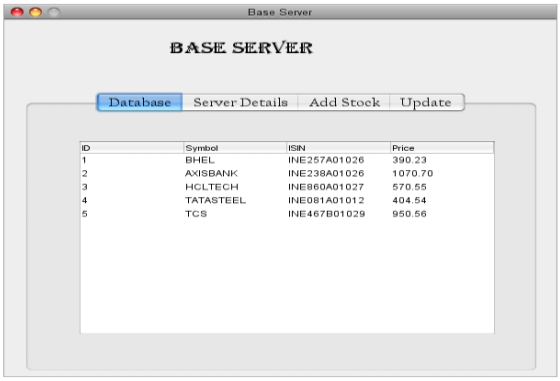


Fig 6: Base Server Database

Client_Id	Symbol	Price	Updated
3	BHEL	420.34	20/09/2012 01:20:00 PM
1	BHEL	420.34	20/09/2012 01:20:00 PM

Fig 7: Middle server Log

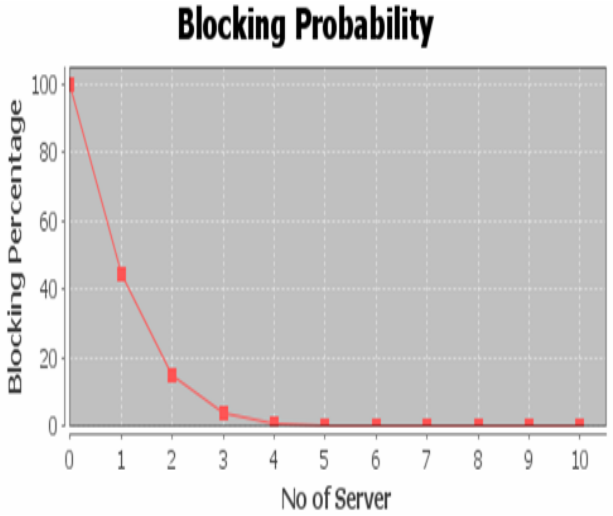
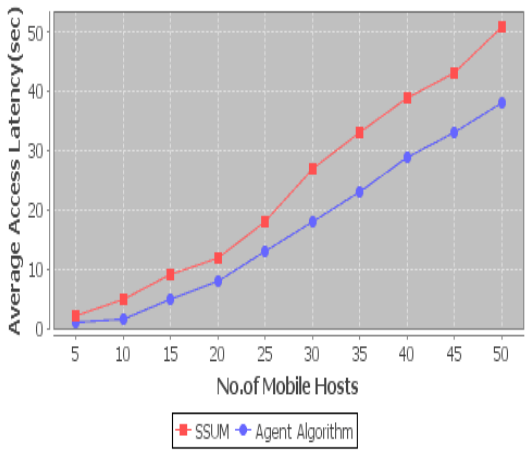


Fig 8: Number of server vs Blocking percentage

Above the graph it shows increase the number of servers, the blocking percentage of server services was reduced



.

Fig 9 : Comparing Agent Algorithm with SSUM Algorithm

5. Conclusion

In this paper, the proposed technique of multiple server queuing system reduces the blocking percentage of server services. key features as stated earlier, are low cost for communication, using a Agent based Algorithm which is easy to maintain the cache consistency and reduction of network traffic . Simulation results shows that the Agent based algorithm gives a significantly better performance than earlier approaches.

Reference

[1] "Niraj tolia and adam wolbach, improving mobile a database Access over wide area networks without degrading consistency ACM 2006

[2] "Alok Madhukar, Reda Alhajj, "An Adaptive Energy Efficient Cache Invalidation Scheme for Mobile Databases", ACM 2006, April 23-27, 2006

[3] "Khaleel Mershad ,Hassan Artail, "Smart Server Update Mechanism for Maintaining Cache consistency in Mobile Environments" IEEE Transaction on Mobile Computing Vol 9, No 6 June 2010

[4] Huaping Shen, Mohan Kumar, Sajal K.Das, "Energy Efficient Caching and Prefetching with Data Consistency in Mobile Distributed Systems", 0-7695-2132-0/04/2004, IEEE.

[5] Yu Huang, Jiannong cao, Beihong Jin, Xianping Tao. "Flexible Cache consistency Maintence over Wireless Adhoc Networks" IEEE Transaction on Parallel and Distributed System Vol 21,No.8,August 2010

[6] C.-Y. Chang and M.-S. Chen, "Exploring Aggregate Effect with Weighted Transcoding Graphs for Efficient Cache Replacement in Transcoding Proxies", Proc. 18th Int'l Conf. Data Eng., Feb. 2002.

[7] Elsen, F. Hartung, U. Horn, M. Kampmann, and L. Peters, "Streaming Technology in 3G Mobile Communication Systems", Computer, 34, No. 9, pp. 46-52, 2001.

[8] S. Hosseini-Khayat, "On Optimal Replacement of Nonuniform Cache Objects", IEEE Trans. Computers, 47, No. 4, pp. 445-457, 2000.

[9] T. Imie linski and B.R. Badrinath, "Wireless Graffiti— Data, Data Everywhere Matters", Proc. 28th Int'l Conf. Very Large Data Bases, 2002.

[10] A. Kahol, S. Khurana, S.K.S. Gupta, and P.K. Srimani, "A Strategy to Manage Cache Consistency in a Distributed Disconnected Wireless Environment", IEEE Trans. Parallel and Distributed System, 12, No. 7, pp. 686-700, 2001.

[11] Mark Kai Ho Yeung, Yu-Kwong Kwok, "Wireless Cache Invalidation Schemes with Link Adaptation and Downlink Traffic", IEEE Transaction On Mobile Computing, 4, No 1, January/February 2009

First Author : G. Shanmugarathinam has been teaching in Engineering College for more than 12 years. He is a Research Scholar(external mode) in Bharathiar university in Coimbatore, Tamil Nadu,India He has been the author of various books and papers which he has published more than 5 articles in international journals. His fields of interest are Mobile Computing, Cisco networking, Management of information system, wireless network and network security.

Second Author : K.Vivekanandan is a Professor , school of management at Bharathiar University, He has 5 years industry experience and 26 years of teaching Experience .His research interests include Management Information System, E-marketing, Quantitative Methods , Data Mining and mobile computing. He has published more than 30 articles in national/international journals and he has successfully guided 16 PhD's and is currently guiding 7 students.

A Novel Robust Backstepping Control for Nonaffine Nonlinear Processes and Application to An Active Magnetic Bearing System

Shuanji Zhang[1], Yida Jiang[1] and Dezhi Xu[2]

[1] Physics and Electronic Information School, Luoyang Normal University, Luoyang, 471022, Henan, China

[2] College of Automation Engineering, Nanjing University of Aeronautics and Astronautics Nanjing, 210016 , Jiangsu, China

Abstract

In this paper, a novel continuous-time robust nonlinear control scheme, is proposed for nonaffine nonlinear systems with unknown uncertainties/disturbance, which is based on backstepping methodology and sliding mode control technique. Firstly, a novel approximation technique is developed to approximate the nonaffine nonlinear dynamic systems. Then, the robust backstepping control for nonaffine nonlinear systems is proposed via the novel approximation technique. In the controller design procedure, the sliding model control method is introduced to avoid the possibility of the over parameterization problem and deal with the unknown uncertainties/disturbance. And, a second-order sliding mode integral filter is employed to facilitate the development of the derivation of the virtual control input with uncertainty terms included. Finally, the designed robust control strategy is applied to three-pole active magnetic bearing system, and simulation results are provided to demonstrate the effectiveness of the theoretic results obtained.

Keywords: *Nonaffine Nonlinear Systems, Robust Backstepping Control, Active Magnetic Bearing System, Simulation.*

1. Introduction

In the past decade, there has been significant progress in the area of control design for nonlinear plants. Isidori [1] developed important results related to the geometric approach for analysis and control design of nonlinear plants. An overview of available nonlinear control techniques is given by [2]. Many of these results have been extended to the case of nonlinear plants with uncertainty [1-2]. Up to now, few research articles related to the nonaffine nonlinear systems [3-4], in addition to the intelligent control algorithms. However, intelligent control algorithms require a lot of expertise or modeling data. For some plants, expertise or modeling data is not easy to obtain.

The problem of controlling the plants characterized by models that are nonaffine in the control input vector is a difficult one [5]. An approach widely used in practice is that based on linearization of the nonlinear plant model around an operating point. In some controls, the nonlinear model of dynamics is generally nonaffine in input u and is commonly linearized around a trim point, that is, an operating point dependent on the current states.

Based on the aforementioned works, this paper develops a novel robust backstepping control (NRBC) methodology for nonaffine nonlinear dynamic systems. A novel approximation technique is firstly employed to approximate the nonaffine nonlinear dynamic system. Then, based on backstepping control, NRBC is proposed. However, in the NRBC controller design procedure, the sliding model control technique is introduced in the backstepping procedure so as to develop an easy-implemented controller, as well as to avoid the possibility of the overparameterization problem and deal with the unknown uncertainties/disturbance. And, a second order sliding mode integral filter is introduced to facilitate the development of the derivation of the virtual control input with uncertainty terms included. Finally, the proposed strategy is also applied to three-pole AMB system suffering from unknown uncertainties/disturbance. The tracking performances of three-pole AMB system could also be well guaranteed.

2. Problem Formulation

Consider the nonaffine nonlinear MIMO system which is represented by the following set of differential equations:

$$\dot{x}_1 = x_2 \tag{1}$$

$$\dot{x}_2 = f[\bar{x}(t), u(t)] + d \tag{2}$$

where $x_1 \in \Re^{n_1}$ and $x_2 \in \Re^{n_2}$ are the state vectors, and $\bar{x}(t) = (x_1, x_2)^T \in \Re^{n_1+n_2}$, $u(t) = (u_1, u_2, \cdots, u_m)^T \in \Re^m$ is the input vector of the system, respectively. $f = (f_1, f_2, \cdots, f_n)^T$, $f_i : \Re^n \times \Re^m \to \Re^1$ are known smooth nonlinear functions whose first derivatives with

respect to $\bar{x}(t)$ and $u(t)$ exist. $d \in \Re^{n_2}$ denotes the function uncertainty with $\|d\| \le \psi$, which is due to the modeling errors and external disturbances.

Most of the nonlinear control methods developed in this context are applicable to nonlinear plant models that are linear in unknown parameters and affine in the control input vector u, that is, characterized by appearing linearly in the equation. However, for nonaffine nonlinear MIMO system, the problem of controlling the plants characterized by models that are nonaffine in the control input vector is difficult one. Without any effective methods to solve this problem. One nonlinear approach to this problem is that based on directly inverting the nonlinear function of on domain. Although the existence of an inverse function can be guaranteed by the implicit function theorem [2], it is generally difficult to prescribe technique to actually obtain such an inverse. However, in the proposed NRBC, such time consuming algorithms are totally avoided and thus the controller design is greatly simplified. Further speaking,for the continuous time nonaffine nonlinear systems, robust control research has not been studied.

In order to convenient unfold the following work, short assumption is given as following

Assumption 1: The input vector u of the system must be measurable or available.

3. Novel NRBC Algorithm Nonaffine Nonlinear Dynamic Systems

A novel NRBC algorithm is proposed here using newly developed nonaffine nonlinear approximation for continue -time systems, which not only avoids complex control development and intensive computation, but also overcomes the shortcomings of other existing methods. unique and rigorous stability proof will be given and its superior performance will be demonstrated in later simulations.

3.1 Novel Nonaffine Nonlinear Approximation

For the nonaffine nonlinear model (1), the Taylor expansion of the nonlinear function $f[\bar{x}(t),u(t)]$ with respect to $u(t)$ around $u(t-\tau)$ can result in

$$\dot{x}_1 = x_2 \tag{3}$$

$$\dot{x}_2 = f(\bar{x}(t),u(t-\tau)) + f_d(\bar{x}(t),u(t-\tau))\Delta u(t) + R_p + d \tag{4}$$

where

$$\Delta u(t) = \begin{bmatrix} u_1(t)-u_1(t-\tau) \\ u_2(t)-u_2(t-\tau) \\ \cdots \\ u_m(t)-u_m(t-\tau) \end{bmatrix}$$

$$f_d(\bar{x}(t),u(t-\tau)) = \left.\frac{\partial f(\bar{x}(t),u(t))}{\partial u(t)}\right|_{u(t)=u(t-\tau)}$$

$$f_{dd} = \left.\frac{\partial^2 f(\bar{x}(t),u(t))}{\partial^2 u(t)}\right|_{u(t)=\zeta}$$

$$R_p = \frac{[\Delta u(t)]^T G_{dd} \Delta u(t)}{2}$$

$\zeta = [\zeta_1,\zeta_2,\cdots,\zeta_m]^T$ with ζ_j being a point between $u_j(t)$ and $u_j(t-\tau)$. Let $0 \le \|f_{dd}\| \le r_p, r_p$ is finite positive number, thus $\|R_P\| \le \frac{r_p\|\Delta u(t)\|^2}{2}$. The parameter $\tau > 0$ is the updating input, It may be chosen as the sampling-time in sampled-data control system, or as an integer multiple of the sampling-time. A better choice of the parameter τ is the sampling because a larger τ may lead to an inaccurate approximation when the system function $f[\bar{x}(t),u(t)]$ varies quickly.

It is easy that (3)-(4) can be representation as the following form.

$$\dot{x}_1 = x_2 \tag{5}$$

$$\dot{x}_2 = f_n(\bar{x}(t),u(t-\tau)) + f_d(\bar{x}(t),u(t-\tau))u(t) + d_\xi \tag{6}$$

Where $d_\xi = R_p + d$, and

$$f_n(\bar{x}(t),u(t-\tau)) = f(\bar{x}(t),u(t-\tau)) - f_d(\bar{x}(t),u(t-\tau))u(t)$$

To approximation accuracy, control input must satisfy the following assumption.

Assumption 2: $|\Delta u(t)| \in [0,\delta]$ and $0 < \left|\frac{\partial f}{\partial u(t)}\right| \le \beta$, δ and β are two finite positive vectors.

In Assumption 2 : $0 < \left|\frac{\partial f}{\partial u(t)}\right| \le \beta$ means that the system (1) has a well defined relative degree [4]. $|\Delta u(t)|$ should not be too large in order to limit the approximation error of the model (5)-(6) for a computed u(t). In many actual process control systems and flight control systems, $|\Delta u(t)| \in [0,\delta]$ is a physical restriction of many practical systems because their states and outputs (actuators) cannot change too fast because of system 'inertia'.

Remak 1: If there is control input saturation constraints, $u(t-\tau)$ must be the actual control input of τ times before,

rather than control input command of τ times before.

3.2 Nonlinear Controller and Stability Analysis

In this subsection, based on above proposed novel non-affine nonlinear approximation algorithm, the robust back-stepping procedure and the sliding model control techniques are introduced so as to develop a NRBC, whose function is to track the reference signal with an acceptable accuracy. The following assumptions are used in the design and analysis procedure.

Assumption 3:: The reference signal x_{1d} , virtual input x_{2d} and their first order derivatives are piecewise continuous and bounded, they are

$$\|x_{1d}\| \le \|x_{1d}\|_{\max} = \Delta_1$$
$$\|x_{2d}\| \le \|x_{2d}\|_{\max} = \Delta_2$$
$$\|\dot{x}_{1d}\| \le \|\dot{x}_{1d}\|_{\max} = \dot{\Delta}_1$$
$$\|\dot{x}_{2d}\| \le \|\dot{x}_{2d}\|_{\max} = \dot{\Delta}_2$$

Moreover, the function uncertainty is assumed to be unknown. However, the upper boundary of its magnitude is known as

$$\|d_{\xi}\| \le \Psi_{\xi}$$

Before we start, respectively, the state tracking error variables e_1 and e_2 as follows

$$e_1 = x_1 - x_{1d} \tag{7}$$
$$e_2 = x_2 - x_{2d} \tag{8}$$

where x_{1d} and x_{2d} are the desired trajectories of x_1 and x_2 , respectively. Note that x_{1d} is given by command signals and x_{2d} will be defined later. From (5) and (7), we have

$$\dot{e}_1 = \dot{x}_1 - \dot{x}_{1d} = x_2 - \dot{x}_{1d} \tag{9}$$

We further assume that x_{2d} is the virtual input to (9), and the desired virtual control is

$$x_{2d} = -E^{-1}(C_1 e_1 - \dot{x}_{1d}) \tag{10}$$

where C_1 is a designed positive diagonal matrix, and E is a unit matrix.

As stated previously, with the inclusion of the uncertainty or disturbance in the virtual input (10), it is difficult in finding, its derivatives because the signal may not be practically differentiable due to noises and/or disturbances, and the problem of overparameterization will occur with the increase of steps as well.

In view of this, a second-order sliding model integral filter is presented in this paper so as to eliminate the analytic computation of $\dot{x}_{2d}$, which will be used as reference in the backstepping procedure. It is worth stressing that the proposed filter works also for the high-order backstepping procedures, just using the output of the $(i-1)$th filter as the input to the ith filter, $i = 1,2,\cdots,n.$ The proposed integral filters are presented as follows

$$\dot{\hat{\lambda}}_1 = -\frac{\hat{\lambda}_1 - x_{2d}}{\varepsilon_1} - \frac{Q_1(\hat{\lambda}_1 - x_{2d})}{\left\|\hat{\lambda}_1 - x_{2d}\right\| + \varsigma_1} \tag{11}$$

$$\dot{\hat{\lambda}}_2 = -\frac{\hat{\lambda}_2 - \dot{\hat{\lambda}}_1}{\varepsilon_2} - \frac{Q_2(\hat{\lambda}_2 - \dot{\hat{\lambda}}_1)}{\left\|\hat{\lambda}_2 - \dot{\hat{\lambda}}_1\right\| + \varsigma_2} \tag{12}$$

where ε_i is the time constant of the filter, Q_i and ς_i are the designed constants, $i = 1,2.$

Obviously, with Q_i assumed to be zero, the proposed filters are reduced to a classical integral filters. It should be pointed out that, with the inclusion of the sliding model control component, the fast convergence of the estimation error produced by the proposed integral filters is guaranteed, which will be analytically studied during the stability analysis. Similar integral filters associated with different control schemes can also be found in varies applications [6-7], and the performance demonstrates their feasibility within the backstepping procedure.

Let us take x_2 as the virtual control variable of x_1 -subsystem, and select $x_{2d} \approx x_2$ as the ideal control input. It is noted that, in this step, the task is to stabilize (7) with respect to the Lyapunov function.

$$V_1 = \frac{1}{2} e_1^T e_1 + \frac{1}{2}(\hat{\lambda}_1 - x_{2d})^T(\hat{\lambda}_1 - x_{2d}) \tag{13}$$

Obviously, the third term in (13) is used to stabilize the estimation error of the proposed filters. Consequently, evaluating its time derivative along the solutions of the system(9), results

$$\dot{V}_1 = e_1^T \dot{e}_1 + (\hat{\lambda}_1 - x_{2d})^T(\dot{\hat{\lambda}}_1 - \dot{x}_{2d}) \tag{14}$$

Substituting (9), (10) and (11) into (14) yields

$$\dot{V}_1 \le -C_1\|e_1\|^2 - \left\|(\hat{\lambda}_1 - x_{2d})^T\right\|\left(\frac{Q_1(\hat{\lambda}_1 - x_{2d})}{\left\|(\hat{\lambda}_1 - x_{2d})\right\| + \varsigma_1} - \|\dot{x}_{2d}\|\right) \tag{15}$$

According to Assumption 3, the parameter Q_1 can be designed as $Q_1 = \vartheta_1 \dot{\Delta}_2$, where $\vartheta_1 > 1$. Hence, we have

$$\dot{V}_1 \le -C_1\|e_1\|^2 - \left\|(\hat{\lambda}_1 - x_{2d})^T\right\|\dot{\Delta}_2\left(\frac{\vartheta_1(\hat{\lambda}_1 - x_{2d})}{\left\|(\hat{\lambda}_1 - x_{2d})\right\| + \varsigma_1} - 1\right) \tag{16}$$

Apparently, if $\lambda_1 - x_{2d} \neq 0$ and the following relation is satisfied, the time derivative of the Lyapunov function will be rendered to negative

$$\left\|\lambda_1 - x_{2d}\right\| > \frac{\varsigma_1}{\vartheta_1 - 1}$$

With the preceding condition, the system will be bounded stable at the origin (i.e., $e_1 = 0, e_2 = 0$), and also, with such condition, the actual estimation error of the proposed filter can be guaranteed within a compact set determined in the form:

$$\left\|\lambda_1 - x_{2d}\right\| > \frac{\varsigma_1}{\vartheta_1 - 1} \tag{17}$$

Obviously, the estimation error of the filter can be adjusted sufficiently small by choosing ς_1 appropriately, and with the inclusion of the sliding mode control component, (17) can be arrived in finite time.

Next, from (8), we have

$$\dot{e}_2 = f_n(\bar{x}(t), u(t-\tau)) + f_d(\bar{x}(t) - u(t-\tau))u(t) + d_\xi - \dot{x}_{2d} \tag{18}$$

The candidate Lyapunov function in this case is defined as

$$V_2 = V_1 + \frac{1}{2}e_2^T e_2 + \frac{1}{2}(\hat{\lambda}_2 - \dot{\hat{\lambda}}_1)^T(\hat{\lambda}_2 - \dot{\hat{\lambda}}_1) \tag{19}$$

Then the time derivative of V_2 is given by

$$\dot{V}_2 = \frac{\partial V_2}{\partial e_1}\dot{e}_1 + \frac{\partial V_2}{\partial e_2}\dot{e}_2 + (\hat{\lambda}_2 - \dot{\hat{\lambda}}_1)^T(\dot{\hat{\lambda}}_2 - \ddot{\hat{\lambda}}_1) \tag{20}$$

Consequently, the parameter Q_2 can be designed as $Q_2 = \vartheta_2\Delta_3$, where $\vartheta_2 > 1$. We have

$$\dot{V}_2 \le -C_1\|e_1\|^2 + e_1^T e_2 + e_2^T(f_n(\bar{x}, u_{-\tau}) + f_d(\bar{x}, u_{-\tau})u + d_\xi - \dot{x}_{2d}) - \left\|(\hat{\lambda}_2 - \dot{\hat{\lambda}}_1)^T\right\|\Delta_3\left(\frac{\vartheta_2(\hat{\lambda}_2 - \dot{\hat{\lambda}}_1)}{\left\|(\hat{\lambda}_2 - \dot{\hat{\lambda}}_1)\right\| + \varsigma_1} - 1\right) \tag{21}$$

where $\left\|\dot{\hat{\lambda}}_1\right\|_{\max} = \Delta_3$. Then, then we can design the global NRBC law as

$$u(t) = -f_d^{-1}(\bar{x}, u_{-\tau})(C_2 e_2 + Ee_1 + f_n(\bar{x}, u_{-\tau}) + r - \hat{\lambda}_2) \tag{22}$$

where C_2 is a designed positive diagonal matrix. r is a robust term designed to cancel the function uncertainty, and

$$r = \begin{cases} \dfrac{e_2\psi_\xi}{\|e_2\|^2} & e_2 \neq 0 \\ 0 & e_2 = 0 \end{cases} \tag{23}$$

Hence, in this case

$$V_2 \le -C_1\|e_1\|^2 - C_2\|e_2\|^2 - \left\|(\hat{\lambda}_2 - \dot{\hat{\lambda}}_1)^T\right\|\Delta_3\left(\frac{\vartheta_2(\hat{\lambda}_2 - \dot{\hat{\lambda}}_1)}{\left\|\hat{\lambda}_2 - \dot{\hat{\lambda}}_1\right\| + \varsigma_2} - 1\right) \tag{24}$$

In the same way as (16), in order to render $V_2 < 0$, we must have

$$\left\|\hat{\lambda}_2 - \dot{\hat{\lambda}}_1\right\| > \frac{\varsigma_2}{\vartheta_2 - 1}$$

Consequently, with such controller, the estimation error of the filter can be guaranteed within the set determined in the form

$$\left\|\hat{\lambda}_2 - \dot{\hat{\lambda}}_1\right\| > \frac{\varsigma_2}{\vartheta_2 - 1}$$

Therefore, the proposed control system is overall asymptotically stable in its origin ($e_1 = e_2 = 0$), and the estimated errors of the filters are all bounded and converge exponentially to a predetermined set. Also, since the included designed parameters do not depend on each other, the size of the set can be made sufficiently small by adjusting the corresponding parameters $\varsigma_i (i = 1,2)$ appropriately.

In summary, we have the following results.

Theorem 1: Under Assumptions (1-3), using the NRBC controller (10) and (22) with the robust term (23) for nonaffine nonlinear dynamic systems (1). The solutions of error system (7-8) are UUB (Uniformly Ultimately Bounded) for $t \to \infty$.

Remark 2: Strictly speaking, when the dimension of control inputs is not equal to that of state variables, the inverse matrix of f_d is in the nonexistence. Thus, in this study,the generalized matrix inverse of f_d can be also obtained as $f_d^T(f_d f_d^T)^{-1}$. If $f_d f_d^T$ is well-conditioned, the inverse of f $f_d f_d^T$ exists. However, $f_d f_d^T$ may be ill-conditioned. A diagonal matrix is defined as $\alpha = diag(a_1, a_2, \cdots, \alpha_m)$ with α_j being a given small positive number and thus matrix $f_d f_d^T + \alpha$ is invertible. Based on the approximation model (6), a global NRBC law (an approximation solution of (6) can then be determined as follows.

$$u(t) = -f_d^T(f_d f_d^T + \alpha)^{-1}(C_2 e_2 + Ee_1 + f_n + r - \hat{\lambda}_2)$$

(25)

Remark 3: In order to obtain the smooth signal r, the unknown d_ξ can be approximately estimated by

$$r = \frac{e_2 \psi_\xi}{\|e_2\|^2 + \beta} \tag{26}$$

β is a given small positive number.

Remark 4: If the nonlinear dynamic characteristics of the process plant can be accurately described by the mathematical model (1), the robust compensation term (23) is not needed. Then, the NRBC can be degraded into a novel backstepping control (NBC) for certain nonaffine nonlinear dynamic systems.

4. Illustrative Example

In this section the objective to evaluate the performance of the NRBC. The evaluation is carried out on the three-pole AMB system [8-9]. First, the three-pole AMB mathematical model is described. It will be shown that the three-pole AMB is a nonaffine nonlinear system.

With the configuration of Fig.1, a magnetic circuit is given in Fig.2, assuming that the reluctance exist only on air gaps, the differential equations the three-pole AMB system is given by

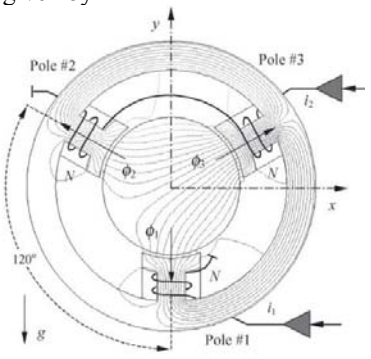


Fig. 1 Nonlinear control of a 3-pole AMB system.

$$\ddot{x}_r = \frac{4}{3}\gamma m \Phi_1 \Phi_2$$
$$\ddot{y}_r = \frac{2}{3}\gamma m(\Phi_1^{\,2} - \Phi_2^{\,2}) - g \tag{27}$$

where (x_r, y_r) is the position of the rotor center, m is the rotor mass and g is the gravitational acceleration. $\gamma = \mu A N^2$, μ is the magnetic permeability of the air, A is the pole face area and N is the coil turns. The relationship

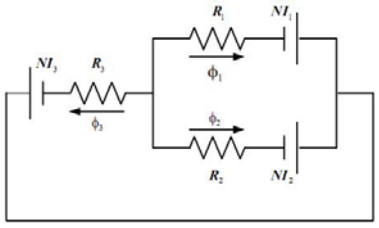


Fig. 2. Magnetic circuit for the 3-pole AMB system.

between (Φ_1, Φ_2) and (i_1, i_2) can be expressed in a matrix form by

$$\begin{bmatrix} \Phi_1 \\ \Phi_2 \end{bmatrix} = -\frac{1}{L}\begin{bmatrix} x_r & \sqrt{3}(2l_0 + y_r) \\ 2l_0 - y_r & \sqrt{3}x_r \end{bmatrix}\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} \tag{28}$$

where $L = 4l_0^{\,2} - (x_r^{\,2} + y_r^{\,2})$ is always positive due to physical constrain $x_r^{\,2} + y_r^{\,2} \le l_0^{\,2}$. The determinant of the matrix in (28) is $-\sqrt{3}L$, which is always non-zero. Therefore, there is a one-to-one correspondence between (i_1, i_2) and (Φ_1, Φ_2).

Define the states of three-pole AMB system (27) are $\bar{x} = (x_r, y_r, \dot{x}_r, \dot{y}_{r)}$, and control input are $u = (i_1, i_2)$. The nominal values of the three-pole AMB system parameters are defined in Table 1.

Table 1 Nominal values of 3-pole AMB model parameters

Description	Value
rotor mass, $\hat{m}$	$0.635/kg$
nominal air gap, $\hat{l}_0$	$9.45\times10^{-4}/m$
Magnetic permeability of the air, μ	$4\pi\times10^{-7}H/m$
pole face area, A	$4\times10^{-4}/m^2$
coil turns, N	300

Suppose that there is uncertainty caused by two parameters: the nominal air gap l_0 and the lumped parameter c_0. Let υ_l and υ_c denote the percentage of the variations in l_0 and c_0 respectively, i.e., $l_0 = \hat{l}_0(1+\upsilon_l)$ and $c_0 = \hat{c}_0(1+\upsilon_c)$. Two levels of parameter variations are considered: $(\upsilon_l = 0; \upsilon_c = 0)$ and $(\upsilon_l = 1.2\%; \upsilon_c = -1.2\%)$, The uncertainty case use the same initial states $x = (2\times10^{-4}, 2\times10^{-4}, 0.015, 0.02)$. In order to verify the proposed control algorithm robustness. NRBC and NBC are designed and implemented for nonaffine nonlinear systems.

(1) The desired tracking commands are $x_{1d} = [0.0]^T, /m$. The other parameters are selected as $C_1 = C_2 = 40, \alpha = 10^{-5}$. According to Assumption 2, $|\Delta u(t)|$ should not be too large in order to limit the

approximation error of the model (5)-(6) for a computed $u(t)$, the parameters of robust compensation term (23) are selected as $\psi_{\xi} = 0.001$ and $\beta = 10^{-7}$.

(2) Compared to NRBC law, NBC has a similar design process NRBC apart from robust term (23). So the controller parameters are also selected as $C_1 = C_2 = 40, \alpha = 10^{-5}$. With the above controller parameters, the control currents i_1 and i_2 of the above controller are all within the range $(-2A, 2A)$.

Fig.3-5 show the rotor trajectories in the case of uncertainty. As can be seen from Fig.3-5, NRBC and NBC all can stabilize the system. Although both NBC and NRBC can stabilize the system in this uncertain case, the latter achieves better performance. Fig. 6 shows the control currents using NRBC law in the case of uncertainty, and Fig.7 shows the control currents using NBC law.

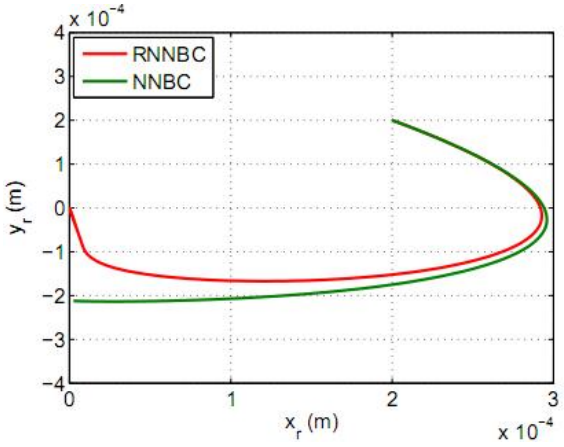


Fig. 3. Rotor trajectory with NRBC and NBC controller.

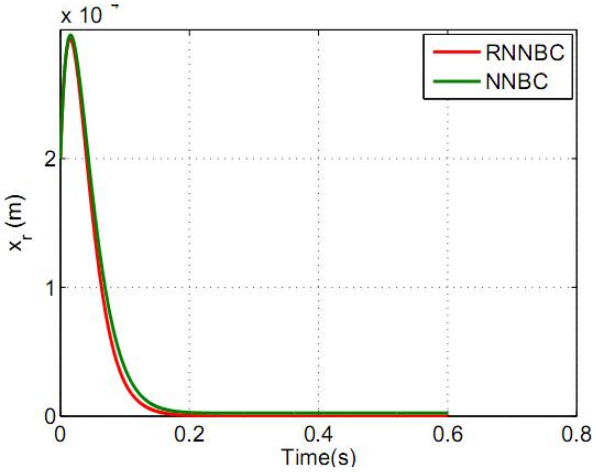


Fig.4.Rotor x_r displacement with NRBC and NBC controller.

5. Conclusion

A continuous-time nonaffine nonlinear controller design scheme for a class of nonlinear systems is presented in this paper. The strategy combines sliding mode and backstepping technique based on a novel approximation technique. The NRBC controller is designed to track the state commands against unknown uncertainties/ disturbance. It is shown that, if the controller is applied, the tracking errors exponentially converge to a compact set and the size of the set can be made arbitrarily small by tuning the design parameters, and its stability is analyzed using Lyapunov theory. The proposed approach is then applied to three-pole AMB system, and simulation results demonstrate and illustrate the effectiveness and capabilities of our scheme.

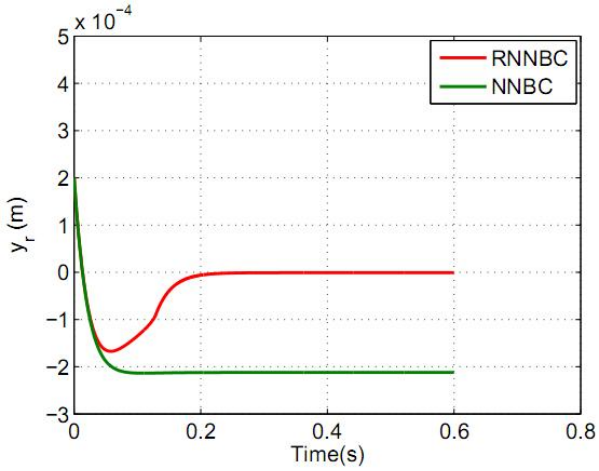


Fig.5. Rotor Y_r displacement with NRBC and NBC controller.

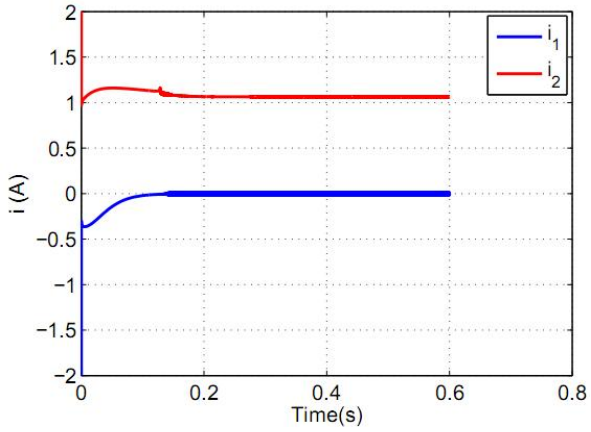


Fig. 6. Input coil current with NRBC controller.

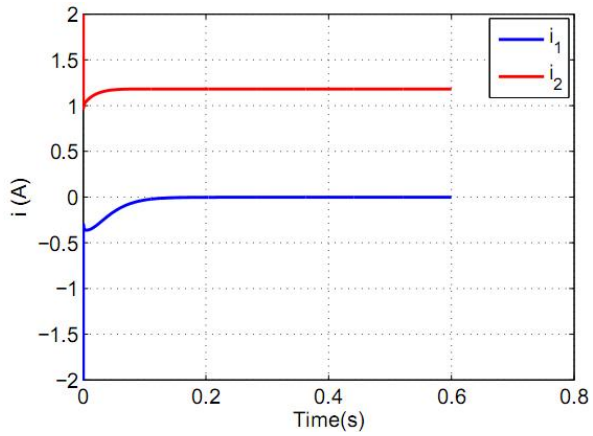


Fig. 7. Input coil current with NRBC controller.

References

[1] A. Isidori. Nonlinear control systems, 3rd ed., Springer-Verlag, New York, 1995.
[2] H, Nijmeijer, V. Schaft. Nonlinear dynamical control systems, Springer-Verlag, New York, 1990.
[3] X. Yan, C. Edwards, S. Spurgeona. Output feedback sliding mode control for non-minimum phase systems with non-linear disturbances.International Journal of Control, 77(15), 1353-1361, 2004.
[4] M. Krstic, I. Kanellakopoulos, and P. Kokotovic. Nonlinear and adaptive control design, Wiley, New York, 1995.
[5] J. Boskovic, L. Chen, and R. Mehra. Adaptive control design for nonaffine models arising in flight control. Journal of Gaidance, Control, and Dynamics, 27(2): 209-217, 2004.
[6] A. Stotsky, J. Hedrick, P. Yip. The use of sliding modes to simplify the backstepping control method. In: Proc. American Control Conference, 1997.
[7] X. Lu, J. Karl, Integral filters from a new viewpoint and their application in nonlinear control design. In: Proc. IEEE International Conference on Control Applications, 2000.
[8] M. Jang, C. Chen, Y. Tsao. Sliding mode control for active magnetic bearing system with flexible rotor. Journal of the Franklin Institute.342, 401-419, 2005.
[9] J. Hung, N. Albritton, F. Xia. Nonlinear control of a magnetic bearing system. Mechatronics, 13, 621-637, 2003.

Shuanji Zhang finished Master Degree in School of computer Science and Engineering, Hebei University of Technology in 2005, is an Associate Professor in Physics and Electronic Information School, Luoyang Normal University, He has been finished three fund projects of Henan province science and technology research, His main research interests include Control theory, automatic device , embedded system.

Yida Jiang received a Maser Degree in Lanzhou University of Technology in 2010, is a teacher in Luoyang Normal University, he has more than 10 research publications in national core journals, His research interests cover Control theory and Power electronics.

Dezhi Xu received a B.Sc. in automatic control from the North University of China, Taiyuan, in 2007, and an M.Sc. in automatic control from the Lanzhou University of Technology, Lanzhou, China, in 2010. He is currently a Ph.D. candidate with the College of Automation Engineering in the Nanjing University of Aeronautics and Astronautics. His research interests include fault diagnosis and fault-tolerant control for complex systems as well as data-driven control.

PERFORMANCE ANALYSIS OF VISION-BASED DEEP WEB DATA EXTRACTION FOR WEB DOCUMENT CLUSTERING

[1]M. Lavanya and [2]M. Usha Rani

[1]Assistant Professor [SL], Department of Master of Computer Applications
Sree Vidyanikethan Engineering College, A.Rangampet, Tirupati, Andhra Pradesh, INDIA-517102

[2]Associate Professor, Department of Computer Science
Sri Padmavati Mahila Viswavidyalayam, (SPMVV Woman's' University), Tirupati
Andhra Pradesh, INDIA-517501

Abstract

Web Data Extraction is a critical task by applying various scientific tools and in a broad range of application domains. To extract data from multiple web sites are becoming more obscure, as well to design of web information extraction systems becomes more complex and time-consuming. We also present in this paper so far various risks in web data extraction. Identifying data region from web is a noteworthy crisis for information extraction from the web page. In this paper, performance of vision-based deep web data extraction for web document clustering is presented with experimental result. The proposed approach comprises of two phases: 1) Vision-based web data extraction, where output of phase I is given to second phase and 2) web document clustering. In phase 1, the web page information is segmented into various chunks. From which, surplus noise and duplicate chunks are removed using three parameters, such as hyperlink percentage, noise score and cosine similarity. To identify the relevant chunk, three parameters such as Title word Relevancy, Keyword frequency-based chunk selection, Position features are used and then, a set of keywords are extracted from those main chunks. Finally, the extracted keywords are subjected to web document clustering using Fuzzy c-means clustering (FCM). The experimentation has been performed on two different datasets and the results showed that the proposed VDEC method can achieve stable and good results of about 99.2% and 99.1% precision value in both datasets.

KEYWORDS: FEATURES, RISKS, PROBLEMS, VDEC, FRAMEWORK, POSITION FEATURES, FUZZY C-MEANS CLUSTERING (FCM)

1. Introduction

Now-a-days, information can be retrieved from web as a main source for the required applications. The content of information from the web is in the form of unstructured text, a huge amount of semi-structured objects, called data records, are enclosed on the Web [5]. Vision-based Web Data Extraction system can be done with various web sources using different techniques and extract the data regions stored in the deep web page. Consider, if the source is a HTML Web page, the extracted information could consist of elements in the page as well as the full-text of the page itself. The deep web data region has to be again convert into a Structured format.[Zhao 2007; Irmak and Suel 2006].Vision-based web data extraction has useful data extraction from the deep web pages which are hidden web pages. The consequence of Vision based Web Data Extraction systems depends large (and quickly growing) amount of information is continuously produced, shared and consumed online: Web Data Extraction systems allow to efficiently collect this information with a limited human effort. Huge web information is presented in the form of a Web record, which consist of whole pages as well as catalog pages. Combining all the information which is extracted from the web, but many web pages may provide the same or related information using entirely diverse formats or syntaxes, which makes the integration of information a challenging task.

For instance, a NITL web page contains details of liberal information in the center of the page, which is the main content of this page. Also, there are advertisements, navigation bars, and others, situated around the main content, which are called as noise blocks [2].

Many of the noisy items are required by web site owners; they will obstruct the web data mining and decrease the performance of the search engines [14], [15]. Based on their different levels of abstraction, Web noise can be classified into two categories: Global noise and Local noise (intra-page). **Global noise** is the surplus objects with large granularities, which are in no means smaller than the individual

pages. Global noises are in mirror sites, replica Web pages and obsolete Web pages that are need to be deleted. **Local (intra-page) noise** is the irrelevant items inside a Web page. Normally, local noise is irrational with the primary content of the page. Such noise encompass banner commercials, navigational guides, garnishing images, etc [16], [17], [18]. Hence, having a method that automatically discovers the information in a web page and allots substantial measures for different areas in the web page is of an immense advantage [19], [20]. It is imperative to distinguish relevant information from noisy content because the noisy content may deceive users' concentration within a solitary web page, and users only pay attention to the commercials or copyright when they search a web page. Thus, different information within a web page will have diverse significance weight based on its location, occupied regions, subject, and more [19], [20].

Within web, semantically related content is usually grouped together and the whole page is divided into regions for diverse contents by means of explicit or implicit visual separators namely lines, blank areas, images, font sizes, colors, and more [4]. Web pages are often chaotic with disquieting features around the body of an object that distract the interest of the user from the actual content they are interested in. These "features" may comprise pop-up advertisement, showy banner advertisements, search and filtering panel, superfluous images, or links scattered around the screen. However, these noisy data are present in various patterns in diverse Web sites. Such irrelevant items should be removed for extracting only the significant information [3].

In this paper we are experimenting on an approach [10] to extract data items from the deep web pages automatically. It consists of two stages of execution: (1) Identification and Extraction of the data extraction for deep web page (2) Web clustering using FCM algorithm. Firstly in a web page, the irrelevant data such as advertisements, images, audio, etc are removed using chunk segmentation operation. The result we will obtain is a set of chunks. From which, the surplus noise and the duplicate chunks are removed by computing the three parameters, such as *Hyperlink percentage*, *Noise score* and *cosine similarity*. For each chunk, three parameters such as *Title word Relevancy*, Keyword *frequency based chunk selection* and *Position feature* are computed. Using these parameters, the sub-chunk weightage of each and every sub-chunk is calculated. The weightage of one or more sub-chunks will be greater than other sub-chunks. These sub-chunks consider as the main chunk and the keywords are extracted from those main chunk. To cluster documents, we have to select right the type of the characteristics or attributes (e.g. words, phrases or links) of the documents on which the clustering algorithm will be based and their representation.

The vision-based web document clustering approaches characterize each document according to its content, i.e. the words (or sometimes phrases) contained in it. The basic idea is that if two documents contain many common words then it is likely that the two documents are very similar. The vision-based approaches can be further classified according to the clustering method used into the following categories: *partitional*, *hierarchical*, *graph-based*, *and neural network-based* and *probabilistic*. Fuzzy clustering approaches, on the other hand, are non-exclusive, in the sense that each document can belong to more than one clusters. Fuzzy algorithms usually try to find the best clustering by optimizing a certain criterion function. The fact that a document can belong to more than one clusters is described by a *membership function*. The membership function computes for each document a membership vector, in which the I i^{th} element indicates the degree of membership of the document in the i-th cluster. The most widely used fuzzy clustering algorithm is Fuzzy c-means (Bezdek, 1984), a variation of the partitional k-means algorithm. In fuzzy c-means each cluster is represented by a *cluster prototype* (the center of the cluster) and the membership degree of a document to each cluster depends on the distance between the document and each cluster prototype. The closest the document is to a cluster prototype, the greater is the membership degree of the document in the cluster.

The paper is organized as follows. Section 2 presents the various risks to overcome in web data extraction methods. The features of visual deep web page is described in section 3 and vision –based deep web data extraction for web document clustering which is an approach is presented in section 4. An efficient approach web document clustering based on vision-based deep web is discussed in section 5. The experimental results are reported in Section 6. Section 7 explains conclusion of the paper.

2. Various Risks To Overcome In Web Data Extraction Methods

Web Data Extraction Systems has various perceptions and it influences on various disciplines like Machine Learning, Logic and Natural Language Processing in design and implementation. Many aspects are taken into consideration, where some of the aspects are independent of particular domain to plan to extract web data in the design of the web data extraction method. Other factors, depend on some of the instead, heavily depend on the exacting characteristics of the application domain where some of the technological solutions which appear to be effective in some application contexts are not suitable in other ones. Some of the approaches use static HTML web pages where tags containing a page level-by-level organizing to retrieve information. All the techniques are only to increase the correctness over huge document collections, where the results obtained are not redundant. The risk involved is extracting data from the web is tough due to number of requirements has to be met.

2.1 The major problems raised in the design of a Web Data Extraction system can be listed as follows:

Specialist help is used for Web Data Extraction techniques. A problem comprises of providing a elevated extent of mechanization by sinking human efforts as much as possible. Specialist feedbacks, however, may involve key role in increase the intensity of accuracy accomplished by a Web Data Extraction system. A associated problem is, therefore, to make out a sensible between the need of building highly automated Web Data Extraction procedures and the requirement of achieving highly accurate performance. Web Data Extraction techniques should be able to process large volumes of data in relatively short time. Such a need is particularly urgent in the field of Business and Competitive Intelligence because a firm needs to perform timely analysis of market conditions. In some of the issues, a Web Data Extraction tool has to regularly extract data from a Web Data source which can evolve over time. Web sources are continuously budding and structural changes happen with no notification thus are irregular. Eventually, in real-world situations it appear the need of maintaining these systems, that might stop functioning correctly if lacking of edibility to detect and face structural modifications of related Web sources. Searching for information on the Web is not an easy task. Searching for personal information is sometimes even more complicated. Below are several common problems we face when trying to get personal details from the web: Majority of the Information is distributed between different sites.

- It is not updated.
- Multi-Referent ambiguity – two or more people with the same name.
- Multi-morphic ambiguity which is because one name may be referred to in different forms.

In the most popular search engine Google, one can set the target name and based on the extremely limited facilities to narrow down the search, still the user has 100% feasibility of receiving irrelevant information in the output search hits. Not only this, the user has to manually see, open, and then download their respective file which is extremely time consuming. The major reason behind this is that there is no uniform format for personal information.

3. Features Of Visual Deep Web Pages

- Users use information from Web pages which are useful for various applications.
- The designers of web page associate different types of information with distinct visual characteristics to make the information on Web pages easy to understand.
- Visual features are important for identifying special information on Web pages.
- Deep Web pages are special Web pages that contain data records retrieved from Web databases, and we imagine that there are some distinct visual features for data records and data items.
- Based on observation based on a large number of deep Web pages is consistent with this hypothesis.
- The main visual features in this section and show the statistics about the accuracy of these features at the end. Position features (PFs). These features indicate the location of the data region on a deep Web page. PF1: Data regions are always centered horizontally.PF2: The size of the data region is usually large relative to the area size of the whole page. Since the data records are the contents in focus on deep Web pages, Web page designers always have the region containing the data records centrally and conspicuously placed on pages to capture the user's attention.

- By investigating a large number of deep Web pages, we found two interes ting facts. First, data regions are always located in the middle section horizontally on deep Web pages. Second, the size of a data region is usually large when there are enough data records in the data region. The actual size of a data region may change greatly because it is not only influenced by the number of data records retrieved, but also by what information is included in each data record.
- Therefore, our VDEC [10] approach uses the ratio of the size of the data region to the size of whole deep Web page instead of the actual size.

4. Vision-Based Deep Web Data Extraction For Web Document

We present new approach for deep web clustering based capture the actual data of the deep web pages. We achieve this in the following two phases. (1) Vision based Data relevant identification (2) Deep web pages clustering[11].

In the first phase,

- A data extraction based measure is also introduced to evaluate the importance of each leaf chunk in the tree, which in turn helps us to eliminate noises in a deep Web page. In this measure, remove the surplus noise and duplicate chunk using three parameters such as hyperlink percentage, Noise score and cosine similarity. Finally, obtain the main chunk extraction process using three parameters such as Title word Relevancy, Keyword frequency based chunk selection, Position features and set of keywords are extracted from those main chunks.

In the second phase,

- By using Fuzzy c-means clustering (FCM), the set of keywords were clustered for all deep web pages.

5. Performance analysis of Vision-Based Deep Web Data Extraction For Web Document

Information extraction from web pages is an active research area. Recently, web information extraction has become more challenging due to the complexity and the diversity of web structures and representation. This is an expectable phenomenon since the Internet has been so popular and there are now many types of web contents, including text, videos, images, speeches, or flashes. The HTML structure of a web document has also become more complicated, making it harder to extract the target content. Until now, a large number of techniques have been proposed to address this problem, but all of them have inherent limitations because they are Web-page-programming-language dependent. In this paper, we present new approach for detection and removal of noisy data to extract main content information and deep web clustering that is both fast and accurate[10]. The two phases and its sub-steps are given as follows.

- **Phase 1:** Vision-based deep web data identification
  - Deep web page extraction
  - Chunk segmentation
  - Noisy chunk Removal
  - Extraction of main chunk using chunk weightage
- **Phase 2:** Web document clustering
  - Clustering process using FCM

5.1 Phase 1: Vision-Based Deep Web Data Extraction

1) Deep Web Page Extraction

The Deep web is usually defined as the content on the Web not accessible through a search on general search engines. This content is sometimes also referred to as the hidden or invisible web. The Web is a complex entity that contains information from a variety of source types and includes an evolving mix of different file types and media. It is much more than static, self-contained Web pages. In our work, the deep web pages are collected from Complete Planet (www.completeplanet.com), which is currently the largest deep web repository with more than 70,000 entries of web databases.

2) Chunk Segmentation

Web pages are constructed not only main contents information like product information in shopping domain, job information in a job domain but also advertisements bar, static content like navigation panels, copyright sections, etc. In many web pages, the main content information exists in the middle chunk and the rest of page contains advertisements, navigation links, and privacy statements as noisy data. Removing these noises will help in improving the mining of web. To assign importance to a region in a web page (W_P), we

first need to segment a web page into a set of chunks. Hence, to clean a web page, a preprocessing step called Chunk Splitting Operation (fig.2) is performed.

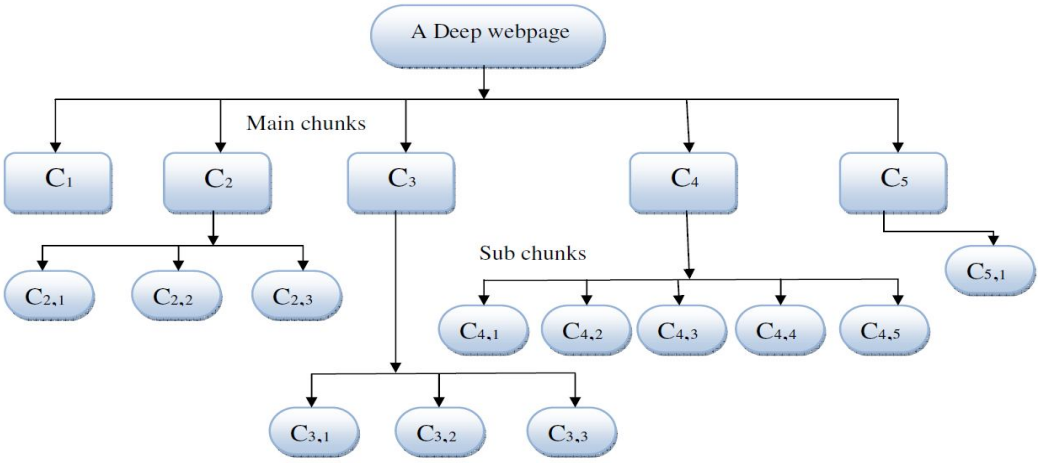


Fig. 1. The tree model of the deep web page

Basically, the layout of many web pages follows a similar pattern in such a way that the main content is enclosed in one big or <td> element which is HTML tags. In our paper, we are concentrating only the content inside the "div" tag. The tag defines a division or a section in an HTML document and it is often used to group chunk-elements. In our approach tag is consider as chunk. Normally, a tag separated by many sub tags based on the content of the deep web page. If there is no tag in the sub tag, the last tag is consider as leaf node. The Chunk Splitting Process aims at cleaning the local noises by considering only the main content of a web page enclosed in div tag. The main contents are segmented into various chunks. The resultant of this process can be represented as follows:

$$C = \{C_1, C_2, C_3, \cdots, C_n\},\ C \in DW_P$$

Where, C → A set of chunks in the deep web page DW_P

n → Number of chunks in a deep web page DW_P

In fig.1 we have taken an example of a tree sample which consists of main chunks and sub chunks. The main chunks are segmented into chunks C_1, C_2 and C_3 using Chunk Splitting Operation and sub-chunks are segmented into $C_{2,1}, C_{2,2} \ldots C_{5,1}$ in fig 2.

3) Noisy Chunk Removal

Surplus Noise Removal: A deep web page W_P usually contains main content chunks and noise chunks. Only the main content chunks represent the informative part that most users are interested in. Although other chunks are helpful in enriching functionality and guiding browsing, they negatively affect such web mining tasks as web page clustering and classification by reducing the accuracy of mined results as well as speed of processing. Thus, these chunks are called noise chunks. Removing these chunks in our research work, we have concentrated on two parameters; they are Hyperlink Percentage (HL_p) and Noise score (N_s) which is very significant. The main objective for removing noise from a Web Page is to improve the performance of the search engine.

4) Extraction of Main Chunk

Chunk Weightage for Sub-Chunk: In the previous step, we obtained a set of chunks after removing the noise chunks and duplicate chunks present in a deep web page. Web page designers tend to organize their content in a reasonable way: giving prominence to important things and deemphasizing the unimportant parts with proper features such as position, size, color, word, image, link, etc. A chunk importance model is a function to map from features to importance for each chunk, and can be formalized as:

$$\langle chunk\ \ features \rangle \Rightarrow chunk\ \text{importance}$$

The preprocessing for computation is to extract essential keywords for the calculation of Chunk Importance. Many researchers have given importance to different information inside a webpage for instance location, position, occupied area, content, etc. In our research work, we have concentrated on the three parameters Title word relevancy, keyword frequency based chunk selection, and position features which are very significant. Each parameter has its own significance for calculating sub-chunk weightage. The following equation computes the sub-chunk weightage of all noiseless chunks.

$$C_w = \alpha T_k + \beta K_f + \gamma PF_r \quad (1)$$

Where,

$\alpha, \beta, \gamma \rightarrow$ Constants

For each noiseless chunk, we have to calculate these unknown parameters T_K, K_f and PF_r. The representation of each parameter is as follows:

5.2 Phase II: Deep Web Documnet Clustering Using FCM

Let DB be a dataset of web documents, where the set of keywords is denoted by $k = \{k_1, k_2, ..., k_n\}$. Let $X = \{x_1, x_2, ..., x_N\}$ be the set of N web documents, where $x_i = \{x_{i1}, x_{i2}, ..., x_{in}\}$. Each $x_{ij} (i = 1,...,N; j = 1,...,n)$ corresponds to the frequency of keyword x_i on web document. Fuzzy c-means [29] partitions set of N web documents in R^d dimensional space into c $(1 < c < n)$ fuzzy clusters with $Z = \{z_1, z_2, ..., z_c\}$ cluster centers or centroids. The fuzzy clustering of keywords is described by a fuzzy matrix μ with n rows and c columns in which n is the number of keywords and c is the number of clusters. μ_{ij}, the element in the i^{th} row and j^{th} column in μ, indicates the degree of association or membership function of the i^{th} object with the j^{th} cluster. The characters of μ are as follows:

$$\mu_{i,j} \in [0,1] \qquad \forall i = 1,2,...,n; \quad \forall j = 1,2,...,c; \quad (6)$$

$$\sum_{j=1}^{c} \mu_{ij} = 1 \qquad \forall i = 1,2,...,n; \quad (7)$$

$$0 < \sum_{i=1}^{n} \mu_{ij} < n \quad \forall j = 1,2,...,c; \quad (8)$$

The objective function of FCM algorithm is to minimize the Eq. (9):

$$J_m = \sum_{j=1}^{c} \sum_{i=1}^{n} \mu_{ij}^m d_{ij} \quad (9)$$

Where

$$d_{ij} = \| k_i - z_j \| \quad (10)$$

in which, m(m >1) is a scalar termed the weighting exponent and controls the fuzziness of the resulting clusters and d_{ij} is the Euclidian distance from k_i to the cluster center z_i.The z_j,centroid of the j^{th} cluster, is obtained using Eq. (11)

$$z_j = \frac{\sum_{i=1}^{n} \mu_{ij}^m k_i}{\sum_{i=1}^{n} \mu_{ij}^m} \quad (11)$$

The FCM algorithm is iterative and can be applied.

6. Results and Discussion

6.1 Experimental set up

The experimental results of the proposed method for vision-based deep web data extraction for web document clustering are presented in this section. The proposed approach has been implemented in java (jdk 1.6) and the experimentation is performed on a 3.0 GHz Pentium PC machine with 2 GB main memory. For experimentation, we have taken many deep web pages which contained all the noises such as Navigation bars, Panels and Frames, Page Headers and Footers, Copyright and Privacy Notices, Advertisements and Other Uninteresting Data. These pages are then

applied to the proposed method for removing the different noises. The removal of noise blocks and extracting of useful content chunks are explained in this sub-section. Finally, extracting the useful content keywords are clustered using Fuzzy c-means clustering.

6.2. Data Sets

GDS: Our data set is collected from the complete planet web site (www.completeplanet.com). Complete-planet is currently the largest depository for deep web, which has collected the search entries of more than 70,000 web databases and search engines. These Web databases are classified into 42 categories covering most domains in the real world. GDS contains 1,000 available Web databases. For each Web database, we submit five queries and gather five deep Web pages with each containing at least three data records. **SDS:** Special data set (SDS). During the process of obtaining GDS, we noticed that the data records from two-thirds of the Web databases have less than five data items on average. To test the robustness of our approaches, we select 100 Web databases whose data records contain more than 10 data items from GDS as SDS.

6.3. Performance Measures

1) Data extraction evaluation

Precision is the percentage of the relevant data records identified from the web page.

$$\text{Precision} = \frac{DR_c}{DR_e}$$

Recall defines the correctness of the data records identified.

$$\text{Recall} = \frac{DR_c}{DR_r}$$

Where,

DR_c is the total number of correctly extracted data records

DR_e is the total number of data records on the page

DR_r is the total number of data records extracted

Revision is defined to be the percentage of the Web databases whose data records or data items are not perfectly extracted, i.e., either precision or recall is not 100 percent.

$$\text{Revision} = \frac{WDB_t - WDB_c}{WDB_t}$$

Where,

WDB_c is the total number of web sites whose precision and recall are both 100%.

WDB_t is total number of web sites processed.

2) Clustering evaluation

Clustering Accuracy, $CA = \frac{1}{N}\sum_{i=1}^{T} X_i$

Where, $N \rightarrow$ Number of data points in the dataset

$T \rightarrow$ Number of resultant cluster

$X_i \rightarrow$ Number of data points occurring in both cluster i and its corresponding class.

6.4 Experimental results

The sample of results obtained by the proposed approach is given in this sub-section. A sample of deep webpage considered for experimentation is shown in Fig. 2. Then, the sample web page is subjected to the proposed approach to identify the relevant web data region. Data region having description of some products is extracted by our data extraction method after removing the noises. The filtered data region is shown in Fig. 3.

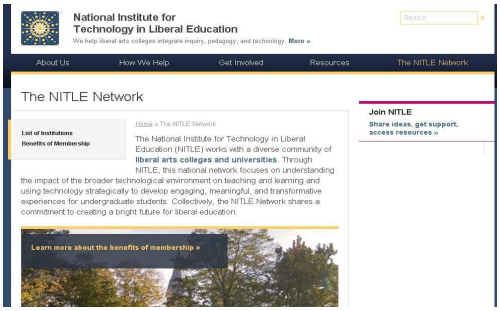


Fig. 2 A Sample Deep Web Page

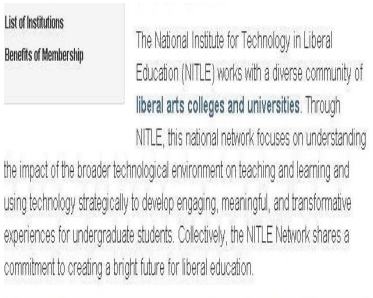


Fig. 3 Filter Data Region

6.5 Performance analysis of phase 1 of our technique

The performance analyses of the three methods on GDS and SDS datasets are presented in this section. Table 1 shows the experimental results on both GDS and SDS. Totally, VDEC performs significantly better than MDR on both GDS and SDS. But in another case, VDEC slightly get slipped in performance evaluation with ViDRE. *Precision:* The precision values of three methods are plotted as a graph shown in Fig 5, in which our proposed method VDEC performs better precision value (99.2% and 99.1%) compared with MDR in both datasets. *Recall:* The recall values obtained for three different methods are plotted in the figure 6, in which our VDEC performs better recall value (98.4% and 97.4%) compared with MDR in both datasets. *Revision:* By analyzing the figure 7, VDEC is better revision value (12% and 8%) compared with MDR in both datasets.

Table 1: Performance comparison of ViDRE, MDR and VDEC on two data sets.

	Data set		Precision	Recall	Revision
ViDRE	GDS		98.7%	97.2%	12.4%
	SDS		98.5%	97.8%	10.9%
MDR	GDS		85.3%	53.2%	55.2%
	SDS		78.7%	47.3%	63.8%
VDEC	GDS	$\alpha = .5, \beta = .5, \gamma = .5$	92.2%	91.9%	24%
	SDS	$\alpha = .5, \beta = .5, \gamma = .5$	93.6%	90.1%	16%
	GDS	$\alpha = .5, \beta = .5, \gamma = .1$	99.2%	97.4%	8%
	SDS	$\alpha = .5, \beta = .5, \gamma = .1$	99.1%	98.4%	12%

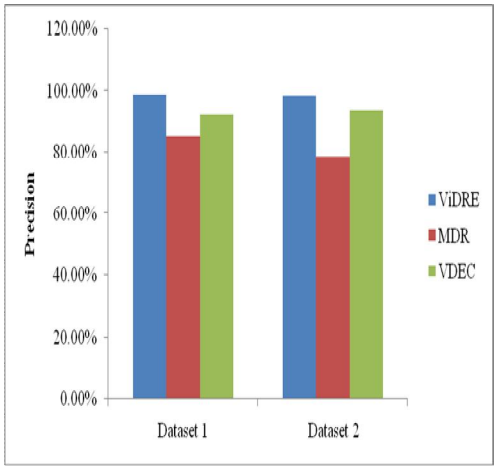


Figure 4.Precision graph of dataset 1 and dataset 2 for parameters $\alpha = .5, \beta = .5, \gamma = .5$

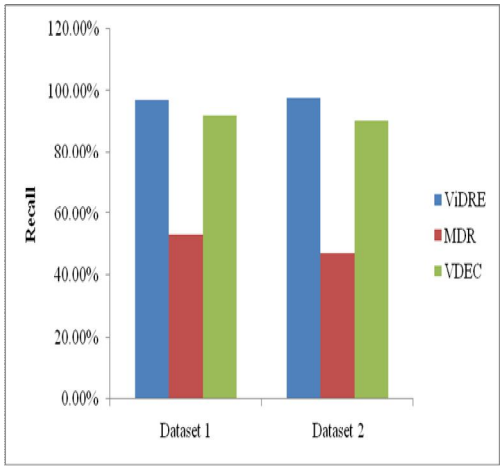


Figure 5.Recall graph of dataset 1 and dataset 2 for parameters $\alpha = .5, \beta = .5, \gamma = .5$

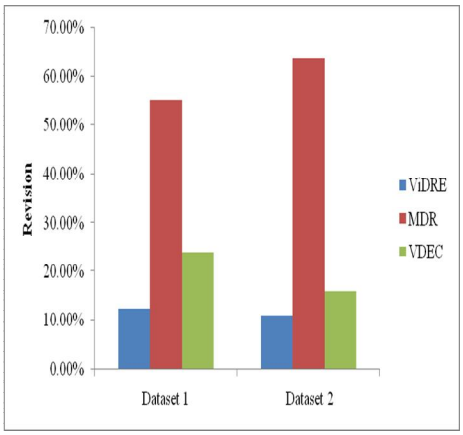


Figure 6.Revision graph of dataset 1 and dataset 2 for parameters $\alpha = .5, \beta = .5, \gamma = .5$

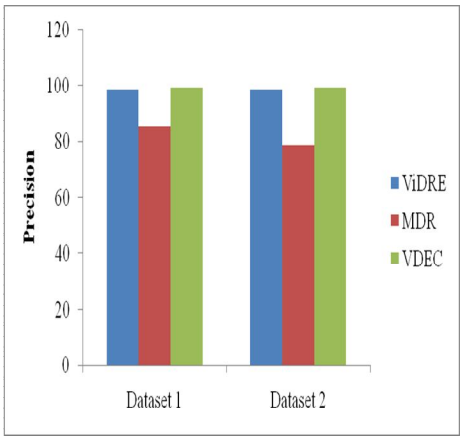


Figure 7.Precision graph of dataset 1 and dataset 2 for parameters $\alpha = .5, \beta = .5, \gamma = .1$

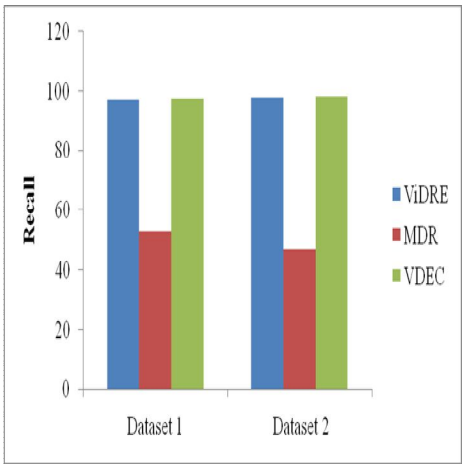


Figure 8.Recall graph of dataset 1 and dataset 2 for parameters $\alpha = .5, \beta = .5, \gamma = .1$

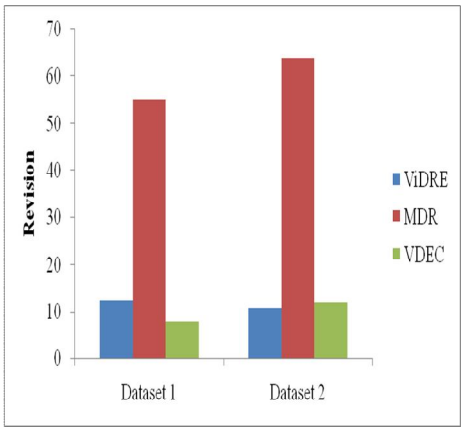


Figure 9.Revision graph of dataset 1 and dataset 2 for parameters $\alpha = .5, \beta = .5, \gamma = .1$

8. Conclusion

In this paper, we have implemented a new approach called vision-based deep web data extraction for web document clustering. In this paper, an approach to vision-based deep web data extraction is proposed for web document clustering. The proposed approach comprises of two phases: 1) Vision-based web data extraction, and 2) web document clustering. In phase 1, the web page information is classified into various chunks. From which, surplus noise and duplicate chunks are removed using three parameters, such as hyperlink percentage, noise score and cosine similarity. To identify the relevant chunk, three parameters such as Title word Relevancy, Keyword frequency-based chunk selection, Position features are used and then, a set of keywords are extracted from those main chunks. Finally, the extracted keywords are subjected to web document clustering using Fuzzy c-means clustering (FCM). Our experimental results showed that the proposed VDEC method can achieve stable and good results of about 99.2% and 99.1% precision value in both datasets.

References

[1] P S Hiremath, Siddu P Algur,"Extraction of data from web pages: a vision based approach," International Journal of Computer and Information Science and Engineering, Vol.3, pp.50-59, 2009.
[2] Jing Li, "Cleaning Web Pages for Effective Web Content Mining, "In Proceedings: DEXA, 2006.
[3] Thanda Htwe,"Cleaning Various Noise Patterns in Web Pages for Web Data Extraction," International Journal of Network and Mobile Technologies,vol.1,no.2,2010.
[4] Yang, Y. and Zhang, H., "HTML Page Analysis Based on Visual Cues," In 6th International
Conference on Document Analysis and Recognition, Seattle, Washington, USA, 2001.

[5] Longzhuang Li, Yonghuai Liu, Abel Obregon," Visual Segmentation-Based Data Record Extraction from Web Documents,"IEEE International Conference on Information Reuse and Integration, pp.502 – 507, 2007.
[6] Qingshui Li; Kai Wu; "Study of Web Page Information topic extraction technology based on vision,"IEEE International Conference on Computer Science and Information Technology (ICCSIT), vol.9, pp.781-784, 2010.
[7] R. B. Yates and B. R. Neto,"Modern Information Retrieval," Addison-Wesley, New York, 1999.
[8] 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.
[9] Chen Hong-ping; Fang Wei; Yang Zhou; Zhuo Lin; Cui Zhi-Ming; "Automatic Data Records Extraction from List Page in Deep Web Sources, "Asia-Pacific Conference on Information Processingvol.1,pp.370-373, 2009.
[10] M.Lavanya, Dr.M.Usha rani. "vision-based deep web data extraction for web document clustering" Global Journals Inc..,March 2012.

[11] M.Lavanya, Dr.M.Usha rani. "A Frame Work For Vision-Based Deep Web Data Extraction For Web Document Clustering ", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 7, September - 2012 ISSN: 2278-0181.
[12] Wei Liu, Xiaofeng Meng, Weiyi Meng,"ViDE: A Vision-Based Approach for Deep Web Data Extraction,"IEEE Transactions on Knowledge and Data Engineering, vol.22, no.3, pp.447-460, 2010.
[13] Ashraf, F.; Ozyer, T.; Alhajj, R.; "Employing Clustering Techniques for Automatic Information Extraction from HTML Documents,"IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol.38, no.5, pp.660-673, 2008.
[14] Manisha Marathe, Dr. S.H.Patil, G.V.Garje,M.S.Bewoor, "Extracting Content Blocks from Web Pages", International Journal of Recent Trends in Engineering, Vol .2, No. 4, November 2009.
[15] Sandip Debnath,Prasenjit Mitra,C. Lee Giles, "Automatic Extraction of Informative Blocks from WebPages", In Proceedings of the ACM symposium on Applied computing, Santa Fe, New Mexico, pp. 1722 – 1726,2005.
[16] Lan Yi ,Bing Liu, "Web page cleaning for web mining through feature weighting", In Proceedings of the 18th international joint conference on Artificial intelligence, pp. 43-48 , August 09 - 15 ,Acapulco, Mexico, 2003
[17] A. K. Tripathy , A. K. Singh , "An Efficient Method of Eliminating Noisy Information in Web Pages for Data Mining", In Proceedings of the Fourth International Conference on Computer and Information Technology, pp. 978 – 985, 2004.
[18] Zhao Cheng-li and Yi Dong-yun, "A method of eliminating noises in Web pages by style tree model and its applications", Wuhan University Journal of Natural Sciences, Wuhan University, co-published with Springer Vol.9, No.5, pp. 611-616, 2004.
[19] Ruihua Song,Haifeng Liu,Ji-Rong Wen, Wei-Ying Ma, "Learning Block Importance Models for Web Pages", Proceedings of the 13th international conference on World Wide Web, pp. 203 - 211 , New York, NY, USA, 2004.
[20] Ruihua Song, Haifeng Liu, Ji-Rong Wen, Wei-Ying Ma, "Learning Important Models for Web Page Blocks based on Layout and Content Analysis", ACM SIGKDD Explorations Newsletter, Vol. 6 , No. 2,pp. 14 - 23 ,2004.
[21] Liu, B., Grossman, R. and Zhai, Y," Mining Data Records in Web Pages," KDD-03, pp. 49-55, 2003.
[22] Zhai, Y., Liu, B,"Web Data Extraction Based on Partial Tree Alignment," Proceedings of the 14th international conference on World Wide Web,pp.76-85,2005.
[23] J. Hammer, H. Garcia Molina, J. Cho, and A. Crespo. "Extracting semistructured information from the web". In Proc.of the Workshop on the Management of Semi-structured Data, 1997.
[24] Hesam Izakian, Ajith Abraham,"Fuzzy C-means and fuzzy swarm for fuzzy clustering problem,"Computer and Information Science, vol.38, no.3, pp.1835-1838, 2011.
[25] Webb, A." Statistical pattern recognition," New Jersey: John Wiley & Sons, 2002.
[26] Tan, P. N., Steinbach, M., & Kumar, V." Introduction to data mining, "Boston: Addison-Wesley, 2005.
[27] Pang, W., Wang, K., Zhou, C., and Dong, L." Fuzzy discrete particle swarm optimization for solving traveling salesman problem," In Proceedings of the fourth international conference on computer and information technology,pp. 796–800, IEEE CS Press,2004.
[28] Hathway, R. J., & Bezdek, J."Optimization of clustering criteria by reformulation," IEEE transactions on Fuzzy Systems, 241–245, 1995.
[29] J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics, Vol. 3, pp.32-57, 1973.

Ms. M. Lavanya obtained Bachelor's degree in Sciences (Computer Science) from S.V.University, Tirupathi. Then she obtained her Master's degree in Computer Applications from S.V.University. She is working as Assistant Professor[SL] in the Department of Master of Computer Applications at Sree Vidyanikethan Engineering College, A.Rangampet, Tirupati. She is pursuing her Ph.D. in Computer Science in the area of Data Warehousing and Data Mining. She is in teaching since 2003. She presented many papers at National and Internal Conferences and published articles in National & International journals.

Dr. M. Usha Rani is an Associate Professor in the Department of Computer Science and HOD for MCA, Sri Padmavati Mahila Viswavidyalayam (SPMVV Woman's' University), Tirupati. She did her Ph.D. in Computer Science in the area of Artificial Intelligence and Expert Systems. She is in teaching since 1992. She presented many papers at National and Internal Conferences and published articles in national & international journals. She also has written 4 books like Data Mining - Applications: Opportunities and Challenges, Superficial Overview of Data Mining Tools, Data Warehousing & Data Mining and Intelligent Systems & Communications. She is guiding M.Phil. and Ph.D. in the areas like Artificial Intelligence, Data Warehousing and Data Mining, Computer Networks and Network Security etc.

WiMAX Based Audio/Video Transmission

Irfanullah[1], Amjad Ali[1], Abdul Qadir Khan[1], Rehanullah Khan[1], Akhtar Khalil[2]

[1] Sarhad University of Science & Information Technology (SUIT), Peshawar, Pakistan
[2] University of Engineering and Technology, Peshawar

Abstract

In this article, we present the idea of a WiMAX based audio/video transmission. The data transmitter will transmit an audio/video signal, which will be received by the receiver side. The hardware of the data transmitter consists of a CCD camera and audio mic. The CCD camera will capture the image and give the analog video signal to a fixed filter of 12 MHz. The filter is designed in such a way that it can only pass a signal of 12 MHz and 4 KHz audio signal. The analog signal is given at the base of C1815 transistor, which will amplify this analog signal. To remove the unwanted signals, we have used capacitors. The amplified signal is then passed through a capacitor of capacitance 504 pf. the output signal at the capacitor is than pass through an RF amplifier. The RF amplifier will give strength and power to the signal. In RF amplifier section we have used a variable tuned circuit. The inductor used at the RF amplifier section act as antenna that will transmit the video signal and audio to the air.

1. Introduction

We have designed and implemented a WiMAX transmitter and receiver in which we have used a CCD Camera and mic that will capture the video signal and audio signal, this signal is amplified using transistor and then it is transmitted in the air. On the receiver side we have used a 3 GHZ demodulator with which a monopole antenna is connected to receive the signal. The signal is than amplified using video amplifier. Finally the signal is given to the television where we see the effect in real time.

WiMAX stands for worldwide interoperability for microwave access that enables the actual broadband wireless network with high speed. It operates same like Wi-Fi but Wi-Fi operates with some limitations, like it is baseband technology and covers only 100 feet radius with slow speed. WiMAX covers radius of 50km and work with speed of 70mbs. It is replacement of wired broadband. In near future, all the intelligent systems will be incorporated with WiMAX technology and a user will be connected to Internet even if he is driving a car with speed of 120km [4].

2. WiMAX Transmitter

This is the one of the advanced technology of the world. In Wimax transmitter we are using highest band of the radio frequency. Its frequency range starts up to from 2.4GHZ. In this transmitter we have use very small antenna and achieve a long distance. Wimax transmitter provide wide bandwidth due to which we load heavy data on Wimax transmitter to transmitting in it voice frequency is 4KHZ and video frequency is 12MHZ. Wimax data can transmit heavy signal. In it we can transmit audio & video data at the same time due wide bandwidth due to high frequency the size of antenna decrease and due to small antenna we get speed of faster than usual.

For example; if a video is transmitting from a TV transmitter it takes several seconds to receive. But in WiMAX data is received in real time. Wimax transmitter has no distortion effect, like gravity, electromagnetic field and other effects. Due to which we receive a signal very clear.

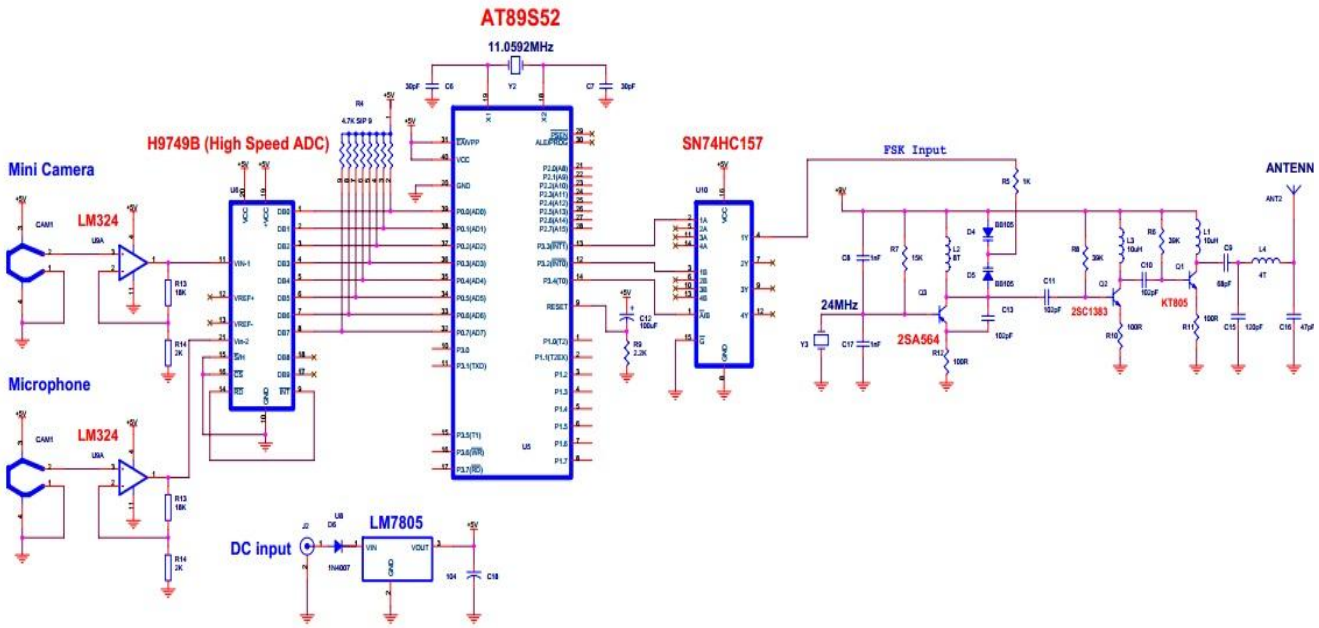


Figure 1: Circuit diagram of transmitter

3. WiMAX Receiver

The function of the receiver is opposite to that of the transmitter. At the transmitter side the analog signals are transmitted and the receiver receives those signals. Due to no distortion effects e.g. gravity, electric magnetic field and other effects clear signals can be received. The receiver will receive the amplified analog video signal, and its effects can be seen at the television in real time.

The hardware consists of small size monopole antenna, demodulator, video amplifier, and tuner, step down transformer and TV. The monopole antenna is connected to a 3 GHz demodulator. A tuner is also connected to the demodulator, which can be tuned between 2.5 GHz and 3 GHz. The monopole antenna can receive the analog signals having the frequency range - 2.4 GHz to 3 GHz.The demodulator is than connected with a video amplifier to amplify the signals. The signals are than given to television through lead from amplifier. On the screen of the TV we can see the image in the real time. In the video amplifier there is an IC L7805 whose function is to provide 5 volt to the demodulator.

4. Operation

In our project we have used Wimax based Audio/video transmitter and receiver shown in Figure 1 and Figure 2 respectively. The data transmitter will transmit an audio & video signal, which will be received by the receiver side. The hardware of the data transmitter consists of a CCD camera and audio mic. The CCD camera will capture the image and mic will capture audio and give the analog signal to a fixed filter of 12 MHz and 4 KHz. The filter is designed in such a way that it can only

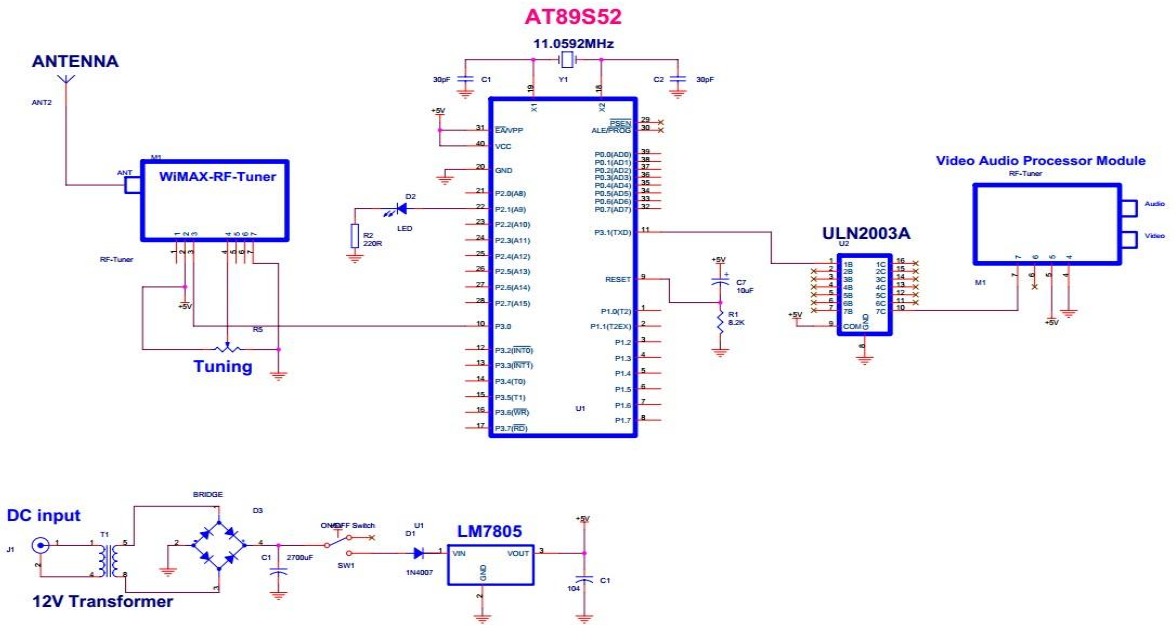


Figure 2: Circuit Diagram of Receiver

pass a signal of 12 MHz and 4 KHz. The analog signal is given at the base of C1815 transistor, which will amplify this analog signal. To remove the unwanted signals we have used capacitors as to remove unwanted signal coming from surrounding. The amplified signal is than pass through a capacitor of capacitance 504 pf. the output signal at the capacitor is than pass through an RF amplifier. The RF amplifier will give strength and power to the signal. In RF amplifier section we have used a variable tuned circuit. The inductor used at the RF amplifier section act as antenna that will transmit the video signal to the air. The receiver will receive the amplified analog video signal and its effect can be seen at the television in real time. The hardware consists of small size monopole antenna, demodulator, video amplifier, and tuner, step down transformer and TV. The monopole antenna is connected to a 3 GHz demodulator. A tuner is also connected to the demodulator, which can be tuned from 2. 5 GHz to 3 GHz. the monopole antenna can receive the analog signal having the frequency range from 2. 5 GHz to 3 GHz. THE demodulator is than connected with a video amplifier to amplify the signal. The signal is than given to television through lead from amplifier. In the video amplifier there is an IC L7805 whose function is to provide 5v voltage to the demodulator. On the screen of the TV we can see the image in the real time and listen the sound.

Acknowledgements

We are thankful to Sarhad University of Science and IT (SUIT) and iFahja* Limited for research and development support. We also thank Kamran Khan (University of Peshawar, Pakistan) and Bilal Syed (SUIT), Shahinshah Khan (SUIT) for necessary guidance and the help. All the algorithms of WIMAX and the related algorithms belong to their respective researchers and authors.

**** www.ifahja.com***

References

[1] Wongthavarawat, K., 2003, Ganz, A.; Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems, International Journal of Communication Systems, Volume 16, Issue 1

[2] E. K. Wesel,Wireless Multimedia Communications, Addison-WesleyWireless Communications Series, MA, 1997

[3] Wireless Networks and Information Systems, 2009. WNIS '09. International Conference on "Mobile WiMAX-partII: A Comparative Analysis ", WiMAX Forum, 2006.

[4] http://www.computerfreetips.com/Ds_Broadband /wireless_dsl.html

Irfan Ullah: Irfan Ullah Marwat did his Graduation in Electrical Engineering from SUIT. He possesses good knowledge of Satellite Communication, and Image Processing. Currently, he is serving Electrical Department in the SUIT.

Amjad Ali: The author was born on March 24, 1977 in Pakistan (KPK). He has received B.Sc Electrical Engineering degree from University of Engineering and Technology Peshawar, Pakistan in September 2002, and began his graduate career in esteemed institutions at the national level. He returned to the same university for Masters in Electrical engineering in 2004. He achieved his doctoral degree from Information and Communication Engineering department of Beijing University of Posts and Telecommunications China. His research interests focus on Communication, Pattern Recognition, Image processing and Biometrics.

Abdul Qadir: Abdul Qadir completed his Graduation in Electrical Engineering from SUIT. He is currently serving in Telecommunication industry.

Rehanullah Khan: Rehanullah Khan completed his B.E and M.Sc from UET, Peshawar, Pakistan and PhD from Tu-Wien, Austria. Currently, he is working as an Associate Professor at Sarhad University of Science and IT, Peshawar. His research interests include Pattern Recognition, Image Processing and Machine Learning.

Akhtar Khalil: is working as an Assistant Professor at Univesity of Engineering & Technology, Peshawar. He has experience of teaching, research and industry. He has worked on some European Commission projects and has extensive experience in technology consulting. His research and development interest lies in the areas of Mobile Applications development, Network and Information Security, Wireless Sensor Networks..

Research on the Model of Secure Transmission of SOAP Messages

Haixia Zhao[1], Yaowei Li[2], Mingchuan Zhang[1], Ruijuan Zheng[1], Qingtao Wu[1]

[1] Electronic & Information Engineering college, Henan University of Science and Technology, LuoYang, 471003, P.R. China

[2]LuoYang Electronic Information Equipment Testing Center. China, LuoYang, 471003, P.R. China

Abstract

SOAP as the basis application of Web Services, and, SOAP messages are closely related to the heterogeneous Web services. Secure transmission of SOAP messages play a vital role for the applicability of Web Services. The main challenges to the secure transmission of SOAP messages includes: confidentiality, authentication, integrity, both-party nonrepudiation, and single sign-on. We analyzed and took advantage of the existing technologies and solutions related to SOAP and Web Services, and proposed a model of secure transmission of SOAP messages, which adopting technologies like XML Signature, XML Encryption, and X.509 Certificate. The analysis in this paper indicates that for the basic requirements towards secure transmission of SOAP messages our model are fulfilled and for the high-level security and efficiency our model are acquired.

Keywords: *SOAP messages, Web Services, secure transmission model, both-party nonrepudiation, single sign-on.*

1. Introduction

Web Services are already a reality for many organizations and are just around the corner for most of the rest of us. One of the core specifications on which Web Services rely heavily is SOAP (Simple Object Access Protocol). In terms of a services-oriented architecture, SOAP is used to send data from one application to another. Web Services make use of SOAP (of course, together with other technologies) to tie heterogeneous business systems together, and as a result, companies can now create and deploy distributed applications without regard to the hardware platform, OS, programming language, or network topology of either party wishing to communicate with the chosen Web Services application. Just like all other network technologies, security is the bedrock for Web Services to enjoy widespread deployment. Without a convincing security model, the Web Services framework would be next to useless[10].

SOAP is an XML-based, simple information exchange protocol applied in dispersed or distributed environment. SOAP's main advantage is loosely coupled[1]. Seen in terms of a service-oriented architecture, SOAP allows for applications to bind to other applications in order to make use of their functionality. SOAP can either be used for messaging between applications (called "Document-based SOAP") or for Remote Procedure Calls (called "RPC SOAP"). Both of messaging and RPCs are the important aspects of SOAP, but in most cases, messaging is preferable to RPC, since it means that applications do not have to share an object model, or rely on a synchronous always-on connection[3]. SOAP is defined as an enveloping protocol, so it is sometimes seen as a messaging protocol as well as a means of using functionality that is published by a remote application.

One of the goals of SOAP designing is simplicity, so security was not taken into account by the SOAP specification. SOAP messaging security relies on the established security concepts and technologies, such as , encryption, digital signature, authentication, and data integrity. This paper is to study the secure transmission of SOAP messages.

2. Related Work

SOAP, which is a messaging protocol based on XML, is about sending messages, meaning that it specifies a way to send XML-based messages from one process to another, usually from one machine to another[8]. More specifically, SOAP is a protocol that specifies an enveloping mechanism for sending data (via XML). Furthermore, it specifies how to send these messages to a final destination, and the processing model that applies if that message goes through several

intermediaries. And, it specifies how to do this over HTTP.

The SOAP specification describes four major components: formatting conventions for encapsulating data and routing directions in the form of an envelope, a transport or protocol binding, encoding rules, and an RPC mechanism. The envelope defines a convention for describing the contents of a message, which in turn has implications on how it gets processed. A protocol binding provides a generic mechanism for sending a SOAP envelope via a lowerlevel protocol such as HTTP. Encoding rules provide a convention for mapping various application datatypes into an XML tag-based representation. Finally, the RPC mechanism provides a way to represent remote procedure calls and their return values. As to the structure, a SOAP message consists of an envelope containing an optional header and a required body, as shown in Figure 1. Envelope, the topmost container, comprises the SOAP message; Header contains additional blocks of information about how the body payload is to be processed; and Body contains the actual message to be processed. Each element contained by the Header is called a header block. The purpose of a header block is to communicate contextual information relevant to how the message is to be processed. This includes routing and delivery settings, authentication or authorization assertions, and transaction contexts. XML elements and attributes for the purpose of SOAP security are just placed inside the SOAP header. The body contains the actual message to be delivered and processed. Anything that can be expressed in XML syntax can go in the body of a message.

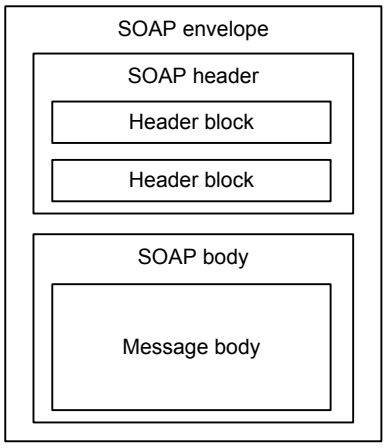


Fig. 1 SOAP message structure.

A SOAP message can be anything: a purchase order, a request for a current stock price, a query for a search engine, a listing of available flights, or any number of other pieces of information that may be relevant to a particular application.

While a SOAP message is fundamentally a one-way transmission of an envelope from a sender to a receiver, that message may pass through various intermediate processors that each in turn do something with the message. The set of intermediaries that the message travels through is called the message path. Every intermediary along that path is known as an actor. SOAP dose specify a mechanism of identifying which parts of the SOAP message are intended for processing by specific actors in its message path. This mechanism is known as “targeting”. Targeting can only be used in relation to header blocks, and the body of the SOAP envelope cannot be explicitly targeted at a particular node. The value of the actor attribute is the unique identifier of the intermediary being targeted. Intermediaries that do not match the actor attribute must ignore the header block[11].

The construction of a message path (the definition of which nodes a message passes through) is not covered by the SOAP specification. Various extensions to SOAP, such as Microsoft's SOAP Routing Protocol (WS-Routing) have emerged to fill that gap. WS-Routing defines a standard SOAP header block for expressing routing information. Its role is to define the exact sequence of intermediaries through which a message is to pass.

3. Proposed Model of Secure Transmission of SOAP Messages

In an enterprise application scenario, along with the involvement of purchase order, services providing, and payment, the information integration among enterprises extends security boundary from intranet to internet. Naturally, the risk of security increases evidently.

3.1 Security Analysis of SOAP messages transmission

The division of information security into logical components makes it easier to understand, and therefore easier to deploy[10]. These logical components, each of which maps a challenge to the security of SOAP messages transmission, are confidentiality, authentication, integrity, and nonrepudiation.

Confidentiality is used to refer to the requirement for data in transit between two communicating parties not to be available to third parties that may try to snoop on the communication. And, confidential information in a SOAP message should remain confidential over the course of a number of SOAP hops[4].

Authentication is an identity-authenticating process. In the web services world, answering the following questions is vitally important:

Who am I?

How do I prove who I am?

Why should you trust me when I tell you who I am?

Who are you?

How can I prove that you are who you say you are?

Why should I trust you when you tell me who you are?

Authentication is just a standard method to ask and answer these questions. And in multiple hops message transit, so called single sign-on is necessary. "Single sign-on", also called "federated trust", means the challenge of providing such functionality: enabling a user to sign on once, and then, without having to sign on again, access different domains that would normally be outside the scope of the primary sign-on domain[12].

Integrity has a special meaning in the field of information security. It does not mean that information cannot be tampered with. It means that if information is tampered with, this tampering can be detected. In an untrusted network, it may be impossible to ensure that the data is tamper-proof when it is in transit to its destination. So, knowledge about the fact that tampering has occurred is the next best thing[9].

Nonrepudiation literally means that the originator of a message cannot claim not to have sent a given message[7]. Nonrepudiation, which promises that malicious message sender cannot deny the fact he has sent the message, and so promises that the constructor and sender of the message is same, is vitally important to B2B applications. Furthermore, the nonrepudiation is a both-party concept in the messaging in B2B applications. Besides the attacks launched by the sender and the malicious third-party, malicious receiver attack is to be protected to fulfil both-party nonrepudiation.

3.2 Technologies and Solutions that Address the Security of SOAP Messages Transmission

SOAP does not yet have a standard binding for reliable messaging. The security provided by HTTPS cannot satisfy the more and more complicated requirement of SOAP message security. A number of technologies and solutions have been developed for the security of SOAP message transit. Several vendors offer reliable messaging solutions[6].

XML Encryption provides not only a way of encrypting portions of XML documents, but also a means of encrypting any data and rendering the encrypted data in XML format. XML Encryption is ideal for confidentiality. The ability to selectively encrypt XML data makes XML Encryption very useful for Web Services. By selectively encrypting data in the SOAP message, certain information may be hidden from SOAP intermediaries as it travels from the originator to the destination Web Services[12].

XML Signature explains how to express the digital signature of any data as XML, as well as explaining how to digitally sign portions of an XML document. The power of XML Signature for Web Services is the ability to selectively sign XML data. For example, if a single SOAP parameter needs to be signed but the SOAP message's header needs to be changed during routing, an XML Signature can be used that only signs the parameter in question and excludes other parts of the SOAP message. If the SOAP request passes through intermediaries en route to the destination Web Service, XML Signature ensures end-to-end integrity[12].

Security Assertions Markup Language (SAML) provides a means of expressing information about authentication and authorization, as well as attributes of an end user in XML format. SAML does not provide authentication, but can express information about an authentication event that has occurred in the past. By authenticating once, being authorized, and effectively reusing that authorization for subsequent Web Services, single sign-on for Web Services can be achieved. If an entity is authorized based on the fact that they were previously authorized by another system, this is called "portable trust[10]".

The XML Key Management specification (XKMS) enables PKI services such as trustworthily registering, locating, and validating keys through XML-encoded messages. PKI is a system that allows public keys to be trusted by providing key signing and key validation services. Although accepted as an important, even vital, technology, PKI has a reputation for being notoriously difficult to implement. By leveraging the benefits of XML and by learning from past experiences with pre-XML PKI architectures, XKMS makes PKI practical for common use[10].

Microsoft's Passport technology takes a different approach to single sign-on. The user authenticates to the passport infrastructure, either directly through www.passport.com or through an affiliate site that makes use of functionality provided by passport.com. Once the user is authenticated and authorized by Passport, their authentication status is also available to other Web Services that use Passport[10].

Another industry proposal for the single sign-on on the Web is the Liberty Alliance Project, championed by Sun. The Liberty Alliance Project aims to enable a non-centralized approach to single sign-on, termed a "federated network identity." It appears the Passport proposal by Microsoft may be taking a similar tack to the Liberty Alliance Project[10].

WS-Security, which has emerged as the de facto method of inserting security data into SOAP messages, is primarily for securing SOAP messages. WS-Security explains how technologies such as XML Signature, XML Encryption, and SAML are used for Web Services security in particular. WS-Security defines placeholders in the SOAP header in order to insert security data, how to add encryption and digital signatures to SOAP messages, how security tokens are contained in SOAP messages, and how XML Security specifications are used to encrypt and sign these tokens. In practice, this means defining the XML elements and attributes that are used to enclose tokens into SOAP messages, and the means to enclose XML Signature and XML Encryption into SOAP[5].

3.3 The Architecture and Mechanism of the Secure Transmission Model

A model of secure transmission of SOAP message is developed here to fulfill the security requirement. The building blocks of the model includes: confidentiality, authentication, integrity, both-party nonrepudiation, and single sign-on. Security of the model is achieved through inserting security blocks into SOAP header, as well as adopting technologies such as XML Encryption and XML Signature. Figure 2 is the architecture of the model.

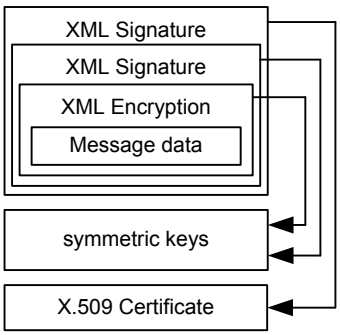


Fig. 2 Architecture of the model

The arrowed lines in Figure 2 represent the reference to the keys or token.

The basic idea is: encrypting the body data using the symmetric keys, signing the encrypted data using the symmetric keys again, and then signing again the signed data, making use of the private key provided by the X.509 certificate of the recipient.

Firstly, XML Encryption is implemented upon the message data, to realize the confidentiality of message data. The result of the encryption to a resource forms EncryptedData, which will replace the original resource being encrypted. How many resources are there to be encrypted, as many EncryptedDatas will be generated. Here, encryption to message data adopts symmetric keys, which are produced randomly every time.

Secondly, XML Signature is implemented to realize the integrity of message data. It includes three steps to construct an XML Signature: to make digest of the object to be signed, to sign the digest using the signature method, and to encrypt the digest, still using symmetric keys. Through decryption, digest verification and signature verification, the recipient can verify the integrity of message.

Finally, another XML Signature is implemented upon the result of first signature, adopting the private key provided by the digital certificate defined in the security block. This additional XML Signature using the digital certificate of the recipient is the key of the model to implement sender's nonrepudiation and single sign-on. Authentication is realized inside the process of single sign-on as part of the latter. The recipient's signing the response message using the digital certificate of the sender is the key of the model to implement recipient's nonrepudiation.

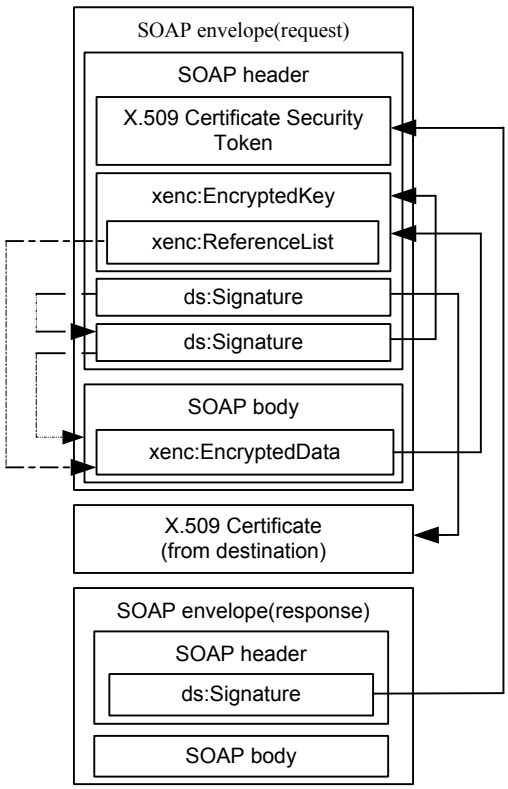


Fig. 3 Mechanism of the model

After being received, the second XML Signature is decrypted and verified. If the integrity is available, the keys, signature method, digest method, and encryption key can be obtained credibly. Thus, the original SOAP information is securely transmitted from the sender to the recipient. The client X.509 certificate and server X.509 certificate supply the asymmetric keys which are necessary in secure transit of the symmetric keys used in XML Encryption and XML Signature.

Figure 3 shows how the security mechanism of the model is established. In Figure 3, the arrowed solid lines represent the reference to the keys or token, well the arrowed dashed lines represent the secure operation.

An example of the implementation of the preceding process is listed as following.

```
<?xml version="1.0" encoding="utf-8"?>
<SOAP-ENV:Envelope xmlns:
SOAP-ENV="http://www.w3.org/2001/12/soap-
envelope"
 xmlns:ds="http://www.w3.org/2000/09/xmldsig#"
 xmlns:xenc="http://www.w3.org/2001/04/xmlenc"

xmlns:wsu="http://schemas.xmlsoap.org/ws/2002/07/ut
ility">
 <SOAP-ENV:Header>
  <wsse:Security

xmlns:wsse="http://schemas.xmlsoap.org/ws/2002/sece
xt">
    <wsse:BinarySecurityToken wsu:Id="X509token"
     ValueType="#X509v3"
     EncodingType="#Base64Binary">
     ……
    </wsse:BinarySecurityToken>
    <xenc:EncryptedKey wsu:id="userSysmetricKey">
     <xenc:EncryptionMethod
      Algorithm="……"/>
     <ds:KeyInfo>
      <wsse:SecurityTokenReference>
       <wsse:Reference URI="#userSysmetricKey"
       ValueType="....."/>
      </wsse:SecurityTokenReference>
     </ds:KeyInfo>
     <xenc:CipherData>
      <xenc:CipherValue>…...</xenc:CipherValue>
     </xenc:CipherData>
     <xenc:ReferenceList>
      <xenc:DataReference URI="#DataBeEncrypted
"/>
     </xenc:ReferenceList>
    </xenc:EncryptedKey>
    <ds:Signature wsu:id="originSignature">
     <ds:SignedInfo>
      <ds:CanonicalizationMethod
Algorithm="……"/>
      <ds:SignatureMethod Algorithm="……"/>
      <ds:Reference URI="#BodyData ">
       <ds:DigestMethod Algorithm="……"/>
       <ds:DigestValue>…...</ds:DigestValue>
      </ds:Reference>
     </ds:SignedInfo>
     <ds:SignatureValue>…...</ds:SignatureValue>
     <ds:KeyInfo>
      <wsse:SecurityTokenReference>
       <wsse:ReferenceURI="#userSysmetricKey"
       ValueType="....."/>
      </wsse:SecurityTokenReference>
     </ds:KeyInfo>
    </ds:Signature>
    <ds:Signature>
     <ds:SignedInfo>
      <ds:Reference URI="#originSignature">
       <ds:DigestMethod Algorithm="……"/>
       <ds:DigestValue>…...</ds:DigestValue>
      </ds:Reference>
     </ds:SignedInfo>
     <ds:SignatureValue>…...</ds:SignatureValue>
     <ds:KeyInfo>
      <wsse:SecurityTokenReference>
       <wsse:Reference URI="X509token"
      </wsse:SecurityTokenReference>
     </ds:KeyInfo>
    </ds:Signature>
  </wsse:Security>
 </SOAP-ENV:Header>
 <SOAP-ENV:Body wsu:Id="BodyData">
<xenc:EncryptedData
    wsu:Id="DataBeEncrypted"
    type="……">
    <xenc:EncryptionMethod Algorithm="……"/>
    <CipherData>
     <CipherValue>…...</CipherValue>
    </CipherData>
  </xenc:EncryptedData>
 </SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

4. Security Analysis of the Secure Transmission Model

XML Encryption to the SOAP message body realizes the confidentiality of the data, and XML Signature to the encrypted data realizes the integrity of the data. Evidently, the encrypting(decrypting) speed of symmetric keys is much faster than the encrypting(decrypting) speed of asymmetric ones. The practice that the symmetric keys used in XML Encryption and XML Signature are produced randomly is securer than the symmetric keys produced using hashing method, because as for the latter, once the keys were captured, succedent transmission would lose security. Through introducing the both-party X.509

certificate (those of client and server), which contain both the asymmetric keys and the identity information of the entity, the preceding security transmission model of SOAP messages acquires a high-level transmission security, as well as enjoys the benefit of high efficiency. The solution to single sign-on is to include information about the end user in the SOAP message itself. Furthermore, by making use of the identity information of the entity, the transmission mechanism realizes both-party nonrepudiation and single sign-on. The model itself is simple and light, but its running requires the support of certificate release of requester and responser. That is, the main load of the whole work is borne by certificate infrastructure. This maybe represents the shortcoming of it.

5. Conclusion and Expectation

Aiming at the challenges that SOAP messages transmission faces in Web Services applications among enterprises, a simple and light transmission model is developed, based on existing technologies. Analysis indicates that the model fulfils the security requirement of SOAP messages transmission: confidentiality, authentication, integrity, both-party nonrepudiation, and single sign-on, enjoying well advantage and efficiency. The cost of this kind of advantage and efficiency is the deployment of X.509 digital certificate on applications communicating via SOAP messages.

It is important to keep the entire security context of the Web Service in mind. This includes properly configured firewalls, the use of patched and locked-down Web servers, and (especially if digital certificates are used) the use of an adequate security policy document. It would be foolish to address just the new security challenges posed by Web Services and leave a system open to attack through more traditional channels.

There is a lot of work to do to strive for higher security of SOAP messages transmission, or even Web Services. To heighten the security and efficiency of the model, a particular block can be inserted into the SOAP header. Add a mustUnderstand="true" attribute to the header block, and require that the recipient must understand it. If this flag is present, and the recipient does not understand the block to which it is attached, the recipient must reject the entire message. In addition, the model developed in this paper should be strengthened to avoid the risk of reply attack.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. U1204614, No. 61003035 and No. 61142002, and in part by the Plan for Scientific Innovation Talent of Henan Province under Grant No. 124100510006, and in part by the Science and Technology Development Programs of Henan Province under Grant No. 112102210187, and in part by the Youth Foundation of Henan University of Science and Technology under Grant No. 2011QN51.

References

[1]David Chappell,Tyler Jewell, Java Web Services, O'Reilly, March 2002, 28-50.
[2]Dongxi Zheng, Shaohua Tang,Shaofa Li, "XML Web Services Security Technology Overview", Computer Engineering and Application, 2004.7, 38-41.
[3]Doug Tidwell, James Snell, Pavel Kulchenko, Programming Web Services with SOAP, O'Reilly, December 2001, 39-61.
[4]IBM developerWorks. http://www.ibm.com
[5]International Business Machines Corporation, Microsoft Corporation,VeriSign, Inc., Web Services Security (WS-Security) Version 1.0, April, 2010.
[6]Jian Jin, Hong Zhang, Jiahua Liang, Hualin Qian, "Analysis of Web Services Security", Micro-electronics and Computer, 2004.3, No3, Vol 21, 19-24.
[7]Jimei Wang, Lianfu Jin, "RESEARCH AND RESOLUTION ON WEB SERVICE SECURITY", Computer Applications and Software, February, 2004, No 12, Vol 21, 91-93.
[8]Keith Ballinger, .NET Web Services: Architecture and Implementation, Addison Wesley, February, 2003, chapter 9.
[9]Luciano Baresi, Elisabetta Di Nitto, Test and Analysis of Web Services, Springer, March 2011, 395-440.
[10]Mark O'Neill et. Web Services Security, McGraw-Hill/Osborne , 2003, chapter 3,4,5,9.
[11]Xiaoning Xu, "Security Study on transport of SOAP messages on Web Services", Information Security, 2011, No 11-3, Vol 22, 115-117.
[12]ZDNetChina community. http://www.zdnetchina.com

Haixia Zhao received her B.S. degree from South West Normal University in 1998 and M.S degree from National University of Defense Technology in 2005. She works as a Lecturer in Henan University of Science and Technology from 1998 to now. In particular, her research interests include wireless sensor networks, Internet of Things, cognitive network, database theory and technology etc.

Yaowei Li received his B.S. degree from National University of Defense Technology in 1998 and M.S degree from National University of Defense Technology in 2004. He works as a Engineer in LuoYang Electronic Information Equipment Testing Center from 1998 to now. In particular, his research interests include Information security, Internet of Things, cognitive network etc.

Mingchuan Zhang received his B.S. degree from Luoyang Institute of Technology in 2000 and M.S degree from Harbin Engineering University in 2005. He works as a Lecturer in Henan University of Science and Technology from 2005 to now. In particular, his research interests include ad hoc network, Internet of Things, cognitive network and future Internet technology.

Ruijuan Zheng received her B.S. degree from Henan University in 2003, studied in Harbin Engineering University from 2003 to 2008, and received Ph.D. degree. She works as an Associate Professor in Henan University of Science and Technology from 2008 to now. In particular, her research interests include bio-inspired networks, Internet of Things, future Internet and computer security.

Qingtao Wu received his Ph.D. degree from East China University of Science and Technology. He works as an Associate Professor in Henan University of Science and Technology from Mar 2006 to now. His research interests include component technology and future Internet security.

Smart dynamic software components enabling decision support in Machine-to-machine networks

Alexander Dannies[1]*, Javier Palafox-Albarrán[1], Walter Lang[1] and Reiner Jedermann[1]
[1] Institute for Microsensors, -actuators and -systems, University of Bremen
Bremen, Bremen, Germany

Abstract
The future Internet of Things will be extended by machine-to-machine communication technologies in order to include sensor information. The overwhelming amount of data will require autonomous decision making processes which are directly executed at the location where data is generated or measured. An intelligent sensor system needs to be able to adapt to new parameters in its surrounding unknown at the time of deployment. In our paper we show that Java enables software updates on mobile devices and also that it is possible to run algorithms required for decision making processes on wireless sensor platforms based on Java.

Keywords: *Machine-to-Machine communication, Internet of Things, autonomous logistics, Java, dynamic updates, OSGi.*

1. Introduction

Classical machine-to-machine (M2M) communication focuses on the supervision of large, expensive machinery, and on remote monitoring in a centralized point. But in the future, according to the vision of the internet of things (IOT), in which smaller physical objects will interact with each other through the use of the M2M concept, communication will become more and more ubiquitous.
The advancement of M2M communications from single machines to supervision of a network of objects will not be made by simply increasing the number of existing system and hardware solutions. After defining the requirements for integration of M2M into the IOT [1], adequate communication and software structures have to be found and programmed onto the hardware. In this paper we will discuss and demonstrate by our prototype implementation how ubiquitous M2M can be enabled by combining and reprogramming system components which are available in the market.

1.1 Combining cellular and infrastructureless networks

Nowadays, M2M communication is typically implemented by cellular radio networks (CRN) technologies, such as GSM and UMTS. The infrastructure of a commercial network operator consists of fixed base stations to cover large geographical areas. In order to make M2M technologies more ubiquitous, devices have to collect information from a high number of devices distributed in the environment. For such a detailed supervision CRN are rather disadvantageous for the following reasons:

- Communication costs have to be kept as low as possible.
- Network protocols have to be optimized for transmission of small packets of sensor data consuming as little energy as possible instead of enabling global communication.
- In many applications, such as the monitoring inside large buildings or rural regions, the supervised area will not be fully covered by the CRN of an external operator.

If wireless connectivity is required, local infrastructureless networks are the better solution for spatial monitoring of an area. Typical Ad-Hoc wireless sensor networks (WSN) using the Zigbee or the underlying 802.15.4 protocols, meet the requirements mentioned above. They provide coverage of even difficult areas by forwarding messages over multiple hops inside the network.

But on the other hand, pure WSN lack the ability to connect to global networks. Therefore, we suggest using a heterogeneous network combining infrastructure and infrastructureless technologies to enable future M2M networks which will not only supervise single machines, but be aware of their environment.

1.2 Local intelligence by Java-based dynamic software frameworks

The vast amount of data, provided by a distributed M2M network needs dedicated processing. According to the concept of cloud computing, the required computation resources can be provided as service by the network. The resources can be hosted by a stationary server farm as in [2] or [3], but in the case of M2M networks a more direct approach is to move the “cloud” into the network by processing collected data directly on the sensors. This approach entails advantages in regard to costs and robustness:

- The costs for the transmission of large amounts of unprocessed sensor data cannot be neglected, they have to be either paid directly to an external operator, or have to be calculated as service costs, for the current that the radio draws from the batteries. The communication volume decreases dramatically, if an intelligent processing directly on the sensor transmits only summaries, conclusions or warning messages about unexpected situations instead of the full raw data over the network.
- If the infrastructure or part of the WSN fails, processing can be continued by the remainder of the local network.

Immutable software, which is programmed in a static way and transferred to the sensor before distribution in the field, is unsuitable. The sensor software has to adapt to new situations, tasks and application fields, which were unknown at the time of deployment. This results in a permanent need for software updates. As a consequence the network nodes have to be equipped with an adequate operating system or software framework. In order to reduce the size of update files, the software should be structured in a modular way, whereby it is possible to update only single components of the software.

In [4], over-the-air and differential reprogramming in WSN is made; however, the applicability of the solution is limited to some hardware devices.

Java is the most common language that meets this requirement because dynamic class loading is one of its intrinsic features. It has penetrated more and more the realm of embedded systems. Optimized virtual machines have become available for several embedded controllers [5], [11].

1.3 Testing Java-enabled wireless sensor nodes and M2M platforms

Because of the enormous spreading and pervasiveness of Java, we focused on this language to implement an intelligent sensor node. Several Java-enabled wireless sensor nodes and CRN-enabled M2M devices were tested. Depending of the available resources of the device, two different software frameworks for handling of code updates were installed. The wireless devices were tested in regard to their ability to execute and update complex software algorithms within their computation and battery capability.

By measurement of required CPU time we could show that there are several Java-enabled wireless sensors platforms, which are capable of running complex algorithms as well as frameworks for automated software updates. Differences in the performance of the tested types of WSN hardware are evaluated by measurement of execution time for benchmark tests and example sensor data processing algorithms.

Our test bed shows how a combined network of infrastructure CRN and infrastructureless WSN can be installed. By local pre-processing on wireless nodes the network can provide a new quality of information to the end-user. World-wide access to the M2M system from the internet is provided by a web interface.

Furthermore, we could demonstrate by our prototype implementation the advantages of such an intelligent network for a logistic supervision and decision support tasks.

In section 2 we give an overview of the theory and background related to our paper. Subsequently, section 3 introduces the platforms for dynamic software updates and section 4 contains the performance measurements of the selected wireless sensor platforms. Section 5 describes the topic of software updates. Finally we summarize in section 6.

2. Background and vision of M2M and IOT

The initial idea behind the creation of the Internet of Things was to interconnect real-world objects globally. It emerged under a logistic point of view in which the items would be tracked over the existing internet. Its development of the communication technology has been built on top of it.

IOT means the connection of clearly recognizable physical objects (Things) with a virtual representation in an internet-like structure. Participants in the IOT are not only of human nature but also machines.
When the IOT concept was created, passive RFID (Radio Frequency Identification) and barcode were already mature technologies for item identification and tracking; the identification on the internet was made by manual inventory. RFID was used however because it does not require line of sight and requires less human intervention than barcode. Automated identification with the help of RFID is often considered as the foundation of the IOT. Its target is the minimization of the information gap between real world and virtual world.

Because some applications require communicating without human intervention, concepts such as
M2M communications came into mind as possible extensions of the IOT concept. M2M was an already existing technology which allowed automated exchange of

information among terminal devices like vending machines, vehicles or containers with a central point. M2M technology connects information and communication technology to build the Internet of Things (IOT).

2.1 Definition of M2M and available technical solutions

A M2M system consists of three basic components. A data end point (DEP) which can be a machine extended by a sender module with the main task of providing data. The second component is a communication network; this can be either wireless or wired. The data integration point (DIP) mainly plays the role of the gateway. It receives information from the DEPs and redirects it to a central point. Several commercial technical solutions for M2M can be found in the fields of industrial automation, transportation, smart energy and logistics.

M2M can be classified according to the physical transmission media it uses: The media can be either based on wired (Ethernet or optical), cellular (for example GSM, GPRS, UMTS, LTE-M and WiMAX) or "capillary" short-range technologies (for example Bluetooth, ZigBee, IEEE 802.15.4).

All of them offer advantages and disadvantages: wired communication offers the best reliability and highest data rates, but is expensive, complicated to install and not scalable. To bridge long distances the communication standards of 2G (GSM) or for higher data rates 3G (UMTS) can be used, but are expensive in maintenance and need a fixed infrastructure. Wireless sensor networks using protocols such as 802.15.4 at 2.4GHz are cheap, scalable, do not need infrastructure, but have drawbacks such as low coverage, security and data-rates and energy constraints.

2.2 The impact of M2M to logistic processes

The supervision of supply chains and logistic tasks is one of the main application fields of the IOT. The IOT differs from the idea of autonomous control in logistics [6]. A fully autonomous object requires basically only communication with its near-by neighbors and not necessarily internet-like networked structures.

Communication between RFID tags and a central point as in conventional M2M allowed complying with a ubiquitous necessity. Tracking the position of the identified items globally, seemed to solve it, however one consideration was missing: In the process of transportation damages such as spoilage or breakages may appear and it is required to know not only the position of the item but also whether its quality is acceptable.

That is the reason why the emergent technology of WSN, together with RFID, can be seen as the enabling concept of IOT. In WSN's the nodes have sensing, communication and processing capabilities and use M2M to communicate with each other and with the gateway.

Conventional M2M solutions require four basic steps: Data collection, transmission, assessment and response. But the gathering and transmission of all the available data alone can lead to a flood of information and asks too much of a human operator and is extremely costly. The entire decision making should be done autonomously, at best in the same location where the data is collected.

Our vision of an intelligent sensor network, from M2M to the IOT, proposes a change of paradigm in which the assessment (data processing) is performed locally in the wireless sensor nodes or on the gateway device. The concept of the intelligent container [7] includes the introduction of a decision support tool (DST) which can, as the name suggests, support humans in making decisions based on the sheer abundance of data occurring every second. The quality of perishable goods like fresh fruits or meat has to be monitored to ensure that the food reaches the end-consumer in the best state possible. On the other hand the economic aspect of the supplier benefits from the monitoring because losses due to reduced shelf-life caused by broken cool chains can be absorbed by intervening in logistics processes. Moreover, as mentioned in [1] the DST should provide device control, which includes activating, deactivating or updating the devices over the air.

Supervision of logistics processes is often limited to data-logging during the transportation and analyzing this data afterwards. Reactions to unexpected situations can only be triggered with long time delay or in the worst case an intervention is not possible anymore.

In order to be able to react to these events on time, it is necessary to monitor the cargo objects in a pervasive way, which means anywhere anytime. The "Internet of Things" can solve this problem. By creating a network of pervasive systems it becomes possible to collect real-time information with simultaneous consideration of the decision making process.

The objects or things use M2M communication to access the real-time data without human intervention. They can for example send an e-mail or SMS to a human with the condition information to be acted on at a reasonable price. Due to the increasing processing power (according to Moore's Law) on the one hand and the decreasing costs for hardware in general on the other hand, the feasibility for an implementation of omnipresent data processing by an advanced internet of things rises. The condition of the cargo, that may be for example signs of degradation, is

calculated by intelligent data processing algorithms in wireless sensor nodes.

As mentioned in [8], M2M enabled intelligent devices like the ones visualized in the concept of the IOT will impact the logistic process mainly in three ways: Self-aware products, delivery by product characteristics and proactive tendering.

In self-awareness, the cargo or thing is able to react to self-related problems as soon as they arise. The problems can be for example deterioration caused by fluctuations of environmental parameters such as temperature or humidity, leading to quality degradations such as decrease in aesthetic appeal. Algorithms to estimate the quality of the goods by biological models may be used. The Gompertz model will be introduced as example in section 4.2.2.

Quality of perishable goods such as fruit and vegetables are highly dependent of failures in the cold-chain. If the quality decreases, the delivery of the good has to be re-planned according to the actual product characteristics. New routes and alternate suppliers or buyers have to be found in accordance with the actual price of the cargo and with the aim of increasing the profitability.

In proactive tendering early information about the quality of the cargo will lead to take actions to reduce waste or to replan the supply orders. Prediction algorithms, such as Feedback-Hammerstein (section 4.2.1) can be applied to compute a model for temperature changes.

3. Platforms for dynamic software management of embedded devices

As mentioned before, our vision of an intelligent sensor network includes device control to react on dynamic changes in the environment. The sensor nodes must not only be able to update the software modules over-the-air, but also to do it dynamically during run-time. WSN is still an emergent technology; the research focus has been mainly on energy efficient algorithms. The question arises whether it is feasible to implement the mentioned solution in an energy-efficient way on resource-limited devices.

Typically, WSN nodes are programmed once, in native code such as nesC without taking into consideration neither modularity, over-the-air (OTA) programming or dynamic features.

In [9] three execution environments for software update management in sensor networks are compared: monolithic (TinyOS), modular and virtual machines (VM). VMs interpret symbolic or intermediate code instead of directly transferring and executing machine code. The size of a program in an intermediate code such as Java class files is between five and ten times smaller than the same program in machine code. Han [9] concludes that VM is the best one regarding the energy costs of network transmission. He also concludes that if VM is combined with a modular environment, the energy costs of updating a task are very low. The only disadvantage of using a VM environment is the cost of interpretation. As mentioned before, IOT should combine the best of capillary and cellular data transmission. Specialized VMs, which are written to run on sensor nodes, such as Maté [10], only cover the first one mentioned (short range) and are not suitable for cellular networks.

Beside these WSN-specificVMs, Java is a mature technology to run intermediate code on a VM. Java implements the concept of "write once, run anywhere". It is the most common language that meets the requirement of the ability to extend software with dynamic code segments through the use dynamic class loaders.

In this section we will discuss the advantages and hardware requirements of virtual machines and software frames for enabling dynamic updates.

3.1 Native Code versus virtual machine

On the selection of the software platform for dynamic updates, a series of figures of merit have to be taken into account. The software must support updates, be fast and able to run on diverse hardware platforms. Basically, there are two types of programming languages: the so called high-level and the interpreted ones, each one of them with their advantages and disadvantages.

When speaking about workstations, the high-level or native languages such as C or C++ have faster execution times and allow memory management but the code has to be compiled according to the hardware. On the other hand the interpreted languages such as Java or C# are platform-independent but the execution time is in general not optimal. This disadvantage can be compensated by Just-in-time compilers, which translate only those parts of the code to machine instructions that are most critical for the executions speed.

With the development of WSN as an emergent technology, it was clear that the solutions were not suitable to be used on the first sensor nodes available to the market because of their very constrained resources. Initially, native code such as nesC was used but they are not able to update software or run on different platforms.

Different VMs have been implemented for sensor nodes. One example is the above mentioned Maté [10]. It allows executing high level instructions by an interpreter. New application scripts can be sent over the air, requiring only

very small communication volume and user memory on the microcontroller.

In the recent years there have been lots of efforts to provide VMs for high level languages such as Java on sensor nodes [11]. Depending on the memory and CPU recourses on the sensor node, either the Java Micro or Standard Edition is supported.

3.2 The Java Micro Edition on sensor nodes

Because of the hardware-constrained nature of microcontrollers, there is not enough space on them to install a full operating system which can be used as a base for a virtual machine. In contrast microcontrollers are using a virtual machine which runs on bare metal. The kilo VM (kVM) requires only a few 100 kBytes of memory and runs on ARM processors. One example implementation is the Squawk VM used by Oracles SunSPOT [11]. It includes the functionally of the Java Micro Edition (JavaME) as part of the Connected Limited Device Configuration (CLDC). New software components can be uploaded in the form of software suites containing MIDlets (see section 3.4.3).

A further example for the implementation of a kVM is the Preon32 sensor node by Virtenio [12]. Its kVM does not exactly cover the whole CLDC standard but is close to it. As consequence it is not possible to install MIDlets on Preon32 nodes.

Floating point and double precision data types are not part of the original JavaME but were introduced in CLDC 1.1. Although the Java SE Math Library is not available by default, manufacturers, such as Oracle and Virtenio, have implemented their own library for mathematical functions.

3.3 The Java Standard Edition on sensor nodes

There is a variety of VMs supporting the Java Standard Edition (JavaSE) on the market, both open-source and commercial ones. We selected JamVM as a representative of the open-source type and JamaicaVM from AICAS as a representative of the commercial ones. Both of them are able to run on workstations or on embedded devices.

JamVM makes use of the GNU Classpath [5]. Their implementation is more suitable for sensor nodes, is extremely small but still able to support the full specification including, class-unloading and native support. It can be installed on several operating systems like Linux, Mac or Solaris as well as different hardware architectures like PowerPC, ARM or AMD64.

JamaicaVM of AICAS [13] provides Hard Realtime Execution, Realtime Garbage Collection, dynamic loading, multi-core support, and native support. It can be installed for diverse operating systems like Linux and Windows, and several architectures like x86 and ARM. Besides it offers the possibility of combining all files relevant for the application (a set of class files) and the Jamaica VM into a single executable file. The implementation offers a trade-off between run-time performance and code size.

JavaSE allows replacing the system class loader by user defined class loaders. This feature, which is not available in JavaME, is essential to control mutual access between different dynamic components (see section 3.4.1).

3.4 Java frameworks for dynamic code

Although it is possible to handle the dynamic loading of new software components by basic features of the Java VM, it is more efficient to use additional Java features or a software framework to handle updates:

- JavaME allows installing and executing new software components during runtime in the form of so called MIDlets [14]. It is commonly used for mobile devices such as cell phones, but also supported by up-to-date WSN hardware such as the SunSPOT sensor node [15].
- Agent platforms, such as MAPS, JADE [16] or Agilla [17], provide a framework which enables migration of software agents between different local platforms. The migration is in general based on an internet connection but can be adapted to the needs of WSN and CRN technologies.
- Whereas software agents have their focus on artificial intelligence and research, the Open Services Gateway initiative (OSGi) framework originates from industrial automation and building maintenance. New components can be installed as software bundles without the need to stop or restart the machine to perform a software update. Furthermore, OSGi provides methods to exchange information and services between different bundles.

Due to the dynamic features, efforts have been made to run OSGi on resource limited devices; OSGi has been tested in pervasive environments [18, 19 and 20] but not in sensor networks context, yet.

There is a major difference in the concepts behind agents and OSGi. OSGi components can be organized in a hierarchical structure. It is even possible to update components which are currently in use by another component. Software agents, on the other hand, are organized in a flat structure. They can exchange messages

for communication, but services are only provided by the framework, without any means to update or modify their implementation. Because we consider this feature to update components in a hierarchical structure as crucial for intelligent objects, we focus on OSGi as second example framework for dynamic code.

3.4.1 Inter component communication

The major task of a framework for handling dynamic software components is to install, upload and run different components in parallel and independently. But it is also necessary to provide some kind of communication between different components. The access to the code and memory of other components has to be restricted in order to avoid that one malfunctioning component can crash the whole system. The concepts range from full protection such as in Android, where apps can only access methods provided by the operating system, to controlled accesses by one of the following solutions:

- One solution to provide inter-component communication is the use of shared memory. As example we consider a system containing the three components “decision unit”, “sensor driver” and “radio driver”. The decision unit has to read sensor data and to communicate the result over the radio. Sensor and radio data are written to the memory by one component and polled by another component.
- A more efficient solution allows that certain parts of the code of one component can be invoked by another component. Only methods that are explicitly published as service can be accessed from the outside. If the component is updated, the framework has to redirect the service request to the new component. The decision unit calls services provided by the sensor and radio component.
- In a third solution, components can exchange messages or events by registering for a special service provided by the framework. For example, the sensor component informs the decision unit by an event if new measurement data is available. This third solution is the standard way of agent communication.

3.4.2 OSGi

OSGi was introduced in the 90’s to manage controller units for building maintenance remotely. The original idea was that a human operator can start, stop and update software components without being on site. But OSGi can also be used in a M2M way: a central unit or machine can control a remote unit by calling system functions provided by the framework. An agent-like migration is also possible: a component uploads its own code to another platform, starts it, and stops its own execution on the first platform.

OSGi provides two ways of inter-component communication. Components can call services published by other components. Or they can send an event to a blackboard. Other components can register as listeners for a certain type of event.

OSGi runs on top of a Java VM. The mutual access to code of different components by services is handled by user defined class loaders. Unfortunately, this feature is only available in the Java Standard Edition. As consequence, it is not possible to run OSGi on the SunSPOT platform. Furthermore, a typical OSGi framework needs at least 32 MByte of user memory, which is also not available on the SunSPOT.

There has been some effort to make OSGi services and dynamic uploads available for JavaME by a so called OSGiME framework [21]. This approach keeps the core features of OSGi like dynamic software updates but being compliant to Java ME CLDC means that user-defined class loaders are not available. To be able to fit on resource-constrained platforms the OSGi technology is simplified by removing unnecessary and semantically complex features like dynamic and optional imports.

A further minimal OSGi implementation by ProSyst, which requires only eight MByte of memory [23], does also not provide dynamic uploads.

3.4.3 JavaME and MIDlets

Recently JavaME has been integrated into WSN technology, for example the SunSPOT nodes use a proprietary Java VM, to offer a sensor node that can be programmed over-the-air, but only inside the sensor network. Updates over external global networks are not supported by the original software package.

shows the execution environment to enable MIDlets. On top of the host operating system sits the configuration which is extended by a profile and optional packages. As a configuration the Connected Limited Device Configuration (CLDC) is used, which contains a very small virtual machine (KVM). On top the Mobile Information Device Profile (MIDP) enables the execution of MIDlets. Writing MIDlets limits the programmer to the functions of JRE 1.3. After installation the MIDlet can be started, paused and destroyed by calling an interface function. The transition to the destroyed-state is irreversible.

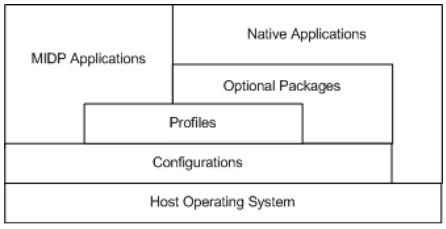


Figure 1: Structure of a MIDP environment

The SunSPOTs have modularity limitations inherent from any JavaME deployment because the MIDlets are unable to communicate among themselves. However, it is possible to overcome this disadvantage by using Record Management Storage (RMS) or programming them within the same suite. In theory Inter-MIDlet-Communication (IMC) is possible in MIDP 3.0, but currently a SDK is missing so that the current status is that MIDP2.0 is still used.

4. Java performance on WSN platforms

The interpretation of dynamic code by a VM requires some overhead, compared to precompiled "C" code. In order to decide whether this overhead hinders the application of Java VMs on WSN nodes we carried out performance measurements on different hardware platforms.

4.1 The hardware platforms

All hardware platforms we used are available off-the-shelf. They can be classified in two categories: telematics units and wireless sensor nodes. Table 2 shows selected characteristics for each hardware platform.

The telematics units are equipped with extended communications possibilities like 3rd generation mobile telecommunications (GSM / UMTS) and wireless LAN. Additionally, it is possible to get geodata via the global positioning system (GPS) and hard disk storage allows extensive data-logging.

The VTC 6200 from Nexcom is used as a reference platform. It is equipped with an Atom processor (1.6 GHz) and 2 GB of main memory.

The DuraNAV serves as an exemplary platform for lower power consumption. It utilizes an ARM architecture CPU (400 MHz) and 64 MB of RAM. Both can run different Java VMs (JamVM, Jamaica) and different OSGi implementations (Prosyst, Equinox).

Table 1: Telematics platforms

	DuraNAV	VTC6100
CPU (MHz)	PXA255 (400)	N270 (1600)
RAM	64 MB	1 GB
OS	Linux	Linux
Java Edition	SE	SE

In the category of wireless sensor nodes we chose three products, which are listed with its properties in Table 2.All these platforms enable the usage of the high-level programming language Java.

Table 2: 802.15.4 Wireless sensor platforms

	Imote2	Sun SPOT	Preon32
CPU (MHz)	PXA 271 (416)	SAM 9G20 (400)	Cortex-M3 (72)
RAM	32 MB	1 MB	64 kB
OS	Linux	None	None
JVM	any	Squawk	Custom
Java Edition	SE	ME CLDC 1.1	ME almost CLDC 1.1

4.2 Example algorithms

As example algorithms we utilized two synthetic standard benchmarks (Dhrystone and LINPACK), the inversion of a 20 by 20 matrix of double values as well as two real-world application algorithms.

Dhrystone is a synthetic benchmark using integer operations, whereas LINPACK uses floating point operations. The latter one calculates the average speed of floating point operations during solving a n by n system of linear equations Ax = b. In addition to the abstract LINPACK benchmark for matrix operations, we also tested the inversion of a 20 by 20 matrix in double precision as example for the computation needs of a more

complex sensor evaluation task. For the matrix inversion the functions of the JAMA library [24] were used.

4.2.1 Temperature prediction

The two real-world application algorithms can be applied in the field of logistics. The Feedback-Hammerstein-algorithm is used in the context of the transportation of bananas in a refrigeration container. Bananas are living organisms with a metabolism so they emit heat and gases such as CO_2 and C_2H_4 - a phytohormone which is responsible for the ripening process. The algorithm identifies the parameter for a model to predict the temperature curve during cooling. By taking the generated heat of the bananas into account the model accuracy is improved compared with a simple model that is based only on thermal time constants. The structure of the applied model is shown in Figure 2.

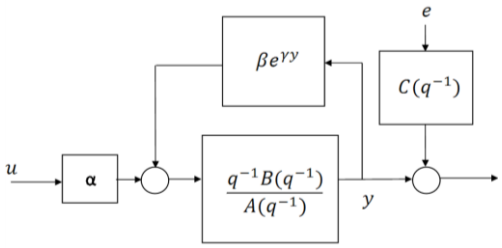


Figure 2: Feedback-Hammerstein Model

4.2.2 Gompertz-model

The Gompertz-model's application lies also in the carriage of fresh produce in the cold chain – transport of meat. A statement about the quality of meat can be done by using the bacteria count in a meat sample. The speed of bacteria growth is in general calculated as a function of temperature according to the law of Arrhenius for reaction kinetics. The bacteria count over time functions shows a lag phase during which the bacteria growth is delayed for a certain period of time. A combined model includes the Gompertz model to describe the lag phase and the Arrhenius model to describe to temperature dependency. By using several fixed values which are specific for a certain type of meat the number of bacteria can be estimated, which is correlated to the quality and the shelf-life [25]. The update of the model state for each measurement interval requires the calculation of three exponential and two logarithmic functions. If the temperature is the same as in the previous interval, the calculation of the logarithmic functions can be skipped.

4.3 Test results (execution speed and feasibility of algorithm)

The following Table 3 contains the results of the different benchmarks for the chosen reference platform (gateway device, Telematics unit VTC).

Table 3: Results of the chosen benchmarks on the reference platform (VTC)

Benchmark	Result
Dhrystone	523 ms
Linpack	45,778 Mflops/s
Feedback-Hammerstein	7 ms
Matrix inversion 20 by 20	1 ms
Gompertz model (single interval)	0,7 ms
Gompertz model (3450 intervals)	9.1 ms

In what follows all diagrams displaying the results of the measurements are relative values in comparison with the reference platform. Figure 3 shows the results of the Dhrystone-benchmark: A correlation between the processing power of the platforms and its available CPU and RAM is obvious. Even though Imote2 and SunSPOT have the same clock-rate of the CPU, the execution time seems to be linked to the available RAM of the systems. Consequently the order of the performances from slower to faster is the same as the amount of memory in ascending order.

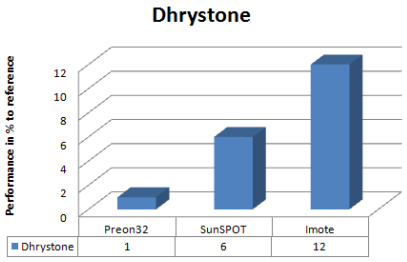


Figure 3: Results of the Dhrystone 2.1 benchmark

Unlike the first benchmark, Figure 4 shows a different correlation for floating point operations. In this case the influence of the amount of RAM at hand on the performance is less significant. The performance of the SunSPOT is better than that of the Imote2 for the LINPACK benchmark and for the matrix inversion. An explanation for this result could be the newer CPU architecture of SunSPOT, which seems to have improved floating-point processing power. The highest difference is observed for the inversion of a large 20 by 20 matrix. In this case the SunSPOT is eight times faster than the iMote at the same clock speed.

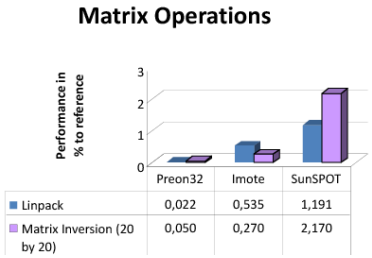


Figure 4: Benchmarks for LINPACK and matrix inversion

The results of the first tested exemplary real-world algorithm is shown in Figure 5. The Feedback-Hammerstein algorithm was executed with different orders. Similar to the previous benchmark the SunSPOT is three to four times faster than the Imote2 because of the newer CPU architecture.

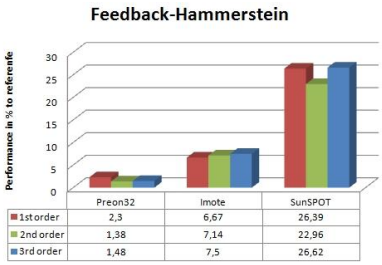


Figure 5: Results of the Feedback-Hammerstein algorithm

The results of the execution time of the second exemplary real-world-application algorithm are depicted in Figure 6. The Imote2 takes about five times more time for the execution than the SunSPOT.

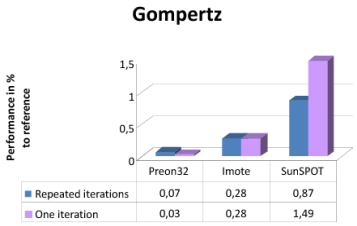


Figure 6: Results of the Gompertz-algorithm

In all five benchmarks the performance of the Preon32 was by far the lowest. But this result is no wonder because of the very resource-constrained platform. The CPU runs at only 72 MHz and only 64 kByte of RAM are available. In comparison to that the SunSPOT has a clock-rate of 400 MHz and 1 MB of RAM.

Even though the results in comparison with the other platforms at first appearance seem to be not that good, it is feasible to run these algorithms on the platforms: For example, the time required to execute 72 iterations of FH-algorithm required for three days of hourly samples is about two seconds. The Gompertz-algorithm takes about three to twelve seconds depending on the amount of temperature changes, which is also fast enough taking into account that a change in temperature is measured only once a minute.

4.4 Comparison of framework performance

In the previous section we have shown that all tested hardware platforms are capable of executing typical algorithms for sensor data evaluation. But it still has to be questioned, how much overhead is created by using a framework to manage software updates. To answer this question, the execution time for different algorithms was compared for a) direct execution as a Java class file and b) running them as software bundle inside the Equinox OSGi framework.

The only platforms able to perform these tests are the gateway devices and one of the sensor nodes – the iMote2. Because of the similar processing power we compared it with the DuraNAV system.
Figure 7 compares the execution time on Imote2 and DuraNAV of the benchmarks as a class-file or an OSGi-bundle.

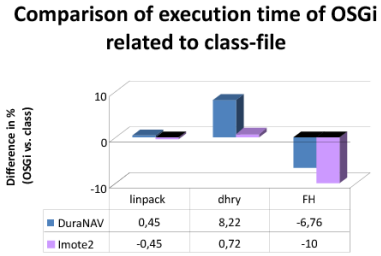


Figure 7: Difference in execution time in % between OSGi and class

The result of our test is that the execution time of the synthetic benchmarks was of the order of eight percent slower when the algorithm was run in the OSGi-environment instead of a direct execution of a class file. A completely different result is generated when comparing the real-world-example-algorithm of Feedback-Hammerstein. In this case the execution time of the OSGi-bundle was approximately seven to ten percent faster than running the algorithm directly from a class file.

From these results one can infer that the use of a framework can make sense. Not only the ability for dynamic software updates becomes possible but also the execution time is not increased in a way that the advantage is negated – even an improvement of execution speed is possible, depending on the type of algorithm.

Of course it remains to be seen if and how a framework can be introduced to resource-constrained wireless sensor nodes in the future and if the behaviour of this implementation will be similar.

5. Software updates over multi modal networks

Oracle offers for its sensor node SunSPOT a graphical user interface (Solarium) and also a command line tool based on ANT to deploy software connected to the computer via USB or over the air. The term over the air (OTA) programming means the distribution of new software updates to wireless devices without the need of a cable connection. By using one sensor as a base station it becomes possible to communicate with other sensors of the WSN in range of this platform.

The disadvantage of this approach is that in order to be able to use this software the user needs to install it on his machine. In contrast to that, our approach is of a web-based nature. As a result anybody can use it from a remote location via the internet.

5.1 Our test and demonstration system

The concept we used in our demonstration system is depicted in Figure 8.

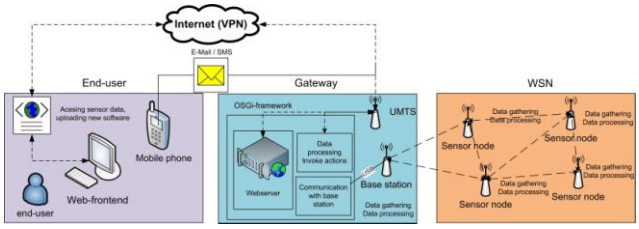


Figure 8: Concept of the demonstration system

From a remote internet connection the user opens a web browser and accesses a web interface, which is provided by the gateway device. This becomes possible by running an OSGi-Framework on the gateway device which hosts a bundle that publishes a web interface. The connection between the end-user and the gateway becomes possible due to the use of a virtual private network (VPN) connection which connects gateway device and end user in one space. The gateway device is connected to a sensor node (SunSPOT) that acts as a base station. In this way a connection through multimodal networks from the end-user (internet) to a sensor node (802.15.4 network) is possible.

The user can get information from any sensor in the network as well as update software on a specific sensor node remotely without being on site.

6. Summary

The Internet of Things is a concept in which objects use the infrastructure of the internet to communicate with each other in a global way. An essential part of the IOT concept is to enable objects to exchange data between each other autonomously, i.e. without human intervention. Autonomous communication between the objects requires sensing, evaluating and communicating. Environmental parameters are sensed, intelligent algorithms run on the sensor node using this acquired data and the result is transmitted wirelessly to a gateway. This leads to the ability to create autonomous decision making or supporting functionality to disburden human operators who otherwise would have to battle their way through a flood of raw data. The gateway which is the connection point to the outer world should have M2M communication capabilities with the sensors and with the internet infrastructure to allow pervasiveness. The impact of the use of all these technologies in the logistic process is mainly in three ways: Self-aware products, delivery by product characteristics and proactive tendering. However, the available technological solutions that make the IOT concept possible have their Achilles' heel in the sensor end-point. Over-the-air dynamic data programming is possible with off-the-shelf components. On the one hand the wireless sensor node SunSPOT from Oracle can be used as base station as well as part of the WSN. On the other hand commercial telemetric units such as DuraNAV and VTC, with Linux as operating system, can serve as a gateway device. In combination with OSGi as software framework a user can remotely update software in the WSN from any location using a web interface. Java on the sensor nodes is useful, because the communication volume for updating software bundles is lower than in the case of monolithic software. However, JavaME running on sensor nodes does not yet allow communication between MIDlets therefore the modularity is limited due to missing communication between different modules.

Acknowledgments

The research project “The Intelligent Container” is supported by the Federal Ministry of Education and Research, Germany, under reference number 01IA10001. The current study is also supported
by International Graduate School in Dynamics in logistics at Bremen University.

References

[1] S. Tompros(Ed.),Internet-of-Things Architecture IOT-A Project Deliverable D3.1 - Initial M2M API Analysis

[2] Integrating K. Ahmed, M. Gregory, Integrating Wireless Sensor Networks with Cloud Computing, Seventh International Conference on Mobile Ad-hoc and Sensor Networks,2011, pp. 364-366.

[3] W. Kurschl, W. Beer, Combining cloud computing and wireless sensor networks. In Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services (iiWAS '09). ACM, New York, NY, USA,2009, pp.512-518.

[4] N.F. Shafi, Efficient Over-the-air Remote Reprogramming of Wireless Sensor Networks,Master thesis,Queen's University Kingston, Ontario, Canada,2011.

[5] R. Lougher, 2010, JamVM [Online]. Available: http://jamvm.sourceforge.net/ [Accessed 20.07.2012].

[6] K. Windt, M. Hülsmann, Changing Paradigms in Logistics - Understanding the Shift from Conventional Control to Autonomous Cooperation and Control. In: Understanding

Autonomous Cooperation and Control - The Impact of Autonomy on Management, Information, Communication, and Material Flow, (M. Hülsmann, K. Windt, eds.) pp. 4-16, Springer, Berlin, 2007

[7] W. Lang, R. Jedermann, D. Mrugala, A. Jabbari, B. Krieg-Brückner, K. Schill, The Intelligent Container - A Cognitive Sensor Network for Transport Management. In: IEEE Sensors Journal Special Issue on Cognitive Sensor Networks, 11(2011)3, 688-698

[8] H.Sundmaeker, M. Würthele, S.Scholze, Challenges for Usage of Networked Devices Enabled Intelligence, Vision and Challenges for Realising the Internet of Things,CERP-IoT – Cluster of European Research Projects on the Internet of Things,2010, pp. 93-103. Available from http://docbox.etsi.org/tispan/open/IoT/CERP-IOT_Clusterbook_2009.pdf

[9] S. Han, R. Rengaswamy, R. S. Shea, M. B. Srivastava, Sensor Network Software Update Management: A Survey, Internation Journal of Network Management . 15 (2005) 283-294

[10] P. Levis, D. Culler, Maté: a tiny virtual machine for sensor networks. In Proceedings of the 10th international conference on Architectural support for programming languages and operating systems (ASPLOS-X). ACM, New York, NY, USA, 2002.

[11] D.Simon,, C. Cifuentes, D. Cleal, J.Daniels, D.White, Java(TM) on the bare metal of wireless sensor devices: the squawk Java virtual machine. In: Proceedings of the 2nd international conference on Virtual execution environments, Ottawa, Ontario, Canada, ACM, 2006. (doi: 10.1145/1134760.1134773)

[12] VIRTENIO. 2012. Available: http://www.virtenio.com/de/produkte/hardware/preon32.html [Accessed 20.07.2012].

[13] F.Siebert, Hard Realtime Garbage Collection. aicas GmbH, Karlsruhe, 2002.

[14] U. Breymann, M. Heiko, JAVAME Anwendungsentwicklung für Handys, PDA und Co. 2008

[15] Sun SPOT World. 2012 Available from http://www.sunspotworld.com/

[16] F. Bellifemine, G. Caire, A. Poggi, ,G. Rimassa, Jade - A White Paper. In: "EXP in search of innovation - Special Issue on JADE" TILAB Journal,3, (2003).

[17] F. Aiello, A. Carbone, G. Fortino, S. Galzarano, Java-based Mobile Agent Platforms for Wireless Sensor Networks, Proceedings of the 2010 International Multiconference on Computer Science and Information Technology (IMCSIT),2010 ,pp.165-172

[18] A.Ibrahim, L. Zhao, Supporting the OSGi Service Platform with Mobility and Service Distribution in Ubiquitous Home Environments. Comput. J. 52, 2 (2009) 210-239 DOI=10.1093/comjnl/bxn032 http://dx.doi.org/10.1093/comjnl/bxn032

[19] M. Desertot, S. Do, D. Donsez, M. Bui,Mobile Agents Platforms over OSGi,Proc. of 4th International Conference on Computer Sciences, Research Innovation and Vision for the Futur, RIVF'06, 2006.

[20] S. K. Lee, J. H. Lee, OSGi based service mobility management for pervasive computing environments. In Proceedings of the 24th IASTED international conference on Internet and multimedia systems and applications (IMSA'06), 2006, pp.159-164.

[21] A. Bottaro, F. Rivard, OSGi ME An OSGi Profile for Embedded Devices. OSGi Community Event 2010.

[22] PROSYST. 2010. The World's smallest OSGi Solution [Online]. Available:http://www.prosyst.com/index.php/de/html/news/details/18/smallest-OSGi/ [Accessed 20.07.2012].

[23] National Institute of Standards and Technology (NIST) 2005, JAMA : A Java Matrix Package. Availabe at http://math.nist.gov/javanumerics/jama/

[24] Determination of the shelf life of sliced cooked ham based on the growth of lactic acid bacteria in different steps of the chain J. Kreyenschmidt, A. Hubner, E. Beierle, L. Chonsch, A. Scherer and B.Petersen Faculty of Agriculture, Institute of Animal Science, University of Bonn, Bonn, Germany

Alexander Dannies received his Diploma in Electrical Engineering and Information Technology with specialisation in microelectronics / micro system technology from the University of Bremen in 2010. Since February 2011 he is a research associate of the Institute for Microsensors, -actors and –systems at the University of Bremen. There he is currently involved in the project "Intelligent container" and is researching the topic "Data interpretation in sensor networks".

Javier Palafox-Albarran has a Master of Science degree in information and automation engineering from the University of Bremen. Previously he has earned several years of industry experience working in industrial Automation. Currently he is pursuing a PhD in the Institute for Microsensors, -actuators and –systems (IMSAS). His research topic is on the analysis and prediction of sensor and quality data in food transportation supervision. He is also a member of the International Graduate School for Dynamics in Logistics.

Walter Lang studied physics at Munich University and received his Diploma in 1982 on Raman spectroscopy of crystals with low symmetry. His Ph.D. in engineering at Munich Technical University was on flame-induced vibrations. Science 2003 he is the head of the Institute for Microsensors, -actors and –systems at the University of Bremen. His research focus includes the manufacturing of miniaturized sensor components and the automated processing of sensor data.

Reiner Jedermann finished his Diploma in Electrical Engineering 1990 at the University of Bremen. After two employments on embedded processing of speech and audio signals, he became in 2004 a research associate in the Department of Electrical Engineering at the University of Bremen. He finished his Ph.D. thesis on automated systems for freight supervision end of 2009. His current research focus is the analyses of spatial temperature profiles and the implementation of automated decision tools for container supervision.

Robust Support Vector Machines for Speaker Verification Task

Kawthar Yasmine ZERGAT[1], Abderrahmane AMROUCHE[1]

[1] Speech Com. & Signal Proc. Lab.-LCPTS
Faculty of Electronics and Computer Sciences,
USTHB, Bab Ezzouar, 16 111, Algeria.

Abstract

An important step in speaker verification is extracting features that best characterize the speaker voice. This paper investigates a front-end processing that aims at improving the performance of speaker verification based on the SVMs classifier, in text independent mode. This approach combines features based on conventional Mel-cepstral Coefficients (MFCCs) and Line Spectral Frequencies (LSFs) to constitute robust multivariate feature vectors. To reduce the high dimensionality required for training these feature vectors, we use a dimension reduction method called principal component analysis (PCA). In order to evaluate the robustness of these systems, different noisy environments have been used. The obtained results using TIMIT database showed that, using the paradigm that combines these spectral cues leads to a significant improvement in verification accuracy, especially with PCA reduction for low signal-to-noise ratio noisy environment.

Keywords: *SVM, Noisy environment, LSF, MFCC, PCA.*

1. Introduction

A typical speaker verification system usually consists of two phases: an enrollment phase, and an Authentication phase. In the enrollment phase, the system extracts speaker-specific information from the speech signal to be used to build a model for the speaker [1], where the purpose of the testing phase is to determine whether the speech samples belong to the person that claims his/her identity or not.

In all audio processing, the speech input is converted into a feature vector representation [2]. Linear Prediction Cepstral Coefficients (LPCC) and Perceptual Linear Prediction Cepstral Coefficients (PLPCC), the Mel Frequency Cepstral Coefficients (MFCC) [3] approach has been the most employed for feature extraction. For modeling, Support Vector Machine (SVM) [4] represents a discriminative classifier which has achieved impressive results in several pattern recognition tasks. Indeed, the SVMs are interesting because they discriminate between classes (speaker/impostor) and could be used to train non-linear decision boundaries in an efficient manner. In this paper, for speaker verification, we investigate on SVM classifier based on Principal Component Analysis (PCA) [5] to get the efficiently reduced dimension of feature vectors. First, the MFCC, Line Spectral Frequency (LSF) features are extracted from the speech voice sample. The concatenation of these features vectors (MFCC-LSF) is made. Secondly, the new feature vectors with reduced dimension are obtained by applying PCA dimensionality reduction to each speaker vectors. Finally, theses transformed feature vectors are used as input to the SVM system for text independent speaker verification task. To validate the influence of PCA dimensionality reduction, we have evaluated the robustness of both SVM and PCA-SVM systems in different noisy environments at different levels of SNRs. The rest of the paper is as follows. In sections 2, we describe the Feature Extraction process used and discuss the principles of SVM in section 3. Section 4 and 5 are the experimental setup and the results of the experiments conducted on a subset of TIMIT Database. Finally, we conclude in Section 6.

2. Feature Extraction

2.1 Mel Frequency Cepstral Coefficient (MFCC)

MFCCs were introduced in early 1980s for speech recognition applications. The key steps involved in computing MFCC features are shown in Fig. 1. The speech signal is first pre-emphasized by applying the following filter [1],

$$x(t) = y(t) - ay(t-1), \text{Where } a \in [0.95, 0.98] \quad (1)$$

The goal of the filter is to enhance the high frequencies of the spectrum, which is diminished during the speech production process. Following the pre-emphasis stage is a windowing step, the speech samples are weighed by a suitable windowing function, The Hamming window is extensively used in speaker verification to taper the original signal on the sides witch reduce the side effects [6].

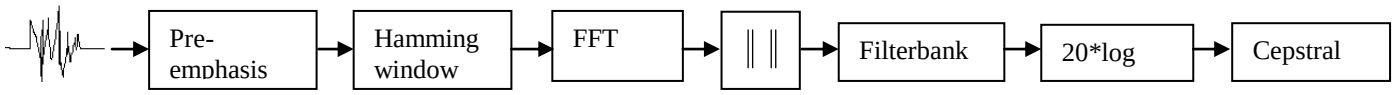


Fig. 1 MFCC features extraction.

The result of windowing the signal is shown below:

$$W(n) = 0.54 - 0.46\cos\left[\frac{2\Pi n}{N-1}\right], \quad 0 \le n \le N-1 \qquad (2)$$

Once the speech signal has been windowed its fast Fourier transform (FFT) is calculated. Finally, the modulus of the FFT is extracted and a power spectrum is obtained [6]. The obtained spectrum is then multiplied by filterbank.

The filters that are generally used in MFCC computation are triangular filters, and their center frequencies are chosen according a logarithmic frequency scale also known as Mel-frequency scale witch conforms to response observed in human auditory systems. The localization of the central frequencies of the filters is given by [1]:

$$f_{MEL} = 1000.\frac{\log(1 + f_{LIN}/1000)}{\log 2} \qquad (3)$$

An additional transform, is to obtain the spectral vectors by taking the log of the spectral envelope and multiply each coefficient by 20 in order to get the spectral envelope in dB [6]. Finally, the cosine discrete transform (DCT) is applied to the spectral vectors witch yields cepstral coefficients frequencies, and is given by [6]:

$$c_n = \sum_{k=1}^{K} S_k \cos[n(k-\frac{1}{2})\frac{\pi}{k}], \; n = 1,2,....,L \qquad (4)$$

Where K is the number of log-spectral coefficients calculated in previous step, S_k are the log-spectral coefficients, and L is the number of cepstral coefficients to calculate.

2.2 Line Spectral Frequency Cues (LSF)

The starting point for deriving the LSF's is the response of the prediction error filter.

$$A(z) = 1 - \sum_{k=1}^{P} a_k z^{-k} \qquad (5)$$

Where P represents the prediction order, and a_k are the LPC filter coefficients. In the LPC the mean squared error between the linearly predicted speech sample and the actual one is minimized over a finite interval [7]. The transfer function of the LPC filter with a gain G is given by:

$$H(z) = \frac{G}{1 + \sum_{k=1}^{P} a_k z^{-k}}, \qquad (6)$$

From $H(z)$, a symmetric polynomial $S_{P+1}(z)$ and an antisymmetric polynomial $\theta_{P+1}(z)$ are calculated by adding and subtracting the time-reversed system function.

$$S_{P+1} = A_p(z) + z^{-(p+1)}A_p(z^{-1}), \text{ And}$$
$$\theta_{p+1} = A_p(z) - z^{-(p+1)}A_p(z^{-1}), \qquad (7)$$

The polynomials contain trivial zeros for even values of p at $z = -1$ and at $z = 1$. These roots can be removed in order to derive the following quantities [7]:

$$\tilde{S}(z) = \frac{S_{P+1}(z)}{(1+z)} = \omega_0 z^p + \omega_1 z^{p-1} + ... + \omega_p, \quad \text{And}$$
$$\tilde{\theta}(z) = \frac{\theta_{P+1}(z)}{(1-z)} = \sigma_0 z^p + \sigma_1 z^{p-1} + ... + \sigma_p, \qquad (8)$$

The LSFs are the roots of $\tilde{S}(z)$ and $\tilde{\theta}(z)$ and alternate with each other on the unit circle [7].

3. Support Vector Machine

Support Vector Machine (SVM) is a binary linear classifier in its basic form. It has been recently adopted in speaker recognition task. Given a set of linearly separable two-class training data, there are many possible solutions for a discriminative classifier [8]. An SVM seeks to find the Optimal Separating Hyperplane (OSH) between

classes by focusing on the training cases that lie at the edge of the class distributions, the support vectors, with the other training cases effectively discarded [8]. Formally, the discriminant function of SVM is given by:

$$f(x) = class(x) = sign\left[\sum_{t=i}^{N} \alpha_i y_i K(x, x_i) + b\right] \quad (9)$$

Here $y_i \in \{-1,+1\}$ are the ideal output values,

$\sum_{t=i}^{N} \alpha_i y_i = 0$ and $\alpha_i \geq 0$. The support vectors x_i, their

corresponding weights α_i and the bias term b.

To calculate the classification function class (x) we use the dot product in feature space that can also be expressed in the input space by the kernel function $K(\cdot,\cdot)$. Among the most widely used cores we find:

RBF kernel: $k(x, y) = e^{-\gamma\|x-x_i\|2}$

Polynomial kernel: $k(x, y) = (x^T . y + 1)$

Finally, the classification of data is made as follow:

$$x \in \begin{cases} Classe\ 1 \\ classe\ 2 \end{cases} \text{ if } \begin{cases} X \succ 0 \\ Otherwise \end{cases}$$

4. Experimental Protocol

4.1 Description of the Database

The corpus used in this work is issued from the TIMIT database. This database includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16 kHz speech file for each utterance and is recorded in ".ADC" format, where each sentence has 3s of length spoken in English language. We have selected a set of 90 speakers, for both training and testing phases, each of whom reads 5 phonetically rich sentences for training task and 3 utterances for testing task. To simulate the impostors, 50 unknown speakers (25 female and 25 male) are used from the same database (TIMIT) and are different from the 90 speakers used previously, with five utterances spoken by each unknown speaker.

4.2 Parameterization Phase

In this work, we have included as many speakers' characteristics as possible. So, our first feature space is made with 12 MFCC coefficients plus Energy parameter and the first and second derivatives, which yield 39-dimensional feature vector, extracted from the middle window every 10ms. A voice activity detector is used to eliminate silence and noise frames from the training and testing signals in order to avoid modeling and detecting the environment rather than the speaker. In the second part, 12 LSF coefficients were extracted. Finally, the first feature space (MFCCs +E+ $\Delta + \Delta\Delta$) was combined with 12 LSF coefficients to constitute a multi-dimensional feature set. The dimension of the combined vectors is then equal to 51. Once the feature vectors have been calculated, they can be centered, using Cepstral Mean Subtraction (CMS), this is carried out by estimating a mean vector for the extracted set of cepstral features and subtracting it from all the feature vectors. In the other hand, the size of the vectors of parameters is an important problem that arises when adding parameters. To address this, technique to reduce the number of parameters was used, these include PCA dimensionality reduction.

4.3 Modeling Phase

For the enrolment phase, we did however make use of a RBF kernel function for the SVM classifier. In order to evaluate the performance of the system, two types of additive noise produced by a Speech Babble and a Subway noises reaching high levels of SNR and derived from the NOISEX-92 database (NATO: AC 243/RSG 10) are added to the test speech signal of the TIMIT database. In speaker verification task, there are two types of errors; *false acceptance* (FA) and *false rejection* (FR). The Equal Error Rate (ERR) is the point where the rate of FR's is equal to the rate of FA's. For classification, The Detection Error Tradeoff (DET) curve is a popular way of graphically representing the performance of speaker verification system.

5. Experiment Results

5.1 Speaker Verification using Original Speech Waveforms

To show the effectiveness of the proposed method, we performed two experiments using speech data without and using the PCA dimensionality reduction for the SVM classifier in speaker verification task. In both experiments we used Equal Error Rate (EER) as the performance criterion. For the first experiment a comparative study shows the contribution of the concatenation between the MFCC and the LSF features. As shown in Fig. 2, the concatenated feature vector brings the less EER equal to 0.54% against the LSF and MFCC feature vectors with an EER equal to 7.39% and 3.69% respectively.

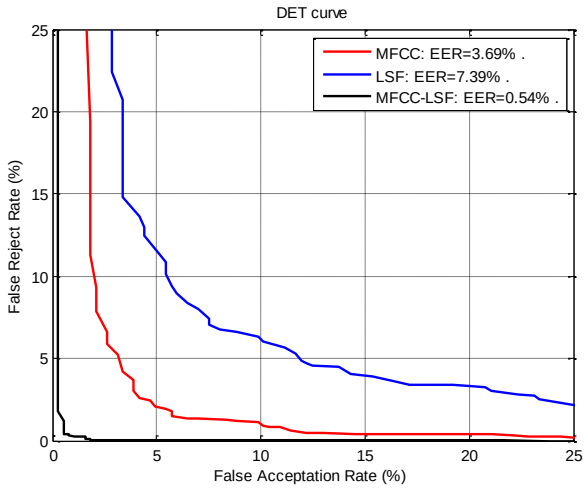


Fig. 2 Performance evaluation for SVM based speaker verification task.

Speaker verification experiment with PCA based SVM classifier has been performed too with these various feature space components. From the following fig. 3, it's clearly seen that PCA improves significantly the recognition accuracy, until an EER=0.51% for the concatenated (MFCC, LSF) feature vectors.

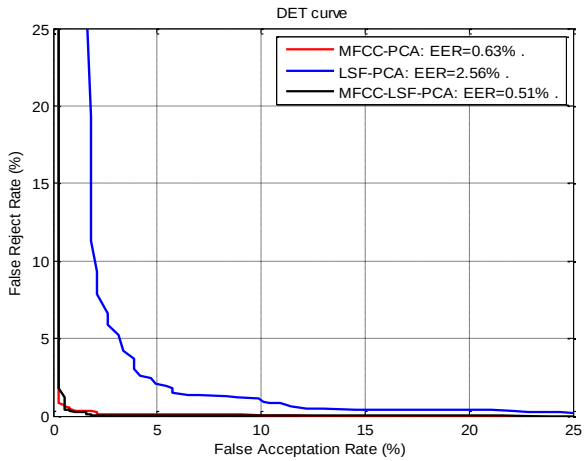


Fig. 3 Performance evaluation for PCA-SVM based speaker verification task.

5.2 Verification Accuracy under Noisy Environments

The main goal of the experiments done in this section is the study of the verification performances of both SVM and PCA-SVM systems in different noisy environments, for this, two noisy environments which are Speech Babble and Subway noises were used. We evaluated the error rate by applying dimensionality reduction by PCA algorithm on the concatenated MFCC-LSF feature vectors. The results are shown in the following figures.

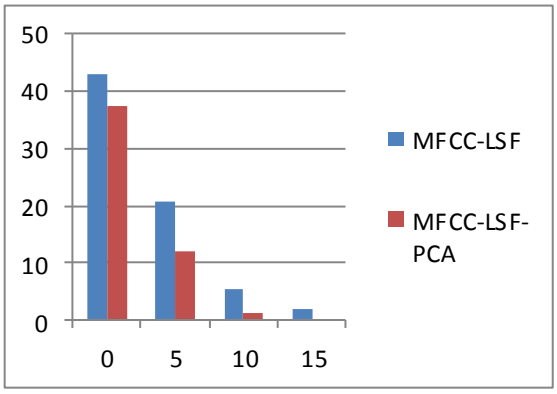


Fig. 4 Performances evaluation for SVM and PCA-SVM in noisy environment corrupted by Babble speech noise.

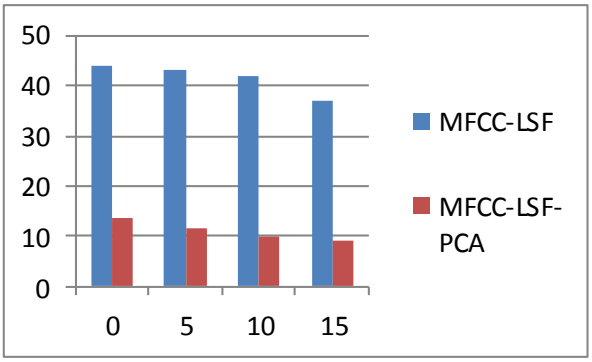


Fig. 5 Performances evaluation for SVM and PCA-SVM in noisy environment corrupted by Subway noise.

As shown in the above figures, it is clearly seen that applying PCA algorithm on feature vectors leads to an interesting increase of speaker verification accuracy. Quantifying the input data by other Algorithms such as LBG and k-means, quantify all the data including the insignificant and repeated items presented in the speech signal [9], by cons, when using PCA dimensionality reduction, we project the data into lower dimensional space, where the low variance components are eliminated. The obtained results confirm the effect of PCA, for example, in case of noisy environment corrupted by Babble speech noise at SNR = 0dB, the EER decrease from 42.87% to 37.39% which represents an interesting improvement in bad conditions.

6. Conclusion

This paper has presented and evaluated a text-independent speaker verification systems based on SVM classifier. To attain better performance, two systems were trained, the SVM and the PCA-SVM systems. In this study, we have examined the influence of feature vectors and PCA dimensionality reduction on speaker verification rate in both clean and noisy environments. Carried out on TIMIT database, it is noticed that, the combination between MFCC and LSF outperforms the conventional MFCC parameters. In the other hand, for the PCA-SVM model, the recognition accuracy has increase significantly comparing to the SVM model, especially in hard conditions (SNR= 0dB). We are currently continuing the effort towards the optimization of this system using other dimensionally reduction method.

References

[1] C. Turnera, A. Josephb, M. Aksuc, and H. Langdonda, "The Wavelet and Fourier Transforms in Feature Extraction for Text-Dependent, Filterbank Based Speaker Recognition", *Complex Adaptive Systems*, Vol. 1, 2011, pp. 124-129.

[2] T. Mazibuko, and D. Marshao, "Feature extraction and dimensionality reduction in SVM speaker recognition", *Southern African Telecommunications and Applications Conference (SATNAC 2006)*, 2006.

[3] S. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Transaction on Acoustics, Speech, and Signal Processing*, 1, 1980, Vol. 28, pp. 357-366.

[4] R. Dehak, N. Dehak, P. Kenny, and P. Dumouchel, "Linear and non linear kernel GMM supervector machines for speaker verification", *Proc. Interspeech, Antwerp*, 2007, pp. 302–305.

[5] C. Hanilci, and F. Ertas, "VQ-UBM based speaker verification through dimension reduction using local PCA", *19th European Signal Processing conference (EUSIPCO)*, 2011.

[6] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. M. Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcıa, D. Petrovska-Delacretaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification", *EURASIP Journal on Advances in Signal Processing*, 2004, pp. 430-451.

[7] D. Addou, S. A. Selouani, K. Kifaya, M. Boudraa, and B. Boudraa, "A noise-robust front-end for distributed speech recognition in mobile communications", *International Journal of Speech Technology Springer Science*, Vol. 10, 2007, pp. 167-173.

[8] W. M. Campbell, D. E Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation", *Proceedings of ICASSP*, 2006.

[9] T. Kinnunen, and H. Li, "An overview of text-independent speaker recognition: From features to supervectors", *Speech Communicatio*, Vol. 52, 2010, pp. 12-40.

Kawthar Yasmine Zergat Received her Master II degree in Communication and Multimedia from the University of Science and Technology Houari Boumedienne (USTHB), Algiers in 2010. Currently, she is pursuing the Ph.D. degree in, Telecommunications and Information Processing in the Communication Systems and Speech Processing Laboratory, USTHB. Her current research concentrates on robust speaker recognition and speech processing.

Abderrahmane Amrouche Was born in Algeria. He received his “diplome d’ingenieur” (engineer degree) in Electronics from the National Polytechnic school of Algiers in 1980. He received his “Magister” degree in 1995 and Doctorat d’Etat” (Ph.D) in Real Time Systems in 2007 from the University of Science and Technology Houari Boumedienne (USTHB). He is an Assistant Professor in Communication Systems and Speech Processing Laboratory, USTHB. His research interests include pattern recognition, speech processing, Multilingual speech recognition, neural networks, prosodic modelling.

Cover Optimization for Image in Image Steganography

Nidhal K. El Abbadi
University of Kufa
Najaf, Iraq

Abstract

This paper develops techniques for discriminating between images which used as steganography cover. Algorithm is based on the hypothesis that a particular message embedding scheme leaves statistical evidence or structure that can be exploited for detection with the aid of proper selection of image features analysis. We pointed out the features of image that should be taken more seriously into account in the design of more successful steganography, weight for each of these features determined by using Analytic Hierarchy Process (**AHP**) which helps to maximize some of the features and gives weight according to the relation between these features. The proposed algorithm tested by using LSB image steganography, stego-image compared with the origin one which gives the promised results.

Keywords*: steganography, features, AHP, information hiding, image*.

1. Introduction

Steganography is the art and science of hiding information by embedding data into media. Steganography (literally meaning covered writing) have been used since ancient time.
Electronic steganography techniques use digital ways of hiding and detecting processes. Normally the detection process is working inversely of the hiding process. Steganography is different from cryptography and watermarking although they all have overlapping usages in the information hiding processes. Steganography security hides the knowledge that there is information in the medium cover, where cryptography revels this knowledge but encodes the data as cipher-text and disputes decoding it without permission; i.e., cryptography concentrate the challenge on the decoding process while steganography adds the search of detecting if there is hidden information or not. Watermarking is different from steganography in its main goal. Watermarking aim is to protect the cover medium from any modification with no real emphasis on secrecy. It can be observed as steganography that is concentrating on high robustness and very low or almost no security [6].
Steganography techniques use different carriers (cover medium in digital format) to hide the data, these carriers may be network packets, hard drive, amateur radio waves, or generally any computer file types such as text, image, audio and video. Restrictions and regulations are thought of in using steganography due to the threat from law and rights enforcing agencies and the need of organizations aiming to secure their information. Many easy to use steganography tools are available to hide secret messages on one side of communication and detect hidden info. on the other side. Steganography uses cover to embedded secret data, this cover chooses randomly and for the same secret data every one can choose different cover without a prior knowledge which one is better, because there are no rules or measurements use for choosing suitable cover.
In this work, we propose many features that can be used to choose the best cover among many suggested covers for embedded secret data (image in image steganography). It also used the Analytical Hierarchy Process (AHP) to determine the weight for each feature. Unfortunately there are no studies about this problem. As best of my knowledge there are only two studies related to choosing cover, the first one presented by Mehdi [6] which studied the cover selection problem through three scenarios in which the secret data either no knowledge, partial knowledge, or full knowledge of the steganalysis technique. Hedieh [4], also presented a technique to compute steganography capacity as a property for image cover selection. This technique used different steganlyzer units, which help to determine the maximum size of embedded that can embedded in cover.

2. Methodology

The aim of this algorithm is to find the best cover for an embedded secret data, it focus on image in image steganography, for that many images features chooses to be scale to select best cover among many suggested covers, the weight for each feature can achieve by using (AHP method). These features will be modified in a way suitable with the aim of this paper. The features suggested to use are:

{Note: subscript (c): mean cover image, (e) mean embedded image, and (Pg) mean probability for color (g) in image = N (g)/M where:
N (g): number of pixels with color g, M: total number of pixels in image }

1. Entropy:

The entropy is a measure of image information content, which tells us how many bits we need to code the image data, and is given by [2].

$$Entropy = -\sum_{g=0}^{L-1} P_{(g)} \log_2 [P_{(g)}] \quad \ldots\ldots \quad (1)$$

Where L: Number of color in image
As the pixel values in the image are distributed among more color level, the entropy increases.

$$0 \leq \text{entropy} \leq \log_2 (L)$$

Coding redundancy occurs when the data used to represent the image are not utilized in an optimal manner. For cover and embedded entropy it is better that

$$Ent_c \geq Ent_e$$

The number of colors (NC) used in cover should be more than number of colors in embedded.
Number of colors in image is

$$NC = 2^{\text{entropy}}$$

Max colors different in an image (256 colors) are
$NC_c - NC_e$ equal to 256-1=255
Then the percent of difference in the number of colors (DNC) is

$$ENT = ((NC_c - NC_e)/255)*100 \quad \ldots\ldots\ldots\ldots \quad (2)$$

Note if $NC_e > NC_c$ then DNC will be negative and subtracted from final result.

2. Capacity:

This term refers to the amount of data that can be hidden in the medium. It is defined as "*the maximum message size that can be embedded subject to certain constraints*"[7].
There are restrictions of data rate that can be embedded in a certain image. The worst case of embedded data is 1 bit in each byte (8 bits) as in LSB which represents (12.5%) of cover size as a maximum.
If the size of data embedded in the cover increased to more than the capacity of cover, then its transparency will be affected; i.e. with very high capacity, the steganography is not strong to keep transparent from eavesdroppers.

To check the capacity you should follow the following steps:

(a) $(size_e/size_c) \leq 0.125$

b) if the result in step (a) is false then we calculate the percent of capacity compatibility (CC) between cover and embedded is

$$CC = 100 - ((Size_e/size_c)/0.125)*100 \quad \ldots\ldots \quad (3)$$

3) Mean:

The mean is the average value which tells us something about the general brightness of the image.
A bright image will have a high mean (more than 127) and dark image will have low mean.

$$Mean = \sum_r \sum_c I(r, c)/m$$

The max difference in mean is 255.
% of mean similarity (MS):

$$MS = 100 - ((abs(g_c' - g_e')/255)*100 \ldots\ldots (4)$$

Where: g': color value mean

4) Variance:

Which tells us something about the contrast, it describes the spread in the data, so a high contrast image will have a high variance, and a low contrast image will have a low variance [17].

$$V_{(g)} = \sqrt{\sum_{g=0}^{L-1} (g - g')^2 p(g)} \quad \ldots\ldots\ldots\ldots\ldots \quad (5)$$

Max variance is when there are just two colors one equal to zero and other equal 255, then the mean is equal to (127.5) and the max variance is (127.5).
It is recommended that V_e approach to zero. Variance similarity (Vs) is calculated as a percent

$$\%VS = ((V_c - V_e)/127.5)*100 \quad \ldots\ldots\ldots\ldots (6)$$

5) Histogram:

Histogram analysis may be required before embedding to prevent the histogram attack [8].
Histogram matching between cover and embedded is done by comparing each color in cover histogram with the corresponding color in embedded histogram, if the number of pixels at that color is more than number of pixels in embedded for the same color then counter increases with one.
% color matching (CM):

$$CM = (counter/256)*100 \ldots\ldots\ldots\ldots .. (7)$$

6) Energy:

The energy measurement tells us something about how the colors distributed [17].

$$Energy = \sum_{g=0}^{L-1} (P_{(g)})^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots (8)$$

The energy measurement has a maximum value of (1) for an image with one color.

The larger this value is the easier to compress the image data. Energy indicates the region of image with identical color value, increasing energy mean increasing the size of this region, and the capability of compression will be increased.
The best distribution is when all colors (g) have the same frequency. (x: number of pixels have the same color g)

$$Energy = \sum_{g=0}^{L-1} x^2 / (size_c)^2$$

$$= \sum_{g=0}^{L-1} x^2 / (x*256)^2$$

$$= 256 * x^2 / (x^2 *(256)^2) = 1 / 256$$

Well, this value of energy (1/ 256) represent (100%) of distribution. Then when the energy value increases, the energy percent will decrease (inverse relation)

$$\% \text{ distribution (DS)} = 1 / (energy_c *256))*100 \quad \ldots\ldots\ldots\ldots (9)$$

7) Robustness

Robustness (R) can only be achieved by redundant information encoding which will degrade the cover heavily and possibly alter probability distribution P_s.
An embedding algorithm will be consider a robust if the embedded message can be extracted after an image has been manipulated without being destroyed.
The more randomness that exists in an image the more evenly the color levels distributed and the more bits per pixels are required to represent the data. This also correlates to information more randomness implies each individual value is less likely which means more information is contained in each pixel value so we need more bits to code each pixel value and more robustness. Best robustness is when
($P = x/ size_c$)
$X = size_c/256$
$P = (size_c /256)/size_c=1/256$

$$Entropy = - \sum_{g=0}^{255} P_c (g) \log_2 P_c (g)$$

$$= - \sum_{g=0}^{255} (1/256) \log_2 (1/256)$$

$$= \log_2 (1/256)$$

$\%R= - (entropy_c / (\log_2 (1/256))*100$ This can be simplified as

$$\% R = (entropy_c/8)*100 \quad \ldots\ldots\ldots\ldots (10)$$

8) Expected Secrecy

Secrecy is one of the most important criteria. The secrecy is the ability to hide information in cover image, and is determined as a magnitude (ε) by comparing the cover image and stego- image according to relative entropy [10].

$$D (Pc||Ps) = \sum_{g=0}^{L-1} P_c (g) \log_2 (P_c (g) / P_s (g)) \quad \ldots\ldots\ldots (11)$$

The relative entropy between two distributions is always non-negative, and is zero if and only if the distributions are equal. We modify this equation to get a new relation that can determine the expected secrecy (the worst secrecy) without needing the existence of stego or hiding algorithm.
If we use LSB then the number of bytes (NB) that should be modified in covering it equals the number of embedded bits. Then
$NB = size_e \times 8$
The number of bytes from each color in cover should be changed depending on probability for each color.
$Prop (g) = freq_{(g)} / size_c$
where: freq= means number of color (g)
The number of bytes change for each color will be:
$NB_{(g)} = 8 \times size \times (ferq_{(g)}/size_c)$
That means each color (g) in cover will reduce with quantity of NB (g) and will increase with quantity of $NB_{(g-1)}$
Then the number of bytes of color (g) in stego will be
a) When(g) odd

$$SNB (g) = freq_{(g)} - NB_{(g)} + NB_{(g-1)} \quad \ldots\ldots\ldots\ldots (12)$$

b) when (g) even

$$SNB (g) = freq_{(g)} - NB_{(g)} + NB_{(g+1)} \quad \ldots\ldots\ldots\ldots (13)$$

Then according to first equation

$$Estimated\ Secrecy = \sum_{g=0}^{255} P_c (g) \log_2 (P_c (g) / P_s (g))$$

$$= \sum_{g=0}^{255} (freq_c (g) / size_c) \log_2 ((freq_c(g) / size_c) / (freq_s(g) / size_s))$$

If we know that $Size_c = Size_s$

$$Estimated\ secrecy\ (ES) = (1/ size_c) \sum_{g=0}^{255} freq_c(g) \log_2 (freq_c (g) / SNB(g))$$

Percent will determined according to

$$\varepsilon = 2^{-secrecy}$$

$$\%es = \varepsilon *100 \quad \ldots\ldots\ldots (14)$$

3. Analytic Hierarchy Process (AHP)

The Analytic Hierarchy Process (AHP) is a mathematical technique for multi-criteria decision

making [11]. It enables people to make decisions involving many kinds of concerns including planning, setting priorities, selecting the best among a number of alternatives, and allocating resources. AHP uses for relative criticality weighting of indicators, and relative criticality weighting of evaluators.

The Analytic Hierarchy Process (AHP) is a structured technique for dealing with complex decisions. Rather than prescribing a "correct" decision, the AHP helps the decision makers find the one that best suits their needs and their understanding of the problem.

Based on mathematics and psychology, it was developed by Thomas L. Saaty in the 1970s and has been extensively studied and refined since then. The AHP provides a comprehensive and rational framework for structuring a decision problem, for representing and quantifying its elements, for relating those elements to overall goals, and for evaluating alternative solutions. It is used around the world in a wide variety of decision situations, in fields such as government, business, industry, healthcare, and education.

Several firms supply computer software to assist in using the process.

Users of the AHP first decompose their decision problem into a hierarchy of more easily comprehended sub-problems, each of which can be analyzed independently. The elements of the hierarchy can relate to any aspect of the decision problem tangible or intangible, carefully measured or roughly estimated, well or poorly understood anything at all that applies to the decision at hand.

Once the hierarchy is built, the decision makers systematically evaluate its various elements by comparing them to one another two at a time. In making the comparisons, the decision makers can use concrete data about the elements, or they can use their judgments about the elements' relative meaning and importance. It is the essence of the AHP that human judgments, and not just the underlying information, can be used in performing the evaluations [12].

The AHP converts these evaluations to numerical values that can be processed and compared over the entire range of the problem. A numerical weight or priority is derived for each element of the hierarchy, allowing diverse and often incommensurable elements to be compared to one another in a rational and consistent way. This capability distinguishes the AHP from other decision making techniques.

In the final step of the process, numerical priorities are calculated for each of the decision alternatives. These numbers represent the alternatives' relative ability to achieve the decision goal, so they allow a straightforward consideration of the various courses of action.

As can be seen in the material that follows, using the AHP involves the mathematical synthesis of numerous judgments about the decision problem at hand. It is not uncommon for these judgments to number in the dozens or even the hundreds. While the math can be done by hand or with a calculator, it is far more common to use one of several computerized methods for entering and synthesizing the judgments. The simplest of these involve standard spreadsheet software, while the most complex use custom software, often augmented by special devices for acquiring the judgments of decision makers gathered in a meeting room.

The procedure for using the AHP can be summarized as:

1. Model the problem as a hierarchy containing the decision goal, the alternatives for reaching it, and the criteria for evaluating the alternatives.
2. Establish priorities among the elements of the hierarchy by making a series of judgments based on pair-wise comparisons of the elements
3. Synthesize these judgments to yield a set of overall priorities for the hierarchy.
4. Check the consistency of the judgments.
5. Come to a final decision based on the results of this process.
6.

We conduct AHP in three steps:

1. Perform pair-wise comparisons
2. Assess consistency of pair-wise judgments
3. Compute the relative weights
4.

- **Pair Wise Comparisons**

AHP enables a person to make pair wise comparisons of importance between decision elements (e.g., *child indicators* influencing a parent indicator, *evaluators* evaluating a leaf indicator) with respect to the scale shown in the following Table.

Table 1: Scale for pair wise comparison

Comparative Importance	Definition	Explanation
1	Equally important	Two decision elements (e.g., indicators) equally influence the parent decision element.
3	Moderately more important	One decision element is moderately more influential than the other.
5	Strongly more important	One decision element has stronger influence than the other.
7	Very strongly more important	One decision element has significantly more influence over the other.
9	Extremely more important	The difference between influences of the two decision elements is extremely significant.
2, 4, 6, 8	Intermediate judgment values	Judgment values between equally, moderately, strongly, very strongly, and extremely.
Reciprocals		If v is the judgment value when i is compared to j, then 1/v is the judgment value when j is compared to i.

- **Computing the Relative Weights**

AHP computes a weight for each decision element based on the pair-wise comparisons using mathematical techniques such as Eigenvalue, Mean Transformation, or Row Geometric Mean. We employ the Eigenvalue technique for computing the weights under AHP.

4. Implementation and the results

For implementing this algorithm we did the following:

4.1 Choose (8) images randomly as covers fig (1), all with the same size fig (2).

4.2 Choose (2) images as secret data (embedded image) fig (1), both with the same size fig (2).

4.3 Determine the features for all images (covers, and embedded).

Fig. 1: The covers and secret images used in experiment

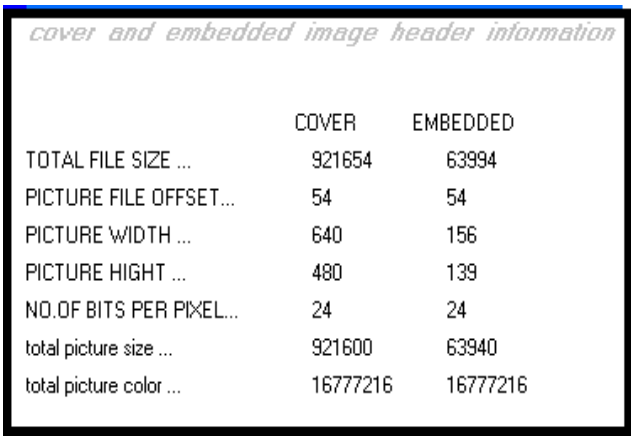


Fig. 2: cover and embedded images specification

4.4 Features are organized according to priorities which are suggested by the user, for this work we suggested the following priorities:

a. ES (Estimated Secrecy).
b. R (Robustness).
c. ENT (Entropy).
d. CC (Capacity).
e. VS (Variance).
f. Ds (Energy).
g. CM (Histogram).
h. MS (Mean).

4.5 Determine the weight for each feature by using AHP process, as following:

Y3 =POWER((D3*F3*H3*J3*L3*N3*P3*R3);(1/8))

	C	D	F	H	J	L	N	P	R	Y	Z	AA
1	1	2	3	4	5	6	7	8	9	10	11	12
2	Parameters	CM	CC	VS	MS	ENT	DS	SEC	R	Eigenvalue	Priority Vector	weight
3	CM	1.000	0.250	0.333	2.000	0.200	0.500	0.125	0.142	0.362	0.031	3.12
4	CC	4.000	1.000	2.000	5.000	0.500	3.000	0.250	0.333	1.223	0.105	10.52
5	VS	3.000	0.500	1.000	4.000	0.333	2.000	0.200	0.250	0.818	0.070	7.03
6	MS	0.500	0.200	0.250	1.000	0.200	0.500	0.111	0.125	0.277	0.024	2.38
7	ENT	5.000	2.000	3.000	5.000	1.000	4.000	0.333	0.500	1.778	0.153	15.30
8	DS	2.000	0.333	0.500	2.000	0.250	1.000	0.166	0.200	0.522	0.045	4.49
9	SEC	8.000	4.000	5.000	9.000	3.000	6.000	1.000	2.000	3.884	0.334	33.42
10	R	7.000	3.000	4.000	8.000	2.000	5.000	0.500	1.000	2.759	0.237	23.74
11										0.000		
12										0.000		
13										0.000		
14	sum	30.500	11.283	16.083	36.000	7.483	22.000	2.685	4.550	11.623	1.000	100.00
15												
16												
17			λ=	8.23685	CI=	0.034	CR=	0.024				

Fig. 3: Priorities and weight of features

a. The value in each field in fig (3) for any row is calculated by comparing feature (parameters) in the row with each feature in the columns one by one, two at each time, and assigned value according to suggested priorities in section 4.4, and table (1).

b. Determine the Eigenvalue = ($\prod$ features values in each row)$^{1/n}$
where (n) is number of features in row.

c. Determine the priority vector where,
Priority for feature [i] = (Eigenvalue for feature [i]) / $\sum_{i=1}^{n}$ *Eigenvalue [i]*

d. Weight of feature [i] = priority [i] × 100

e. Inconsistent matrices typically have more than 1 eigenvalue. To check the consistency of the judgments, we have to measure the consistency ratio which should be less than one.

f. $\lambda_{max} = \sum_{i=1}^{n} sum_i \times priority_i$

g. CI (consistency index) = $(\lambda_{max} - n) / (n-1)$

h. CR (consistency ratio) = CI / RI (**should be < 1**)

Random Consistency Index (RI) is obtained from Table 2 [12].

Table 2: consistency index

n	RI		n	RI
1	0		**6**	1.25
2	0		**7**	1.35
3	0.52		**8**	1.4
4	0.89		**9**	1.45
5	1.11		**10**	1.49

4.6 The final weight for each cover (when embedded images (1 and 2)) determined according to features weight calculated in AHP above where:
Final weight = CC+ENT+MS+VS +CM+DS+R+ES
The final results sorted in descending order, where the highest weight represents the best cover for embedding the specific image as shown in fig. 4.

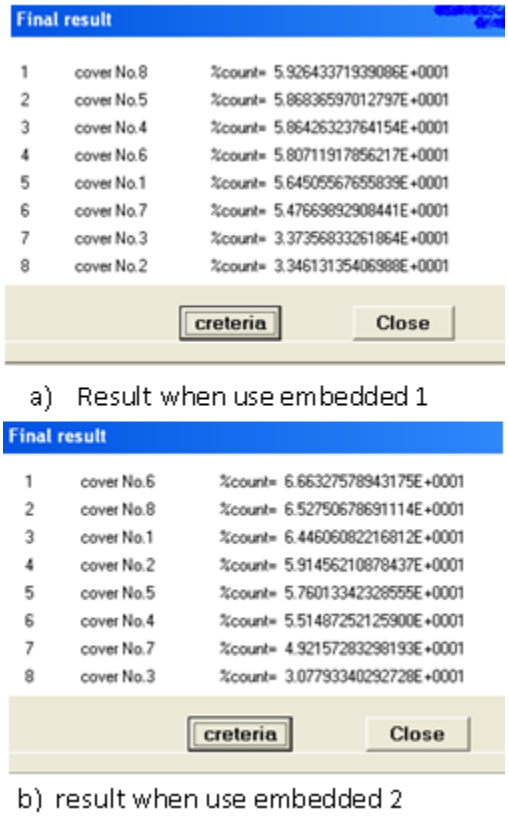


Fig 4: Final weight when calculate features with both embedded 1 and embedded 2

5. Prove the Results

Perfect steganography is when we get stego-image similar to original cover by both perceptual and computer reading. This may be impossible to reach.
In our work we hope to choose cover, give the closest features to original cover when it changes to stego-image.
To prove this we try to apply the following step, which helps us to evaluate our work

5.1 First convert each cover to stego-object (by hiding each embedded image in all covers) by using (LSB) hiding technique.

5.2 Determine the perceptual difference between the origin cover and stego image fig (5).

5.3 Determine the histogram for origin image and stego image fig. (6).

5.4 Determine the similarity between the cover image and the corresponding stego-object. Formally, similarity can be defined via similarity function [3].
Let c be nonempty set.
Function Sim: $c^2 > [-\infty, 1]$ is called similarity function on c,
if for $(x, y) \in c$ $Sim(x, y) = 1$ iff $x = y$
For $x \neq y$, sim $(x, y) < 1$
Perfect similarity ≈ 1
In the case of digital images the correlation between two images can be used as similarity function. Therefore most practical steganographic systems try to fulfill the condition
Sim (cover, stego) =1
Similarity determine by comparing both of cover and stego image.

5.5 Determine the security for stego-object by using Eq. (11).
Perfect security = **0.**

5.6 Determine the PSNR.

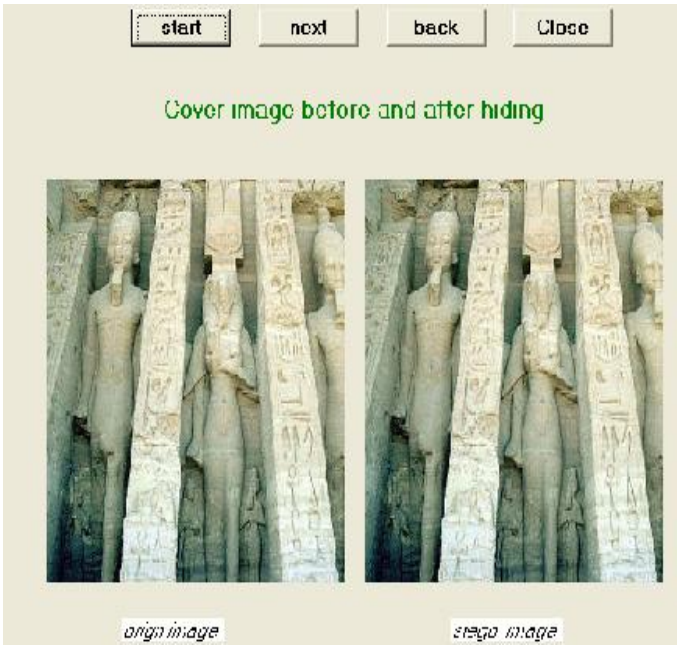


Fig. 5: comparing cover image before and after hiding embedded 1

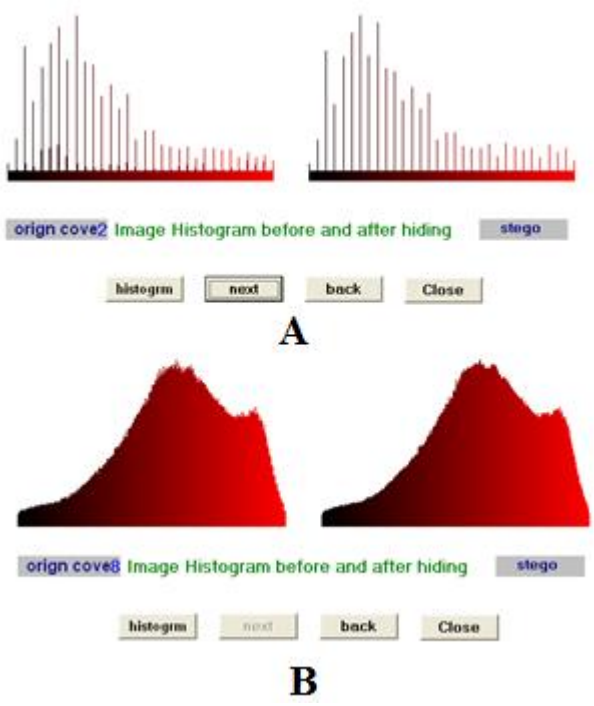


Fig 6: Histogram for both origin and stego images for covers (1 and 8) when hiding embedded image 1

Table 3: Comparing result when hiding embedded 1 in covers.

cover	similarity	secrecy	PSNR
1	0.90610	0.00803	34.310
2	0.90143	0.15486	34.239
3	0.90303	0.15538	34.262
4	0.90604	0.00708	34.311
5	0.90633	0.00800	34.311
6	0.90607	0.00778	34.317
7	0.90555	0.01082	34.300
8	0.90687	0.00203	34.318

Table 4: Comparing result when hiding embedded 2 in covers.

cover	similarity	secrecy	PSNR
1	0.86755	0.64115	33.812
2	0.86750	0.64115	33.816
3	0.86120	0.64118	33.745
4	0.86054	0.64114	33.738
5	0.86835	0.64026	33.821
6	0.86793	0.62098	33.832
7	0.86349	0.63030	33.769
8	0.86551	0.64118	33.766

It is clear from the results above the following

- **A.** There is no perceptual difference between origin and stego image.
- **B.** Histogram of origin and stego image is almost the same.
- **C.** The values of (similarity, security, and PSNR) confirm the result in fig. 4 for both cover 8 when embedding embedded image1 in it, and cover 6 when embedded the embedded image 2 in it. Almost both of them give the best result.

6. Conclusions

This paper introduced a novel algorithm to choose cover from many suggested covers; it is the first algorithm discusses this problem.
The algorithm proved by using LSB image in image steganography, and measuring the perceptual and computer reading similarity, PSNR, security, and histogram to prove the efficiency of the algorithm.
Tables (3, 4) proved the results in fig. (4) and the best cover in fig. (4) get the best result when comparing stego-image with the cover images, at the same time the cover with the minimum weight gets worst result in comparing stego-image with cover image.
AHP algorithm used to count the weight of each feature. Final results may change if the features priorities will be changed, due to change of weight.
From all the results, we can say, that we proposed and built dependable algorithm, and by using other images features, we can develop this algorithm to become more accurate.
We suggest for future works, determine the features for each channel of the image color (Red, Green, and Blue).

References

[1] Abbas Cheddad, JoanCondell, KevinCurran, PaulMcKevitt, "Digital image steganography: Survey and analysis of current methods", Journal signal processing, Volume 90, Issue 3, 2010

[2] Gerhard X. Ritter; Joseph N. Wilson, Handbook of Computer Vision Algorithms in Image Algebra, CRC Press LLC , 1996

[3] Gonzalez R.C. and Woods R.E, Digital Image Processing, 3rd edition , Prentice Hall, 2008.

[4] Hedieh Sajedi, M. Jamzad ,"Contourlet-Based Steganography Using Cover Selection", International Journal of Information Security, Springer, vol. 9, no.5, 2010, pp. 337-345.

[5] İsmail Avcıbaş "Image Quality Statistics and their use in Steganalysis and Compression", PhD thesis, Boğaziçi University, Istanbul, Turkey, 2001

[6] Katznbeisser S., Petitcolas F., Information Hiding Teqniques for Steganography and Digital Watermarking, artech house, 2000.

[7] K B Shiva Kumar , K B Raja, R K Chhotaray, Sabyasachi Pattnaik, "Steganography Based on Payload Transformation", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011

[8] Maurice Maes " Twin Peaks: The Histogram Attack to Fixed Depth Image Watermarks", Information hiding second international workshop proceedings, 1998, vol. 1525 of lecture notes in computer science Springer pp 290-305

[9] Mehdi Kharrazi a, Husrev T Sencar b, Nasir Memon, "Cover Selection for Steganographic Embeddeing", IEEE International Conference on Image Processing, 2006.

[10] Ross J. Anderson "Stretching the Limits of Steganography" , proceedings of the first international workshop on information hiding, 1996, springer-Verlag, London, UK, pages 39-48.

[11] Saaty, Thomas. L., Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process, RWS Publications, Pittsburgh, PA. 1996

[12] **Saaty, Thomas L**. "Relative Measurement and its Generalization in Decision Making: Why Pairwise Comparisons are Central in Mathematics for the Measurement of Intangible Factors - The Analytic Hierarchy/Network Process". *RACSAM (Review of the Royal Spanish Academy of Sciences, Series A, Mathematics)* **102** (2): 251–318., 2008

[13] V. Lokeswara Reddy, A. Subramanyam, P. Chenna Reddy, "Implementation of LSB Steganography and its Evaluation for Various File Formats", Int. J. Advanced Networking and Applications, Volume: 02, Issue: 05, Pages: 868-872, 2011.

[14] Weiming Zhang, Shuozhong Wang, and Xinpeng Zhang," Improving Embedding Efficiency of Covering Codes for Applications in Steganography", IEEE communicatons letters, vol. 11, no. 8, 2007

[15] Westfeld, A. "High Capacity Despite Better Steganalysis: F5 – a Steganographic Algorithm." *Proceedings of the 4th Information Hiding Workshop, Lecture Notes in Computer Science 2137*, 2001, pages 301-314.

[16] Yong Xu, Hui Ji, "Viewpoint Invariant Texture Description Using Fractal Analysis", International Journal of Computer vision, Volume 83, Issue 1, pages 85-100, 2009

[17] Zöllner, J., H. Federrath, etl. "Modeling the Security of Steganographic Systems." Proceedings of the 2nd Workshop on Information Hiding, Lecture Notes In Computer Science. 1998, Springer-Verlag, pages 344-354.

Nidhal El-Abbadi received BSc in chemical engineering, BSc, MSc, and PhD in computer science, worked in industry and many universities, he is general secretary of colleges of computing and informatics society in Iraq, Member of Editorial Board of Journal of Computing and Applications, reviewer for a number of international journals, has many published papers and three published books (programming with Pascal, C++ from beginning to OOP, Data structures in simple language), his research interests are in image processing, biomedical, and steganography, He's Associate Professor in Computer Science in the University of Kufa – Najaf, IRAQ.

A New Image Fusion Technique to Improve the Quality of Remote Sensing images

A. El Ejaily[1], F. Eltohamy [2] , M.Y. EL Nahas [3] and G. Ismail [2]

[1] MTC
Cairo/ Egypt

[2] Egyptian Armed Force
Cairo/ Egypt

[3] Alazhar University
Cairo/Egypt

Abstract

Image fusion is a process of producing a single fused image from a set of input images. In this paper a new fusion technique based on the use of independent component analysis (ICA) and IHS transformation is proposed. A comparison of this new technique with PCA, IHS, and ICA-based fusion techniques is given. Quick Bird data are used to test these techniques, the output was evaluated using subjective comparison, statistical correlation, information entropy, mean square error, and standard deviation. The results of the proposed technique show higher performance compared to the other techniques.

***keywords*:** *Image fusion, principle component analysis(PCA), Independent component analysis(ICA), Intensity Hue Saturation(IHS).*

1. Introduction

Image fusion is the process of combining relevant information from two or more images into a single image. The resulting image will be more informative than any of the input images [1].

Many techniques and software tools for fusing images have been developed [2-4]. From the well-known methods the Brovey method, the intensity–hue–saturation (IHS) , principal component analysis (PCA), and discrete wavelet transform (DWT). Although these methods (IHS, PCA, and Brovey transform)
provide us with fused images with high spatial resolution, but they suffer from spectral distortion that introduced during the fusion process.
The invention of multi-resolution analysis tool like wavelet transform allow us to further improve the quality of the fused image, but it needs a proper selection of decomposed level according to the ratio between the PAN and MS images .The former studies show that we can use independent component analysis (ICA) to achieve a fused image with a high spatial resolution and minimum spectral distortion [5,6].
The ICA-based remote sensing image fusion is performed by transform the MS image into its independent components then the most significant IC is replace by PAN image and then an inverse ICA is performed on the new combination to retain the fused image. the author in [7] used wavelet transform to improve the performance of ICA-fusion technique.
In this paper we propose a new approach based on the use of IHS transform and ICA to improve the ICA performance in the fusion of remote sensing images. First the MS and PAN images are fused using a classical ICA method to get intermediate fused image then we use IHS transform to transform the resultant fused image together with the original MS image to IHS color space
in the last step inverse IHS is applied to color components - H & S- of MS image and the intensity component of the intermediate fused image.

2. Concept of Independent Component Analysis.

ICA is a computational method for separating a multivariate signal into additive subcomponents supposing the mutual statistical independence of the non-Gaussian source signals[8].
The estimation of the data model of independent component analysis is usually performed by formulating an objective function and then minimizing or maximizing it. Therefore, the properties of the ICA method depend on both the objective function and the optimization algorithm. assume that there is an M-dimensional zero mean vector $S = (S_1, S_2, \ldots, S_M)^T$,

whose components are mutually independent distributions

$$p(s) = \prod_i^M p_i(S_i) \qquad (1)$$

a data vector $x = (x_1, x_2, \dots, x_N)^T$ is observed vector at point t , such that

$$s(t) = Ax(t) \qquad (2)$$

where, A is an N*M scalar matrix which is called mixing matrix.
The goal of ICA is to find a linear transformation *W* of the correlative signal *x* that makes the outputs as independent as po

$$u(t) = Wx(t) = WAs(t) \qquad (3)$$

Where, u is an estimate of the sources.

3. Image fusion algorithm based on ICA-IHS

Although the main significant IC contains the main content of color image, the spectral and spatial information of the color image is not completely separated. Some of spatial information are also located in the other ICs. Therefore, the ICA- based fusion method tends to produce a fused image with high spatial detail, but also may lead to spectral distortion in some local regions. To minimize this spectral distortion we use IHS transform to further process the resultant fused image and replacing its color components with that of original MS image.
This is the central idea of the improved ICA fusion method. Figure 1 outlines the general procedure to fuse MS and PAN images using the improved ICA method based on IHS, and the detailed steps of ICA-IHS fusion method are as follows:
(1) Registering the MS and the PAN images with the error pixel.
(2) Applying ICA to MS bands to transform the correlated band to a linear combination of independent components.
(3) The most significant component of step (2) is replaced with PAN image, then the intermediate fused image is obtained by inverse ICA.
(4) To overcome the spectral distortion of the intermediate fused image, we transform this fused image and the original MS image to IHS color space and we replace the color components (hue & saturation) of the intermediate fused image with that of original MS image

(5) At last the final fused image is obtained by inverse IHS of the new combination of step (4).

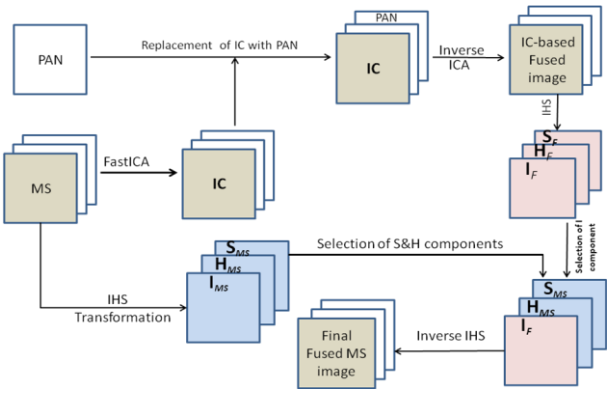


Fig. 1. the flow diagram of the proposed image fusion method.

3. Quality Assessment Criteria

Quality refers to both the spatial and spectral quality of images [9]. The goal of image fusion is to increase the spatial resolution of the MS images while preserving their spectral contents. In this paper we use correlation and discrepancy to assess the spectral quality of the fused image, while for the measurement of spatial quality we use the high-pass correlation and entropy.

3.1 spectral quality assessment:

For a good spectral fidelity of the fused image, the low spatial frequency information in the high-resolution image shouldn't be transferred to the fused image. The following measures are used to test the spectral quality of the fused image:

3.1.1 correlation coefficient(CC):

CC measures the correlation between the original MS image and the fused images. The higher the correlation between the fused and the original images, the better the estimation of the spectral values [10]. The ideal value of correlation coefficient is 1.
The formula to compute the correlation between two images A&B is given by:

$$Corr(A/B) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(A_{i,j} - \bar{A})(B_{i,j} - \bar{B})}{\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M}(A_{i,j} - \bar{A})\sum_{i=1}^{N}\sum_{j=1}^{M}(B_{i,j} - \bar{B})}} \qquad (4)$$

Where $\bar{A}$ and $\bar{B}$ are the mean values of the corresponding data set, (i,j) denotes a given pixel, N&M are the image dimension.

3.1.2 discrepancy:

The spectral quality is measured by the discrepancy D_k, at each band as follows:

$$D_k = \frac{1}{M * N}\sum_{i=1}^{M}\sum_{j=1}^{N}\left|F_{K,i,j} - MS_{K,i,j}\right|$$

$$K = R, G, B \qquad (5)$$

Where $F_{K,i,j}$ and $MS_{K,i,j}$ are the pixel values at position(*i,j*) in the *Kth* band of the fused and original MS image, respectively. It is known that the spectral quality of the image increases as *DK* decreases[11].

3.1 Spatial quality assessment:

For a good spatial fidelity of the fused image, the high spatial frequency information in the high-resolution image(that represent image edges) should be transferred to the fused image. The indexes which can reflect the spatial fidelity of fused image include:

3.2.1 High pass correlation:

The spatial quality is obtained by computing the correlation coefficient, *CK*, between the high pass component of the fused image bands and the original PAN image. The high pass component is used in the evaluation of this criterion because the spatial information is mostly concentrated in the high frequency domain. The spatial quality of the image is directly proportional to *CK*[12]. To extract the high frequency data we apply the following convolution mask to the images:

$$mask = \begin{vmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{vmatrix} \qquad (6)$$

3.2.1 Entropy:

The Entropy(EN) can show the average information included in the image and reflect the detail information of the fused image [10]. The greater the Entropy of the fused image is, the more abundant information included in it, and the greater the quality of the fusion is.
According to the information theory of Shannon, The image entropy is given by:

$$EN = -\sum_{g=0}^{L-1} p(g)\log_2 p(g) \qquad (7)$$

Where p(g) is the probability of grey level g, and the range of g is [0,...,L-1]. And L is the maximum value of the gray levels.

4. Experimental results and discussion

The QuickBird data were selected for carrying out the experimental work of this study, which were acquired on June 2006. the site were located in Tripoli, Libya. The images are composed of MS image with spatial resolution of 2.44m and PAN image with spatial resolution of 0.6 m. A sub-images of 1000 x 1000 pixels have been considered. Figure 2(a) and figure 2(b) shows the MS and PAN image respectively.

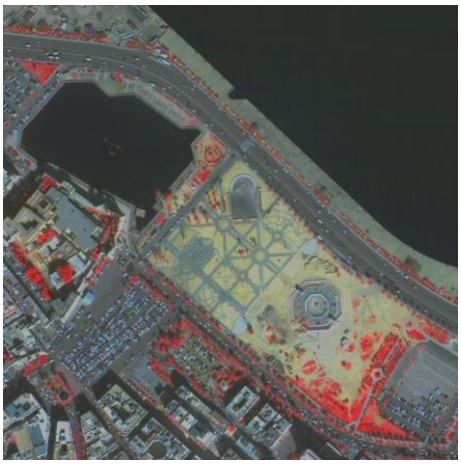
(a)

(b)

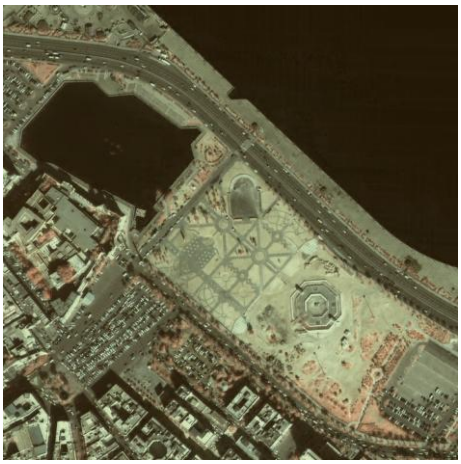
Fig. 2. QuickBird data of Tripoli (a) MS image; (b) PAN image

In order to examine the performance of the proposed image fusion method, we compare it with other fusion methods like IHS, PCA, and ICA. The resultant fused images are shown in figure 3.

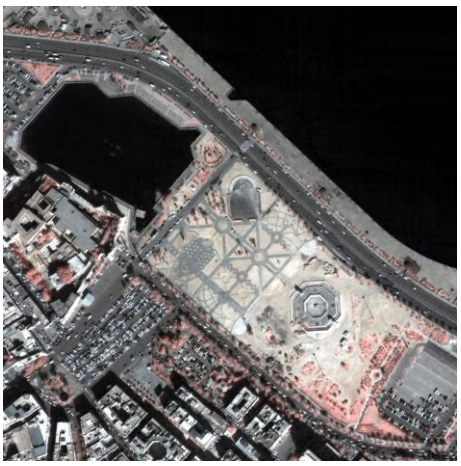
(a)



(b)



(c)



(d)

Fig. 3. The results of different fusion methods using QuickBird data (a)fused image by IHS; (b) fused image by PCA; (c) fused image by ICA; (d) the fusion result of applying the proposed method.

The visual inspection of the obtained fused images shows that all fusion methods retained an acceptable amount of spatial detail, the PCA method preserves fewer details than IHS. we get a maximum spatial detail in the fused image when we use ICA, but the distortion of the original color is very obvious . The improved ICA method proposed in this paper can solve this problem effectively, and we can achieve a fused image with a better balance between spectral characteristic preservation and high spatial resolution

To objectively assist the spatial quality of the obtained fused images we measure the high-pass correlation coefficients(C_k) and the information entropy (*EN*), the results of those two measure are shown in table 1, the results show that a fused image with a maximum spatial details is achieved when we used the improved ICA method. To assess the spectral quality we use the measurements of spectral discrepancy(D_k) and correlation coefficients (CCs). Table 2 shows the results which indicates that the improved ICA and IHS methods have almost the same spectral quality better than the other fusion techniques.

Table 1. The results of spatial quality assessment.

Fusion method	C_k			EN		
	R	G	B	R	G	B
PCA	0.7838	0.8185	0.7640	6.7397	6.6563	6.7061
IHS	0.9343	0.8969	0.8933	7.4069	7.4019	7.3456
ICA	0.8376	0.9309	0.8616	6.9958	7.1663	7.0015
proposed method	0.9269	0.9756	0.9550	7.5223	7.4881	7.5161

Table 2. The results of spectral quality assessment.

Fusion method	D_k			CC		
	R	G	B	R	G	B
PCA	0.1343	0.1330	0.1213	0.9314	0.9298	0.9338
IHS	0.0685	0.0690	0.0679	0.9706	0.9710	0.9680
ICA	0.1144	0.1152	0.1182	0.9527	0.9454	0.9083
Proposed method	0.0677	0.0902	0.0787	0.9714	0.9517	0.9552

5. Conclusion

In this paper we present an improved ICA image fusion method. In the convenient way of ICA-based image fusion, the most significant IC is replaced with PAN image then the fused image is obtained through an inverse ICA. This fused image is further processed by transform it to IHS color space. Then we select the intensity component together with color components (H and S) of the original MS image to produce a new fused image using inverse IHS transformation. Such new fused image will contain the original color with a high spatial detail. The QuickBird images were used to show the performance of the proposed method. The results show that the proposed method can preserve abundant spectral information and higher spatial resolution, and has a better performance compared with other techniques.

References

[1] C. Pohl, and V. Genderen,"Multisensory image fusion in remote sensing: concepts, methods and application", in int. j. remote sensing, Vol. 19, No. 5, 1998,pp823- 854.

[2] M. Gonzalez, and J.L.Saleta,"Fusion of Multispectral and Panachromatic Images Using Improved IHS and PCA Mergers Based on Wavelet Transform", IEEE Transaction on Geoscience and Remote Sensing, Vol. 42, No. 6,Joun 2004.

[3] M. Choi "A new Intensity-Hue-Saturation Fusion Approach to Image Fusion With Tradeoff Parameter", IEEE Transaction on Geoscience and Remote Sensing, Vol. 44, No. 6,Joun 2006.

[4] Li. Kwok ,and Y.Wang "Using the Discrete Wavelet Transform Frame Transform to Merge Landsat TM and SPOT Panchromatic Images", information fusion,2002, pp 17-23.

[5] H. Wang, Lu. Wei. Name, and Cai he "An adaptive Image Fusion Algorithm Based on ICA" in International Conference on Computer Mechatronic, Control and Electronic Engineering (CMCE) 2010.

[6] M. Wang, and J. Yang, "Multi-sensor Image Fusion with ICA Based Region Rule",in 10t International Conference on Control, Automation, Robotics and Vision, December 2008.

[7] F. Chen, F. Qin, G. Peng, and S. Chen "Fusion of Remote Sensing Images Using Improved ICA Mergers Based on Wavelet Decomposition", in International Workshop on Information and Electronic Engineering (IWIEE), 2012

[8] A. Hyvarinen, J. Karhunen, and E. Oja, "Independent Component Analysis", New York: John Wiely & Son, Inc, 2001.

[9] L. Wald, T. Ranchin, and M. Mangolinii" Fusion of Satellite Images of Different Spatial Resolution: Assessing the Quality of Resulting Images" in Photogrammetric Engineering & Remote Sensing, Vol. 63, No. 6, 1997, pp. 691-699.

[10] S. Han, H.T. Li, and H.Y. Li "The Study of Image Fusion for High Spatial Resolution Remote Sensing Images", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science Vol. XXXVII, Part B7, Beijing 2008.

[11] Z. Zhenhua, Z. Jing, and X. yang, "Color Transform Based Remote Sensing Image Fusion Using Non-Separable Wavelet Frame Transform", Pattern Recognition Letters, Vol. 26, Issue. 13, 2055, pp. 2006-2014.

[12] D. Zhou, J. Civo, and A. Silander "wavelet Transform Method to Merge Landsat TM and Spot Panchromatic Data", in Int. J . Remote Sensing, Vol. 19,No. 4, 1998, pp. 743-757.

Video-based multiclass vehicle detection and tracking

Zhiming Qian, Hongxing Shi , Jiakuan Yang and Lianxin Duan

Chuxiong Normal University
Chuxiong, 675000, China

Abstract

This paper presents a real time multiclass vehicle detection and tracking system. The system uses a combination of machine learning and feature analysis to detect and track the vehicles on the road. Multiclass SVM and PCA methods are utilized to create multiclass training samples. The online classifiers are trained using these samples to achieve detection and classification of vehicles in video sequences of traffic scenes. The detection results provide the system used for tracking. Each class vehicle is tracked by SIFT method. The system combines the advantages of both multiclass detection and tracking in a single framework. Experimental results from highway scenes are provided which demonstrate the effectiveness of the method.

Keywords: *Vehicle detection, Vehicle tracking, Online learning, Feature analysis.*

1. Introduction

Video based intelligent transportation systems (ITS) are getting large attention as an attractive field, not only because they are easy to install and operate, but also because they have the potential to provide a much richer description about vehicle. As the basic parts, detection and tracking of vehicle is a fundamental problem in ITS. For this task, we need to first detect the vehicle and segment them from the video images, and then track them across different frames while maintaining the correct identities.

Robust detection and tracking of vehicles on the road based on video is a challenging problem. Roads are dynamic environments, with the illumination and background changes. The sizes and the locations of vehicles on the road are diverse. There is high variability in the appearance of vehicles with viewpoint, illumination, and possible articulation. Moreover, partial occlusion of vehicles of interest by other vehicles or objects on the road is also an important factor influencing detection and tracking.

For the last two decades researchers have spend quality time to develop different methods that can be applied in the field of video based vehicle detection and tracking [1-3].

In the following section, we will present a brief overview of recent related works in video vehicle detection and tracking.

Video vehicle detection is a process of detection the presence or absence of a vehicle in the sequences. The result of detection is used as initialization process for tracking. There are four main approaches to detect vehicle regions, they are：

1. Frame differencing method [4,5]: this method detects moving vehicle regions by subtracting two consecutive image frames in the image sequence. It works well in case of uniform illumination conditions, otherwise it creates non-vehicular region and also frame differencing method does not work well if the time interval between the frames being subtracted is too large.

2. Background subtraction method [6,7]: this method is one of the widely used methods to detect moving vehicle regions. It subtracts the generated background image from the input image frame to detect the moving vehicle regions. This difference image is then thresholded to extract the vehicle regions. The problem with the stored background frame is that they are not adaptive to the environment changes which may create non-existent vehicle regions and also works for stationary background.

3. Feature based method [8,9]: this method made use of sub-features to detect moving vehicle regions. These features are grouped by analyzing their motion between consecutive frames. Thus a group of features segments a moving vehicle from the background. The advantages of this method is that the problem of occlusion between the vehicle regions can be handled well, the feature based methods have less computational complexity compared to background subtraction method. But the disadvantage is that if the features are not grouped accurately, then there may be failure in detecting vehicles correctly.

4. Motion based method [10,11]: this method assumes that vehicles tends to move in a consistent direction over time and that foreground motion has different saliency. It is less sensitive to noise and very effecting on small moving objects, but the disadvantage is that calculation of motion information consumes time, and it can not be used to

detect static obstacles which can represent a big threat to detection task.

After vehicle detection, ITS will carry out the task of vehicle tracking. Vehicle tracking is a process that generates the trajectory of the vehicle over time by locating its position in every frame of the video sequences. The existing tracking approaches may be classified into four major categories:

1. Region based method [12,13]: this method subtracts image frame containing vehicles from the background frame which is then further processed to obtain vehicle regions (blobs). Then these vehicle regions are tracked. It can work well in free flowing traffic conditions, but the disadvantage is that it has difficulty in handling shadows and occlusion.

2. Active contour based method [14,15]: this method represents vehicle by bounding contour of the object and dynamically update it during the tracking. The advantage of active contour tracking over region-based tracking is the reduced computational complexity. But the disadvantage of the method is their inability to accurately track the occluded vehicles and tracking need to be initialized on each vehicle separately to handle occlusion better.

3. Feature based method [16,17]: this method extracts suitable features from the vehicle regions and these features are processed to track the vehicles correctly. The method has low complexity and also can handle occlusions well. The disadvantage is the recognition rate of vehicles using tow-dimensional image features is low, and the problem that which set of sub features belong to one object is complex.

4. Model based method [18,19]: this method tracks vehicle by matching a projected model to the image data. The advantages of model based vehicle tracking is it is robust to interference between nearby images and also be applied to vehicle classification. But the method has high computational cost and they need detailed geometric object model to achieve high tracking accuracy.

Above approaches can effectively accomplish detection and tracking tasks. However, these approaches need more system computation and have certain application conditions. In order to reduce calculation time and to improve system efficiency, learning-based approaches have been adopted by many researchers to detect and track video vehicles efficiently.

Based on how the learning takes place over time, the learning-based approaches can be categorized as offline learning and online learning. Offline learning requires all the training data to be available from the beginning of learning process. These kind of approaches try to produce results which are consistent with all the collected data samples. On the other hand, online learning requires the training data to arrive sequentially over time and additionally. These kind of approaches provide the machine with the ability to learn continuously and adapt all the time to its inputs.

There are some offline learning methods to detect and track video vehicle which have obtained good results. Sun et al [20] employed support vector machines to learn Haar wavelet features for vehicle detection. Junior and Nunes [21] used multilayer feed forward neural network based approach to detect vehicles. Khammari et al [22] utilized a gradient analysis and Adaboost classification to accomplish rear vehicle detection. Negri et al [23] presented an algorithm for the on-board vision vehicle detection problem using a cascade of boosted classifiers. Withopf and Jahne [24] presented a learning algorithm for real-time vehicle tracking in video sequences which uses an improvement of a feature selection method. Chen et al [25] proposed a framework for spatiotemporal vehicle tracking using unsupervised learning-based segmentation and object tracking.

In offline learning methods, large amount of training samples could be required for obtaining a generic detector. The quality and quantity of the training samples directly determine the detection and tracking performance of the system. In order to resolve the problem, online learning methods have been an area of great recent interest in the vehicle detection and tracking. Nguyen et al [26] employed online boosting algorithm for car detection from high resolution aerial images. Chang and Cho [27] presented a real time vision based vehicle detection system using an online Adaboost algorithm. Sivaraman and Trivedi [28] proposed a general active-learning framework for on-road vehicle detection and tracking.

In real world, the vehicle type is various. Comparing the strategy that all vehicles are categorized as single class, multiclass vehicle detection and tracking have great practical significance and applicable value great practical importance. In this paper, a framework for video multiclass vehicles detection and tracking is introduced. The proposed framework has the following characteristics: (1)It has multiclass vehicles detection ability; (2)It can be update based on new training samples which come from video images to adapt new environment; (3)It can track vehicles accurately in real-time environment. The

proposed framework in this paper has been validated with video vehicle sequences from real-world traffic scenes.

2. The proposed framework

2.1 Overall structure

Given an input of a video sequence taken from roadway vehicles, system first outputs the types and locations of the vehicles in the images, then a feature information description of the detected vehicles is obtained, and finally this description is used to match the detected vehicles in the next frame. The framework contains three main processes: vehicle classification, vehicle detection, and vehicle tracking. In the vehicle classification process, using offline learning to create multiclass classifier, once the created multiclass classifier recognizes a potential vehicle in an image, the system generates a train sample for a corresponding vehicle detector. The vehicle detectors were then trained by online learning based on these generated train samples. In the vehicle detection process, using the trained vehicle detectors to classify and locate vehicles from video sequence, while at the same time the vehicle detectors will continue to be trained to improve detection ability. In the vehicle tracking process, the tracker analyzes the feature information of the detected vehicles in the previous image frames and matches the feature information of the detected vehicles in the current image. If the matching result is accurate, the tracker outputs the label information for the detected vehicle. A general overview of the system framework can be seen in Fig. 1.

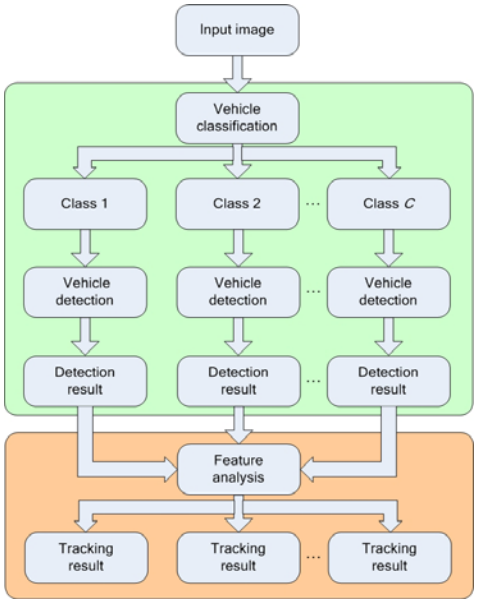


Fig. 1 Overview of our framework.

2.2 Vehicle classification

In order to achieve vehicle classification task, multiclass SVM is employed to our framework. The SVM has been introduced as one of the most efficient learning algorithms in computer vision. While many challenging classification problems are inherently multiclass, the original SVM is only able to solve binary classification problems. Due to significant appearance variation across different vehicles, a direct solution of vehicle classification using single SVM module should be avoided. The better method is to use a combination of several binary SVM classifiers to classify vehicles. The "one against one" and the "one against all" are the two most popular methods for multiclass SVM. Hsu and Lin [29] had compared the performance of the two methods with a large set of different problems. Experiments show that the "one against one" method may be more suitable for practical use.

To be useful, the task of vehicle classification should categorize vehicles into a sufficiently large number of classes. However as the number of class increases, the processing time required also increases. Therefore, a simple classification method is needed which can quickly categorize vehicles at a coarse level. Based on the application, further classification can be done. In the paper, we use the "one against one" method in the LibSVM [30] to learn Haar wavelet features for vehicle classification.

The one-against-one method constructs an SVM for every pair of classes by training it to discriminate the two classes. If k is the number of classes, then $k(k\text{-}1)/2$ classifiers are constructed and each one trains data from two classes. The decision function for class pair ij is defined by

$$f_{ij}(x) = (\phi(x) \cdot w^{ij}) + b^{ij} \tag{1}$$

It is found by solving the following optimization problem:

$$\min \frac{1}{2}\left\|w^{ij}\right\|^2 + C\sum_n \xi_n^{ij} \tag{2}$$

$$\begin{cases} \phi(x_n)\cdot w^{ij} + b^{ij} \geq 1 - \xi_t^{ij}; & \xi_t^{ij} \geq 0, \quad x_t \text{ in the ith class} \\ \phi(x_n)\cdot w^{ij} + b^{ij} \leq \xi_t^{ij} - 1; & \xi_t^{ij} \geq 0, \quad x_t \text{ in the jth class} \end{cases} \tag{3}$$

Finally, the "max wins" voting strategy is used to determine the class of a test pattern in this approach. Fig. 2 shows the flowchart of vehicle classification.

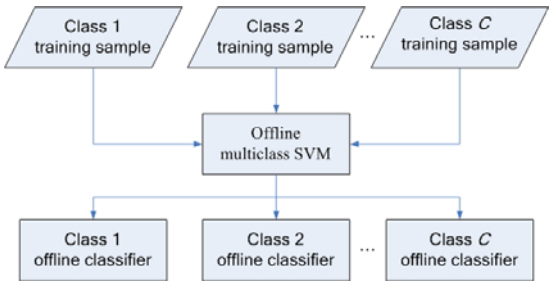


Fig. 2 Flowchart of offline classification.

2.3 Sample creation

Due to various complexities, the classification results using multiclass SVM may be inconsistent with the expected results. However, the classification results are very important for online learning, and its accuracy can directly affects the performance of the detection system. In order to eliminate these false results, we consider using eigenvehicle method to filter the classification results as post processing. Eigenvehicle method is based on the well-known method eigenface [31]. However as the method is used for vehicle detection we named it as eigenvehicle method. The main idea is to decompose vehicle images into a small set of characteristics feature images called eigenvehicle, which may be thought of as the principal components of the original images. The eigenvehicle function as the orthogonal basis vectors of a subspace called vehiclespace. For each class of vehicle, we prepare M=50 vehicle images as the train set. Each image in the train set is transformed into a vector of size N and placed into the set:

$$S = \{\Gamma_1, \Gamma_2, \Gamma_3, \ldots\ldots, \Gamma_M\} \tag{4}$$

The average matrix is calculated, then subtracted from the original samples and the result stored in the variable Φ_i:

$$\Psi = \frac{1}{M}\sum_{n=1}^{M}\Gamma_n \tag{5}$$

$$\Phi_i = \Gamma_i - \Psi \tag{6}$$

In the next step the covariance matrix C is calculated according to

$$C = \frac{1}{M}\sum_{n=1}^{M}\Phi_n\Phi_n^T = AA^T \tag{7}$$

Calculate the eigenvectors and eigenvalues of the covariance matrix:

$$\lambda_k = \frac{1}{M}\sum_{n=1}^{M}(u_k^T\Phi_n)^2 \tag{8}$$

$$u_l^T u_k = \delta_{lk} = \begin{cases} 1 & l = k \\ 0 & otherwise \end{cases} \tag{9}$$

Finally, the eigenvehicle will be obtained

$$L = A^T AL_{nm} = \Phi_m^T\Phi_n \tag{10}$$

$$u_l = \sum_{k=1}^{M} v_{lk}\Phi_k \qquad l = 1, \ldots\ldots, M \tag{11}$$

where L is a $M \times M$ matrix, v are M eigenvectors of L.

While a new sample detected by multiclass SVM coming into, it is transformed into its eigenvehicle components. First we compare sample image with mean image of the same class and multiply their difference with each eigenvector of the L matrix. Each value would represent a weight ω and would be saved on a vector.

$$\omega_k = u_k^T(\Gamma - \Psi) \qquad l = 1, \ldots\ldots, M \tag{12}$$

$$\Omega_{new}^T = [\omega_1, \omega_2, \ldots\ldots, \omega_M] \tag{13}$$

Calculate the average Euclidean distance of between the new sample and all the eigenvehicle of the same class. If the value D is bellow an established threshold θ, the input sample is consider to belong to a vehicle image of the corresponding class.

$$D = \frac{1}{M}\sum_{k=1}^{M}\left\|\Omega_{new}^T - \Omega_k\right\| \tag{14}$$

2.4 Online learning

With video sequences as input, a series of training samples are collected by the system and then fed into the boosting learning algorithm. Boosting is one of the mostly applied methods in vehicle detection. Boosting for vehicle detection as described in the previous section most works offline. Hence, all training samples must be given in advance, which is not the case for vehicle detection in video environment.

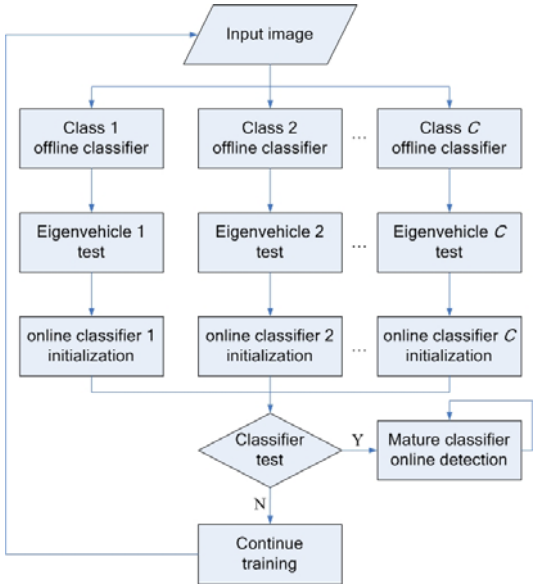


Fig. 3 Flowchart of online detection.

Since for online learning each training sample is discarded directly after an update all steps have to be online. In this paper, we select Haar-like features as the weak classifier, and use Grabner et al's [32] online boosting method creating vehicle detector. The main steps of online learning are briefly described below:

A selector $s_n(x)$ can be considered a set of w weak classifiers $\{h_1(x),\ldots,h_w(x)\}$ that are related to a subset of features $F_n = \{f_1,\ldots,f_w(x)\}$, where F is the full feature pool. At each time the selector $s_n(x)$ selects the best weak hypothesis according to the estimated training error.

To start the learning process a fixed set of n selectors $s_1,\ldots,s_n$ is initialized randomly. Whenever a new training sample (x, y) arrives the selectors are updated. These updates are performed with respect to the importance weight λ of the current sample, which is initialized with $\lambda = 1$.

To update the selector s_n first all weak classifiers $h_{n,m}(x)$ are estimated by evaluating the feature $f_{n,m}$ on the sample image x and the corresponding errors:

$$\varepsilon_{n,m} = \frac{\lambda_{n,m}^{wrong}}{\lambda_{n,m}^{corr} + \lambda_{n,m}^{wrong}} \tag{15}$$

are computed. The weights, $\lambda_{n,m}^{corr}$ and $\lambda_{n,m}^{wrong}$ are estimated from the correctly and wrongly classified examples seen so far. Then, the selector s_n selects the weak classifier h_{n,m^+} with the smallest error $\varepsilon_n = \varepsilon_{n,m^+}$, where $m^+ = \text{argmin}_m(\varepsilon_{n,m})$:

$$s_n(x) = h_{n,m^+}(x) \tag{16}$$

According to the error ε_n the voting weight α_n and the importance weight λ are updated:

$$\alpha_n = \frac{1}{2}\ln\left(\frac{1-\varepsilon_n}{\varepsilon_n}\right) \tag{17}$$

$$\lambda = \begin{cases} \lambda \dfrac{1}{2(1-\varepsilon_n)} & s_n(x) = y \\ \lambda \dfrac{1}{2\varepsilon_n} & otherwise \end{cases} \tag{18}$$

The importance weight λ is passed to the next selector s_{n+1}. In order to increase the diversity of the classifier pool F_n and to adapt to changes in the environment the worst weak classifier h_{n,m^-}, where $m^- = \text{argmax}_m(\varepsilon_{n,m})$, is replaced by a classifier randomly chosen from the feature pool F. Finally, a strong classifier is computed by a linear combination of N selectors:

$$H(x) = sign\left(\sum_{n=1}^{N} \alpha_n . s_n(x)\right) \tag{19}$$

After all online classifiers are constructed, we will obtain C different vehicle classifiers. When a new image entering, it will be analyzed use these classifiers based on the "max wins" voting strategy, so that achieve the task of vehicle detection and classification.

$$R = \max_i sign(H_i(x)) \qquad i = 1,\ldots\ldots,C \tag{20}$$

Fig. 3 shows the flowchart of online detection.

2.5 SIFT feature analysis

SIFT(Scale Invariant Feature Transform)is a well-established local feature descriptors, which was proposed in 1999 by Lowe [33]. Duo to SIFT feature descriptor is invariant to uniform scaling, orientation, and partially invariant to affine distortion and illumination changes, it has been widely applied to object tracking and image matching. For multiclass vehicle tracking, we need a kind of feature which can describe different vehicles accurately, the SIFT feature is very suitable in the circumstance. The SIFT algorithm includes four steps: scale-space extrema detection, feature point localization, orientation assignment and generation of feature point descriptors. Main process is as follows:

Interest points for SIFT features correspond to local extrema of difference-of-Gaussian filters at different scales. Given a Gaussian-blurred image described as the formula

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{21}$$

where L is the scale space of an 2D image, $I(x,y)$ is the gray value of input image in the coordinates (x,y), $G(x,y,\sigma)$is a variable scale Gaussian, whose result of convolving an image with a difference-of-Gaussian filter is given by

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{22}$$

which is just be different from the Gaussian-blurred images at scales σ and $k\sigma$. Interest points are identified as local maxima or minima of the DoG images across scales. Each pixel in the DoG images is compared to its 8 neighbors at the same scale, plus the 9 corresponding neighbors at neighboring scales. If the pixel is a local maximum or minimum, it is selected as a candidate feature point. Remove the low contrast candidate points and eliminated the edge response, then use Hessian matrix to compute the principal curvatures and eliminate these feature points that have a ratio between the principal curvatures greater than the ratio.

Finally, an orientation histogram was formed from the gradient orientations of sample points within a 4×4 region with 8 orientations around the feature point in order to get an orientation assignment. So the descriptor of SIFT that was used is 4× 4 × 8 = 128 dimensions.

2.6 Feature matching and updating

For each vehicle detected from multiclass detection framework, extract SIFT feature and establish vehicle information database (VID). The VID consists of four parts: vehicle class, vehicle number, vehicle location (rectangle coordinates) and SIFT feature point descriptor (feature priority, feature point coordinate, orientation and scale), each vehicle detected from multiclass detection framework is tracked in a new video frame sequences by separately comparing its feature point with the same class vehicle from the VID. The Euclidean distance is introduced as a similarity measurement of feature characters.

Suppose N_i as the feature number of the current vehicle matching the ith vehicle of the VID, N as the total feature number of the current vehicle, the matching rate between the current vehicle and the ith vehicle of the VID can be defined as $P_i=N_i/N$. Set the threshold T for the matching parameters. When P_i is greater than T, the current vehicle is considered equivalent to matching the ith vehicle. Supposing that M_j is the number of the jth class vehicle of the VID, $\{P_{ij}(j=1,\ldots,M_j)\}$ is the matching results of the current vehicle and all vehicles of the VID with the same class, and n is the number of elements in the set $\{P_{ij} | P_{ij} > T, j=1, \ldots, M_j\}$. When $n=1$, the ith vehicle is matching with the jth vehicle of the VID with the same class; when $n>1$, we select $\max(p_i)$ as the matching result.

The VID stores the data of vehicle which appears in the recent video sequences. It needs to be updated after one frame, input the current vehicle data and delete the data of the long term unmatched vehicle. We set a feature priority for each feature point of the VID in the vehicle information update process.

Suppose R_{ij} as the feature priority of the jth feature point of the ith vehicle, the specific update process is as following:

(1) Add new vehicle: if the current vehicle is not matching all the vehicle of VID with the same class, this vehicle will be considered as a new vehicle, add its information into the VID, and set its feature priority of all feature points $R=R_{\max}$.

(2) Update feature priority: if the current vehicle matches the ith vehicle of the VID, the information of the ith vehicle will be update, set the feature priority of these matching feature points $R_{ij}= R_{\max}$, and use new coordinate of these matching feature points to replace original coordinate. In addition, the feature priority of unmatched feature points between the current vehicle and the ith vehicle is replaced with $R_{ij} = R_{ij}$ -1, the new feature points of unmatched feature is added into the VID. After all matches of the current frame are finished, if there are no matching vehicles to be found from the VID, all the feature priority of these vehicles will be replace with $R_{ij} = R_{ij}$ -1. When a frame image is completely processed, the feature point whose feature priority is equal to zero will be removed from the VID.

(3) Delete vehicle: When a frame image is completely processed, the vehicle whose feature priority of feature point meets the following condition will be deleted from the VID.

$$R_1 + R_2 + \ldots + R_{All} < \theta \quad (23)$$

Fig. 4 shows the flowchart of vehicle tracking.

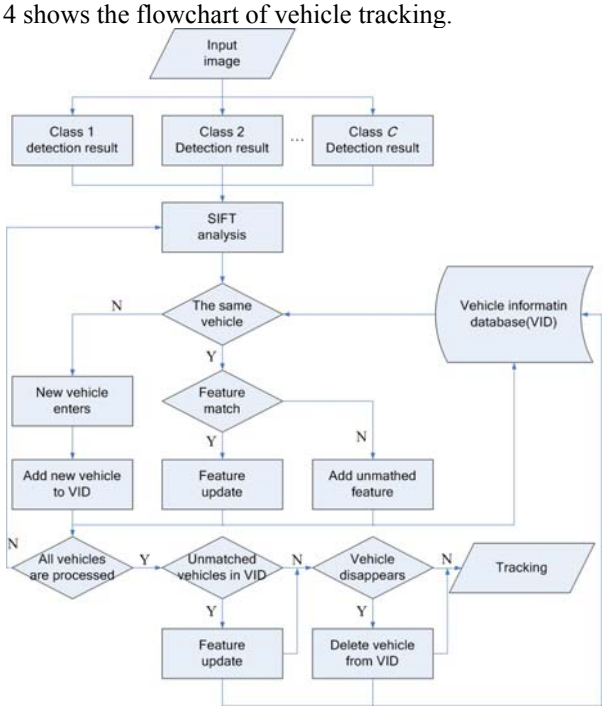


Fig. 4 Flowchart of vehicle tracking.

3. Experiment

We consider the samples from a profile viewpoint for vehicles, and all video sequences which are achieved a frame rate of about 20 fps were generated by shooting around Chuxiong city under highways conditions. All our experiments shown below on a standard PC (Intel Core2 Duo E7500 2.93GHz with 2 GB RAM). The strong classifier consists of 50 selectors and the shared feature pool provides 250 weak classifiers. Set the threshold θ=0.2, the number of class C=4 (motorcycle, bus, truck, and car).

In the training phrase, the data set is the image segmentation data, where each class is a vehicle type collected from a 32×16 region of a vehicle image. The training set consists of 500 samples per class. Some training images are shown in Fig. 5. In the test phrase, the data set is the video sequences, which consists of more than 1 hour of RGB video taken on city highways during the day. The test is divided into two parts, namely the detection test and the tracking test.

Fig. 5 A subset of the training samples for the four classes.

In the detection test, if the classifiers obtain detection result which gives the desired location and classification, the result will be considered to include in the detection rate; if all the classifiers do not obtain detection results or the detection results give the incorrect classification, the detection result will be considered to include in the error rate. The online training result as shown in Fig. 6, the result indicates that, with the increasing of sample size, detection rate increases continuously and finally, it fluctuates smoothly in some ranges. We use the classifier with half hour of training as the final vehicle classifier. In order to evaluate performance of the proposed method, we make a comparison of detection rate and error rate with and offline boosting classifiers. Establish a classifier for each vehicle class using 1200 positive samples and 1500 negative samples, and use the same dataset to test two methods. The experimental results are shown in Table 1. It clearly shows that our method performs better than the other method. More significantly, we create online multiclass classifiers which are suitable for video sequence with small training samples. Some detection results in the video sequences are shown in Fig. 7.

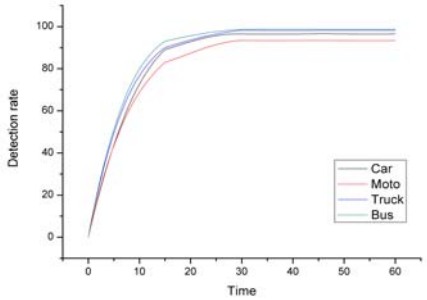


Fig. 6 Detection rate versus the time.

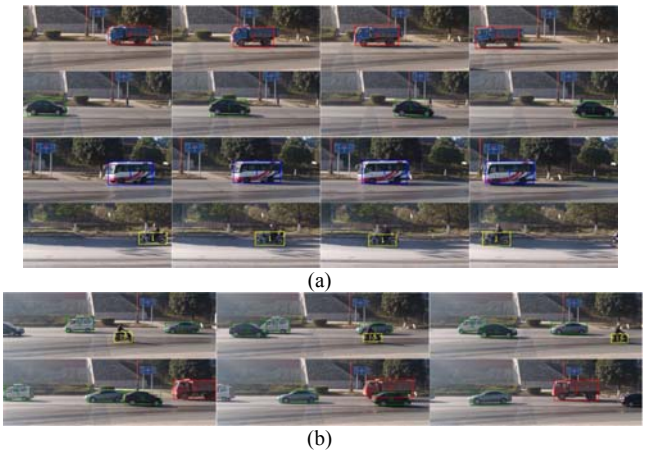


Fig. 7 (a)Single class detection results in the experimental sequences. (b) Multiclass detection results in the experimental sequences.

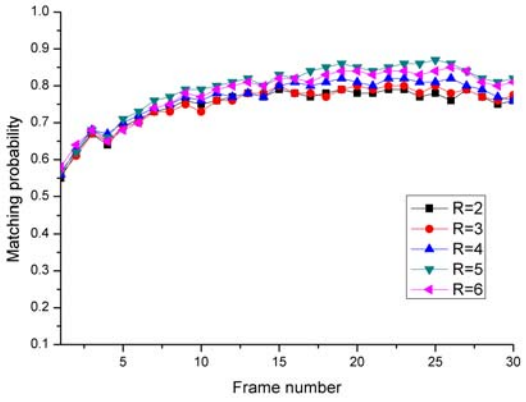


Fig. 8 Influence of R_{max} on matching probability.

In the tracking test, if the classifiers obtain detection result which gives the desired location and identifier, the result will be considered as the correct tracking in current frame, otherwise the result will be considered as the incorrect tracking in current frame. Since there are no suitble methods to compare the multiclass tracking effect, we just test our method on test data. Table 2 shows the tracking results for our method.

Our research also shows the performance of matching algorithm when the parameters R_{max} takes different values. The experimental results as shown in Fig. 8. It shows that while the SIFT features of vehicle were progressively increased with the vehicle packs with the target area between the first and ten frames. Between the ten and twenty frames, the matching algorithm achieves stability while the vehicle appears utterly. A proper value for the parameter R_{max} depends on the scene being modeled. In case of a simple scene, a small value for R_{max} is sufficient.

For complex scenes, more feature information is needed to match the vehicles. The proximity value R_{max} for feature matching is easy to find by experimenting with different values. Values between 4 and 5 gave good results for all of our test sequences. It should be noticed that the bigger the value of R_{max} , the slower the processing, and the greater the memory requirements.

Table 1: Comparison of detection results of two methods.

Class	Offline boosting detection rate	Offline boosting error rate	Proposed method detection rate	Proposed method error rate
Motocycle	71%	36%	92%	13%
Bus	85%	21%	98%	8%
Truck	78%	31%	96%	10%
Car	82%	27%	95%	12%

Table 2: Tracking results on video sequences.

Class	Tracked number	Vehicles not tracked	Average number of frames during tracking
Motocycle	71	6	35
Bus	85	3	46
Truck	78	2	42
Car	82	4	31

4. Conclusion

We have proposed a real-time vision framework that detects and tracks multiclass vehicles in video sequences. The method by learning a small number of labeled offline samples and a large number of unlabeled online samples to establish the vehicle classifier, and by analyzing the SIFT feature of detected vehicles to achieve vehicle tracking. The framework is able to run in real time with simple, low-cost hardware. Our experimental results demonstrate effective, multiclass vehicle detection and tracking in real traffic environments by applying the proposed framework. If new classes of vehicles or unfamiliar environments are encountered, the proposed framework can adapt itself to the changes and detect vehicles successfully.

Acknowledgements

This work was supported by the Natural Science Foundation of Yunnan Province, China (No. 2011FZ187).

Reference

[1] Z. H. Sun, G. Bebi, R. Miller, On-road vehicle detection: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (5) (2006) 694–711.

[2] M. M. Trivedi, T. Gandhi, J. McCall, Looking-in and looking-out of a vehicle: computer-vision-based enhanced vehicle safety, IEEE Transactions on Intelligent Transportation Systems 8 (1) (2007) 108–120.

[3] B. T. Morris, M. M. Trivedi, Learning, modeling, and classification of vehicle track patterns from live video, IEEE Transactions on Intelligent Transportation Systems 9 (3) (2008) 425–437.

[4] D. Dailey, F.W. Cathy, S. Pumrin, An algorithm to estimate mean traffic speed using uncalibrated cameras, IEEE Transactions on Intelligent Transportation Systems 1 (2) (2000) 98–107.

[5] T. N. Schoepflin, D. J. Dailey. Dynamic camera calibration of roadside traffic management cameras for vehicle speed estimation, IEEE Transactions on Intelligent Transportation Systems 4 (2) (2003) 90–98.

[6] C.Wern, A. Azarbayejani, T. Darrel, A. Petland, Pfinder, real-time tracking of human body, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 780–785.

[7] X. Gao, T. Boult, F. Coetzee, V. Ramesh, Error analysis of background adaption, IEEE Conference on Computer Vision and Pattern Recognition, 2000, pp. 503–510.

[8] D. Koller, K. Danilidis, H. H. Nagel, Model-based object tracking in monocular image sequences of road traffic scenes, International Journal of Computer Vision 10 (3) (1993) 257-281.

[9] S. M. Smith, J. M. Brady, ASSET-2: real-time motion segmentation and shape tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (8) (1995) 814-820.

[10] A. Bainhridge-Smith, R. G. Lane, Deremining optical flow using a differential method, Image and Vision Computing 15 (1) (1997) 11-22.

[11] L. Wilson, Detecting salient motion by accumulating directionally-consistent flow, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 774–780.

[12] D. J. Dailey, F. W. Cathy, S. Pumrin, An algorithm to estimate mean traffic speed using uncalibrated cameras, IEEE Transactions on Intelligent Transportation Systems 1 (2) (2000) 98– 107.

[13] S. Gupte, O. Masoud, R. F. K. Martin, N. P. Papanikolopoulos, Detection and classification of vehicles, IEEE Transactions on Intelligent Transportation Systems 3 (1) (2002) 37-47.

[14] 14. D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, S. Russell, Toward robust automatic traffic scene analysis in real-time, IEEE Conference on Pattern Recognition, 1994, pp. 126–131.

[15] N. Paragios, R. Deriche, Geodesic active contours and level sets for the detection and tracking of moving objects, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (3) (2000) 266–280.

[16] L. Huang, M. Barth, Real-time multi-vehicle tracking based on feature detection and color probability model, IEEE Conference on Intelligent Vehicles Symposium, 2010, pp. 981–986.

[17] B. Coifman, D. Beymerb, P. McLauchlan, J. Malik, A real-time computer vision system for vehicle tracking and traffic surveillance, Transportation Research Part C: Emerging Technologies 6 (4) (1998) 271–288.

[18] M. Haag and H. Nagel, Combination of edge element and optical flow estimate for 3D model-based vehicle tracking

in traffic image sequences, International Journal of Computer Vision 35(3) (1999) 295–319.
[19] T. N. Tan, K. D. Baker, Efficient image gradient based vehicle localization, IEEE Transactions on Image Processing 9(8) (2000) 1343–1356.
[20] Z. Sun, G. B. D. Dimeo, A real-time precrash vehicle detection system, IEEE Conference on Application of Computer Vision, 2002, pp. 171–176.
[21] O. Ludwig, U. Nunes, Improving the generalization properties of neural networks: an application to vehicle detection, IEEE Conference on Intelligent Transportation Systems, 2008, pp. 310–315.
[22] A. Khammari, Vehicle detection combining gradient analysis and adaboost classification, IEEE Conference on Intelligent Transportation Systems, 2005, pp. 66–71.
[23] P. Negri, X. Clady, S. Hanif, L. Prevost, A cascade of boosted generative and discriminative classifiers for vehicle detection, EURASIP Journal on Advances in Signal Processing 2008 (2002) 1-12.
[24] D. Withopf, B. Jahne, Learning algorithm for real-time vehicle tracking, IEEE Conference on Intelligent Transportation Systems, 2006, pp. 516–521.
[25] S. C. Chen, Spatiotemporal vehicle tracking: the use of unsupervised learning-based segmentation and object tracking, IEEE Robotics and Automation Magazine 12(1) (2005) 50–58.
[26] T. Nguyen, H. Grabner, B. Gruber, On-line boosting for car detection from aerial, IEEE Conference on Research, Innovation and Vision for the Future, 2007, pp. 87–95.
[27] W. C. Chang, C. W. Cho, Online boosting for vehicle detection, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 40(3) (2010) 892-902.
[28] S. Sivaraman, M. M. Trivedi, A general active-learning framework for on-road vehicle recognition and tracking, IEEE Transactions on Intelligent Transportation Systems 11 (2) (2010) 267–276.
[29] C. W. Hsu, C. J. Lin, A comparison of methods for multi-class support vector machines, IEEE Transactions on Neural Networks 13 (2) (2002) 415-425.
[30] C. C. Chang, C. J. Lin, LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (3) (2011) 1–39.
[31] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.
[32] H. Grabner, H. Bischof, On-line boosting and vision, IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 260–267.
[33] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91-110.

Zhiming Qian received a Master Degree in Engineering from Yunnan University, Kunming, China, in 2009. He is now an associate professor at Chuxiong Normal University, Chuxiong, China. His current areas of interest include image processing and pattern recognition.

Hongxing Shi received a Master Degree in Science from Yunnan University, Kunming, China, in 2007. He is now an associate professor at Chuxiong Normal University.

Jiakuan Yang received the B.E. degree from Yunnan Normal University, Kunming, China, in 2003. Now he is a lecturer at Chuxiong Normal University.

Lianxin Duan received the B.E. degree from Yunnan Normal University, Kunming, China, in 1996. Now he is an experimentalist at Chuxiong Normal University.

Research and realization of Resource Cloud Encapsulation in Cloud Manufacturing

Zhang Ming[1], Hu Chunyang[2]
1 Department of Teaching and Practicing, Guilin University of Electronic Technology
Guilin, 541004, China
2 School of Mechanical&Electrical Engineering, Guilin University of Electronic Technology
Guilin, 541004, China

Abstract

Resource Cloud Encapsulation (RCE) is one of the key technologies of recourse accessing in Cloud Manufacturing (CM) and the foundation of resource virtualization. Under the analysis of the technologies of resource virtualization in Cloud Manufacturing, a framework of RCE is proposed. With the definition for each layer of the framework, realization of the resource accessing can be achieved. A demonstration with the physical manufacturing resource of stepper motors is enumerated to demonstrate the resource virtual accessing in CM. With the achievement of this article, a prototype of the cloud service platform is developed.

Keywords: *Cloud Manufacturing, resource accessing, Resource Cloud Encapsulation*

1. Introduction

The concept of Cloud Manufacturing (CM) is proposed by LI Bo-Hu who is an academician of Chinese Academy of Engineering. He defines the CM as a new mode of network intelligent manufacturing which is service-oriented, high efficiency and low consumption and knowledge-based, and the development of the concept of Cloud Computing (CC) in manufacturing. CM is becoming the latest network manufacturing mode, with many advantages such as quickly respond to market demand, enhanced competitiveness and collaborative manufacturing[1-3].

Resource Cloud Encapsulation (RCE) which is one of the key technologies of CM based on resource virtualization technology constructs large-scale of virtual manufacturing resource pool, and provides the best service resource meeting the task requirement through service matching. RCE can be used for the interacting and feedback control of manufacturing resources including hardware resources and software resources. It is based on Internet of Things (IoT), computer technology, Computer Virtualization and other technologies. RCE also can reduce the coupling between physical resource and manufacturing application by the transferring from physical resources into logical resources and virtual CM service with high utilization, high agility, high security and high reliability[6].

2. Resource Cloud Encapsulation (RCE)

2.1 A framework of the resource virtualization

A framework of the RCE in CM is proposed as shown in Fig. 1, after analyzing the characteristics of the CM. Resource virtualization and resource service-oriented technology is adopted to specify the forms of manufacturing resource in this framework. Each layer of the framework is relatively independent. Every layer provides services to its upper layer and the upper layer only. Every layer can access to the services of its lower layer through the interface of the lower layer.

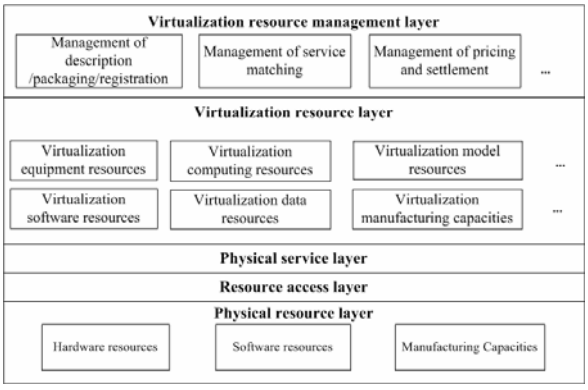


Fig. 1 Framework of the resource virtualization

- Physical resource layer

The bottom layer of RCE framework is a physical resource layer including all kinds of heterogeneous manufacturing resources with various distributions. Each manufacturing resource has its own characteristics

including heterogeneity, resource autonomy, wide distribution, etc.

Generalized manufacturing resources include hardware resources, software resources and resources with manufacturing capacities, etc. Hardware resources include manufacturing equipments and physical resources, such as all kinds of CNC machine tools, simulation equipments and material resources. Software resources include software, database resources and model resources. Manufacturing capacities include experimental capacities, integration capacities, etc. The CM system ensures the autonomy of scattered manufacturing resources to meet the requirements of costumers by the dynamic integration and the sharing of resources. For this purpose, standard protocol is required for the description and packaging of the manufacturing resources.

- Resource access layer

The main role of the resource access layer is the cloud encapsulation of the physical resources, service packaging, etc. This layer is based on Internet of Things (IoT), embedded technology and other technologies to establish the recognition and connecting between manufacturing resources.

- Physical service layer

The service generated by the description and packaging of the manufacturing resources is called physical service. Physical service is web accessing software components that are physically deployed by mapping the physical manufacturing resources to service. Defined interface which is described in WSDL can be called by the customers to access the physical service. Different management strategies are adopted in different physical service management domain. Physical service can be called by customers, applications and other services.

- Virtualization resource layer

Virtualization resource layer known as virtualization resource cloud pool is defined as resource dispatching and task management. Customers can choose and reorganize virtualization resources to meet their task requirements. Meanwhile, virtualization resource is bound to different physical services to meet the flexible requirements of customers.

- Virtualization resource management layer

The virtualization resource management includes description, packaging, registration management and service matching of virtualization resources, pricing and settlement, etc.

2.2 Resource access of RCE

RCE of software and hardware manufacturing resources and resource sharing are key technologies of the resource virtualization in CM. In the framework of RCE, IoT, Cyber Physical Systems (CPS), embedded technology and other technologies are adopted to achieve the accessing and connecting of manufacturing resources. In RCE, physical manufacturing resources are transferred into logical service to reduce the coupling the physical manufacturing resources and manufacturing applications. Therefore virtualized services with high utilization, high agility, high security and high reliability are published. The manufacturing resource adapter based on WSRF which packaging manufacturing resources into manufacturing services with unified structure and standard calling protocol is presented by Wu Lei in the paper of Research on Resource Virtualization in Manufacturing Grid[7].

To achieve the accessing and controlling of the physical resources in CM system, advanced computer information technologies are adopted in resource accessing. With the development of IoT and CC, reliable technical support is provided for the accessing into the CM system of physical manufacturing resources.

2.3 Demonstration model

For the Manufacturing Resource as a Service (MRaaS) mode, Cloud Service Provider (CSP) provides bottom layer resource to the Cloud Service Customer (CSC). Standard interface is called by CSC for the accessing of resources. In this paper stepper motor of 28BYJ-48 is selected as physical manufacturing resources. In the demonstration system, the control circuit is designed to access the control of physical resources. With a software interface designing of the physical resources, remote accessing by CSC can be done.

To demonstrate resource accessing, stepper motors, S3C2440 development board, and host computers are required in the demonstration. The control signals between host computer and client computer are transferred through Serial Port. This verification system

based on embedded system provides Web Service through embedded platform. Control signals from the remote computer client based on UDP protocol is used in the control of stepper motors such as forward rotation, reverse rotation, accelerating and decelerating.

- Resource description

For the MRaaS mode, physical resource is bound to a service which provides the creation and maintenance of the resource through the interface of the physical resource. In this paper, stepper motors are described in XML schemas and its attributes is presented in XML.

```
<?xml version="1.0"?>
< ManufacturingResource>
<ResourceID>00001</ ResourceID >
<Name>28BYJ-48 Stepper Motor </Name>
<Type>Stepper Motor</Type>
<MinTorque>300.00</MinTorque>
<MinFrequency>150.00</MinFrequency>
…
</ ManufacturingResource >
```

- Resource registration

Resource registration is performed on the web page of the CM service resource node to package resource information into XML schemas. The steps of registration are listed as : (1) Administrator of the physical resources input detail information of the resources such as names, attributes and parameters in the web page which will call the application named as “AddProductLine” to submit the information. (2) The submitting process will call the “Add” method of the “MgResourceManager” service which is assigned to manage the CM service resources. (3) “Add” method creates a service resource of “MgResource” with its attributes in XML.

- Service resource matching

After resources are registered in the CM system by the resource provider, the resource will be indexed and recognized by the system. When CSC submits a manufacturing task on the CM system website, the task will be analyzed into manufacturing parameters such as maximum processing size, accuracy and so on by the JavaBeans of the system. Matching calculation is the procedure of searching for the manufacturing resource matching those parameters. After that, the task will be assigned to the task list of the best resource by adding the index number of the task into the task list of the resource.

- Resource packaging

(1) The definition of service interface

The service interface is also called PortType which describes the Web service in Web Services Description Language (WSDL). In the WSDL, “wsrp: ResourceProperties” attribute is used to describe the resource attributes in the PortType element. Resource attributes must be defined in the type attributes of the WSDL, which is listed as follows.

```
<portType name="mgcommandPortType"
wsdlpp:extends="wsrpw:GetResourceProperty"
wsrp:ResourceProperties=
"tns:CommandResourceProperties">
<operation name="command">
<input message="tns:CommandInputMessage"/>
<output message="tns:CommandOutputMessage"/>
</operation>
<operation name="add">
<input message="tns:AddInputMessage"/>
<output message="tns:AddOutputMessage"/>
</operation>
<operation name="getValueRP">
<input message="tns:GetValueRPInputMessage"/>
<output message="tns:GetValueRPOutputMessage"/>
</operation>
</portType>
```

(2) Web service programming in Java

The Web service is designed to create a resource in the system resource node and modify the attributes of the resource such as resource status (working, repairing, scrapped, etc.). The method named as “CreateResource” is called to create a resource through the URI of the Web service with the return of Endpoint Reference (EPR) through which the modification of the resource can be made. With the EPR of the resource, attributes and methods of the resource can be accessed without creating the resource.

(3) WSDD Configuration

The Web Service will be published with the file called publishing description file for the accessing to the CM service. The publishing description file with Web Service Deployment Descriptor (WSDD) format is listed as

follows.

```
<?xml version="1.0" encoding="UTF-8" ?>
<deployment name="defaultServerConfig"
xmlns=http://xml.apache.org/axis/wsdd/
xmlns:java=http://xml.apache.org/axis/wsdd/providers/ja
va xmlns:xsd="http://www.w3.org/2001/XMLSChema">
<service name="mgrid/mgcommand" provider=
"Handler" use="literal" style="document">
<parameter name="className" value=
"mgrid.mgcommand.impl.commandService" />
<wsdlFile>share/SChema/mgrid/mgcommand/mgcomm
and_service.wsdl</wsdlFile>
<parameter name="allowedMethods" value="*" />
<parameter name="handlerClass" value=
"org.globus.axis.providers.RPCProvider" />
<parameter name="SCope" value="Application" />
<parameter name="providers" value="GetRPProvider"
/>
<parameter name="loadOnStartup" value="true" />
</service>
</deployment>
```

3. The CM prototype system

The prototype system includes three roles of users: CM resource providers, CM customers, CM service providers.

(1) CM resource providers

The modules designed for this role include the installation of the system, deployment and configuration, resource designing template, resource virtualization, resource packaging and registration, service deploying, servicing strategy designing and modifying, Portal developing, accessing controlling, etc. The CM resource provider can achieve the missions such as resource rent providing, registration, declaring and monitoring as shown in Fig.2 and Fig. 3.

Fig. 2 Interface of rental reservation

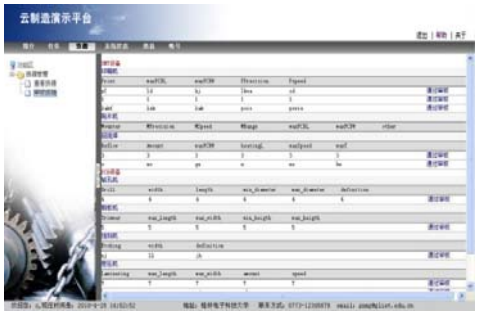


Fig. 3 Confirming interfaces of the resource provider

(2) CM customers

The modules designed for this role include user registration and canceling, Portal, task description, resource search, resource dispatching, task monitoring, service level negotiation (client side), service reserving (client side), notification mechanisms, etc. CM customers can achieve the missions such as task submitting, service level negotiation, task monitoring, resource overview as shown in Fig. 4 and Fig. 5.

Fig. 4 Interface of new task submitting

Fig. 5 Interface of resource listing

(3) CM service providers

The modules designed for this role include renting negotiation, resource registration and canceling, Portal, resource description and description templates, service level negotiation (server side), service reserving (server side) and security management as shown Fig. 6 and Fig. 7.

4. Conclusions and future work

CM is the latest network manufacturing and revolution

Fig. 6 Interface of resource reservation

Fig. 7 Interface of resource template

of manufacturing mode. RCE provides access technology for physical resource into CM system, which is one of the key technologies of CM. After proposing a framework of the RCE, the realization is demonstrated. With our achievements, the prototype system of CM has been constructed. In the future, the virtualization standard of manufacturing resources will be studied.

5. References

[1] L. Zhang, Y. Luo, F. Tao, L. Ren and H. Guo, "Key technologies for the construction of manufacturing cloud", Computer Integrated Manufacturing Systems, Vol. 16, No. 11, 2010, pp. 2510-2520.

[2] H. Ch. Yang, "Cloud manufacturing is a manufacturing service", China Manufacturing Information, No.2, 2010, pp. 22-23.

[3] X. J. GU, J. X. Chen, Y. J. Ji, G. N. Qi and W. Peng, “Group Technology in Cloud Manufacturing”, Group Technology & Production Modernization, Vol. 27, No. 3, 2010, pp. 1-4.

[4] B. Li, B. L. Zhang and S. Wang, "Cloud Manufacturing: a new service-oriented manufacturing mode", Computer Integrated Manufacturing Systems, Vol. 16, No. 1, 2010, pp. 1-8.

[5] Quack, T., H. Bay and L. Gool, “Object Recognition for the Internet of Things”, In Internet of Things 2008, 2008, pp.230-246.

[6] L. Ren, L. Zhang, Y. Zhang, "Resource virtualization in cloud manufacturing," Computer Integrated Manufacturing Systems, Vol. 17, No.3, 2011, pp. 511-519.

[7] Lei Wu, “Research on Resource Virtualization in Manufacturing Grid”, Ph.D. thesis, School of Computer Science and Technology, Shandong University, Ji Nan, China, 2008.

[8] W. Q. Jia, Y. X. Feng, J. R. Tan, X. H. An. Source: Proceedings of the 2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2012, 2012, pp: 757-762.

[9] C. Q. Li, Y. L. Shang and C. Y. Hu, “Research of Structure and Key Technologies for Cloud Manufacturing”, Modular Machine Tool & Automatic Manufacturing Technique, No. 7, 2011, pp. 104-107.

First author: Zhang Ming received Bachelor of Mechanical&Electrical Engineering Technology degree in 2001 and Master degree in Signal Processing in 2006 from Guilin University of Electronic Technology, Guilin, China. He has been working as a faculty of Guilin University of Electronic Technology since 2001. His research interests have been in the area of advanced manufacturing technology.

Second author: Hu Chunyang received Master degree in Mechanical&Electrical Engineering Technology in 2012 from Guilin University of Electronic Technology, Guilin, China.

Color Averaging Technique using Dominant Color for Content Based Image Retrieval

Prabhakar Sharma1, Deepty Dubey2

1 Department of Computer Science and Engineering
Chhatrapati Shivaji Institute of Technology, Durg, C.G., India,

2 Department of Computer Science and Engineering
Chhatrapati Shivaji Institute of Technology, Durg, C.G., India,

Abstract

In many areas of commerce, government, academia, and hospitals, large collections of digital images are being created. Many of these collections are the product of digitization of existing analogue photographs, diagram, drawings, paintings, and prints. Usually these collections were merely searched by means of textual annotations. This huge digital image database causes the need to develop an efficient image retrieval system. Content based image retrieval CBIR has been one of the most important research area in the field of computer science. There were many CBIR techniques have been proposed in the last decade [1]. This paper provides a method for content based image retrieval using only color aspect. In this paper we regard as the dominant pixel intensity values to shrink the feature vector and performed similarity measure between pixel intensities present in both query image and database image for same positions.

Keywords: *CBIR, feature vector*

1. Introduction

In the past access to collection of digital images were provided by librarians, curator and archivists through the manual assignment of textual descriptor and classification code. Automatic assignment of text attributes to images was developed by utilizing captions and transcripts later. Text based image retrieval (TBIR) makes use of text descriptors to retrieve relevant images. Past research shows that some of the useful text descriptor such as time, location, event objects, and aboutness of image content and topical terms are most helpful to users. The advantage of this approach was that it enabled widely approved text information system to be used for visual retrieval systems. However manual assignment is time consuming and costly while automatic assignment may not be possible if the image collections do not have accompanied text [2]. In literature the term content based image retrieval (CBIR) has been used for the first time by Kato et.al. [3], to describe his experiments into automatic retrieval of images from a database by color and shape feature [4].CBIR is an exciting and in-depth area of research, which has garnered much interest over the past few years [5]. Application of World Wide Web (www) and the internet is increasing exponentially, and with it the amount of digital image data accessible to the users. A huge amount of image databases are added every minute and so is the need for effective and efficient image retrieval systems [6]. The relevance of visual information retrieval in many areas such as fashion and design, crime prevention, medicine, law, and science makes this research field one of the important and fastest growing in information technology. Image retrieval has come a long way where it started off with text based retrieval. However, there are many problems associated with retrieving images based on text such as manual annotation of keywords, differences in perception and interpretations, and few others. Due to this researchers came up with CBIR where images are retrieved based on low-level features (human vision related), middle-level features (objects related), or low-level features (semantic related). Among these features low-level features are the most popular due to its simplicity compared to other level of features plus automatic object recognition and classification is still among most difficult problems in image understanding and computer vision [5]. The low-level features are color, texture, shape, and spatial properties. However spatial properties are implicitly taken into account so the main features to

investigate are color, texture and shape. Color feature is one of the most widely used features in low level feature [7]. Compared with shape feature and texture feature, color feature shows better stability and is more insensitive to the rotation and zoom of image. Color not only adds beauty to objects but also more information [8]. Texture generally refers to the presence of a spatial pattern that has some properties of homogeneity. Directional features are extracted to capture image texture information. The four extraction methods available for texture feature retrieval are The Steerable Pyramid; The Contour let Transform, The Gabor wavelet Transform, and The Complex Directional Filter Bank [9].

2. Previous Work

A method proposed by Dr. H.B.Kekre, Sudeep D. Thepade, Akshay Maloo all image pixels are considered as feature vector and Euclidean distance is used in RGB plane to find the best match, which is used to calculate precision and recall. Another method proposed by Dr. H.B.Kekre, Sudeep D. Thepade, and Akshay Maloo row mean of image is calculated to be feature vector and then Euclidean distance is used in RGB plane to find the best match, which is used to calculate precision and recall. In column method proposed by Dr. H.B.Kekre, Sudeep D. Thepade, and Akshay Maloo feature vector is composed of column mean of image is calculated and then Euclidean distance is used in RGB plane to find the best match, which is used to calculate precision and recall. In row and column mean of image method proposed by Dr. H.B.Kekre, Sudeep D. Thepade, Akshay Maloo row and column mean of image are considered together as feature vector and then Euclidean distance is used in RGB plane to find the best match, which is used to calculate precision and recall. In forward diagonal method, backward diagonal mean of image is considered as feature vector and then Euclidean distance is used in RGB plane to find the best match, which is used to calculate precision and recall. In forward diagonal method, forward diagonal mean of image is calculated and then Euclidean distance is used in RGB plane to find the best match, which is used to calculate precision and recall. Both methods were proposed by Dr. H.B.Kekre, Sudeep D. Thepade, and Akshay Maloo. They also purposed another method in which both forward and backward diagonal mean both are considered together as feature vector of image and then Euclidean distance is used in RGB plane to find the best match, which is used to calculate precision and recall[4][10].

3. Proposed Method

The concept behind this proposed method is that if two images are very relevant then in both images at same positions there will be same or very nearest intensity values will present. In this proposed work, we consider the dominant pixel intensity values to reduce the feature vector and performed similarity measure between pixel intensities present in both query image and database image for same positions. We proposed a modified color averaging technique in which we consider the positions of pixel's intensity values present both in query image and database image. In this proposed method we take the average of intensity values present in corresponding location of database image for respective intensity values present in query image. If the average from database image and intensity present in query image are same or near to same, then query image and database image will be similar to each other. In this method we have to first find out the intensity values present in query image with their respective positions in query image. In the next Step for each intensity values present in query image we calculate average of intensity values present in corresponding positions in database image. To perform this calculation it is required to have both images should be of same size. The steps of algorithms are as follows.

Step1: Find out the dominant intensity values present in query image. To find out dominant intensity values present in query image, we determine the frequency of occurrences of all intensity values present in query image and considering only those intensity values whose frequency of occurrences is very high in query image.

Step2: **Find out the position of dominant intensity values present in query image.** Because digital images are represented in the form of matrix, it is difficult to find the location of any intensity values present in image. To resolve this problem we need to represent the image into vector form. In this representation we can easily find the location of an intensity value present in query image.

$$\begin{pmatrix} a_{11}\ a_{12}\ldots\ldots\ldots\ldots a_{1n} \\ \\ \\ a_{m1}a_{m2}\ldots\ldots\ldots\ldots a_{mn} \end{pmatrix} = \begin{pmatrix} a_{11}\ a_{21}\ldots a_{m1}\ldots a_{1n}\ldots a_{mn} \end{pmatrix}$$

(a)Digital image in Matrix form (b) Digital image in Vector form

Fig. 1 Digital image in matrix form and Digital image in vector form

Fig. 1 shows the matrix and vector representation of digital image, in which a_{11}, a_{12}….a_{1n}…a_{m1}, a_{m2}…a_{mn} are intensity value of pixels present in digital image.

Step 3: Find the average of intensity values present in database image for corresponding positions of an intensity value in query image. To perform this operation, it is require representing the database image in vector form. In this way we can easily locate the positions of intensity vales found in query image into database image. But before performing this operation it is necessary to make the database image size same as query image size. For example

1	2	3
1	2	1
3	1	2

(a) Query image

1	3	2
2	1	1
3	1	2

(b) Database Image

Fig. 2 Sample of Query image and Database image

For intensity value '1' in query image positions are 1,2,6,8.For this corresponding positions in database image, average of intensities will be

$$\frac{1+2+1+1}{4} = \frac{5}{4} = 1.25$$

Step 4: Perform the similarity measure between query image and database image by taking difference of an intensity value in query image and average intensity for that intensity in database image. This similarity measure is performed using Euclidean Distance formula

$$D = \sqrt{\sum_{i=1}^{N} (V_{pi}-V_{qi})^2} \qquad (1)$$

Equation (1), calculates the Euclidian distance between two pixel values, where, Vpi and Vqi are the feature vectors of image P and Query image Q respectively with size 'N'.

Step 5: Retrieve those images for which Euclidean distance is minimum.

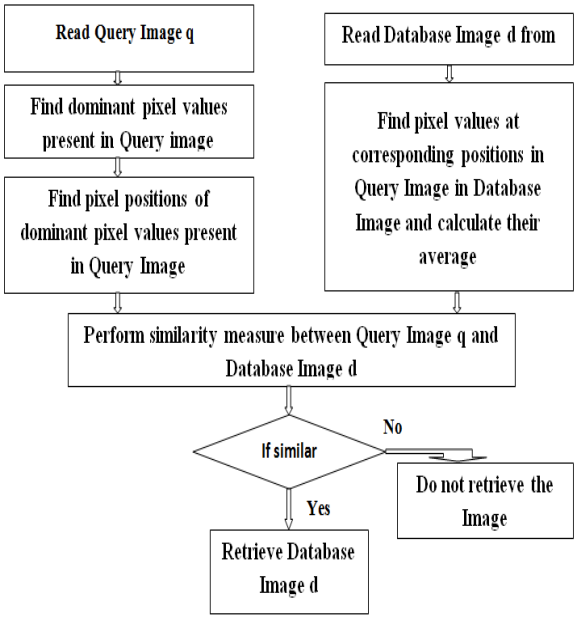


Fig. 3 Flow Diagram of proposed Color averaging technique

4. Result and Discussion

We have used Wang Database for image retrieval. The WANG database is a subset of 1,000 images of the Corel stock photo database which have been manually selected and which form 10 classes of 100 images each.

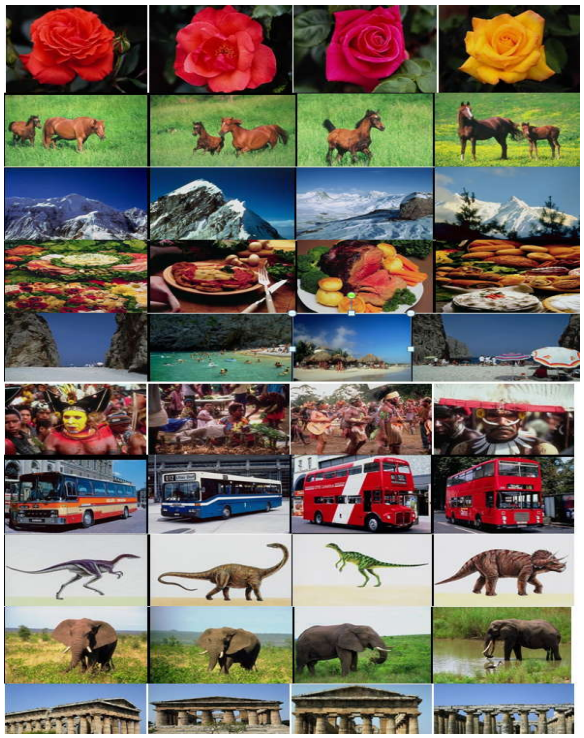
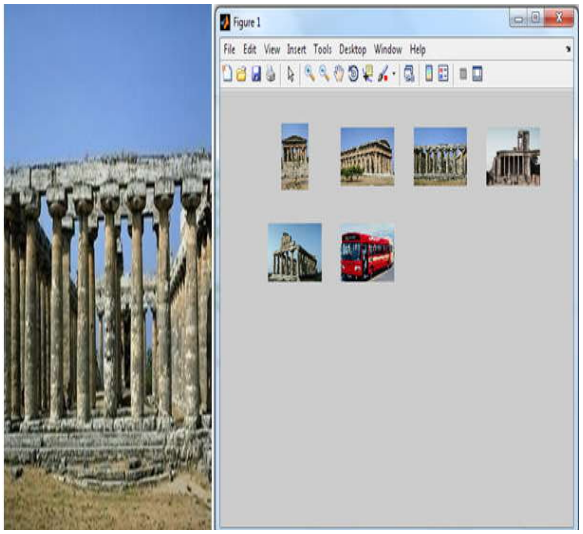


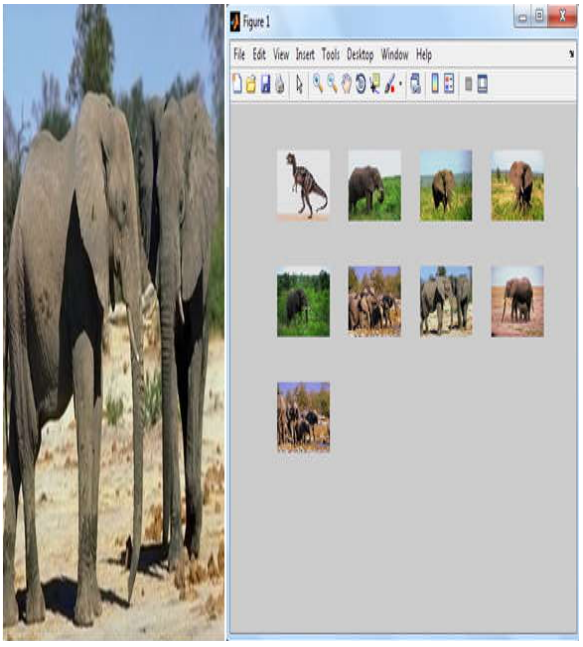
Fig. 4 Database images

The performance evaluation of retrieval system is measured by means of precision and recall values. The precision is defined as number of relevant images retrieved to total number of images retrieved; whereas recall is defined as number of relevant images retrieved to total no of relevant images in database [1.4, 10]. For each category 10 queries were fired. The results are as follows.



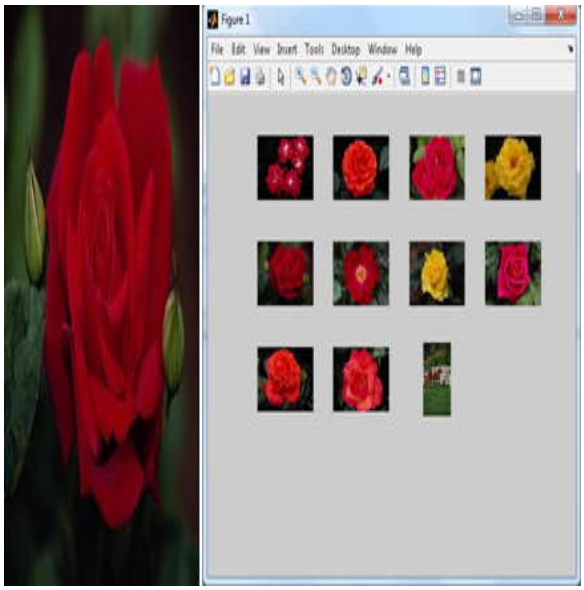
Input image Output image

Fig.5 output for category "Architect"



Input image Output image

Fig.6 output for category "Elephant"



Input image Output image

Fig.7 output for category "Flower"

Table 1: Average precision and recall values category wise

Sr. No.	Category	Average Precision	Average Recall
1	Tribe	92.5	24
2	Architect	78.4848	35.5372
3	Bus	82.9167	28.6364
4	Dinosaur	100	100
5	Elephant	72.0545	58
6	Flower	98.1818	99
7	Horse	71.4535	41
8	Mountain	74.2738	49
9	Meals	62.6667	25
10	Beach	81.1825	48

Table 2: overall precision and recall value of proposed CBIR system

Average precision value	81.2438
Average recall value	52.58326

5. Conclusion

The experimental outcome shows that precision value of proposed CBIR system is **81.2438** and recall value is **52.58326**. From result it has concluded that precision is high and recall is low for proposed CBIR system, which means that the system is able to find good match in concerned class but in a small amount of numbers only. The high precision values indicates that when two similar images of same class consist of small number of different color intensities and very few colors are dominant then a good match found but when two similar images of same category contains blend of large number of different color intensities and every color intensity values are present with approximately same population then it is complicated to find a good match thereby recall value decreases.

References

[1] Prabhakar Sharma, Deepty Dubey," A Modified Color Averaging Technique for Content based Image Retrieval" International Journal of Computer Applications (0975 – 8887) Volume 51– No.20, August 2012.

[2] Christopher C. Yang," Content –based image retrieval:a comparison between query by example and image browsing map approaches", Journal of information Science, 30(3) 2004 pp. 254-267

[3] Hirata K. and Kato T. "Query by visual example – content-based image retrieval", In Proc. of Third International Conference on Extending Database Technology, EDBT'92, 1992, pp 56-71

[4] Dr. H.B.Kekre1, Sudeep D. Thepade2, Akshay Maloo3," Query by Image Content Using Color Averaging Techniques", Dr. H.B. Kekre et al. / International Journal of Engineering Science and Technology.Vol. 2(6), 2010, 1612-1622.

[5] Mas Rina Mustaffa, Fatimah Ahmad, Rahmita Wirza O.K. Rahmat, Ramlan Mahmod." Content-based image retrieval based on color-spatial features" Malaysian Journal of Computer Science, Vol. 21(1), 2008.

[6] P. V. N. Reddy, K. Satya Prasad, " Color and Texture Features for Content Based Image Retrieval", P V N Reddy et al, Int. J. Comp. Tech. Appl., Vol 2 (4), 1016-1020.

[7] Th.Gevers (2001). "Color Based Image Retrieval". *Springer Verlag GmbH*. pp.886-917

[8] Manesh Kokare, B.N. Chatterji and P.K. Biswas, "A survey on current content based image retrieval methods", IETE Journal of Research, Vol. 48, No.3 and 4, May-Aug 2002.

[9] H.B.Kekre, Sudeep D. Thepade, "Rendering Futuristic Image Retrieval System", National Conference on Enhancements in Computer, Communication and Information Technology, EC2IT-2009, 20-21 Mar 2009, K.J.Somaiya College of Engineering, Vidyavihar, Mumbai-77.

[10] Dr. H.B.Kekre, Sudeep D. Thepade, Akshay Maloo," Image Tiling to Improve Performance of Image Retrieval Using Color Averaging Techniques", IJCA Special Issue on "Computer Aided Soft Computing Techniques for Imaging and Biomedical Applications" CASCT, 2010

Twitter Assisted Team Based Learning: Providing a new way of communication in classroom

Sami M. Alhomod [1] and Mohd Mudasir Shafi[2]

[1] King Saud University,
Riyadh 11321, PO Box 231201, Kingdom Of Saudi Arabia

[2] King Saud University,
Riyadh 11321, PO Box 231201, Kingdom Of Saudi Arabia

Abstract

This paper explores the use of twitter in a team based learning scenario. Twitter has been recently used by many educational institutions but most of these have been related to provide the information to the general audience. There has not been much study done to propose twitter as an educational in a classroom scenario. This paper tries to establish the use of twitter with a well establish mode of Team based learning. The paper also demonstrates the use of twitter in student teams as well as individually by the students. The paper also explains teacher – team and teacher – student communication via twitter.

Keywords: *Social Networking; Twitter; Team Based Learning; Communication.*

1. Introduction

Team based learning (TBL) is based on the use of small groups in order to transform them into high performance teams to accomplish complex tasks. According to Fink [2], "Team based learning is a particular instructional strategy that is designed to (a) Support the development of high performance learning teams. (b) Provide opportunities' for these teams to engage in significant learning tasks". There have been a lot of studies that prove that team based learning and teaching have extremely effective to achieve wide range of goals. TBL promotes higher level reasoning, enhances content retention and learning and increases the social support in the classroom. TBL offers an opportunity for an average student to put more effort and enables teams to accomplish tasks which could not have been done by even the excellent students individually [2, 23].

Social networking site like twitter has gained extreme popularity among the internet user over the past few years. These sites were intended for personal communication among individuals but now increasing number of Organizations are using these sites to engage their stakeholders [13]. Twitter is one such site which has seen huge growth since its launch. Twitter offers a means of informal communication among its users [15]. A lot of studies have been conducted recently on the use of twitter in educational sector. These studies have established that twitter can act as a tool of communication in the modern educational system. Today more and more educational institutions are experimenting with twitter as a tool in education.

This paper examines the use of twitter in a Team based learning system and points out the benefits of using twitter in a TBL system. The paper is organized as follows; first we will provide a brief background of TBL and Twitter. Second, we will discuss the related work done in this regard. Next, we will describe how twitter can assist TBL. And at last, we will provide a conclusion to the study.

2. Background

2.1 Team based learning

The term "Team based learning" was first coined by during the 1970's.Team based learning(TBL) in education is a technique in which students work together in teams in order to learn things with better understanding. TBL transforms the traditional lecture based coursework into a more active self-learning and promotes teamwork. It allows students to achieve the levels of higher quality learning which can be hard to achieve when students are working individually [1]. TBL involves making small groups of students and using these groups as instructional strategy. TBL links each learning activity to the next activity in order to achieve deeper understanding among

students and develop the teams of higher performance and understanding [2]. As an example, students can be asked to work in teams so that they can cover a more learning material without having to exert excessive pressure individually [3]. According to Michaelsen, Knight and Fink, 2002, there are two specific purposes of TBL:

1. Form Teams of high learning performance.

2. Participate and gain experience in tasks of educational importance.

Another important factor of TBL is group cohesiveness. As the students start working in teams, the group cohesiveness increases which results in higher level of efficiency and understanding among students. Once a student group is formed there are four stages of transforming it into a team. First the students interact with each other. Second, the students review the resources that are available to them. Third students receive a task and work towards its completion. And at last, performance of each individual member of group is evaluated. Once these stages are completed, the group has transformed into the team [1, 4].

One of the important benefits of team based learning is that it helps students with developing skills. Possessing excellent teamwork skills is one of the important factors for employers in the job market [5]. According to a survey conducted by Wall Street Journal, a teamwork skill is the second most important skill for the business graduates to possess [6]. TBL allows students to organize the problems and devise a solution for each problem accordingly. TBL also allows students to interact with each other on a daily basis and enables students to complete tasks within teams [4].

2.2 Twitter

Twitter is today's most popular micro blogging site which gives free service to users. It was developed by Jack Dorsey and was launched in October, 2006 [7]. Twitter allows users to send and receive tweets from other users. Each message on twitter can be of up to 140 characters. These messages are called "tweets". These tweets are available on users profile page and can be viewed and replied to by people known as followers [10].As of 2011; twitter had 300 million users generating 300 million tweets per day [8]. People use twitter to communicate with each other on constant basis. People also use twitter to get help from other users like asking for directions, advice, support etc. twitter also allows sending message through external applications like smart phones or short message services (SMS).

Probably, one of the first studies on twitter was conducted by Java, Song, Finn & Tseng in 2006. The results of the study found that there are three main categories of twitter users:

1. Information sources: These people tweet news and have a large number of followers.

2. Friends: it's a broad category which covers a large number of users. Friends can include family members, coworkers as well as strangers

3. Information Seekers: There are the people who themselves tweet rarely but follow other's regularly.

Users use twitter for daily chat or to discuss events in their life. People also use twitter to share their thoughts as well as share information and URL's. People also use twitter to comment on current events and also on some news items [12].

3. Related Work

Twitter has been subject of research among many scholars recently. In some earlier work, researchers have mainly focused on the usage areas of twitter. Some researchers have also focused on the network areas of the twitter to find out the usage of twitter by people [7]. For example, Wigand [13] discussed the use of twitter by government organizations. The research explored the use of twitter and web 2.0 technologies by the government departments in US including the congress, NASA and U.S. Air Force. The research argued that twitter can be established as a medium for communicating with citizens and enhance collaboration in a way which might be hard to achieve with other platforms. Another study by Java et al [11] presented an overview of user intentions behind using twitter. The research also studied the geographical and topographical properties of twitter to conclude the user intention of using twitter as information sharing, information seeking and social activity [7, 11]. Honeycut & Herring [12] discussed the conversation and collaboration with respect to @ sign among the users of twitter. The study discovered high degree of conversation among the users. Zhao & Rosson [15] discussed the motivation behind using twitter. The research discussed the relational and personal benefits of using an informal mode of communication. All of these studies revealed that twitter can be used an information platform to send messages to the other users. Junco, Heiberger & Loken [14] Provided and experimental evidence to establish that twitter can be used as an educational tool to engage students and faculty into an active role and increase their participation the educational process. The research concluded that twitter increases student engagement and has a positive effect on student

grades. Grosseck & Holotescu [9] discussed the use of twitter for educational activities. The study presented the use of twitter in educational sector with respect to Romanian Twitosphere. The research presented the benefits, drawbacks and the logistics of using twitter as an educational tool. Goroshko & Samoilenko [17] presented the use of twitter as a platform for global academic academy to engage in a real time dialogue. The paper discussed the potential of twitter as an educational system in both e-learning 2.0 and e – learning3.0. The paper concluded the potential of twitter in global educational world where students and faculties communicate online in both virtual and real classrooms to achieve the idea of whole learning. Borau, Ullrich, Feng & Shen [18] used twitter tool for communication in shanghai Jiao Tong University where English is a foreign language. The results of study indicate that 70% students indicate that it's easy to communicate using twitter while as none of the students disagreed with this fact. Al-Khalifa [19] proposed the use of twitter to send updates to students on mobile phones. The researcher proposed three benefits of sending twitter updates on mobile phones as better connection with students, time saving and timely announcements. The study also found that 76% of students were satisfied with this service with 93% saying that they would prefer this mode of information for future courses

4. Twitter Assisted Team Based Learning

In TBL majority of class time is spent on activities so that students can learn to solve problems which they are likely to face in professional world. According to Michaelsen et al [20], there are three phases of any Team based Learning; preparation phase, application phase, assessment phase. We will discuss how twitter can assist and enhance each of these three phases [21].

4.1 Twitter Assisted preparation phase

In this phase students read the topics before they are discussed in the class. The main aim of this activity is to have a prior knowledge about the topic to be discussed in the class. This phase starts with individual preparation of the topic by each student of the group followed by discussing the topic in the group. The students first run the test individually followed by the same test in the group. Both the tests are graded in class and announced. And at last teacher offers understanding of the concepts that were not understood by the class. This marks the end of the preparation phase.

Twitter can be of extreme importance in this phase. Firstly, students can communicate with teacher as well as among themselves to know about the topics to be discussed in the class before they actually enter the class. The teacher can tweet the topic name on their twitter page and all the students can have a prior knowledge about the topic. Once the topic is provided, the students in the group can discuss topic among each other both prior to coming in the class and during the class. Once the test is conducted individually, the team test can be conducted over twitter with the specific question send to a particular team. Students in the group can communicate with each other by tweeting their thoughts on the question. This will enable students to work individually on the question and discuss it in the group. This will also help teacher to know about the students actively trying to solve the question and also about the students who are not contributing to solve the problem. This will enable teachers to identify the weaker student in the group and possibly put more effort towards the weaker students of the group. Once the tests are conducted individually as well as in groups, the teacher can declare results and provide solution to the question over twitter. This will allow each student to cross examine their result and check the weakness in their answer. Once the results are announced the students can discuss the results among themselves with teacher having an eye on the discussion. Using twitter in this way enables to continue the classroom activity even after the class. Once the class is over the students can continue discussing the question even after the class and can continue discussing the topic even at their home. The teacher can also participate in the discussion even after the school is over. The teacher can add to the topic at any time. The teacher can choose to inform few things about the topic so as to start the discussion and then gradually add to the topic over twitter. This will increase better reasoning among students and can help students to research more about the topic.

Twitter will also allow students an easier way to ask questions. If there is any point in the topic that a student don't understand or need more clarification, he can ask teacher or his team by raising the point on twitter at any time.

4.2 Twitter Assisted Application phase

In this phase students apply the knowledge of the course content they learned during the preparation phase to solve the problems, make predictions or create explanations for complex problems. Each group or team in this phase provides their responses to the problem in the class and the teacher evaluates the responses of each group to provide feedback to every group. At the end of this phase, students learn to work in team to provide solutions as well as form a strong bonding with the other students in the group.

Twitter can help in this phase by connecting students with each other as well as with teacher. The constant contact between students and the teacher create cohesion among

them which is important for student persistence [14]. Twitter can act as a constant medium for student groups and teacher interaction which is an important factor for the success of students [22].

The students in this phase can solve the problems and share their responses via twitter with other members of the group or to the whole class and teacher. The teacher can also form a network on twitter for a particular group or for the whole class. The teacher can post their responses on the twitter and can make it visible to a particular group or student. This will create a secret form of communication between teacher and the students. The teacher can choose to guide a particular group or student if they are not doing well as compared to the class. The student can post their arguments and question anytime on the twitter and the teacher can choose to respond to these questions at any time on twitter. This will promote out of class learning and students and enable anytime / anywhere learning.

4.3 Twitter Assisted Assessment phase

This is the Final phase of team based learning. In this phase teams are required to solve the problems based on the understanding of the course material [21]. This phase also allows students to use the previous studied material and incorporate it with the new material [1]. The responses from each team are evaluated by the teacher and the grades for each student and teams are decided.

Timely feedback is one of the fundamental principles of team based learning. It helps students in content retention and learning which in turn helps in student and team development [2]. Twitter can help in this notion of timely feedback. A teacher can provide feedback to student and teams on twitter as soon as he is done with the assessment. For student it provides an opportunity to readily assess their performance as well as the performance of the team. A team can also share their comments regarding their results and performance.

Besides helping teacher and teams to accomplish their tasks, Twitter can also help in the formation of groups. According to Junco, Heiberger & Loken [14], it's easy to organize students into groups via twitter. A teacher can ask students about their interesting subjects and thus can form groups based on students with similar interests. Twitter can also act as a debate starter. For example, a teacher can post some topic on twitter page and ask teams about their thoughts about the topic. This can encourage out of class learning. Teams can discuss the topic over the twitter before the topic is actually discussed in the class.

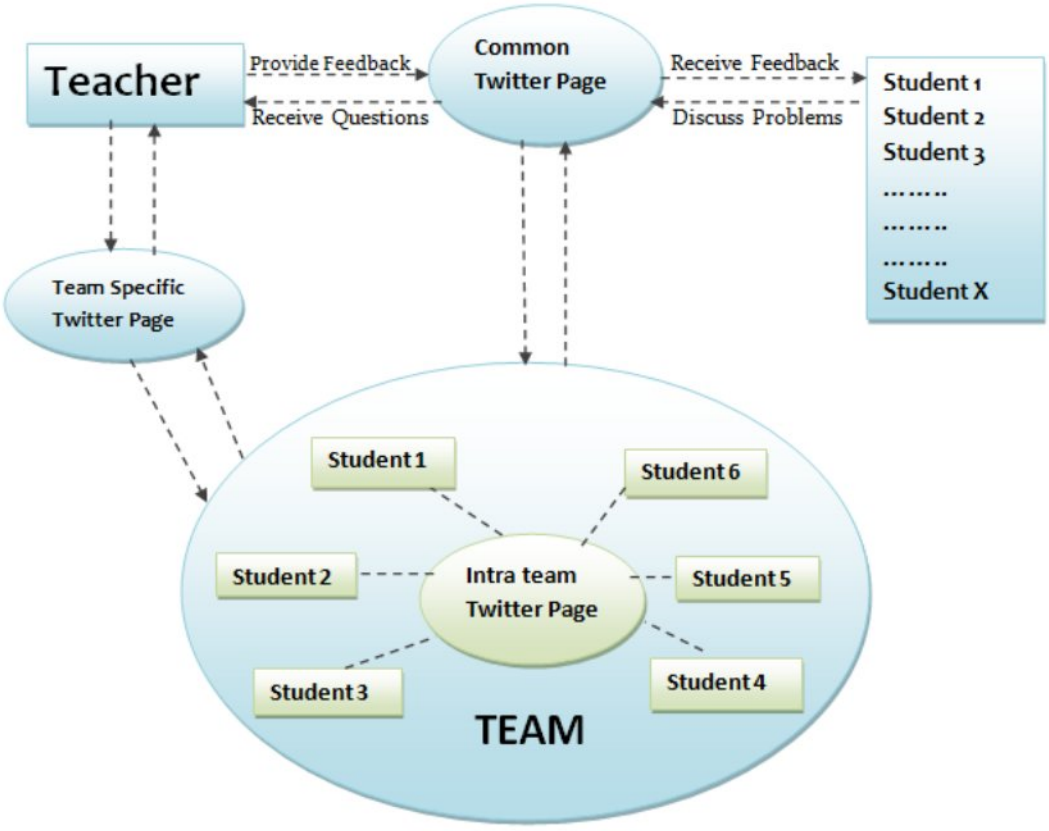


Fig. Twitter Assisted Learning

Figure above shows a general framework of using twitter in a classroom structure. A team can communicate with each other using twitter as well as communicate with teacher and other students in class via twitter. Teams can set up their intra team twitter page and discuss the problem assigned to the in team. Teams can remain in constant touch with each other even outside the class and can increase cohesiveness among the team which is important for the successful completion of the tasks. Students can also communicate individually with groups, teacher and other students via a common classroom twitter page. Teachers can disseminate classroom information on the common twitter page of the classroom. As far as teacher-team communication is concerned, this can be accomplished by forming a team specific page. Each team in the class can for a page which is connected only to their teacher. Teams can post their questions on this page and can receive answers from teacher on this page without the interference of out of team members. This can help to increase the teacher- team communication and can allow teacher to access the progress of each individual team. If the teacher has comments for a particular team, teacher can post his comments on the team specific page without other teams and students getting to know about the comments. This can encourage the teams to work hard without going through embarrassment in front of the class.

5. Conclusion and future work

It is established that TBL can enhance education in multiple ways. Recently twitter has been used by many educational institutions as a tool to enable student achieve the desired outcomes. More and more educational organizations are using twitter in one way or the other. Keeping this in mind, we tried to demonstrate how twitter can further enhance the widely accepted mode of learning i.e. TBL. We tried to establish the high level of inter student, intra team interaction, student – teacher and team – teacher interaction via twitter. This study demonstrated to use twitter at each phase of TBL as well as tried to demonstrate the use of twitter in a general based classroom scenario. The paper tried to suggest the ways in which twitter can be used in a TBL. As the use of twitter in educational institutions grows, we recommend measuring the impact of twitter in a TBL scenario. We also recommend measuring the impact of twitter with other learning scenario as well as explore new ways of integrating twitter into classroom structure.

References

[1] Michaelsen, L. K, Knight, A.B., & Fink, L.D. (Ed.). (2002). “Team-Based Learning: A Transformative Use of Small Groups”. Westport, CT: Praeger Publishers.

[2] Larry K. Michaelsen , Dean X. Parmelee , Kathryn K. McMahon, Ruth E. Levine, Diane M. Billings “Team-Based Learning for Health Professions Education: A Guide to Using Small Groups for Improving Learning”

[3] Lerner, L. D. (1995). “Making Student Groups Work”. Journal of Management Education, 19(1), 123-125.

[4] Allison Brittney Goo, 2011"Team-based Learning and Social Loafing in Higher Education" Online Available: http://trace.tennessee.edu/utk_chanhonoproj/1423/

[5] Chapman, K. J., Meuter, M., Toy, D., & Wright, L. (2006). “Can’t We Pick our Own Groups? The Influence of Group Selection Method on Group Dynamics and Outcomes”. Journal of Management Education, 30(4), 557-569.

[6] Alsop, R. (2004, September 22). “How to get hired”, Wall Street Journal. p.R8

[7] Zhiheng Xu, Rong Lu ,Liang Xiang ,Qing Yang 2011 "Discovering User Interest on Twitter with a Modified Author-Topic Model" International Conferences on Web Intelligence and Intelligent Agent Technology

[8] Taylor, Chris (June 27, 2011). "Social networking 'utopia' isn't coming". CNN. Retrieved December 14, 2011.

[9] Gabriela GROSSECK,Carmen HOLOTESCU 2008,”Can We Use Twitter For Educational Activities?”

[10] F. Dianne Lux Wigand "Twitter in Government: Building Relationships One Tweet at a Time” 2010 Seventh International Conference on Information Technology

[11] Java, A., Song, X., Finn, T., & Tseng, B. (2006, August).”Why we Twitter: Understanding microblogging usage and communities.” Joint 9th WEBKDD and 1st SNA-KDD Workshop ?07, San Jose, CA

[12] Courtenay Honeycut,Susan C. Herring "Beyond Microblogging: Conversation and Collaboration via Twitter" Proceedings of the 42nd Hawaii International Conference on System Sciences - 20

[13] F. Dianne Lux Wigand,2010 "Twitter in Government: Building Relationships One Tweet at a Time "Seventh International Conference on Information Technology

[14] R. Junco, G. Heiberger† & E. Loken, “The effect of Twitter on college student engagement and grades”

[15] D. Zhao and M. B. Rosson. “How and Why people Twitter: The role that microblogging plays in informal communication at work”. In Group, 2009.

[16] Meeyoung Cha,Hamed Haddadi,y Fabicio Benevenutoz, Krishna P. Gummadi 2010 "Measuring User In?uence in Twitter: The Million Follower Fallacy"

[17] Olena I. GOROSHKO Sergei A. SAMOILENKO "Twitter as a Conversation through e-Learning Context"

[18] Kerstin Borau, Carsten Ullrich, Jinjin Feng, and Ruimin Shen "Microblogging for Language Learning: Using Twitter to Train Communicative and Cultural Competence "ICWL 2009, LNCS 5686, pp. 78–87, 2009

[19] Al-Khalifa, H.S.:”Twitter in academia: a case study from Saudi Arabia”. eLearn. 2008(9), 1–1 (2008)

[20] Michaelsen, L. K., Knight, A. B., & Fink, L. D. (Eds.). (2004). “Team-based learning“ Sterling, VA: Stylus.

[21] Haberyan, April, 2007 "Team-based learning in an Industrial/Organizational Psychology course" North American Journal of Psychology Volume: 9 Source Issue: 1

[22]Kuh G.D. (2002) “The National Survey of Student Engagement: Conceptual Framework andOverview of

Psychometric Properties. Center for Postsecondary Research",Indiana university, Bloomington.
Availableat:http://nsse.iub.edu/pdf/psychometric_framework_2002.pdf
[23] McKeachie, W., (1999). "Teaching tips: Strategies, Research, and Theory for College and University Teacher (10th ed)". New York: Houghton Mifflin.
[24] Twitter, http://www.twitter.com

Dr. Sami M. A. Al Homod is a faculty member in the department of Management Information Systems; College of Business Administration in King Saud University. He worked as the Dean of the eLearning and Distance Learning Deanship. He received his PhD in Information Technology from George Mason University; School of Information Technology and Engineering. He received his Master of Science in Information Systems from George Mason University; Information System and Software Engineering Department. He got his B.S. in Computer Information Systems from King Saud University; College of Computer and Information Sciences. He was A Board member of the Saudi Computer Society. He got a Distinguish Degree in Information Systems from George Mason University and Honor Degree from King Saud University. He is a member of many scientific and administrative committees and attended many conferences and scientific seminars.

Mohd Mudasir Shafi was born and raised in Srinagar, Kashmir, India. He has received his Master of Science in Computer Science from Jamia Hamdard (Hamdard University), New Delhi, India in the year 2009. He is currently working as a Researcher in Deanship of distance and Electronic Learning at King Saud University, Kingdom of Saudi Arabia. He has published many research papers in National and International journals. He has also actively attended many international conferences. His areas of interests are in E governance, Mobile governance, Network Privacy and security, Software Engineering, Social Media and E Learning. Apart from that he has previously worked as Quality Analyst and software engineer at various MNC"s in India.

Theoretical Model of Software Process Improvement for CMM and CMMI based on QFD

Yonghui CAO[1,2]
1 School of Management, Zhejiang University
2 School of Economics & Management, Henan Institute of Science and Technology

Abstract

In this paper, we first introduce Software Process Improvement (SPI) and Quality Function Deployment (QFD); then study theoretical model of SPI for CMM and CMMI based on QFD. Through the research, we hope to achieve three goals: first, to develop a method, based on QFD, for the integration and prioritization of requirements from multiple perspectives; second, to map process requirements, including business requirements, to CMM or CMMI with the help of QFD; third, to be able to prioritize software process improvement actions based on process requirements. Finally, we also draw conclusions.

Keywords: *Software Process Improvement; CMM; CMMI; QFD*

1. Introduction

Software Process Improvement (SPI) has is the key to the survival of many software development organizations. Many international SPI models/standards are developed for SPI. The Capability Maturity Model (CMM) and Capability Maturity Model Integrated (CMMI) from the Software Engineering Institute are two SPI models. Like all the other standards and models on software process improvement, CMM and CMMI address the question of "what to do" while leaving "how to do it" to organizations. Therefore, some methodology is needed to transform CMM activities or CMMI Practices into a set of actions that are detailed enough to be followed by software engineers.

In this study, frameworks were developed to help map business and other process requirements of an organization to CMM and CMMI elements, and help develop action plans to satisfy those requirements using Quality Function Deployment (QFD).

QFD was first introduced in Japan by Dr. Yoji Akao in 1966. In 1972, Mitsubishi Heavy Industry put it in practice at Kobe Shipyards. Translating customers' requirements into product design requirements and relative production requirements is the most popular application of QFD. The house of quality is the focus of QFD. The customers' requirements are sometimes called customers' voice. The main idea is: quality of a product is defined by customers, not engineers. This kind of concept is mainstream of today's business practice. A frequent heard term is "injecting customers' voice into the product design". Customer requirements are often stated in non-technical or non-measurable terms. With QFD, these non-technical terms could be analyzed and converted into technical specifications. The structure of QFD is simple. The process of data analysis and converting is a complex and time-consuming one. This is often owing to the subjective nature of data itself and the potential complexities of the QFD charts.

In the traditional approach, sequential product design approach, some design defects will not be found until the final stages. To correct this kind of design defect, the design process has to start over from the early design stage. In QFD, the process requires a mufti-disciplinary team. With a mufti-disciplinary team, design defects that will result in costly prototyping and time re-design can be found and solved in the early stages of design.

QFD is not only a map for product design. It is also a map for quality improvement for current products. With the House of Quality, a design team could see how a company's product met customer requirements and what the market position of company's product regarding to "qualities" was. This will provide directions for market and quality improvement.

Currently, data mining is a hot issue. In today's computer era, every one is flooded by information. There is a great deal of information involved in designing a product. How to present correct information in the correct format becomes one of the key issues in product design. If information is presented in the wrong format, this could result in longer design time or even faulty design. QFD provides a good data-presenting format for product design. QFD is also a good format of data presentation for supporting other kinds of decision-making.

2 .SPI Framework Based on CMM Using QFD

CMM is used in the framework as the reference model because of its popularity in the industry. Although the support for CMM from SEI has discontinued and CMMI has been recommended since then, it takes time for many companies currently using CMM to switch to CMMI.

The framework is designed in such a way that the process requirements can be reflected through the proposed framework all the way down to the action plans. As a result, the priority value of each requirement is adjusted after the impacts from the other requirements are assessed.

The set of requirements with adjusted priorities are related to the key goals in CMM KPAs. The goals are prioritized based on those process requirements. Thus, the goals that achieve higher overall satisfaction of process requirements get higher importance. In order to achieve these goals, CMM has KPs categorized into five common features. Both the common features and the KPs contained in them can have different priorities. The priorities of the common features are determined by their natures in CMM. For instance, "Commitment to Perform" should be considered before "Verifying Implementation." The priorities of KPs in various common features, on the other hand, are determined by their correlations with KPA goals. Thus, the KPs in each common feature are prioritized separately based on the priorities of the goals. KPs that aim to achieve higher overall satisfaction of key goals receive higher importance values. Separate sets of action plans are derived from KPs in each of the common features. The actions that help to support more important KPs receive higher priorities.

As a result, the process requirements are reflected in KPA goals, KPs, and the actions. The actions both follow the process maturity standards in CMM and satisfy the process requirements. Those actions with higher importance values help to achieve higher process requirements satisfaction.

As illustrated in Figure 1, this framework starts with the elicitation and integration of requirements. In this phase, the requirements for the improvement of the organizational process are gathered from various branches/departments, including the business goals from the executive board. For instance, one of the business goals may state that "Our product should lead in the competition," or a software process requirement from the management level may be that "The employee productivity should be increased." Depending on which branches and departments they come from, these software process requirements are grouped into perspectives with each branch/department being a perspective.

In Figure 1, various perspectives are represented as Pl through Pn. Each perspective contains multiple requirements. The software process requirements in perspective 1 are represented as R1-1, R1-2, etc. These perspectives of software process requirements can then be prioritized based on their relative importance within the organization and integrated into one single set of requirements. In Figure1, these integrated requirements are represented as R1 through Rm, where m is the total number of software process requirements from all perspectives. The prioritization ensures that requirements from different perspectives are comparable with each other, and the integration reflects the correlations among requirements from different perspectives. The deliverable of this phase is a set of prioritized and integrated software process requirements, which serves as the input to the next phase.

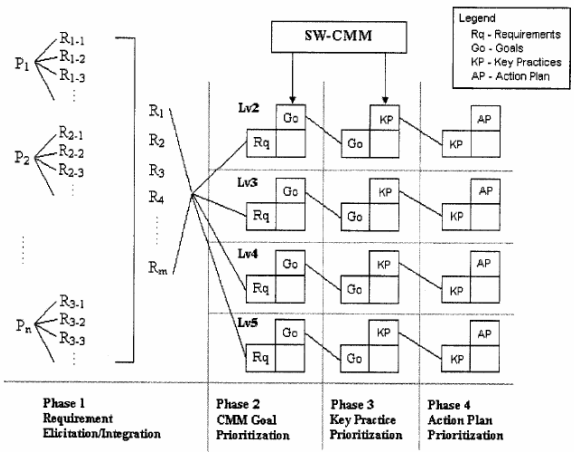


Fig. 1 Software Process Improvement through CMM Using QFD

The second through fourth phases of this framework are applied to Level 2 to Level 5 of the CMM model, The prioritized and integrated requirements from Phase 1 are linked to all KPA goals in each of the four levels in CMM using relationship matrices, These prioritized KPA goals are used as the basis for the prioritization of KPs. Finally, the prioritized KPs are transformed into prioritized action plans using House of Quality (HoQ).

In the second phase, which is "CMM goal prioritization," the goals of all KPAs in a particular CMM level are selected and prioritized based on the requirements from the previous phase. There are two objectives of this framework and this phase is significant in terms of achieving both. First, the organization needs to comply with the CMM standard. At the same time, the

organization needs to ensure that by reaching a particular maturity level, the process is also satisfying the business and other requirements within the organization. In Phase 2, a relationship matrix is used to establish connections between the requirements from the organization and KPA goals in CMM. This matrix demonstrates that complying with the CMM standard also helps satisfy the business and other requirements in the organization. Second, the final set of action plans needs to be prioritized based on the priorities of requirements so that more important actions receive more resources. KPA goals serve as the bridge between requirements and the action plan. By prioritizing KPA goals, requirements from the organization can be transformed to the KPs in the third phase, and finally to the action plans in the final phase. In this way, a set of actions can be executed not only to achieve a specific maturity level in CMM, but also to satisfy organizational process requirements.

The third phase of the proposed framework, which is "key practice prioritization," involves the prioritization of KPs within all KPAs of a specific level. The prioritization is carried out on the basis of the deliverables from Phase 2. According to CMM specifications, all these KPs have to be performed in order to reach that particular maturity level. However, these KPs serve as a bridge between the requirements and the final actions, and it is necessary to know how these KPs reflect the software process requirements. In order to show the connections between the requirements and the final action plans, these KPs have to be prioritized based on KPA goals, which are now reflecting requirements priorities. The mapping between KPA goals and KPs has been provided in Appendix E of the 1995 SEI CMM book, and it can be modified if necessary.

In the fourth phase of the framework, which is "action plan development and prioritization," a set of actions is derived from the prioritized KPs. These actions should reflect the requirements integrated in the first phase. Meanwhile, they also state what needs to be executed in order to reach a particular CMM maturity level. These actions guide the process improvement. Thus, more resources should be assigned to those actions with high priorities.

The above framework addresses the problem that CMM specifies only "what to do" but not "how to do." By incorporating requirements from the organization into action plans through KPA goals and KPs, the connection between the objectives of the organization and CMM maturity levels becomes clear.

3 .SPI Framework Based on CMMI Model using QFD

3.1. SPI framework for CMMI staged model using QFD

SPI framework based on CMMI is gaining popularity in the industry. In addition, QFD is used to help with the SPI based on CMMI.

First, software process requirements, which are from multiples perspectives, are prioritized so that requirements with more and stronger impacts on other requirements can receive higher priority values. Second, business and other requirements within an organization are mapped to CMMI Process Areas and practices. Because a connection is established, the organization can clearly see how CMMI helps with its business goals. Third, in CMMI, QFD helps transform requirements of the organization into process actions through Process Areas (PAs) and Practices. Therefore, the ordering of the actions taken is based on how they are related to both the software process requirements and the corresponding Practices in CMMI.

The SPI framework for CMMI staged model, as shown in Figure 2

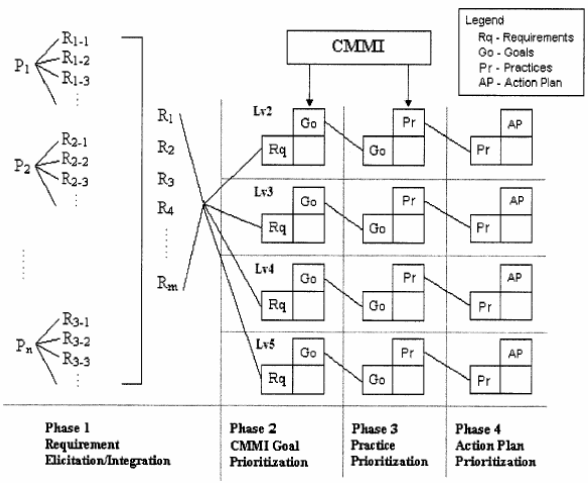


Fig. 2 Software Process Improvement through CMMI Staged Model Using QFD

For each of the four maturity levels, the set of requirements with adjusted priorities are related to the goals. The goals are prioritized based on those process requirements. Thus, the goals that achieve higher overall satisfaction of process requirements get higher importance.

In order to achieve these goals, CMMI staged model has generic practices categorized into four common features as well as the specific practices which correspond to the "Activities Performed" common feature in CMM. The priorities of Practices are determined by their correlations with goals. Thus, the generic practices in each common feature and the specific practices are prioritized separately based on the priorities of the goals. Practices that aim to achieve higher overall satisfaction of goals receive higher importance values. Separate sets of action plans are derived from the generic practices in each of the common features as well as from the specific practices. The actions that help to support more important Practices receive higher priorities.

As a result, the process requirements are reflected in PA goals, Practices, and the actions. The actions both follow the process maturity standards in CMMI staged model and satisfy the process requirements. Those actions with higher importance values help to achieve higher process requirements satisfaction.

Because of the close resemblance between CMMI staged model and CMM, the four phases for the SPI framework based on CMMI staged model as shown in Figure 2.

In Figure 2, phase 1 is exactly the same with the SPI framework based on CMM. Various perspectives are represented as P1 through Pn. Each perspective contains multiple requirements. The software process requirements in perspective 1 are represented as R1-1, R1-2, etc. These perspectives of software process requirements can then be prioritized based on their relative importance within the organization and integrated into one single set of requirements. In Figure 2, these integrated requirements are represented as R1 through Rm, where m is the total number of software process requirements from all perspectives. The prioritization ensures that requirements from different perspectives are comparable with each other, and the integration reflects the correlations among requirements from different perspectives. The deliverable of this phase is a set of prioritized and integrated software process requirements, which serves as the input to the next phase.

The second through fourth phases of this framework are applied to Level 2 to Level 5 of the CMMI staged model. The prioritized and integrated requirements from Phase 1 are linked to all goals in each of the four levels in CMMI staged model using relationship matrices. These prioritized goals are used as the basis for the prioritization of Practices. Finally, the prioritized Practices are transformed into prioritized action plans using House of Quality (HoQ).

In the second phase, which is "CMMI goal prioritization," the goals of all PAs in a particular maturity level are selected and prioritized based on the requirements from the previous phase. This phase helps to achieve two important objectives. First, the organization needs to comply with the CMMI standard. At the same time, the organization needs to ensure that by reaching a particular maturity level, the process is also satisfying the business and other requirements within the organization. In Phase 2, a relationship matrix is used to establish connections between the requirements from the organization and the goals in CMMI. This matrix demonstrates that complying with the CMMI standard also helps satisfy the business and other requirements in the organization. Second, the final set of action plans needs to be prioritized based on the priorities of requirements so that more important actions receive more resources. The goals serve as the bridge between requirements and the action plan. By prioritizing the goals, requirements from the organization can be transformed to the Practices in the third phase, and finally to the action plans in the final phase. In this way, a set of actions can be executed not only to achieve a specific maturity level in CMMI, but also to satisfy organizational process requirements.

The third phase of the framework, which is "practice prioritization," involves the prioritization of Practices within all PAs of a specific level. The prioritization is carried out on the basis of the deliverables from Phase 2. According to CMMI specifications, all these Practices have to be performed in order to reach that particular maturity level. These Practices serve as a bridge between the requirements and the final actions, and it is necessary to know how these Practices reflect the software process requirements. In order to show the connections between the requirements and the final action plans, these practices have to be prioritized based on the goals, which are now reflecting requirements priorities. The mapping between the goals and Practices has been has been clearly shown in CMMI documentation.

The fourth phase of the framework is "action plan development and prioritization". A set of actions is derived from the prioritized Practices. These actions should reflect the requirements integrated in the first phase. At the same time, they also state what needs to be executed in order to reach a particular CMMI maturity level. These actions guide the process improvement. Thus, more resources should be assigned to those actions with high priorities.

As shown in the above theoretical framework, the connection between the objectives of the organization and CMMI maturity levels becomes clear, by incorporating requirements from the organization into action plans through goals and Practices.

3.2. SPI framework for CMMI continuous model using QFD

The SPI framework for CMMI continuous model differs a lot from the staged framework. However, the same techniques of correlation-based prioritization with the help of QFD are used in the framework. In the continuous model of CMMI, the capability levels are assigned to individual PAs. Different PAs can be at different capability levels.

Each PA has two types of goal: 1) generic goals and 2) specific goals. Generic goals try to institutionalize the capability levels in CMMI, with one generic goal for each level. Specific goals describe the practices that must be implemented to satisfy the process area. These goals are satisfied by including generic practices and specific practices. Figure 3 illustrates how the practices and the actions are prioritized in the SPI framework for CMMI continuous model using QFD. The process requirements are used to in the prioritization of both PAs and Practices. The first step is to calculate the priority values of PAs. Then the Practices are prioritized from both the process requirements and PAs. Depending on which PA a Practice is from, the priority value of that Practices calculates from the requirements is multiplied by the PA priority. Finally, the action priority values are calculated from the Practice priority values.

Thus, as illustrated in Figure 3, the PAs are prioritized based on those process requirements and the PAs that help achieve higher overall satisfaction of process requirements get higher importance.

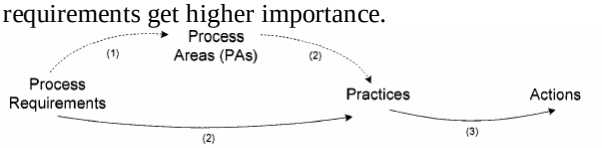


Fig. 3 Priority Calculation in SPI Framework Based on CMMI Continuous Model Using QFD

In order to make improvements on the PAs, generic practices for the generic goals and specific practices for specific goals at various capability levels are prioritized at the next phase. The priorities of Practices at different capability levels are determined by their correlations with the same set of process requirements. Because in CMMI continuous model, different PAs can have different of capability levels, the prioritization of Practices should be done for individual PAs. Thus, in this framework for CMMI continuous model, the Practices in each level of individual PAs are prioritized separately. The Practices that aim to achieve higher overall satisfaction of key goals receive higher importance values. The priority values for each PA calculated in the previous phase are used in the calculation of priorities of practices.

As a result, the process requirements are reflected in PAs, Practices, and the actions. The actions both follow the process capability standards in CMMI and satisfy the process requirements. Those actions with higher importance values help to achieve higher process requirements satisfaction.

In Figure 4, phase 1 is exactly the same with the SPI framework based on CMM. Various perspectives are represented as P1 through Pn. Each perspective contains multiple requirements. The software process requirements in perspective 1 are represented as R1-1, R1-2, etc. These perspectives of software process requirements can then be prioritized based on their relative importance within the organization and integrated into one single set of requirements.

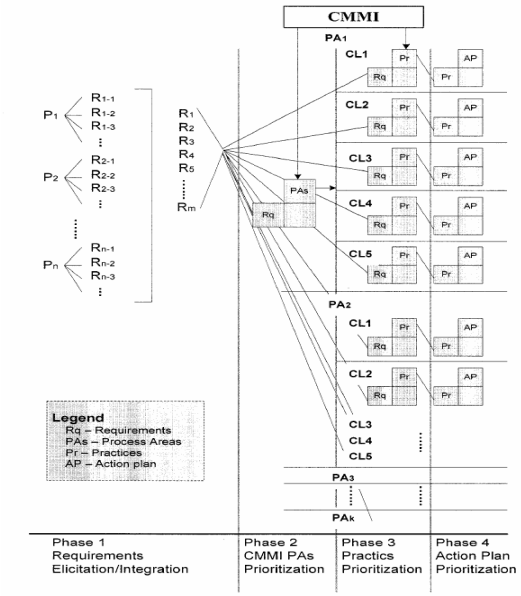


Fig. 4 Software Process Improvement through CMMI Continuous Model Using QFD

In Figure 4, these integrated requirements are represented as Rl through Rm, where m is the total number of

software process requirements from all perspectives. The prioritization ensures that requirements from different perspectives are comparable with each other, and the integration reflects the correlations among requirements from different perspectives. The deliverable of this phase is a set of prioritized and integrated software process requirements, which serves as the input to the next phase.
The second through fourth phases of this framework are applied to the PAs in the CMMI Continuous model. Because in CMMI continuous model, different capability levels are applied to different PAs, the framework for the staged model cannot be applied. Instead of mapping the prioritized and integrated requirements from Phase 1 to all the goals in a particular maturity level, they are linked to each of the PAs in Phase 2 and, depending on the target capability level, linked to each of the Practices in that level in Phase 3 using relationship matrices. In addition to the correlation values between process requirements and Practices, the priority value for each PA also participates in the calculation of the prioritization of Practices in that PA for a particular capability level. Finally, the prioritized Practices are transformed into prioritized action plans using House of Quality (HoQ).

In the second phase, which is "CMMI PA prioritization," all PAs are selected and prioritized based on the requirement priorities derived from the previous phase. This phase helps achieve two important objectives.

- The organization needs to comply with the CMMI standard. At the same time, the organization needs to ensure that by improving process areas to higher capability levels, the process is also satisfying the business and other requirements within the organization. In Phase 2, relationship matrices are used to establish connections between the requirements from the organization and each of the PAs. This matrix demonstrates that complying with the CMMI standard also helps satisfy the business and other requirements in the organization.

- The final set of action plans needs to be prioritized based on the priorities of requirements so that more important actions receive more resources. The PAs serve as the bridge between requirements and the action plan. By prioritizing the PAs, requirements from the organization can be transformed to the Practices in the third phase, and finally to the action plans in the final phase. In this way, a set of actions can be executed not only to reach higher capability levels in various PAs, but also to satisfy organizational process requirements.

The third phase of the proposed framework, which is "practice prioritization," involves the prioritization of Practices for a particular capability level within each PA. The prioritization is carried out on the basis of the deliverables from Phase 2. According to CMMI specifications, all these Practices for a capability level within a PA have to be performed in order for that PA to reach that particular capability level. However, they do not necessarily require the same amount of resources. These Practices serve as a bridge between the requirements and the final actions, and it is necessary to know how these Practices reflect the software process requirements. In order to show the connections between the requirements and the final action plans, these Practices have to be prioritized based on their correlations with requirements as well as the priority values of the Pas they belong to, which are now also reflecting requirements priorities.

In the fourth phase of the framework, which is "action plan development and prioritization," sets of actions are derived from the prioritized Practices for the desired capability levels of various PAs. These actions should reflect the requirements integrated in the first phase. Meanwhile, they also state what needs to be executed in order to reach a particular capability level of a particular PA. These actions guide the process improvement. Thus, more resources should be assigned to those actions with high priorities.
As shown in the above framework, by incorporating requirements from the organization into action plans through the goals and the Practices the connection between the objectives of the organization and PA capability levels becomes clear.

4. Conclusions

QFD is used to help an organization achieve three objectives. First, business and other requirements within an organization are mapped to CMM/CMMI goals and activities. A connection is established so that the organization can clearly see how CMM/CMMI helps with its business goals. Second, software process requirements from multiples perspectives are prioritized so that requirements with more and stronger impacts on other requirements can receive higher priority values. Third, QFD helps transform requirements of the organization into process actions through Key Process Areas (KPAs) and Key Practices (KPs) in CMM/CMMI. Therefore, the ordering of the actions taken is based on how they are related to both the software process requirements and the corresponding KPs in CMM/CMMI.

Acknowledgments

This work is financially supported by the National Natural Science Foundation of China (Project No. 90718038). Thanks for the help.

References

[1] Paulk, Mark C., Bill Curtis, Mary Beth Chrissis, Charles V. Weber. " Capability Maturity Model for Software, Version 1 .1." Technical Report. CMU/SEI-93-TR-024, ESC-TR-93-177, February, 1993.

[2] Paulk, Mark C., Charles V. Weber, Suzanne M. Garcia, Mary Beth Chrissis, Marilyn Bush. "Key Practices of the Capability Maturity Model, Version 1.1." Technical Report. CMU/SEI-93-TR-025, ESC-TR-93-178, February, 1993.

[3] Francois Coallier, "How ISO 9001 Fits Into the Software World," IEEE Software, Vol. 11，No. 1，January 1994, pp. 98-100.

[4] Paulk, Mark C., Charles V. Weber, and Mary Beth Chrissis, "The Capability Maturity Model for Software." In K. El Emam and N. H. Madhavji (eds.), Elements of Software Assessment and Improvement, IEEE CS Press, 1999.

[5] Jennifer Gremba, Chuck Myers, "The IDEAL Model: A Practical Guide for Improvement", Bridge, Issue 3, 1997.

[6] Bamberger J. 1997. Essence of the Capability Maturity Model. Computer 30(6): pp.112-114.

[7] Paulk M., Weber C., Curtis B, Chrissis M. Eds. 1995. The Capability Maturity Model: Guidelines for Improving the Software Process. Reading, MA, Addison-Wesley.

[8] Akao, Yoji, ed., Quality Function Deployment: Integrating Customer Requirements into Product Design, Cambridge, MA, Productivity Press, 1990.

[9] Liu X, Inuganti P., Veera C. 2003. An Integration Methodology for Software Quality Function deployment. Final Project Report to the Toshiba Corporation.

[10] Xiaoqing (Frank) Liu, Yan Sun, Praveen Inuganti, Chandra Sekhar Veera, and Yuji Kyoya. "A Methodology for the Tracing of Requirements in Object-Oriented Software Design Process Using Quality Function Deployment," Software Quality Professional Journal, September 2007, Volume 9, Issue 4.

[11] Akao, Yoji, Glenn H. Mazur. "Using QFD to Assure QS9000 Compliance." 4th International Symposium on Quality Function Deployment, Sydney, 1998.

[12] Zultner, R.E. "Quality Function Deployment (QFD) for Software." American Programmer, 1992.

[13] Akao Y., Hayazaki T. "Environmental Management System on ISO14000 Combined with QFD." Transactions of the Tenth Symposium on QFD. Novi, Michigan. ISBN 1-889477-10-9

[14]Ita Richardson. "Quality Function deployment-A Software Process Tool?" Third Annual International QFD Symposium. Linkoping, Sweden, Oct. 1997.

[15]Ita Richardson, Eamonn Murphy, KevinRyan, "Development of Generic Quality Function Deployment Matrix", Quality Management Journal, Vol. 9, No. 2, APRIL 2002, pp. 25-43

[16] Zultner, Richard E. "Business Process Reengineering with Quality Function Deployment: Process Innovation for Software Development." 7th Symposium on QFD (ISBNI-889477-07-9), 1995.

[17]Song, Ki-won; Kim, Jin-soo. Measurement and Management of the Level of Quality Control Process in SoC (System on Chip) Embedded Software Development, International Journal of Advanced Robotic Systems, APR 5 2012

Author Yonghui CAO received the MS degree in business management from Zhejiang Universityin 2006. He is currently a doctorate candidate in Zhejiang University. His research interest is in the areas of management information systems.

The Perceptual and Statistics Characteristic of Spatial Cues and its application

Heng Wang[1,2], Ruimin Hu [1], Weiping Tu [1] and Cong Zhang[2]

[1] National Engineering Research Center for Multimedia Software,
Wuhan University
Wuhan, China

[2] School of Mathematic & Computer Science
Wuhan Polytechnic University
Wuhan, China

Abstract

In present mobile communication system, low bit rate audio signal is supposed to be provided with high quality. This paper researches the mechanism exists of perceptual and statistics redundancy in spatial cues and establishes a selection model by joint perceptual and statistics characteristic of spatial cues. It does not quantize the values of spatial cues where the frequency bands can't easily be perceived by human ears according to the selection model. Experimental results showed that this method can bring down the parametric bitrate by about 15% compared with parametric stereo, while maintaining the subjective sound quality.

Keywords: *Interaural Level Difference, just notice difference, spatial cues, perceptual characteristic.*

1. Introduction

Spatial audio coding is a method by downmixing stereo to mono and extracting spatial parameters which represent the orientation information of spatial sound field. The most important spatial cues contain: Interaural level difference (ILD), interaural time difference(ITD) and interaural correlation(IC)[1][2].The bitrate of stereo coding can be reduced effectively because the bitrate of spatial parameters is smaller than that of channel signals. With the multichannel audio encoding technology is mature and widely used, the direction of spatial audio coding technology gradually changes from stereo coding to multichannel coding. As the bitrate increases linearly with the increasing number of channels, how to effectively reduce the bitrate of spatial parameters is an important problem in the field of spatial audio coding.

Yang Won Jung in 2006 pointed out that the auditory sensitivity of spatial parameters is associated with the channel configuration, especially ILD. He proposed a quantization method which was used multiple quantization tables corresponding to different channel configuration instead of existing single quantization table to improve quality and efficiency of quantification [3]. K. Kim in 2007 also pointed out that the quantitation method for spatial cues in MPS[4] lacks of theoretical background and appropriates quantization steps and proposed parameter quantization scheme based on position information of virtual sound source for spatial cues [5]. Because the movement of sound source in spatial sound field is generally slow and the spatial cues of adjacent frames have strong correlation, B. Cheng in 2008 proposed a differential coding scheme for ILD that calculated the difference of spatial parameters corresponding frequency band in adjacent frames and only quantized the difference [6].

These methods can reduce the parametric bitrate, but they did not consider the human ear's perceptual and statistics characteristics in different frequencies. There is still redundancy in spatial parameters.

As there exists certain perceptual thresholds for the perception of sound intensity, the perception of spatial orientation change by human ear is also limited. And human ear can perceive the changes of sound image orientation only when the difference of binaural cue reaches a certain threshold value, and this threshold value is known as Just Noticeable Difference (JND)[7]. The main influence factors of JND are frequency, intensity and intensity difference and so on. Scholars have made various measurement and analysis in allusion to these factors.

In 1960, miller measured JND of ILD under 11 tone signals and discovered that there is some relationship between ILD and frequency, especially the JND of ILD reached a maximum at 1000 Hz[8]. Yost in 1988[9] and many other scholars verified this conclusion by their respective experiments. But in 1992[10], When Kaigham measured JND of ILD under narrow band signal, he found

that the change of JND is not obvious with the rising of frequency which is different from miller's conclusions.

In 1969, Hershkowitz[11] researched the influence of sound direction to JND of ILD and found that the JND increases as binaural signal intensity difference increases. It illustrates that the closer is the sound to ear, the more insensitive is human ear to sound location. But the author only measured the signal of 500Hz. In 1988, Yost [9] measured the JND of 5 frequencies and 3 ILD, the results showed that the law of JND for different ILD was basically the same as frequency.

In 2000, Andrew [12] researched the relationship of JND for ILD to signal duration and intensity. He found that JND of ILD decreased with the increasing of signal duration and intensity under constant intensity, but it is very big under crossover frequency and had very less relationship with signal duration and intensity.

In 2008, Chen [13] researched the relationship between sinusoidal tone signal and the JND of binaural clues. There is some great different between her result and abroad scholars, and the reason may be that sound pressure of each frequency was not kept at a constant value, but the sum of signal at each frequency was constant.

This paper is organized as follows: Section 2 researches the perceptual characteristic of spatial cues in different frequency bands. Section 3 analyses the statistics characteristic of spatial cues in different frequency bands. Section 4 proposes the selection model and quantization algorithm. The results from experiments are presented in Section 5. Finally, Section 6 gives some conclusions.

2. Perceptual Characteristics of Spatial Cues

Early perceptual experiments showed that people's perception of spatial orientation is related to the age of listeners, intensity, frequency, listening environment and other factors. As a result, in order to obtain the JND of spatial cues, we need carry out numerous individual experiments and adapt mathematical and statistical analysis. This article take ILD as an example to introduce specific methods, other cues are the same.

2.1 Subjects

There were 24 listeners which are all graduate students in this experiment including 15 male and 9 female, they are all 21-27 years old. These listeners have made many professional training and subject listening test before. In this experiment, every people need to make 32 times audiometry, there is 12 frequency points every time, each frequency point need 5 minutes. Because signal frequency may lead to weary easily, so each listeners need to have a rest after several frequency points. It cost about 2 hours to finish the whole process every time.

2.2 Stimuli

The method in this article used a two-alternative-forced-choice paradigm to measure the JND. Both reference and test signals were 250 ms in duration including 10 ms raised-cosine onset and offset ramps. They were randomly combined into stimulus and separated by 500 ms duration. The Stimuli were create by personal computer and presented to the subjects over headphones (Sennheiser HDA 215) at a level of 70 dB SPL. In order to exclude other factors influence on this experiment, the environment of the entire testing process should be consistent and the intensity of test sound must remain around 70 dB SPL. Meanwhile the ITD should be zero in the whole experiment in order to remove the effect on the result caused by other binaural cues and the sum of energy of left and right channels should remain unchanged.

The reference values of ILD in this experiments are 0 dB, which respond to the midline in the horizontal plane.

The experiment divided the whole frequency domain into 20 sub-bands, and each frequency sub-band satisfied the same perceptual characteristics of human ear. The frequency sub-bands closely mimic the critical band concept and are formed in such a way that each band has a bandwidth, BW(in Hz), which is approximately equal to the equivalent rectangular bandwidth(ERB), following

$$BW = 24.7(0.00437f + 1) \quad (1)$$

with f the(center) frequency given in Hz[14].

The stimuli are pure tones whose frequencies are 75, 150,225,300,450,600,750,900,1200,1500,1800,2100,2400, 2700,3300,4200,5400,6900,10500,15500Hz.

2.3 Method

Discrimination thresholds were estimated with an adaptive procedure. During any given trial, subjects would listen two stimuli by activating a button on a computer screen by mouse-click, with a free number of repeats but the order of two part stimulus changed. The subjects was to indicate which of two stimuli was lateralized to the left relatively by means of an appropriate radio button response in a given response time. Subjects were allowed 1.5 s to respond.

An adaptive, 1-up-3-down method was also used in this article. The difference of ILD in decibels was increased in every one wrong or decreased in every three consecutive correct. The difference between reference and test signals in first trials was the initial variable which was much larger than target JND value, it was changed by an given step according to previous test result.

The step was changed adaptively, it was adjusted by 50% for the first two reversals, 30% for the next two reversals, then liner changed in a small step size for the next three reversals, last step size was the step of expected accuracy for the last three reversals. In a transformed-up-down experiment, the stimulus variable and its direction of change depends on the subjects responses. The direction alternates back and forth between “down” and “up”. Every transform between “down” and “up” was defined as a reversal [15].

Because of heavy workload of these experiments, an adaptive test software was designed to simplify the experiments and the process of data collection and analysis. The software automatically generated test sequences and played one after another. According to the listener’s choice, the software changed ILD values of test stimulus properly, and saved the results to excel sheet until listener hardly distinguished the orientation differences between two sequences. And the value of ILD at this time was the JND value.

2.4 Results

After a subjective listening test for half a year, we get 120 groups of data, each group containing 24 JNDs corresponding to 24 subjects. For every group, we select the data that has the confidence degree of 75% to be JND in that condition. Table 1 gives the JNDs in all conditions.

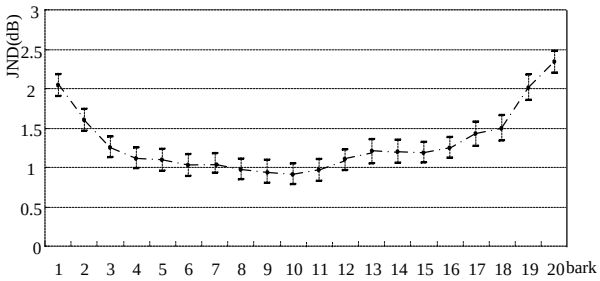


Fig. 1 JND curve of ILD=0.

3. Statistical Distribution of Spatial Cues

Spatial Audio Coding is expected to use perceptual characteristics of spatial parameters in different frequency bands and accurately extract spatial parameters which represent the orientation information of spatial sound field. The "exact representation" consists of two aspects of meaning: The first is to extract spatial parameters which represent the orientation information of spatial sound field from the audio signal; the second is to remove redundant information of spatial parameters as possible as we can.

We had already given the frequency dependence JND curve of ILD. It is clear that the values of ILD can be perceived by human ear when they are larger than JND in a certain frequency band. However, such information that which frequency bands can be perceived must transfer to decoder. We can’t achieve the purpose to enhance coding efficiency and reduce the coding bitrate because the extra information need to code. How to remove the redundant information of ILD and improve the coding efficiency is the goal of spatial audio coding.

We researched statistical distribution characteristics of ILD and obtained the probability that the actual values of ILD can be perceived in stereo signal. It can guide the quantization of ILD in spatial audio coding and improve the coding efficiency.

300 sections of typical stereo music were chosen for the statistical experiment in this article, including Chinese folk music, western musical instruments, natural sounds, popular songs and other music accompaniment material. These materials were selected from the program source CD GSBM61001-89 used for the subjective listening evaluation of national standard, commonly used in domestic audio demo track exhibition and so on. These test sequences were divided into 12 categories. For simplicity, the length of test sequences used in this experiment is 20s and the sampling rate is 48 kHz.

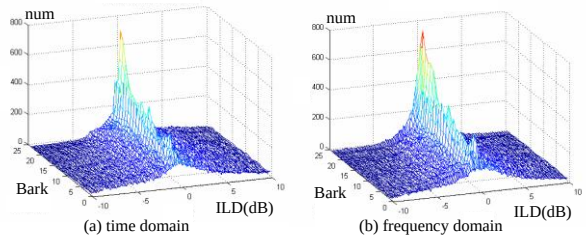


Fig. 2 The statistical distribution of ILD in different domains

From the above distribution picture we can see: ILD statistical distribution in time domain is basically the same as that in FFT domain, which means the domain of

parameters extracting has little impact on the distribution characteristics of ILD parameters. The probability of ILD near zero is the highest and it increases with the frequency increases; ILD distribution characteristics under different domains is insignificant, the distribution characteristics of time-domain can be replaced by that of spatial parameters which we expect to get in this experiment.

We had got the ILD statistics distribution of these sequences by Matlab as follows:

Table 1: The statistical distribution of ILD

section / type	|ILD|<=0.5	0.5<|ILD|<=1	1<|ILD|<=2	|ILD|>2
es01	15.4%	70.0%	9.7%	5.0%
es02	93.7%	2.3%	2.5%	1.5%
es03	92.6%	1.9%	0.8%	4.7%
sc01	18.9%	13.5%	20.6%	47.0%
sc02	15.0%	10.4%	20.2%	54.4%
sc03	17.7%	13.6%	21.6%	47.1%
si01	21.3%	12.8%	16.9%	49.1%
si02	33.5%	20.8%	25.7%	20.0%
si03	7.7%	6.1%	16.9%	69.4%
sm01	9.8%	4.4%	8.5%	77.3%
sm02	29.5%	16.6%	21.7%	32.2%
sm03	11.1%	8.1%	15.5%	65.3%
film music	13.3%	12.5%	21.9%	52.3%
symphonic music	6.2%	6.2%	13.8%	73.8%
light music	15.7%	14.7%	20.9%	48.7%
pop music	15.7%	14.8%	27.3%	42.2%
average	26.0%	14.3%	16.5%	43.1%

First we get the statistical distribution of ILD in all frequency bands for MPEG sequence and other music.

We can see that more than 40% probability of ILD is less than 1 dB, it was proved by more and more experiments. However, the JNDs of ILD are almost larger than 1 dB, so there is too much redundancy which is hard to be perceived by human ear.

According to the JND value of each band in the perceptual experiments, the statistical distribution characteristics of ILD in each band above the JND in time-domain can be obtained. The distribution figure is as follows:

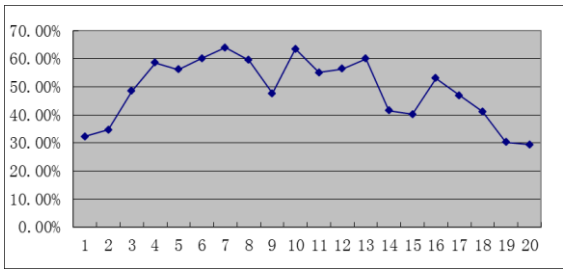


Fig. 3 The statistical distribution of ILD in different frequency bands.

This paper proposed a method that combined with time-domain distribution characteristics and perceptual characteristics of ILD to guide the selection of ILD and improve the coding efficiency.

4. Selection Model for Spatial Cues

Spatial parameter selection is a process to determine whether the spatial parameters of frequency bands should be encoded. The perceptual and statistical distribution characteristics of ILD that we have studied are the basis of the spatial parameter selection.

From the JND curve of ILD we can find: human ears can't easily be perceived when the values of ILD are smaller than the values of JND at corresponding bands. We need to encode such parameters only when the spatial parameters are greater than the values of JND at the same frequency bands. From the statistical distribution curve of ILD for real stereo signals we can find: we don't encode the parameters when the values of ILD appear in low probability range in some conditions.

Based on perceptual and statistical characteristics, the selection model of ILD can be got as following steps:

First of all, according to the experiment result of perceptual sensitivity of ILD, we can divide the 20 frequency bands into the following areas: B1-B3 and B18-B20 are very insensitive areas; B4-B12 and B15-B17 are much sensitive areas; B13-B14 is less sensitive areas.

Secondly, according to the statistical distribution of ILD above the JND in each frequency band, 20 frequency bands can be also divided into the following areas: B1-B3 and B18-B20 are very insensitive regions, B4-B12 and B16-B17 are much sensitive areas; B13-B15 is less sensitive areas. We don't considered B9 here alone as a special case.

Finally, according to the perceptual importance determined by the above two steps, we can find their sensitivity

characteristics are basically the same. In consideration of the two situations, the following ways can be used to extract the parameters:

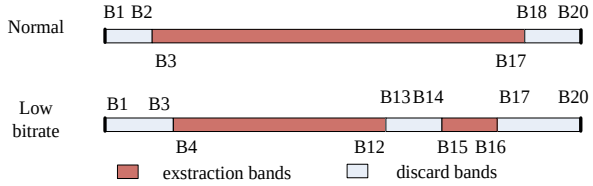


Fig. 4 The selection model of ILD.

The method of parameter extracting proposed in this paper can be simply applied in current spatial audio coding. We just need to add an estimate module on the original framework. Only when the frequency band is the one we need to extract according the selection model, the original extraction and quantitative methods will be used to encode. Otherwise, the encoding index will be zero or the index difference between the current band and the previous band will be zero, so we can reduce the coding bitrate while the complexity will not increase. The selection scheme of spatial parameters proposed in this paper is based on the frequency perceptual characteristics of spatial parameters, it can guarantee the loss will not be heard or easily perceived by human ear.

5. Experiment and Result

The selection model of ILD proposed in this article was mainly applied in spatial parameter extraction. This paper adopted the selection strategy in high quality.

5.1. Bitrate

This experiment first calculated the total number of bits used to quantify ILD according to the method that this article proposed, then compared with the PS quantization method and calculated the percentage of bitrate decline. The proposed method was implemented in Enhanced aacPlus[16] standard code, the bitrate is 32 kbps.

The Table 2 list bitrate decreased for ILD quantization from the MPEG standard test sequences. We can find that the decline percentage of speech is greater than other types. The decline in the percentage of other types is basically the same, which maintain about 15%-18%. The average parameters bit rate has decreased by 18.86%..

Table 2: The percentage of bitrate decline

Seq \ method		*PS (bit)*	*This aticle (bit)*	*Bitrat decline*
Speech	es01	5716	4520	20.92%
	es02	702	556	20.80%
	es03	3102	2207	28.85%
Complex mix sound	sc01	14320	11962	16.47%
	sc02	16307	13424	17.68%
	sc03	12499	10313	17.49%
Single instrument	si01	7483	6403	14.43%
	si02	7560	6344	16.08%
	si03	24011	17373	27.65%
Simple mix sound	sm01	10166	8609	15.32%
	sm02	9355	7966	14.85%
	sm03	14566	12265	15.80%
average		-	-	18.86%

5.2. Quality

The sound quality after decoding was evaluated by subjective test.

Subjective test adopt MUSHRA listening test standards, used the Sennheiser HD215 professional headphone. The subjects were 12 persons aged between 20-30 years in which male to female ratio of 1:1, they all had receive professional training. Figure 5 shows the subjective quality of the quantization methods proposed in this article and PS.

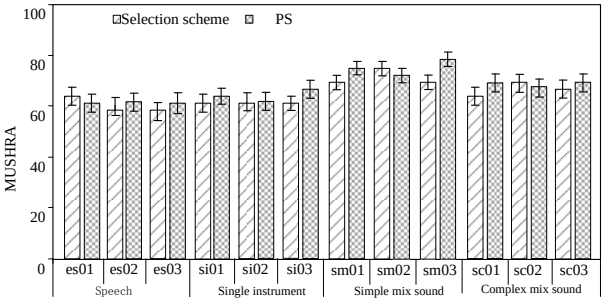


Fig. 5 Subjective quality of two methods

As is shown above figure, the subjective quality scores of the two methods were maintained in the same interval, rose and declined in the same range and level. Therefore, we can consider that of the two methods were equal.

6. Conclusions

We have demonstrated in this work that the mechanism exists of perceptual and statistics redundancy in spatial

parameters and try to remove it by joint perceptual and statistics characteristic. The new quantization strategy merely quantizes the perceived variable quantity of spatial parameters to reduce the coding bitrate. Experimental results show that this method can bring down the parametric bitrate by about 20% compared with parametric stereo, while maintaining the subjective sound quality.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61231015, 61102127, 61272278, 61201340, 61201169), the major national science and technology special projects (2010ZX03004-003-03), the Doctoral Fund of Ministry of Education of China(Grant No.20090141110054) and the Fundamental Research Funds for the Central Universities.

References

[1] J.W. Strutt. "The theory of sound", Dover Publications, 1877.
[2] F. Baumgarte and C. Faller. "Binaural Cue Coding - Part I: Psychoacoustic fundamentals and design principles", IEEE Trans. on Speech and Audio Proc., 2003,Vol. 11.
[3] J. Herre et al.. "The Reference Model Architecture for MPEG Spatial Audio Coding", in Proc. 118th Audio Eng. Soc. Convention, 2005.
[4] Y.W. Jung, H.O. Oh, H. J. Kim, and S. J. Choi. "New CLD quantization method for spatial audio coding", in Proc. 120th AES Conv., Paris, 2006.
[5] K.Kim, S.Beack, J.Seo, D.Jang, and M.Hahn. "Improved Channel Level Difference Quantization for Spatial Audio Coding", ETRI Journal,2007, Vol.29, pp.99-102.
[6] Cheng, B., C.H. Ritz and I.S. Burnett. "Psychoacoustic-based quantization of spatial audio cues", Electronics Letters, 2008,Vol.44.
[7] Zwislocki, J. and R.S. Feldman. "Just Noticeable Differences in Dichotic Phase", J. Acoust. Soc. Am., 1956,Vol.28(5): p. 860-864.
[8] Mills, A.W., "Lateralization of High-Frequency Tones,"J. Acoust. Soc. Am. ,1960,Vol.32,pp.132-134.
[9] Yost, W. A., and Dye, J. R. H.. "Discrimination of interaural differences of level as a function of frequency", J. Acoust. Soc. Am,1988,Vol.83,pp.1846-1851.
[10] Kaigham J.G, "Frequency dependence of binaural performance in listeners",J. Acoust. Soc. Am.,1992, Vol.91,pp. 336-347.
[11] R.M.Hershkowitz and N.I.Durlach, "Interaural Time and Amplitude jnds for a 500-Hz Tone", J. Acoust. Soc. Am., 1969,Vol.46, pp. 1464-1465.
[12] J Oxenham, A.J. and S. Buus, "Level discrimination of sinusoids as a function of duration and level for fixed-level, roving-level, and across-frequency conditions", J. Acoust. Soc. Am. , 2000,Vol.107, pp. 1605-1614.
[13] Chen Shuixian, Hu Ruimin, "Frequency Dependence of Spatial Cues and Its Implication in Spatial Stereo Coding",in International Conference on Computer Science and Software Engineering, 2008, pp. 1066-1069.
[14] Glasberg, B.R. and B.C.J. Moore. "Derivation of auditory filter shapes from notched-noise data", Hearing Research, 1990. Vol.47(1-2): p. 103 - 138.
[15] Levitt, H.C.C.H., "Transformed Up-Down Methods in Psychoacoustics.",Acoustical Society of America Journal, 1971. Vol. 49:p p. 467-477.
[16] 3GPP TS 26.405 : Enhanced aacPlus general audio codec; Encoder Specification Parametric Stereo part.

Mr. Heng Wang is a researcher in the field of audio processing and coding, 3D audio Systems. At present he is working in the Department of School of Mathematic & Computer Science, Wuhan Polytechnic University, Wuhan. He has completed a bachelor's degree in Huazhong University of Science and Technology. Now he is pursuing Ph.D in Wuhan university, Wuhan.

Prof. Ruimin Hu is an eminent researcher in the field of multimedia processing, network communications, security and emergency. At present he is working in the Department of National Engineering Research Center for Multimedia Software, he is the Associate Dean of School of Computer Science, Wuhan University. He has completed Ph.D. in Huazhong University of Science and Technology, Wuhan, China.

Dr. Weiping Tu is a researcher in the field of audio processing and coding and network communications. At present he is working in the Department of National Engineering Research Center for Multimedia Software. He has completed Ph.D. in Wuhan Unversity.

Prof. Cong Zhang is an eminent researcher in the field of multimedia processing and network communications,. At present he is the Dean in the Department of School of Mathematic & Computer Science, Wuhan Polytechnic University, Wuhan. He has completed Ph.D. in Wuhan University, Wuhan, China.

The Analysis of Vibration Characteristics and Motion Stability of the Tracked Ambulance Nonlinear Damping System

Meng Yang, Xinxi Xu, Chen Su

Institute of Medical Equipment, Academy of Military Medical Sciences,
Tianjin, China

Abstract

Considering the impact of the nonlinear stiffness, a 2 DOF dynamic nonlinear vibration model with cubic terms was established according to the structural feature and nonlinear behavior of the tracked ambulance. In the case of primary resonance and 1: 1 internal resonance, multiple scale method was used to obtain a first-order approximate solution for this model. Taking the parameters of the tracked ambulance for instance, the approximate solution was verified and the influence of the parameters on damping effect was investigated. Finally, the motion stability of the damping system was analyzed with singularity theory and the theoretical bases for improving efficiency of the damping system were provided.

Keywords: *Damping System; Cubic Nonlinearity, Multiple Scale Method; Internal Resonance; Stability*

1. Introduction

The tracked ambulance can be delivered through a variety complex terrain and implement first aid to the sick and wounded. In order to achieve the safe transfer and implement first aid on the way, it is often necessary to demand the tracked ambulance has good mobility and meet the special needs of the sick and wounded of comfort. For the tracked ambulance is refitted by crawler chassis, he installation of the vehicle damping system becomes the main way to improve the ride comfort of the sick and wounded.

The tracked ambulance damping system is composed of the carriage, the stretcher base, the chassis and the nonlinear shock absorber. Hence, it can be easily converted into a multi-degree of freedom nonlinear vibration system. The use of the nonlinear vibration system presents various advantages, such as better performance in the inhibition of broadband vibration, especially low-frequency vibration. However, complex mechanical properties usually exist in a nonlinear vibration system such as chaos and bifurcation, which makes it difficult to be analytic calculation and analysis, therefore approximate analytic algorithm widely used. Christopher Lee [4] investigated suspended,elastic cables driven by planar excitation with near commensurable natural frequencies in a 2:1 ratio. The first order analysis shows that there are saturation and jump phenomena and the first order analysis reveals that the cubic nonlinearities disrupt saturation. Li Jian et al [5] applied multiple scales method and Runge-Kutta to study the nonlinear vibration characteristics of the axial movement, multi-layered cylindrical shells made from composites. The results show some nonlinear properties of the system such as the phenomenon of internal resonance and indicate that excitation amplitude, damping and speed can affect the response amplitude, range of interval resonance and soft feature of the system. Zhang Xin et al [6] used average method to analyze piecewise nonlinear characteristics of the viscoelastic shocker absorber and the relationship of amplitude-frequency characteristics and system parameters. Li Xinye [7] used average method to study the possibility of delay feedback control over the gyroscope system under force. Tsuyoshi Inoue [8] investigates the vibration phenomena of the one-degree-of-freedom magnetically levitated system considering the effect of the nonlinearity of the electromagnet, using a shooting method.

In this paper, the differential equations of the 2 DOF tracked ambulance nonlinear damping system, including the cubic nonlinear spring was presented. In the case of primary resonance and 1: 1 internal resonance, multiple

scale method was used to obtain a first-order approximate solution of the differential equations. Taking the parameters of the tracked ambulance for instance, the accuracy of the approximate solution was established by compared to numerical results. The influence of the parameters on damping effect and motion stability was also investigated. Furthermore, the theoretical bases for improving efficiency of the damping system were put forward.

2. Damping System Physical Model

The tracked ambulance damping system is shown in Figure 1. Damping system is mainly composed of rubber damping shock absorber and zero stiffness damper. The linear model is used to describe the stiffness and damping of the rubber damping shock absorber. For zero stiffness damper, the damping is described by linear model and stiffness is described by positive and negative stiffness parallel model[9], shown in Figure 2.

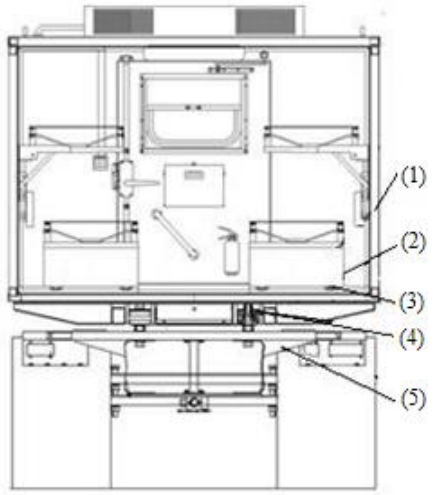


Fig.1 Ambulance damping system

(1) Carriage(2)Stretcher base(3) Zero stiffness damper (4)Rubber damping shock absorber(5) Coach chassis

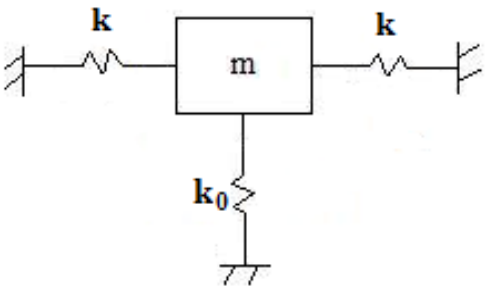


Fig.2 Positive and negative stiffness parallel model

The stiffness, original length and initial deformation of horizontal spring, in Figure1, are defined as k, L and λ. k_0 Stands for the stiffness of vertical spring. The vertical elastic restoring force of the model can be expressed in form

$$f(x) = k_0 x - k[x - \frac{L-\lambda}{\sqrt{1-(x/L)^2}}\frac{x}{L}] \quad (1)$$

Using the Taylor series, I seek a second-order expansion in the form

$$\frac{1}{\sqrt{1-(x/L)^2}} = 1 + \frac{1}{2}(\frac{x}{L})^2 + \frac{13}{24}(\frac{x}{L})^4 + \cdots \quad (2)$$

Substitute the first two into the Eq.(1) result in

$$f(x) = (k_0 - \frac{k\lambda}{L})x + \frac{L-\lambda}{L^3}\frac{k}{2}x^3 \quad (3)$$

Hence, the restoring force of quasi-zero stiffness damper can be expressed in form

$$f(z) = K_s x + \beta K_s x^3 \quad (4)$$

Where, $K_s = (k_0 - \frac{k\lambda}{L})$, $\beta K_s = \frac{L-\lambda}{L^3}\frac{k}{2}$ and β is a small parameter.

According to the occupant of the vehicle ride(lying) comfort evaluation standards, occupant comfort is mainly affected by the vertical vibration acceleration. Ignoring the other two directions of vibration, the 2 DOF model of the tracked ambulance damping system is shown in Figure 3.

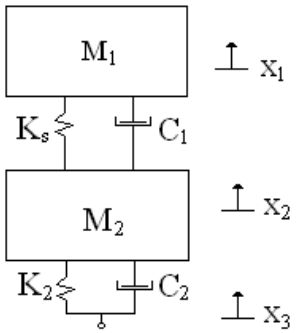


Fig.3 The 2 DOF damping system

Including:

M_1—The quality of stretcher and decubital body;

M_2—The quality of carriage;

K_s—The stiffness of quasi-zero stiffness damper;

C_1—The damping of zero stiffness damper;

K_2—The stiffness of rubber damping shock absorber;

C_2—The damping of rubber damping shock absorber;

x_1、x_2、x_3—Stretcher base vibration displacement, Carriage vibration displacement, Chassis vibration displacement.

The differential equations describing the motion of the damping system are

$$M_1\ddot{x}_1 + C_1(\dot{x}_1 - \dot{x}_2) + K_s(x_1 - x_2) + \beta K_s(x_1 - x_2)^3 = 0 \quad (5)$$

$$M_2\ddot{x}_2 - C_1(\dot{x}_1 - \dot{x}_2) - K_s(x_1 - x_2) - \beta K_s(x_1 - x_2)^3 + C_2(\dot{x}_2 - \dot{x}_3) + K_2(x_2 - x_3) = 0 \quad (6)$$

We rewrite the Eq.(5) and (6) as

$$\ddot{x}_1 + \omega_1^2 x_1 = l_1 x_2 - 2u_1\dot{x}_1 + 2u_1\dot{x}_2 - b_1(x_1 - x_2)^3 \quad (7)$$

$$\ddot{x}_2 + \omega_2^2 x_2 = f\cos\Omega t + 2u_2\dot{x}_1 - 2u_3\dot{x}_2 + l_2 x_1 + b_2(x_1 - x_2)^3 \quad (8)$$

Where $\omega_1^2 = K_s / M_1$, $l_1 = K_s / M_1$, $2u_1 = C_1 / M_1$, $b_1 = \beta K_s / M_1$, $2u_2 = C_1 / M_2$, $\omega_2^2 = (K_2 + K_s) / M_2$, $f\cos\Omega t = K_2 x_3 + C_2\dot{x}_3$, $2u_3 = (C_1 + C_2) / M_2$, $l_2 = K_s / M_2$, $b_2 = \beta K_s / M_2$.

3. Perturbation Analysis

Using multi-scale method, the dynamic response of damping system is solved. The new independent time scales

$$T_n = \varepsilon^n t \qquad n = 0,1,\cdots \quad (9)$$

are introduced where ε represent a small positive parameter and T_n , $n = 0,1,\cdots$ are 'slow' time scales which capture the response due to the nonlinearities, damping and external excitation. And we note that

$$\frac{d}{dt} = D_0 + \varepsilon D_1 + \cdots \quad (10)$$

$$\frac{d^2}{dt^2} = D_0^2 + 2\varepsilon D_0 D_1 + \varepsilon^2(D_1^2 + 2D_0 D_2) + \cdots \quad (11)$$

Where $D_n = \partial / \partial T_n$, $n = 0,1,\cdots$. We expand the time-dependent variable x_1 and x_2 in powers of ε as

$$x_1 = x_{11}(T_0, T_1) + \varepsilon x_{12}(T_0, T_1) \quad (12)$$

$$x_2 = x_{21}(T_0, T_1) + \varepsilon x_{22}(T_0, T_1) \quad (13)$$

Then we substitute Eq.(10)-(13) into the Eq.(7) and (8), equate coefficients of like powers of ε ,and obtain the following:

Order(ε^0):

$$\begin{cases} D_0^2 x_{11} + \omega_1^2 x_{11} = 0 \\ D_0^2 x_{21} + \omega_1^2 x_{21} = 0 \end{cases} \quad (14)$$

Order(ε^1):

$$\begin{cases} D_0^2 x_{12} + \omega_1^2 x_{12} = -2D_0(D_1 x_{11} + u_1 x_{11} - u_1 x_{21}) \\ + l_1 x_{21} - b_1 x_{11}^3 + 3b_1 x_{11}^2 x_{21} - 3b_1 x_{11} x_{21}^2 + b_1 x_{21}^3 \\ D_0^2 x_{22} + \omega_1^2 x_{22} = -2D_0(D_1 x_{21} - u_2 x_{11} + u_3 x_{21}) \\ + l_2 x_{11} + b_2 x_{11}^3 - 3b_2 x_{11}^2 x_{21} + 3b_2 x_{11} x_{21}^2 - b_2 x_{21}^3 \\ + f\cos(\Omega T_0) \end{cases} \quad (15)$$

The solution of Eq. (14) can be expressed as

$$\begin{cases} x_{11} = A_1(T_1)\exp(i\omega_1 T_0) + cc \\ x_{21} = A_2(T_1)\exp(i\omega_2 T_0) + cc \end{cases} \quad (16)$$

To express 1:1 internal resonance and the nearness of the excitation frequency to the first order natural frequency, we introduce two detuning parameter σ_1 and σ_2 defined by $\omega_2 = \omega_1 + \varepsilon\sigma_1$, $\Omega = \omega_1 + \varepsilon\sigma_2$.

Substitution of Eq.(16) and $\omega_2 = \omega_1 + \varepsilon\sigma_1$, $\Omega = \omega_1 + \varepsilon\sigma_2$ into Eq.(15) leads to secular terms. Elimination of these secular terms leads to the two state equations

$$\begin{cases} -2A_1' i\omega_1 - 2u_1 A_1 i\omega_1 + 2u_1 i\omega_2 A_2 \exp(i\sigma T_0) \\ + l_1 A_2 \exp(i\sigma T_0) - 3b_1 A_1^2 \bar{A}_1 + 3b_1 A_2^2 \bar{A}_2 \exp(i\sigma T_0) \\ + 3b_1 A_1^2 \bar{A}_2 \exp(-i\sigma T_0) + 6b_1 A_1 \bar{A}_1 A_2 \exp(i\sigma T_0) - \\ 6b_1 A_1 \bar{A}_2 A_2 - 3b_1 \bar{A}_1 A_2^2 \exp(2i\sigma T_0) = 0 \\ -2A_2' i\omega_2 - 2u_3 A_2 i\omega_2 + 2u_2 A_1 i\omega_1 \exp(-i\sigma T_0) \\ + l_2 A_1 \exp(-i\sigma T_0) + 3b_2 A_1^2 \bar{A}_1 \exp(-i\sigma T_0) - \\ 3b_2 A_2^2 \bar{A}_2 - 3b_2 A_1^2 \bar{A}_2 \exp(-2i\sigma T_0) - \\ 6b_2 A_1 \bar{A}_1 A_2 + 6b_2 A_1 \bar{A}_2 A_2 \exp(-i\sigma T_0) + \\ 3b_2 \bar{A}_1 A_2^2 \exp(i\sigma T_0) + \frac{1}{2} f \exp(i\sigma_2 T_1 - i\sigma_1 T_1) = 0 \end{cases} \quad (17)$$

Where $A_n' = D_1 A_n$, $\sigma = \varepsilon\sigma_1$ .Introducing the polar form

$$A_n = \frac{1}{2} a_n \exp(i\theta_n) \text{ , } n = 1,2$$

into Eq.(17) and separating the equation into real and imaginary parts results in the following four state equations,

$$-a_1'\omega_1 - u_1 a_1 \omega_1 + u_1 \omega_2 a_2 \cos\gamma + \frac{1}{2} l_1 a_2 \sin\gamma + \frac{3}{8} b_1 a_2^3 \sin\gamma + \frac{3}{8} b_1 a_1^2 a_2 \sin\gamma - \frac{3}{8} b_1 a_1 a_2^2 \sin 2\gamma = 0 \quad (18)$$

$$a_1\theta_1'\omega_1 - u_1\omega_2 a_2 \sin\gamma + \frac{1}{2} l_1 a_2 \cos\gamma - \frac{3}{8} b_1 a_1^3 + \frac{3}{8} b_1 a_2^3 \cos\gamma + \frac{9}{8} b_1 a_1^2 a_2 \cos\gamma - \frac{3}{4} b_1 a_1 a_2^2 - \frac{3}{8} b_1 a_1 a_2^2 \cos 2\gamma = 0 \quad (19)$$

$$-a_2'\omega_2 - u_3 a_2 \omega_2 + u_2 a_1 \omega_1 \cos\gamma - \frac{1}{2} l_2 a_1 \sin\gamma - \frac{3}{8} b_2 a_1^3 \sin\gamma + \frac{3}{8} b_2 a_1^2 a_2 \sin 2\gamma - \frac{3}{8} b_2 a_1 a_2^2 \sin\gamma + \frac{1}{2} f \sin\varphi = 0 \quad (20)$$

$$a_2\omega_2\theta_2' + u_2 a_1 \omega_1 \sin\gamma + \frac{1}{2} l_2 a_1 \cos\gamma + \frac{3}{8} b_2 a_1^3 \cos\gamma - \frac{3}{4} b_2 a_1^2 a_2 - \frac{3}{8} b_2 a_1^2 a_2 \cos 2\gamma + \frac{9}{8} b_2 a_1 a_2^2 \cos\gamma - \frac{3}{8} b_2 a_2^3 + \frac{1}{2} f \cos\varphi = 0 \quad (21)$$

where $\gamma = \sigma T_0 + \theta_2 - \theta_1$, $\varphi = \sigma_2 T_1 - \sigma_1 T_1 - \theta_2$. At steady state, $a_n' = \theta_n' = 0$ and the average Eq.(18)-(21) reduced to the form

$$-u_1 a_1 \omega_1 + u_1 \omega_2 a_2 \cos\gamma - \frac{3}{8} b_1 a_1 a_2^2 \sin 2\gamma + \frac{3}{8} b_1 a_1^2 a_2 \sin\gamma + \frac{3}{8} b_1 a_2^3 \sin\gamma + \frac{1}{2} l_1 a_2 \sin\gamma = 0 \quad (22)$$

$$a_1\sigma_2\omega_1 - u_1\omega_2 a_2 \sin\gamma + \frac{1}{2}l_1 a_2 \cos\gamma - \frac{3}{8}b_1 a_1^3 + \frac{3}{8}b_1 a_2^3 \cos\gamma + \frac{9}{8}b_1 a_1^2 a_2 \cos\gamma - \frac{3}{4}b_1 a_1 a_2^2 - \frac{3}{8}b_1 a_1 a_2^2 \cos 2\gamma = 0 \quad (23)$$

$$-u_3 a_2\omega_2 + u_2 a_1\omega_1 \cos\gamma - \frac{3}{8}b_2 a_1^3 \sin\gamma + \frac{3}{8}b_2 a_1^2 a_2 \sin 2\gamma - \frac{1}{2}l_2 a_1 \sin\gamma - \frac{3}{8}b_2 a_1 a_2^2 \sin\gamma + \frac{1}{2}f \sin\varphi = 0 \quad (24)$$

$$a_2\omega_2(\sigma_2 - \sigma_1) + u_2 a_1\omega_1 \sin\gamma + \frac{1}{2}l_2 a_1 \cos\gamma + \frac{3}{8}b_2 a_1^3 \cos\gamma - \frac{3}{8}b_2 a_2^3 - \frac{3}{8}b_2 a_1^2 a_2 \cos 2\gamma - \frac{3}{4}b_2 a_1^2 a_2 + \frac{9}{8}b_2 a_1 a_2^2 \cos\gamma + \frac{1}{2}f \cos\varphi = 0 \quad (25)$$

which provide the steady state amplitudes and phases.

4. Simulation Analysis

To establish the accuracy of the average equations and illustrate the relationship between the tracked ambulance parameters and damping effect, simulation analysis is performed for the parameters: $M_1 = 180kg$, $M_2 = 2000kg$, $K_s = 217582N/m$, $C_1 = 4200N \cdot s/m$, $\beta = 0.1$, $K_2 = 2200000N/m$, $C_2 = 19540N \cdot s/m$, $f = 1500N$. For $\omega_1 = \omega_2 = 34.8rad/s$, the 1:1 resonance may occur. In Figure 4, we show the comparison between numerical solutions, obtained by Runge-Kutta, and perturbation solutions.

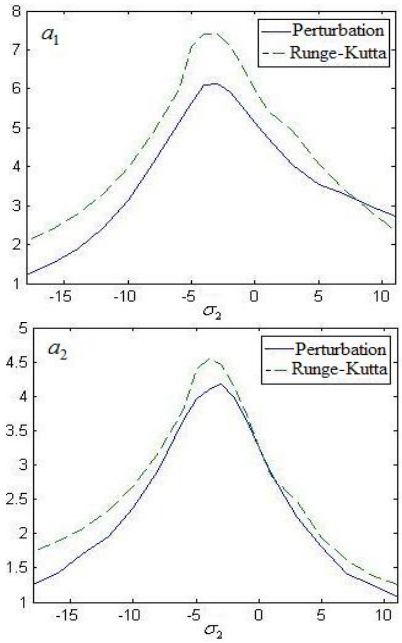


Fig.4 Amplitude-frequency curve

In Figure 4, the trend and resonance position of perturbation and numerical solutions are the same. But the amplitudes are different. Because we only use the first-order approximation, which don't affect our qualitative analysis of the behavior of the system dynamics. The damping system amplitude-frequency diagram is similar to the linear system, where jump phenomenon does not occur.

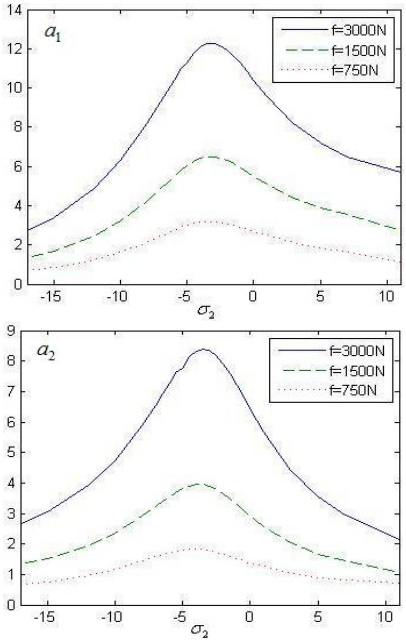


Fig.5 The influence of excitation amplitude

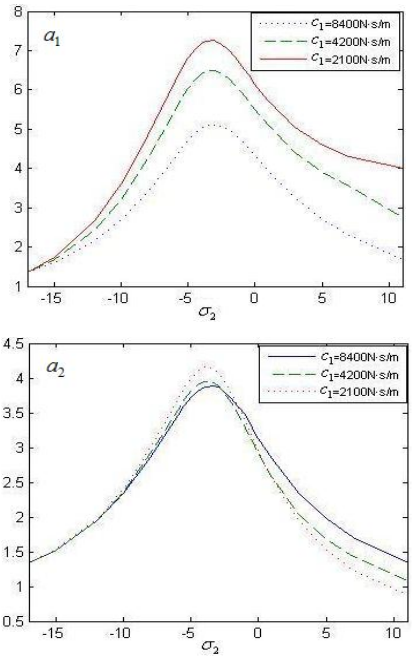
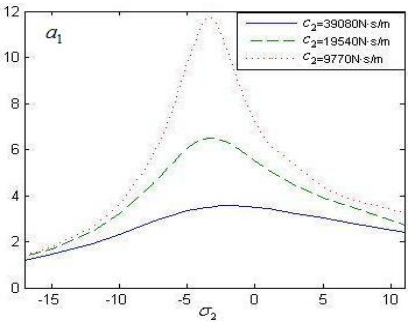


Fig. 6 The influence of C_1



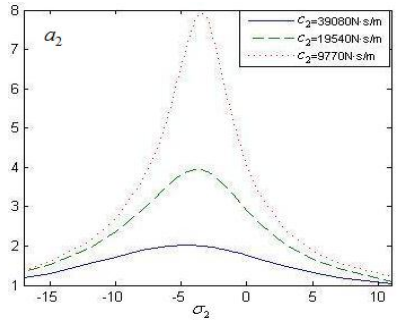


Fig. 7 The influence of C_2

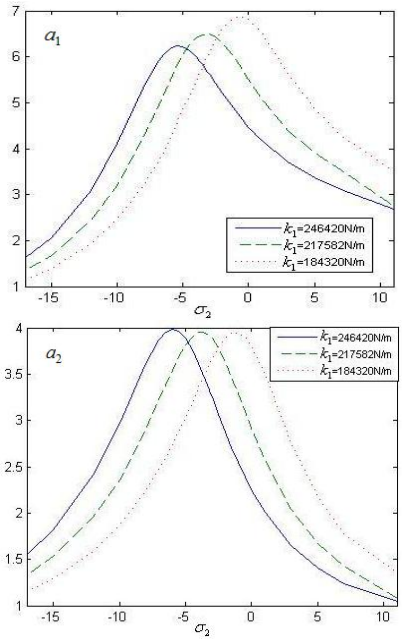


Fig. 8 The influence of K_s

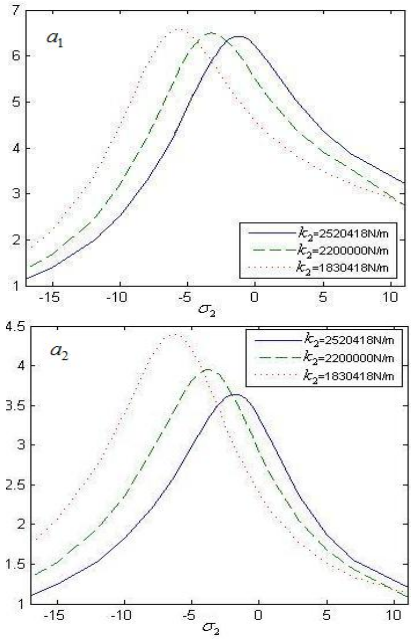


Fig. 9 The influence of K_2

Figure 5-9 show the impact of the tracked ambulance parameters on the system vibration, where a_1 and a_2 represent the vibrating amplitudes of the carriage and the stretcher base. As can be seen from Figure 5, the amplitude of the excitation force has a great impact on a_1 and a_2. Figure 6 and Figure 7 clearly demonstrate that he damping C_1 has a great impact on the amplitude of a_1, but little effect on the amplitude of a_2 and damping C_2 has a great impact on the amplitude of a_1 and a_2.The damping is greater and amplitude is smaller. Hence, increasing the damping, to some degree, is more effective to reduce vibration. Figure 8 and Figure 9 show that K_s only has a major impact on the amplitude of a_1 and K_2 only has a major impact on the amplitude of a_2 on the premise of meeting 1:1 internal resonance approximately. But both K_s and K_2 affect the resonance frequency greatly. Increasing K_2 or decreasing K_2 can increase the resonance frequency, which is beneficial to reduce vehicle vibration[10]. With comprehensive comparison of Figures 6 to 9, damping has a great impact on the amplitude of vibration and stiffness has a great impact on resonance frequency.

5. Stability Analysis

In order to analyze the stability of the system in the primary resonance, we need convert the average equations in polar form into a rectangular form by introducing $p_1 = a_1\cos(\gamma+\varphi)$, $q_1 = a_1\sin(\gamma+\varphi)$, $p_2 = a_2\cos\varphi$, $q_2 = a_2\sin\varphi$ [11-12]result in

$$\dot{p}_1 = -u_1 p_1 + u_1 p_2 - \frac{l_1}{2\omega_1}q_2 - \frac{3}{8\omega_1}b_1(p_2^2+q_2^2)q_2 - \sigma_2 q_1 - \frac{3}{8\omega_1}b_1(p_1^2+q_1^2)q_2 - \frac{3}{8\omega_1}b_1(q_1p_2^2 - q_2^2q_1 - 2p_1p_2q_2) - \frac{3}{4\omega_1}b_1(p_1p_2+q_1q_2)q_1 + \frac{3}{4\omega_1}b_1(p_2^2+q_2^2)q_1 + \frac{3}{8\omega_1}b_1(p_1^2+q_1^2)q_1 \quad (26)$$

$$\dot{p}_2 = -u_3 p_2 + u_2 p_1 - \frac{l_2}{2\omega_2}q_1 - \frac{3}{8\omega_2}b_2(p_1^2+q_1^2)q_1 - (\sigma_2-\sigma_1)q_2 + \frac{3}{8\omega_2}b_2(2p_1q_1p_2 - p_1^2q_2 + q_1^2q_2) - \frac{3}{8\omega_2}b_2(p_2^2+q_2^2)q_1 + \frac{3}{8\omega_2}b_2(p_2^2+q_2^2)q_2 + \frac{3}{4\omega_2}b_2(p_1^2+q_1^2)q_2 - \frac{3}{4\omega_2}b_2(p_1p_2+q_1q_2)q_2 \quad (27)$$

$$\dot{q}_1 = -u_1 q_1 + u_1 q_2 + \frac{l_1}{2\omega_1}p_2 + \frac{3}{8\omega_1}b_1(p_2^2+q_2^2)p_2 + \sigma_2 p_1 + \frac{3}{8\omega_1}b_1(p_1^2+q_1^2)p_2 - \frac{3}{8\omega_1}b_1(p_1p_2^2 - q_2^2p_1 + 2q_1p_2q_2) + \frac{3}{4\omega_1}b_1(p_1p_2+q_1q_2)p_1 - \frac{3}{4\omega_1}b_1(p_2^2+q_2^2)p_1 - \frac{3}{8\omega_1}b_1(p_1^2+q_1^2)p_1 \quad (28)$$

$$\dot{q}_2 = -u_3 q_2 + u_2 q_1 + \frac{l_2}{2\omega_2} p_1 + \frac{3}{8\omega_2} b_2 (p_1^2 + q_1^2) p_1 + \frac{f}{2\omega_2} - \frac{3}{8\omega_2} b_2 (p_1^2 q_1 - q_1^3 + 2 p_1 q_1 q_2) + \frac{3}{8\omega_2} b_2 (p_2^2 + q_2^2) p_1 + (\sigma_2 - \sigma_1) p_2 - \frac{3}{8\omega_2} b_2 (p_2^2 + q_2^2) p_2 - \frac{3}{4\omega_2} b_2 (p_1^2 + q_1^2) p_2 + \frac{3}{4\omega_2} b_2 (p_1 p_2 + q_1 q_2) p_2 \quad (29)$$

Where the average equations become more complex and the exact analytical solution cannot be obtained. At steady state, $\dot{p}_1 = 0$, $\dot{p}_2 = 0$, $\dot{q}_1 = 0$, $\dot{q}_2 = 0$ and we use Newton Method to calculate the value of the equilibrium point of the average Eq.(26)-(29) by repeatedly changing the initial value of the equilibrium point. There are three sets of equilibrium points, as follows

$$\phi_1 = \{-3.9041, -1.2622, 2.0696, 1.8797\}$$
$$\phi_2 = \{12.1124, 7.7186, -13.7048, -10.4106\}$$
$$\phi_3 = \{-10.9811, -6.7073, 11.9741, 8.9831\}$$

The stability of the system at the equilibrium point is governed by the eigenvalue of the Jacobian matrix of Eq.(26)-(29) based on the singularity theory. The eigenvalues are obtained:

$\lambda_1 =$ {-13.5988 +22.7426i, -13.5988 -22.7426i,
-4.0362 + 5.5237i,-4.0362 - 5.5237i}

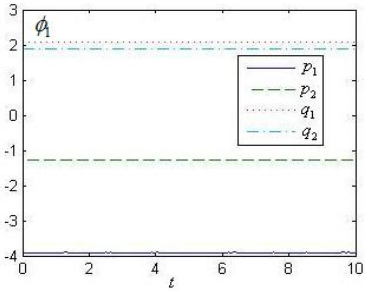
$\lambda_2 =$ {-31.75 + 120.17i,30.18,-31.75 - 120.17i,-1.95}

$\lambda_3 =$ {-28.95 + 103.88i,-1.70, -28.95 - 103.88i,24.33}

The Eq.(30) is the Jacobian matrix of the Eq.(26)-(29) at equilibrium point, where the expressions of n_{ij} ($i = 1 \ldots 4$, $j = 1 \ldots 4$) are given in appendix.

$$A = \begin{bmatrix} n_{11} & n_{12} & n_{13} & n_{14} \\ n_{21} & n_{22} & n_{23} & n_{24} \\ n_{31} & n_{32} & n_{33} & n_{34} \\ n_{41} & n_{42} & n_{43} & n_{44} \end{bmatrix} \quad (30)$$

After singularity analysis, the system is only stable in the first equilibrium point. Since there is only one stable equilibrium point, the jump phenomenon does not occur. Use Runge-Kutta method to validate the singularity analysis. The Figure10 presents the final stable position of the Eq.(26)-(29) whose the initial values are the three equilibrium points.



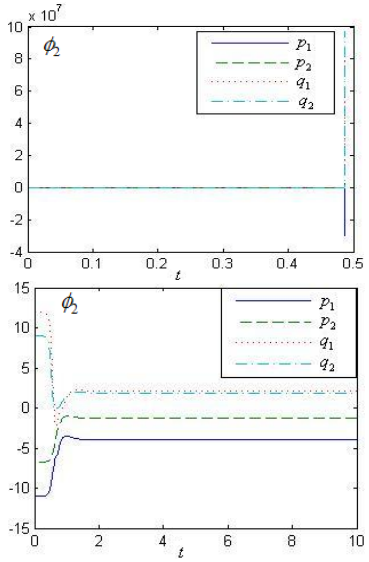


Fig. 10 System stability location

Figure 10 clearly illustrates that the system is only stable in the first equilibrium point, which is in line with the actual system and diverges to infinity(Figure b) or converge to the stable equilibrium point(Figure c) at unstable equilibrium point. Therefore, the system is impossible to maintain a stable state in the unstable equilibrium point.

6. Conclusion

This paper established the dynamics model of a tracked ambulance damping system containing three nonlinear terms. We used Multiple Scales Method to investigate the dynamics model and obtain the average equations. The average equations were verified with the actual parameters. The influence of damping system parameters for the damping effect as well as the stability of the damping system were analyzed. The result explained the reasons that there is no jump phenomenon. This analysis method is suitable for multi-degree-of-freedom bearing motion system, particularly suitable for vehicle. The research results are valuable for the vehicle damping system design as well as forecast the damping system dynamic behavior.

Appendix:

$$n_{11} = -u_1 - \frac{3b_1 q_2 p_1}{4\omega_1} + \frac{3b_1 q_2 p_2}{4\omega_1} - \frac{3b_1 q_1 p_2}{4\omega_1} + \frac{3b_1 q_1 p_1}{4\omega_1}$$

$$n_{12} = u_1 - \frac{3b_1 q_2 p_2}{4\omega_1} - \frac{3b_1 (2q_1 p_2 - 2p_1 q_2)}{8\omega_1} - \frac{3b_1 q_1 p_1}{4\omega_1} + \frac{3b_1 q_1 p_2}{2\omega_1}$$

$$n_{13} = -\frac{3b_1 q_2 q_1}{2\omega_1} - \frac{3b_1 (p_2^2 - q_2^2)}{8\omega_1} - \sigma_2 - \frac{3b_1 (p_1 p_2 + q_1 q_2)}{4\omega_1} + \frac{3b_1 (p_2^2 + q_2^2)}{4\omega_1} + \frac{3b_1 q_1^2}{4\omega_1} + \frac{3b_1 (p_1^2 + q_1^2)}{8\omega_1}$$

$$n_{14} = -\frac{l_1}{2\omega_1} - \frac{3b_1 q_2^2}{4\omega_1} - \frac{3b_1 (p_2^2 + q_2^2)}{8\omega_1} - \frac{3b_1 (p_1^2 + q_1^2)}{8\omega_1} - \frac{3b_1 q_1^2}{4\omega_1} + \frac{3b_1 (p_1 p_2 + q_1 q_2)}{4\omega_1} + \frac{3b_1 q_2 q_1}{2\omega_1}$$

$$n_{21}=u_2-\frac{3b_2p_1q_1}{4\omega_2}+\frac{3b_2(q_1p_2-p_1q_2)}{4\omega_2}+\frac{3b_2p_1q_2}{2\omega_2}-\frac{3b_2p_2q_2}{4\omega_2}$$

$$n_{22}=-u_3+\frac{3b_2p_1q_1}{4\omega_2}-\frac{3b_2p_2q_1}{4\omega_2}+\frac{3b_2p_2q_2}{4\omega_2}-\frac{3b_2p_1q_2}{4\omega_2}$$

$$n_{23}=-\frac{l_2}{2\omega_2}-\frac{3b_2q_1^2}{4\omega_2}-\frac{3b_2(p_1^2+q_1^2)}{8\omega_2}-\frac{3b_2(p_2^2+q_2^2)}{8\omega_2}+\frac{3b_2q_1q_2}{2\omega_2}+\frac{3b_2(p_1p_2+q_1q_2)}{4\omega_2}-\frac{3b_2q_2^2}{4\omega_2}$$

$$n_{24}=\frac{3b_2(-p_1^2+q_1^2)}{8\omega_2}-\frac{3b_2q_1q_2}{2\omega_2}-\sigma_2+\sigma_1+\frac{3b_2q_2^2}{4\omega_2}+\frac{3b_2(p_2^2+q_2^2)}{8\omega_2}+\frac{3b_2(p_1^2+q_1^2)}{4\omega_2}-\frac{3b_2(p_1p_2+q_1q_2)}{4\omega_2}$$

$$n_{31}=\frac{3b_1p_2p_1}{2\omega_1}-\frac{3b_1(p_2^2-q_2^2)}{8\omega_1}+\sigma_2+\frac{3b_1(p_1p_2+q_1q_2)}{4\omega_1}-\frac{3b_1(p_2^2+q_2^2)}{4\omega_1}-\frac{3b_1p_1^2}{4\omega_1}-\frac{3b_1(p_1^2+q_1^2)}{8\omega_1}$$

$$n_{32}=\frac{l_1}{2\omega_1}+\frac{3b_1p_2^2}{4\omega_1}+\frac{3b_1(p_2^2+q_2^2)}{8\omega_1}+\frac{3b_1(p_1^2+q_1^2)}{8\omega_1}+\frac{3b_1p_1^2}{4\omega_1}-\frac{3b_1(p_1p_2+q_1q_2)}{8\omega_1}-\frac{3b_1p_2p_1}{2\omega_1}$$

$$n_{33}=-u_1+\frac{3b_1q_1p_2}{4\omega_1}-\frac{3b_1q_2p_2}{4\omega_1}+\frac{3b_1q_2p_1}{4\omega_1}-\frac{3b_1q_1p_1}{4\omega_1}$$

$$n_{34}=u_1+\frac{3b_1q_2p_2}{4\omega_1}-\frac{3b_1(q_1p_2-p_1q_2)}{4\omega_1}+\frac{3b_1q_1p_1}{4\omega_1}-\frac{3b_1q_2p_1}{2\omega_1}$$

$$n_{41}=\frac{l_2}{2\omega_2}+\frac{3b_2p_1^2}{4\omega_2}+\frac{3b_2(p_1^2+q_1^2)}{8\omega_2}+\frac{3b_2(p_2^2+q_2^2)}{8\omega_2}-\frac{3b_2(p_1p_2+q_1q_2)}{4\omega_2}-\frac{3b_2p_1p_2}{2\omega_2}+\frac{3b_2p_2^2}{4\omega_2}$$

$$n_{42}=\frac{3b_2p_1p_2}{2\omega_2}+\sigma_2-\sigma_1-\frac{3b_2p_2^2}{4\omega_2}-\frac{3b_2(p_2^2+q_2^2)}{8\omega_2}-\frac{3b_2(p_1^2+q_1^2)}{4\omega_2}+\frac{3b_2(p_1p_2+q_1q_2)}{4\omega_2}$$

$$n_{43}=u_2+\frac{3b_2p_1q_1}{4\omega_2}-\frac{3b_2(p_1^2-3q_1^2+2p_1q_2)}{8\omega_2}-\frac{3b_2q_1p_2}{2\omega_2}+\frac{3b_2p_2q_2}{4\omega_2}$$

$$n_{44}=-u_3-\frac{3b_2p_1q_1}{4\omega_2}+\frac{3b_2p_1q_2}{4\omega_2}-\frac{3b_2p_2q_2}{4\omega_2}+\frac{3b_2p_2q_1}{4\omega_2}$$

Acknowledgment

An acknowledgment should be shown to Dr. Weihua Su who helped us during the writing of this paper.

Reference

[1] Shafic S. Oueini, Char-ming Chin, Ali H. Nayfeh, "Dynamics of a Cubic Nonlinear Vibration Absorber", Nonlinear Dynamics, 1999, Vol.20,No.3, pp.283-295.

[2] Shi Peiming, Liu Bin, Jiang Jinshui, "Stability and approximate solution of a relative-rotation nonlinear dynamical system with coupled terms",Acta Physical Sinica, 2009, Vol.58, No.4, pp.2147-2154.

[3] Ali H. Nayfeh, Walter Lacarbonara, "On the Discretization of Distributed-Parameter Systems with Quadratic and Cubic Nonlinearities", Non-linear Dynamics,1997,Vol.13,No.3, pp.203-220.

[4] Christopher L.Lee, Noel C.Perkins, "Nonlinear Oscillations of Suspended Cables Containing a Two-to-One Internal Resonance", Nonlinear Dynamics, 1992, Vol.3, No.6, pp.465-490.

[5] Li Jian, Guo Xinghui, Yang Kun, et al, "Study on The Nonlinear Vibration of Axially Moving Cylindrical Shells Made from Composites",Chinese Journal of Solid Mechanics, 2011, Vol.32, No.2, pp.176-185.

[6] Zhang Xin,Sun Dagang,Song Yang, et al, "Analysis of Damping Vibration Reduction Performance of Viscoelastic Shocker Absorber under Low Frequency and Heavy Loading", China Mechanical Engineering, 2012, Vol.23, No.14, pp.1651- 1656.

[7] LI Xinye, Zhang Lijuan, Zhang Huabiao, "Forced vibration of a gyroscope system and its delayed feedback control", Journal of Vibration and Shock, 2012, Vol.31, No.9, pp.63-68.

[8] Tsuyoshi Inoue, Yukio Ishida, "Nonlinear forced oscillation in a magnetically levitated system: the effect of the time delay of the electromagnetic force", Nonlinear Dynamics, 2008, Vol.52, No.1-2, pp.103-113.

[9] Peng Xian, Zhang Shixiang, "Nonlinear Resonance Response Analysis of a Kind of Passive Isolation System with Quasi-Zero Stiffness", Journal of Human University (Natural Sciences), 2011, Vol.38, No.8, pp.34-39.

[10] Su Chen, Xu Xinxi, Gao Zhenhai, et al, "Analysis on Two-Level Damping Efficiency and Recumbent Comfort for Tracked Emergency Ambulance", Journal of Vibration,Measurement &Diagnosis, 2012, Vol.32, No.5, pp.754-857,869.

[11] Liu Shuang, LI Yangshu, Liu Bin, et al, "Parametric Vibration Analysis and Control in Coupling Rotating Mechanical Drive System", China Mechanical Engineering, 2012, Vol.23, No.12, pp. 1461-1466.

[12] Liu Haoran, Zhu Zhanlong, Shi Peiming, et al, "Stability control of a coupled nonlinear torsional vibration system", Journal of Vibration and Shock, 2011, Vol.30, No.9, pp.140-144.

Meng Yang received the bachelor degree from Tianjin University, Tianjin, China, in 2011 and is now a postgraduate student of Institute of Medical Equipment, Academy of Military Medical Sciences, Tianjin, China. His research interests cover the nonlinear dynamics and

ergonomics.

Xinxi Xu received a master degree from Tianjin University, in 1989 and a Ph.D.degree from Tianjin University in 2008. He is now the research associate and doctoral tutor of Institute of Medical Equipment, Academy of Military Medical Sciences, Tianjin, China. His research interests cover the structural dynamics, vehicle NVH technology and ergonomics.

Chen Su received a master degree form Military Transportation University, in Tianjin, China, in 2008 and a Ph.D.degree from Institute of Medical Equipment, Academy of Military Medical Sciences, in 2011. He is now a research assistant of the Institute of Medical Equipment.

Study of Verification of the Reputation Scaling Module of Trust Management System

Yonghui CAO[1, 2]
1, School of Economics & Management, Henan Institute of Science and Technology, Xin Xiang, 453003 ,China
2, School of Management, Zhejiang University, Hang Zhou,310058 ,China

Abstract
The trust management system (TMS) developed through this research effectively implements a decentralized access and permission management scheme. The Reputation Scaling module (RSM) was the heart of the TMS. Our Reputation Scaling module applied different levels of trust to reports and observations. RSM advanced the current state of the art by introducing a reputation scaling mechanism that maintained a memory of past behavior grades and observers. The RSM used this historical knowledge to apply the observer's current RI to any behavior grade he might have made. In addition to dynamic FI weighting, the RSM's 3Win reputation scaling equation provided a more conservative approximation, allowing smaller fluctuations in the node's reputation than other equations currently in use. This testing concluded that this conservative approach benefited the network because it forced nodes to sustain positive behavior for longer periods than was necessary in the WMA or other reputation management mechanisms to achieve the same positive reputation. Each node gathered and processed feedback to calculate a usable RI for its peers. The TMS implemented an Trust model to represent the reputations that were compiled by a node on each of its peers.

Keywords:*Reputation Scaling, Inter-networking Mobility, Dynamic Collaborative Environment*

1 .Verification testing Goals and Objectives

Verification in engineering or quality management systems, it is the act of reviewing, inspecting or testing, in order to establish and document that a product, service or system meets regulatory or technical standards. Verification determined that each module had correctly transformed its inputs into the expected output. This testing involved isolating each function of each module by the arrangement of input and processing parameters. Specific outputs were then analyzed to check their correspondence to expected results. In general, verification testing revealed that the modules worked in accordance with the requirements. Verification theory is a theory relating the meaning of a statement to how it is verified.

Simulation models are increasingly being used in problem solving and to aid in decision-making. The developers and users of these models, the decision makers using information obtained from the results of these models, and the individuals affected by decisions based on such models are all rightly concerned with whether a model and its results are "correct". This concern is addressed through model validation and verification. Once basic verification was complete, performance boundary analysis was conducted to ascertain under which conditions the module operated best and under which conditions performance was impaired. Once these expectations were met, the modules were combined and the system validated. The following sub-sections provide the analysis of verification testing. These sub-sections follow a standard methodology (Bryce, Dimmock et al. 2005) is that： First，defining the role of each component. Second， analyzing the component to determine how module failure or impaired performance influences overall system functioning.

2. Reputation Scaling

The trust management system (TMS) developed through this research effectively implements a decentralized access and permission management scheme. Each resource owner uses the linked characteristics of identity, reputation, and risk to make access decisions. Because the TMS tracks a user’s behavior, using past behavior as future performance, no a priori user configuration is required. The TMS also offers a unique ability to enforce multiple access levels without the burden of implementing and managing multiple cryptographic keys or hierarchies of roles. A node provides its peers customized views of its contents and services based on its individual trust profile and the peer’s trustworthiness. As peers’ reputations change, their access changes to safeguard the node’s

resources for those peers that have shown themselves to contribute to the node's and the coalition's goals.
Situational Trust described the degree of trust that an individual was prepared to trust any other person in a given situation. This trust was formed upon the intention to extend trust in a particular situation, regardless of what the person knew or did not know about the other party in the situation. It was suggested that this type of trust occurred when the trusting party stood to gain with very little attendant risk. Situational trust was different than System trust because there were no implied structural or system safeguards. It was, in short, an individually conceived situational strategy and did not involve an evaluation of the trustworthiness of the other party.

The Reputation Scaling module (RSM) was the heart of the TMS. The RSM's purpose was to implement a quantitative method for aggregating behavior feedback items (FIs) to generate a reputation value for each associate. A node used this value, called a Reputation Index (RI), as a measure of the trustworthiness he had of a specific network peer based on the peer's previous behavior. The RSM responded to the fluid nature of the Dynamic Collaborative Environment (DCE) by re-evaluating the source of each behavior grade before using the grade as input to the reputation scaling equation. The end result of this equation was a substantiated reputation index (RI) that was provided to the TMS. The RI was then compared against the trust thresholds to determine whether or not the system should extend trust and grant access to the requested resource.
If the RSM failed during operation, the TMS would have no way of processing behavior information or judging an associate's trustworthiness. The TMS would have to abandon a trust-based approach and resort to pre-configured access control methods, such as RBAC or identity-based mandatory access controls (MAC.) Neither method was considered acceptable in a DCE, for reasons discussed elsewhere in this research.

The RSM's approach had the following characteristics. It:

(1) Was node-centric, so that each node calculated only the reputations of the peers it was concerned with;
(2) Weighted FI to emphasize current behavior trends while accounting for past performance;
(3) Merged behavior reports (first-hand experiences) with observations (second-hand remarks);
(4) Aged FI over time to remove outdated behavior information;
(5) Enabled nodes to recover their reputation by demonstrating desirable behavior.

Verification of the RSM required the reputation scaling equation to produce a RI that conservatively estimates the observed peer's actual behavior. Next section compares the RSM's performance to the actual behavior grade and commonly used estimation techniques, such as the exponential weighted moving average.

3 .General Testing

We have developed a system where user nodes cooperated to exchange behavior reports and establish a record of each node's behavior history. This history, based on reports and observations, was expressed as a reputation index (RI). The RI, with evidence in the form of signed FIs, provided an expectation of their partner's behavior before entering into or dissolving an SA. By providing an indication of each other's trustworthiness, nodes avoided misbehaving nodes. The TMS that is installed on each node. In the following sections, this paper discusses how the TMS implements each of McKnight and Chervany's constructs to produce an access control decision. The RSM was tested to verify that the RI output by the module displayed a hysterisis effect with respect to its input, as shown in Figure 1. As the recording node (e.g., Joe) moved through different network conditions, the RSM was expected to produce results that accurately reflected the original input as the recording node (e.g., Joe) moved through different network conditions. In testing the RSM, the 3Win method was compared against the original input and the exponential weighted moving average equation used by Buchegger (Buchegger and Le Boudec 2002b). Comparing 3Win to the actual input and an exponential weighted moving average (WMA), Figure 1 shows how the RSM produced an RI that lagged behind the changes in behavior, as desired.

In the subsequent tests, we wanted to investigate the RMS's response in mobile situations. Interactivity traces were constructed using MATLAB and a Random Waypoint model. A 100 node network was constructed inside a 1000 x 1000 meter area. Each simulation was run for 1000 seconds. Humans developed a concept of reputation as an aggregation of trust information. They used this concept to predict the actions of others based on historical behavior information gained through personal interaction or the shared observations of peers. Researchers pointed out that reputation could be utilized in a virtual society, such as a MANET, to make up for the lack of the physical, interpersonal clues that humans use to determine trustworthiness.

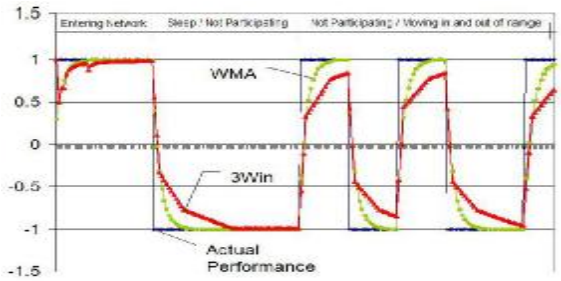


Fig. 1 the Hysterisis Effect of the 3Win Method

Scenarios were designed to test the ability of a RSM to:

- Identify selfish behavior,
- Allow a node to rehabilitate its reputation following a period of poor connectivity, and
- Respond to nodes that try to denigrate other nodes with undeserved negative performance observations.

In addition to these three performance measures, testing also needed to gauge the impact of dynamic FI weighting. Dynamic FI weighting linked and maintained the identity of the reporting node with the observation. The reporting node's current reputation (i.e., the RI value at the time of the calculation) was applied to the observation each time the RI was calculated. This method allowed the reputation scaling method to consider the changes in observers' reputation values during the calculation of the RI.

In each test, a node (e.g., Joe) received FI generated from observations and formal reports. The data set represented a period of approximately one hour of operation in the test bed and was distributed in the interval [-1,1] based on the previously described movement and behavior-based scenarios. Nodes provided observations on a 10-12 second interval. In all of the following graphs, the X-axis represents the passage of time and the Y-axis is the value of the node's RI.

The "unreliable node" scenario, shown in Figure 2, tested the RSM's flexibility in allowing a formerly unreliable node (e.g., Bob) to rehabilitate the reputation value that Joe maintained for him as Bob moved in and out of the Joe's transmission range. After a period of mobility, Bob relocated to a position with more stable connectivity and resumed cooperating with the network. Joe's associates observed and commented on Bob's behavior, providing the behavior grading that Joe fed to his RSM.

Because the periods of positive and negative observations were balanced, it was expected that the RSM would allow Bob to rebuild his reputation as he moved into the operating range of his peers. Figure 2 illustrates the performance of the two reputation scaling methods in the "unreliable node" scenario. Of note is the sharply fluctuating, optimistic curve that is produced by the WMA method. The 3Win mechanism produced an RI curve that was a smoother and more conservative approximation of the input while also allowing reputation rehabilitation.

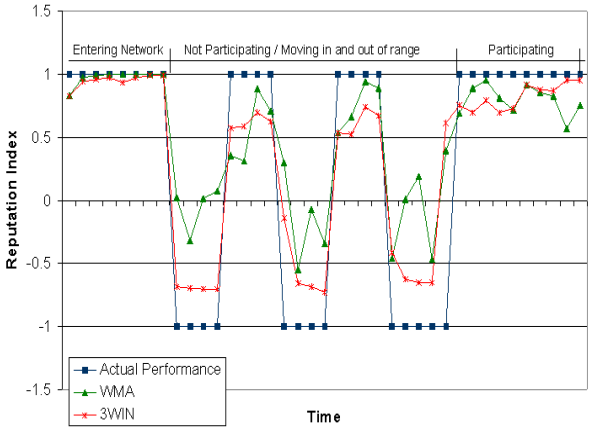


Fig. 2 Performance of Reputation-Scaling Mechanisms in an "Unreliable Node" Scenario

Points where the 3Win curve departed drastically from the WMA showed the effects of dynamically weighting observations. Because of its lack of history, the WMA method could only weight the most current observation and then only at the time it was applied to the reputation calculation. The 3Win method reapplied the weights of the observers at each calculation. As the observers' reputations changed, the value of their recommendations (in the form of FIs), changed as well.

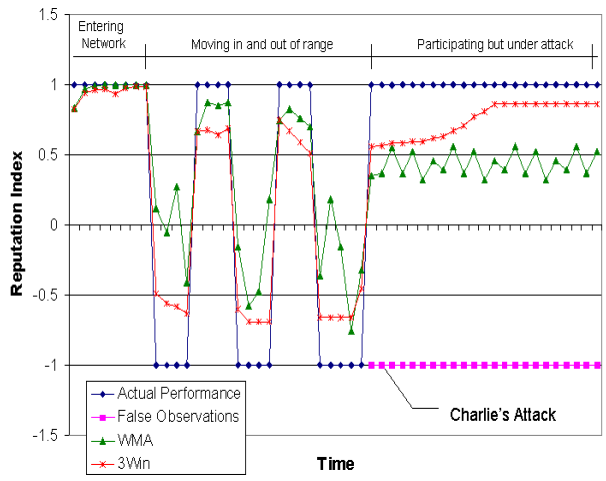


Fig. 3 Performance of Reputation-Scaling Mechanisms in a "Smear" Scenario

The "smear" scenario tested the efficiency of the reputation management mechanisms at resisting attacks on a node's reputation (see Figure 3). The observed node (again we'll use Bob) moved into the observer's (e.g., Joe's) operational area and should have been receiving positive feedback but Charlie tried to smear Bob's reputation by maliciously reporting negative feedback. The results show that the reputation management methods resisted the attack by dynamically weighting the FI, effectively diminishing Charlie's ability to impact Bob's reputation. When Charlie (the smearing node) was determined to be malicious, his reports were discounted and allowed Bob's reputation to recover.

The 3Win method provided a more conservative approximation, allowing smaller fluctuations in the node's reputation than the WMA. This testing concluded that this conservative approach benefited the network because it forced nodes to sustain positive behavior for longer periods than was necessary in the WMA or other reputation management mechanisms.

4 .Inter-networking Mobility Testing

Each node gathered and processed feedback to calculate a usable RI for its peers. The TMS implemented an Trust model to represent the reputations that were compiled by a node on each of its peers. This trust type was node specific, so that the trust of one node to another was direct and not transitive. The following summarizes the trust model:

- Trust was context dependent.
- Trust had positive and negative degrees of trustworthiness.
- Trust was expressed in continuous values, as described by Marsh.
- Trust was based on experiences and observations between individuals.
- Trust information was exchanged between nodes.
- Trust was subjective. Nodes calculated different reputation values for the same observed node.
- Trust was dynamic and was modified, in a positive or negative direction, based on new observations and reports.

Once the reports and observations had been gathered, they were processed to provide a meaningful value that a node used for its trustworthiness evaluation. The reputation value needed to give a conservative approximation of the feedback input. We also wanted to emphasize current behavior while aging older input to diminish its impact on the reputation calculation. As in CORE and CONFIDANT, a node maintained a reputation value for each TP. Nodes entered the network with a reputation value of 0, a basic level of trust. Our expectation was that a node would desire a positive reputation. A node with a negative reputation would be isolated as nodes refused to interact with it.

Our Reputation Scaling module applied different levels of trust to reports and observations. Nodes placed full trust in KMS reports. On the other hand, periodic observations from other peers and friends were weighted using the reporting node's reputation (RIx) before into the reputation calculation. These weighted observations were called Feedback Items (FIs).

The test, displayed in Figure 4, showed the system acting on the test vignettes designed in Appendix A. The scenario began in the Tactical Operations Center (TOC), a medium density network environment made up of "good" users. Joe entered the network and associated himself with peers in the area, such as Alice. The presence of unreliable user behavior increased the trust thresholds but did not require any associations to be dissolved. When a "bad" user, Natasha, joined, she was not extended trust based on the information in the referrals received from Alice and other "good" users.

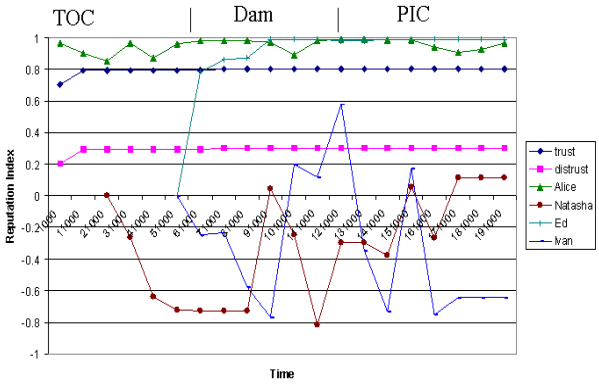


Fig. 4 Inter-Network Mobility Effects on Reputation Scaling

At time mark 61,000, Joe left the TOC and arrived at the Balcony Falls Dam to conduct a site survey. The Dam represented a sparse density, low population network environment of unreliable or resource-constrained users. Joe encountered Ed and, again based on referrals received from other associates, extended him trust while, at the same time, denying trust to Ivan, another "bad" user.

Finally, Joe entered the Public Information Center (PIC), a large, dense network of "bad" users. Having transitioned through two previous network environments, it was

important that Joe's RSM be able to continue to differentiate between desirable and undesirable associates. In other words, the RSM had to remember a sufficient amount of previous activity to re-establish or maintain association with people he met throughout the day. Although the presence of other "bad" users enabled Natasha and Ivan to improve their RIs slightly, their previous behavior still prevented them from gaining access to Joe's resources. At the same time, the "good" associates were maintained.

Figure 5 illustrates a situation where Joe's RSM was presented with a pair of "bad" users (Ivan and Natasha). This pair was actively colluding to subvert the network. The collusion was effected by having Ivan and Natasha only give each other positive observations while, at the same time, either not provide behavior grades or have them provide negative grades when none were warranted. In this manner, their plan was to allow Natasha to gain a foothold in the network by constructing associations with good users and then use these associations to insinuate Ivan, her confederate, into the network.

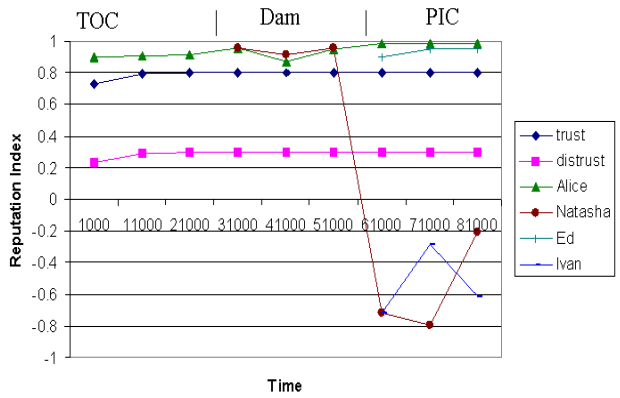


Fig. 5 Inter-Network Mobility with Collusion

Joe entered the network at the TOC, as in the previous test. He arrived at the Dam and was introduced to Natasha, who was performing as a good user. This association continued when Joe got to the PIC but he then observed that Natasha changed her behavior after she introduced her confederate, Ivan. Joe's RSM, recognizing that Natasha's behavior had become undesirable, lowered her RI as well as the RIs of those that she had introduced or referred. While at the PIC, Joe dissolved his association with Natasha but maintained her behavior history in his TS to enable him to weight her previous observations with her now unacceptably low RI. Furthermore, despite the collusion, Joe was still able to access Ed as a trustworthy sort and allowed him access.

These tests supported the validity of the RSM's performance. The RSM demonstrated the ability to maintain RIs on individual associates in varying network environments. Additionally, the use of dynamic weighting sped the process of identifying and isolating "bad" users by applying current reputations to RI calculation rather than depending on the observer's RI from the time of the observation.

Where Interpersonal trust was dependent upon peer behavior trends and System trust was determined through an evaluation of system behavior tendencies, Situational trust was independent of the behavior of other users altogether. This type of trust used the trust store, representing the user's memory of previous peers and situations, to determine what action it would take. A situational trust decision was predicated on remembering a previous decision that had yielded a positive outcome, regardless of the behavior of peers that may or may been involved.

5 .Contributions and Conclusion

Trust management offers the ability to make access control decisions in mobile ad-hoc collaborative environments without the need for pre-configuration or centralized management. By linking a node's identity to observations on its performance, its peers can calculate its reputation and evaluate its trustworthiness. Through a process of introduction, nodes share performance observations and are able to calculate reputations of newly encountered nodes in a peer to peer manner

The RSM advanced the current state of the art by introducing a reputation scaling mechanism that maintained a memory of past behavior grades and observers. The RSM used this historical knowledge to apply the observer's current RI to any behavior grade he might have made. This reevaluation was called dynamic FI weighting and it proved very successful in isolating not only misbehaving nodes but also nodes that might be colluding with them. In addition to dynamic FI weighting, the RSM's 3Win reputation scaling equation provided a more conservative approximation, allowing smaller fluctuations in the node's reputation than other equations currently in use. The RSM performed as expected and provided the TMS with a basis for trust decisions. It maintained correct reputation assessments on associates regardless of the characteristics of the other network users.

References

[1] Laird, J. and R. Wray (2011). Variability in Human Behavior Modeling for Military Simulations. Proceedings of the 2003 Conference on Behavior Representation in Modeling and Simulation (BRIMS), Scottsdale, AZ, Pp. 1-10.

[2] Krishnan, R., M. Smith, et al. (2011). "Economics of Peer-to-Peer Networks." Journal of Information Technology Theory and Application 5(3): 31-44

[3] Keser, C (2010). "Experimental games for the design of reputation management systems." IBM Systems Journal 42(3): 498-506.

[4] Commander Taco. (2010).”Slashdot." Retrieved 15 December, 2005, from http://slashdot.or

[5] KuroShin. (2010). "KuroShin." Retrieved 15 December, 2005, from http://www.kuro5hin.org.

[6] Powazek, D. (2010). "Gaming the system: How moderation tools can backfire." Retrieved 15 December, 2005, from http://designforcommunity.com/essay8.html.

[7] Ben Salem, N., J.-P. Hubaux et al. (2009). Reputation-based Wi-Fi deployment protocols and security analysis. Proceedings of the 2nd ACM International Workshop on Wireless Mobile Applications, Philadelphia, PA, Pp. 29-40.

[8] Bryce, C., N. Dimmock, et al. (2009). Towards an Evaluation Methodology for Computational Trust Systems. Proceedings of the Third International Conference in Trust Management (iTrust 2005), Paris, FR, Pp. 289-304.

[9] Lo Presti, S., M. Butler, et al. (2009). A Trust Analysis Methodology for Pervasive Computing

Systems. Trusting Agents for trusting Electronic Societies. R. Falcone, S. Barber, J. Sabater and M. Singly Springer.

[10]A. Datta, S. Quarteroni, and K. Aberer, "Autonomous Gossiping: A self-organizing epidemic algorithm for selective information dissemination in mobile ad-hoc networks.," Ecole Polytechnique Federale de Lausanne 2010.

[11] W. J. Adams, G. C. Hadjichristofi, and N. J.Davis, "Calculating a Node's Reputation in a Mobile Ad-Hoc Network," presented at the 24th IEEE International Performance Computing and Communications Conference (IPCCC 2005),Phoenix, AZ, 2005.

[12] R. Jain, The Art of Computer Systems Performance Analysis. New York, NY: John Wiley & Sons, 2005.

[13] E. Gray, J.-M. Seigneur, Y. Chen, and C. Jensen,"Trust propagation in small worlds," presented at the First International Conference on Trust Management (iTrust2003), 2003.

[14] S. Buchegger and J.-Y. Le Boudec, "A Robust Reputation System for Mobile Ad-Hoc Networks," EPFL 2003.

[15] M. Prietula and K. Carley, "Boundedly rational and emotional agents - cooperation trust and rumor," in Trust and Deception in Virtual Societies, C. Castelfranchi and Y.-H. Tan, Eds. Norwood, MA: Kluwer Academic Publisher,2001, pp. 169 - 193.

[16] A. Fernandes, E. Kotsovinos, S. Ostring, and B. Dragovic, "Pinocchio: Incentives for honest participation In Global-Scale Distributed trust management," University of Cambridge, Cambridge, UK 2001.

[17] W. Adams, R. Thomas, and N. Davis, "Sizing the Credential Cache in a Trust-based Access Control System," submitted to IEEE Global Telecommunications Conference (GLOBECOM 2005), St. Louis, MO, 2001.

Author Yonghui Cao received the MS degree in business management from Zheji**ang** University in 2006. He is currently a doctorate candidate in Zhejiang University. His research interest is in the areas of management information systems.

Web Service Testing Tools: A Comparative Study

Shariq Hussain[1], Zhaoshun Wang[2], Ibrahima Kalil Toure[3] and Abdoulaye Diop[4]

**[1,2,3,4] School of Computer and Communication Engineering, University of Science and Technology Beijing
Beijing, 100083, China**

Abstract

Quality of Service (QoS) has gained more importance with the increase in usage and adoption of web services. In recent years, various tools and techniques developed for measurement and evaluation of QoS of web services. There are commercial as well as open-source tools available today which are being used for monitoring and testing QoS for web services. These tools facilitate in QoS measurement and analysis and are helpful in evaluation of service performance in real-time network. In this paper, we describe three popular open-source tools and compare them in terms of features, usability, performance, and software requirements. Results of the comparison will help in adoption and usage of these tools, and also promote development and usage of open-source web service testing tools.

Keywords: *Web Services, Performance, Software Testing, Testing Tools, Open-source Software.*

1. Introduction

The success of web service technology is clearly evident from the usage and adoption of this IT technology. A large number of providers from different sectors of industry are shifting to web service technology. Web services are software components accessible through programmatic interfaces and can perform tasks from simple requests to complex processes [1]. The heterogeneous nature of web service technology offers advantages like interoperability, usability, use of standardized communication protocol, deployability, etc. This makes web services technology an ideal candidate for organizations to host and deploy services in order to collaborate with other organizations in a flexible manner.

In order to attain the trust of service users, it is necessary that the system must conform to the performance requirements as it is the most important criteria for evaluating a system. It is therefore necessary to test the system before deployment in order to ensure that the system meets quality of service requirements. Various testing tools have been developed and designed for testing of web services. By using these test tools, web engineers can perform their tasks easily and efficiently, thus improving the quality of the system.

There are commercial as well as open-source test tools available in the market with different features and functionalities. In our study we are focusing on testing of Simple Object Access Protocol (SOAP) [2] web services. SOAP web services use XML language for definition of message architecture and message format. Web Services Description Language (WSDL) [3], an XML language is used to describe operations and interfaces of the web service. HTTP protocol is used for communication due to its wide usage and popularity.

Test tools automate the process of testing and are targeted to a specific test environment such as functional testing, performance testing, load testing, exception testing, etc. With the help of test tools, testers can create, manage and execute tests for a specific test environment for a particular application. The test results are compared with the expected results to evaluate the quality of the product.

Web service testing is a quite challenging area for researchers. The importance of this can also be judged with the ongoing research in this field. Several methods and techniques proposed by researchers as well as development of testing tools. There are commercial as well as open-source test tools available today for testing of web services.

Several studies are available which have compared various web service testing tools from functionalities, features, services, popularity, and so on. To our knowledge, there is still no comparative study on the representative testing tools, such as JMeter, soapUI, and Storm. In this paper, we compare these tools in terms of features, architecture, test environments, software requirements, and provide some observations. The comparison may help in selection of most suitable web service testing tool and promote the development and usage of open-source test tools.

This paper is organized as follows: Section 2 presents an overview of testing tools. Section 3 describes the three selected tools and their comparisons are reported in Section 4. Section 5 introduces related work and Section 6 concludes the paper.

2. Testing Tools

Software testing is the process of executing a program to verify its functionality and correctness [4]. Software testing is mostly deployed by programmers and testers. The aim of testing is to find the problems and to fix them to improve the software quality. Software testing methodology can be divided into two groups. One is manual testing and the other is automated testing. Manual Testing is a process in which testing process are carried out manually by the tester, usually follow a test plan comprised of test cases. On the other hand, automated testing is done with the help of automated test tools. Automated testing uses scripts to test operations of application automatically, reduces the need of human involvement and requires less time.

Software testing tools provide enormous aids to the testers in performing their tasks. Although the scope of testing tools is limited to particular test environments, the advantages associated are quite impressive. Benefits of automated testing includes: a better test coverage, quality improvements and more tests can be completed within a shorter time [5]. Tests can be performed to analyze the behavior of application in repeated executions of same operations. Further, testing tools perform testing faster than human users.

3. Overview of Open-Source Web Service Testing Tools

There is a number of open-source web service testing tools available in the software market. Although the core functions of these tools are similar, they differ in functionality, features, usability and interoperability. Keeping in view the above-mentioned aspects, we have selected three representative web service testing tools for comparison. Among them are, JMeter and soapUI are implemented in Java, and Storm is implemented in F# (F Sharp). A brief description of each of them is presented below.

3.1 JMeter

JMeter [6] is an open-source testing tool developed by Apache Software Foundation (ASF). It is distributed under Apache License. It was originally designed to test Web applications but has been extended to other test functions. The core function of JMeter is to load test client/server application but it can also be used for performance measurement. Further, JMeter is also helpful in regression testing by facilitating in creation of test scripts with assertions. By this way, we can verify that the application returns the expected results.

JMeter supports full multithreading that allows concurrent sampling by many threads and simultaneous sampling of different functions by separate thread groups. JMeter offers high extensibility due to use of pluggable components. These pluggable components include timers, samplers and visualization plugins. JMeter offers user-friendly Graphical User Interface (GUI). Configuration and setting up a test plan requires very little efforts. JMeter offers a number of statistical reports as well as graphical analysis. The latest release is version 2.8.

3.2 soapUI

soapUI [7] is an open-source testing tool for Service Oriented Architecture (SOA) [8] and web service testing. It is developed by SmartBear Software and is provided freely under the GNU LGPL. soapUI facilitates quick creation of advanced performance tests and execution of automated functional tests. The set of features offered by soapUI helps in performance evaluation of web services. Analysis of the test results provides a mean to improve the quality of services and applications.

soapUI offers easy-to-use GUI and is capable of performing variety of tests by offering many enterprise-class features. soapUI received a number of awards: ATI Automation Honors, 2009 [9], InfoWorld Best of Open Source Software Award, 2008 [10] and SOAWorld Readers' Choice Award, 2007 [11]. The latest version of soapUI is 4.5.1.

3.3 Storm

Storm [12] is a free and open-source tool for testing web services. It is developed by Erik Araojo. Storm is developed in F# language and is available for free to use, distributed under New BSD license.

Storm allows to test web services written using any technology (.Net, Java, etc.). Storm supports dynamic invocation of web service methods even those that have input parameters of complex data types and also facilitates editing/manipulation of raw soap requests. The GUI is very simple and user friendly. Multiple web services can be tested simultaneously that saves time, speed up testing schedule. Current stable version is r1.1-Adarna.

4. Comparison of Web Service Testing Tools

In this section, we present a comparison of the three web service testing tools, and then provide our observations. Such a comparison is helpful for the users/researchers to choose the suitable test tool for their needs.

Table 1: Technical overview of web service testing tools

Tool	*Technology Support*	*First release*	*Latest version/ Release date*	*Programming language*	*Operating System Support*	*Requirement*	*License*	*Developed by*	Website
JMeter	Web-HTTP, HTTPS SOAP Database via JDBC LDAP JMS Mail-SMTP(S), POP3(S) and IMAP(S) Native commands or shell scripts	2001	2.8 / Oct 6, 2012	Java	Cross-platform	JRE 1.5+	Apache License 2.0	Apache Software Foundation	http://jmeter.apache.org/
soapUI	Web-HTTP, HTTPS SOAP Database via JDBC JMS REST AMF	2005	4.5.1 / Jun 27, 2012	Java	Cross-platform	JRE 1.6+	GNU LGPL 2.1	SmartBear Software	http://www.soapui.org/
Storm	SOAP	2008	1.1 / Oct 29, 2008	F#	Microsoft Windows	.NET Framework 2.0 F# 1.9.3.14 (optional)	New BSD License	Erik Araojo	http://storm.codeplex.com

4.1 Technical Overview

The three testing tools chosen for comparison are based on different platforms and technologies. A detailed technical overview of them is shown in Table 1.

4.2 Comparison and Evaluation Approach

In order to compare the representative testing tools, we consider sample of three web services. The detail of web services is presented in Table 2.

To test the representative testing tools, each tool need to be configured to run the tests. The configuration includes installation, setting up test environment, test parameters, test data collection, reports analysis, etc. Each tool is configured to test the sample web services and gather test results.

We run the tests on an Intel Core 2 Duo 2.0 GHz processor machine with 3GB RAM, running Microsoft Windows 7 Ultimate, and 2Mbps of DSL Internet connection.

The tests were conducted four times a day at regular intervals to get fair and transparent results. The reason was to minimize the affect of Internet connection's performance on the test results and to obtain realistic measurements. The performance of Internet varies depending on the time of day and other factors such as internet traffic, subscribed users, etc.

Table 2: Sample web services

ID	*Web Service Name*	*Description*	Publisher
W1	TempConvert	Conversions from Celsius to Fahrenheit and vice versa	W3Schools
W2	Weather	Allows to get city's weather	CDYNE Corporation
W3	ZipCode	Returns a list of City+State for a supplied zip code	Ripe Development LLC

Table 3: Minimum and maximum response time of testing tools for web services

Tool	*Web Service ID*	*Response Time (ms)*							
		12:00 AM		6:00 AM		12:00 PM		6:00 PM	
		Min	Max	Min	Max	Min	Max	Min	Max
JMeter	W1	1237	4906	1056	4304	1077	1921	1147	4320
	W2	1880	18276	1121	16087	1595	19056	1523	18984
	W3	954	25660	806	3852	866	10023	912	7052
soapUI	W1	334	1423	300	1158	307	1424	299	4048
	W2	557	60011	315	12062	402	6124	527	16096
	W3	639	7113	534	6750	576	9761	625	7002
Storm	W1	666	3581	577	1482	593	1298	624	1794
	W2	1060	15179	619	99013	718	7318	936	32417
	W3	998	7634	822	2246	852	6895	936	4103

The three selected tools were tested by invoking the sample web services for a pre-defined sample count. The results were collected and compiled for analysis.

4.3 Results and Discussions

In this section we describe different comparative results with testing tools.

Each tool has different architecture and internal processes to carry out tasks. This factor provides basis to compare the tools in terms of response time. Minimum and maximum values for response time at different time intervals are shown in Table 3.

From Table 3 we observe that the response time values taken at 6:00 AM are most optimal. This shows that the performance of Internet connection is better at 6:00 AM which is reflected in response time values.

Further, results of the tests are summarized to calculate average response time of each test tool for each web service. Table 4 shows the average response time of each tool for each web service. This data is also presented in the form of graph as shown in Figure 1.

Table 4: Average response time of testing tools

ID	*Average Response Time (ms)*		
	JMeter	**soapUI**	**Storm**
W1	1359.83	401.44	758.73
W2	3541.33	1193.01	1939.92
W3	1357.25	1046.78	1350.33

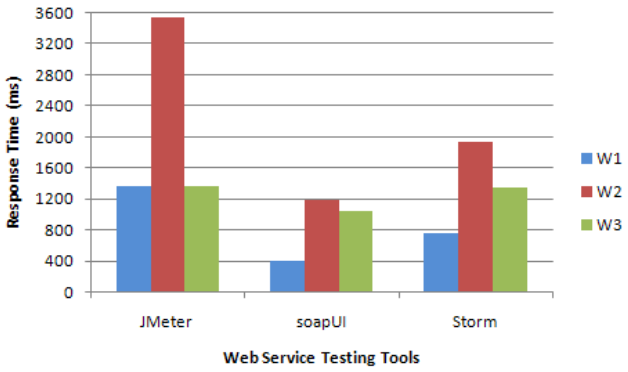


Fig. 1 Average response time of testing tools for sample web services.

From the results, we observe that JMeter is taking more time in responding to web services as compared to other two tools. Storm is behaving better than JMeter but not promising as soapUI. In this test, soapUI outperforms other two testing tools and can be regarded as fastest tool in terms of response time.

The next comparison is based on the average throughput criterion. Throughput is the measure of the number of requests that can be served by web service in a specified time period [13]. Only JMeter and soapUI supports this type of testing and provides information on throughput test results. Average throughput of each tool for each web service is shown in Figure 2.

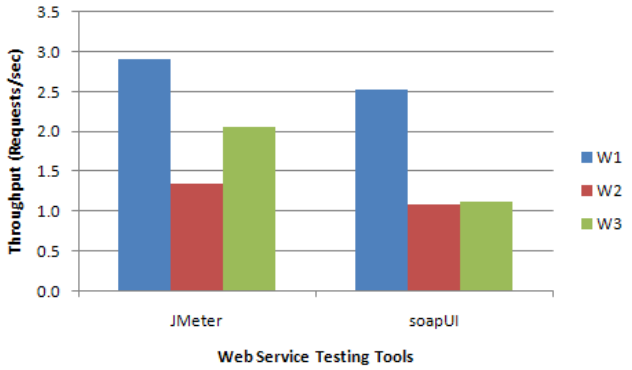


Fig. 2 Average throughput of testing tools for sample web services.

The results of throughput test demonstrate that JMeter has better throughput than soapUI. In case of W3, JMeter shows more than 84% throughput than soapUI, while for W1 and W2, increase is 14.5% and 24% respectively. Therefore, JMeter has better throughput than soapUI.

Another parameter that is observed during testing of web services is number of bytes per second processed by the test. Figure 3 shows the KB/sec values of both tools.

From Figure 3 it is seen that number of bytes processed by JMeter is higher than soapUI. This is in relation to the throughput attribute, as JMeter has better throughput as shown in Figure 2. This means that JMeter has processed more bytes during test as its throughput is better than other tool.

Further useful information related to testing, reported by JMeter and soapUI contains number of assertion errors and number of lost samples (failed request ratio).

Usability is another factor for evaluation of testing tools. The study shows that the Storm has a very simple and easy

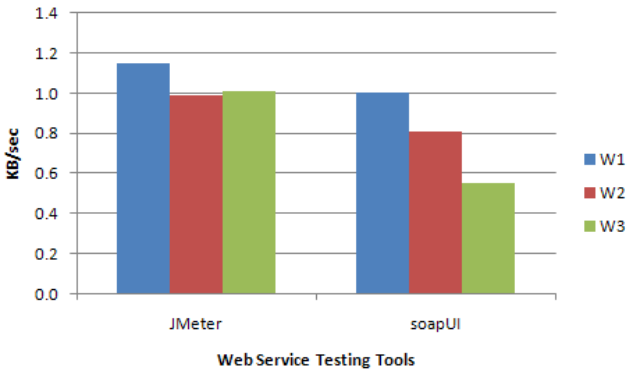


Fig. 3 Average Kilobytes per second of testing tools.

to use interface. Simply by adding WSDL reference, supported methods are displayed for invocation. To perform a test one need to provide input parameters for methods. The result of the web service call along with SOAP response and elapsed time for the test is returned to user. soapUI has attractive graphical user interface with many useful test utilities. Steps to perform a test includes, creation of a new project with name and WSDL reference, addition of web service methods into new request and finally creation of load test. LoadTest window provides statistical information about test along with graphical representation. JMeter has an excellent user interface with an iconic toolbar and a right pane to display the details of each element of test plan. For web service testing, a new test plan has to be created with thread group, loop controller, timer and sampler. Each element need to be placed in a parent child relationship in a hierarchical form. Presently, JMeter doesn't support automatic WSDL handling. There are three options for the post data (soap message): text area, external file, or directory. Different types of listeners are available to show test results in various forms. JMeter supports a lot of different types of test plans.

Comparison of different testing tools is a complex task due to the fact that testing tools may not comply with same test criteria i.e. one tool may have the ability to test throughput (in our case JMeter and soapUI), while another tool i.e. Storm, does not have this criteria. Furthermore, one tool may have better performance in one test case, while poorer in other test criteria. For example, in our study soapUI has better response time but throughput is not as good as JMeter's throughput.

5. Related Work

Since the beginning of web service testing, different approaches have been proposed in literature. In this section, we describe several closely related work.

Performance testing of web services using JMeter is demonstrated in great detail [14]. JMeter is also able to perform load testing of web applications especially J2EE-based web applications [15-16]. In [17], authors proposed test steps of the web service testing tools for testing an orchestration and showed the applicability with soapUI. In some approaches, testing of web services based on WSDL descriptions, soapUI tool is used for derivation of SOAP envelope skeleton [18-19]. soapUI is also used for testing of sample web services developed using SAS BI platform [20]. The study [21] presented a preliminary approach towards an evaluation framework for SOA security testing tools and tested soapUI for suitability assessment. A research study on different test tools and techniques for

testing atomic and composed services is presented along with development of a prototype for automated choreography tests [22]. Security testing is also a challenging task in SOA domain. Altaani and Jaradat [23] analyzed the security requirements in SOA and presented techniques for security testing, validated the results by using soapUI. A comparative study of three web service testing tools with several selected web services is done in which soapUI outperforms other two tools [24].

In the paper, we presented an overview of three open-source web service testing tools and a technical overview of each tool. We also provide a comprehensive comparison of them, which may help researchers in selection of suitable tool.

6. Conclusions

Nowadays we can see that web service technology turn out to be the latest trend and provides a new model of web. The rapid growth of web service market necessitated developing of testing methodologies, hence different methods and different tools proposed to test web services. In this paper, we present a comparative study of open-source web service testing tools with technical overview and features. Comparison is made on several quality factors including response time, throughput, and usability. Tools are evaluated by collecting the sample web services and collecting the test results. The comparison may give researchers an informative overview with potential benefits of open-source testing tools, and also help in promotion and development of open-source testing tools.

Acknowledgments

The work reported in this paper was supported by the National Natural Science Foundation of China (Grant No. 60903003), the Beijing Natural Science Foundation of China (Grant No. 4112037), and the Research Fund for the Doctoral Program of Higher Education of China (Grant No. 2008000401051).

References

[1] D. S. Zhang, "Web services composition for process management in E-business", Journal of Computer Information Systems, Vol. 45, No. 2, 2004, pp. 83-91.

[2] “W3C - SOAP Version 1.2”, http://www.w3.org/TR/soap12-part1/

[3] “W3C - Web Services Description Language (WSDL) 1.1”, http://www.w3.org/TR/wsdl

[4] C. Kaner, J. F. Falk, and H. Q. Nguyen, Testing Computer Software (2nd ed), New York: Van Nostrand Reinhold, 1993.

[5] K. Karhu, T. Repo, O. Taipale, and K. Smolander, "Empirical Observations on Software Testing Automation", in International Conference on Software Testing Verification and Validation (ICST '09), 2009, pp.201-209.

[6] “Apache JMeter”, http://jmeter.apache.org/ (Accessed Jun 2012)

[7] “soapUI – The Home of Functional Testing”, http://www.soapui.org/ (Accessed Jun 2012)

[8] M. P. Papazoglou, P. Traverso, S. Dustdar, and F. Leymann, "Service-oriented computing: state of the art and research challenges", Computer, Vol.40, No.11, 2007, pp.38–45.

[9] “1st Annual ATI Automation Honors on Vimeo”, http://vimeo.com/8078081 (Accessed Jun 2012)

[10] “InfoWorld announces our 2008 Best of Open Source Awards | Open Source - InfoWorld”, http://www.infoworld.com/d/open-source/infoworld-announces-our-2008-best-open-source-awards-065 (Accessed Jun 2012)

[11] “SYS-CON Media Announces 2007 SOA World Readers’ Choice Awards | SOA World Magazine”, http://soa.sys-con.com/node/397933 (Accessed Jun 2012)

[12] “Storm”, http://storm.codeplex.com/ (Accessed Jun 2012)

[13] A. Mani, and A. Nagarajan, Understanding Quality of Service for Web Services, January 2002. http://www-106.ibm.com/developerworks/library/wsquality.html (Accessed Jun 2012)

[14] D. Nevedrov, Using JMeter to Performance Test Web Services. (Accessed Jun 2012) http://dev2dev.bea.com/pub/a/2006/08/jmeter-performance-testing.html

[15] F. A. Torkey, A. Keshk, T. Hamza, and A. Ibrahim, "A new methodology for Web testing", in ITI 5th International Conference on Information and Communications Technology (ICICT 2007), 2007, pp. 77-83.

[16] Q. Wu, and Y. Wang, "Performance Testing and Optimization of J2EE-Based Web Applications", in Second International Workshop on Education Technology and Computer Science (ETCS), 2010, Vol.2, pp. 681-683.

[17] H. Yoon, E. Ji, and B. Choi, "Building test steps for SOA service orchestration in web service testing tools", in Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication (ICUIMC '08), 2008, pp. 555-557.

[18] C. Bartolini, A. Bertolino, E. Marchetti, and A. Polini, "Towards Automated WSDL-Based Testing of Web Services", in Proceedings of the 6th International Conference on Service-Oriented Computing (ICSOC '08), 2008, pp. 524-529.

[19] C. Bartolini, A. Bertolino, E. Marchetti, and A. Polini, "WS-TAXI: A WSDL-based Testing Tool for Web Services", in Proceedings of the 2009 International Conference on Software Testing Verification and Validation (ICST '09), 2009, pp. 326-335.

[20] D. Jahn, Using SAS® BI Web Services and PROC SOAP in a Service-Oriented Architecture, 2008. http://support.sas.com/rnd/papers/sgf2008/soap.pdf

[21] N. Kabbani, S. Tilley, and L. Pearson, "Towards an evaluation framework for SOA security testing tools", in

Proceedings of 4th Annual IEEE International Systems Conference, 2010 , pp. 438-443.

[22] F. M. Besson, P. M. B. Leal, and F. Kon, "Towards verification and validation of choreographies", 2011, Technical Report No: RT-MAC-2011-01.

[23] N. A. Altaani, and A. S. Jaradat, "Security Analysis and Testing in Service Oriented Architecture", International Journal of Scientific & Engineering Research, Vol. 3, No. 2, 2012, pp. 1-9.

[24] S. Azzam, M. N. Al-Kabi, and I. Alsmadi, "Web services testing challenges and approaches", in Proceedings of the 1st Taibah University International Conference on Computing and Information Technology (ICCIT 2012), 2012, pp. 291-296.

Shariq Hussain received his Master's degree in Computer Science from PMAS Arid Agriculture University, Rawalpindi, Pakistan, in 2007. He is now a PhD student in the School of Computer and Communication Engineering, University of Science and Technology Beijing. His main research interests include web service QoS, web service monitoring, web service testing and e-learning.

Zhaoshun Wang is a Professor and the Associate Head of the Department of Computer Science of the University of Science and Technology Beijing. He graduated from Department of Mathematics, Beijing Normal University in 1993. He received his PhD from Beijing University of Science and Technology in 2002. He completed postdoctoral research work at the Graduate School of the Chinese Academy of Sciences from 2003 to 2006. Teaching and research work, research direction has been engaged in the direction of computer software for software engineering, software security, information security, ASIC chip design. In recent years, participated in the national "863", "973", the National Natural Science Foundation of China, National "Eleventh Five-Year" password Fund, Outstanding Young Teachers Fund of the Ministry of Education, Beijing natural science fund of each one; auspices of the National Information Security Standardization Technical Committee Project 2, and 8 Plant Association project topics. He has published more than 60 scientific papers in core computer science journals and international conferences, which retrieve an SCI, EI retrieval of more than 10 articles, ISTP retrieval of more than 10 articles; teaching and research of more than 10 papers. Further achievements include a provincial and municipal Science and Technology Progress Award, two national invention patents, two textbooks, and a monograph.

Ibrahima Kalil Toure received his Master's degree in Programming Analysis from University of Conakry, Conakry, Guinea, in 2000. He is now a PhD student in the School of Computer and Communication Engineering, University of Science and Technology Beijing. His main research interests include web services and composition.

Abdoulaye Diop is currently a PhD student in the School of Computer and Communication Engineering, University of Science and Technology Beijing. His main research interests include wireless sensor networks and security.

Wavelet Based Image De-noising to Enhance the Face Recognition Rate

Isra'a Abdul-Ameer Abdul-Jabbar[1], Jieqing Tan[1] , Zhengfeng Hou[1]

[1] School of Computer and Information, Hefei University of Technology, Hefei 230009, China

Abstract

In this paper a comparison between face recognition rate with noise and face recognition rate without noise is presented. In our work we assume that all the images in the ORL faces database are noisy images. We applied the wavelet based image de-noising methods to this database and created new databases, then the face recognition rate are calculated to them. Three experiments are given in our paper. **In the first** experiment different wavelet methods with different level of decomposition (up to ten decompositions) are used for de-noising the ORL database and the comparison is done when Principal Components Analysis (PCA) is applied to evaluate the verification rate. **In the second** experiment de-noising different sets of ORL database with methods that have best performance in levels (1, 2, 3, and 10) is done (as a result from experiment 1). **In the third** experiment we implement the proposed Haar10 method on PCA, Linear Discriminate Analysis **(**LDA), Kernel PCA, Fisher Analysis (FA) face recognition methods and the recognition rates are evaluated for both the noisy and de-noisy databases.

Keywords: *Image de-noising, Wavelet decomposition, Noisy and de-noisy face recognition rate, False accept rate (FAR), verification rate at 0.1% rate, Face recognition rate.*

1. Introduction

Quality of a biometric sample affects the performance of the recognition algorithm. In literature, several research papers exist on analyzing the effects of quality on the performance of different biometric modalities such as iris and fingerprint. Environmental corruption such as noise, blur, adverse illumination and compression rates (in JPEG and other compression techniques) influence the performance of state-of-art recognition algorithms [1, 3].

An excellent overview of pattern recognition in wavelet domain can be found in [5]. It would also be worthwhile to mention at this point that the most papers use wavelets as part of the face recognition system.

The work presented in [6] is along those lines of thought. Sabharwal and Curtis [7] used Daubechies 2 wavelet filter coefficients as input into PCA. The experiments were performed on a small number of images and the number of wavelet decomposition was increased in each experiment (up to three decompositions). The observed recognition rate increased mostly around 2 %.

Garcia et al. [8] performed one standard wavelet decomposition on each image from the FERET database. This gave four bands, each of which was decomposed further (not only the approximation band). In this way there are 15 detail bands and one approximation.

Similar idea can be found in [9] as well. However, in this paper several wavelets are tested (Daubechies, Spline, Lemarie), and finally Daubechies 4 is chosen to be used in a PCA-based face recognition system. The HH subband after three decompositions was used as input to PCA and recognition rate increase of ≈ 5%.

Xiong and Huang [10] performed one of the first explorations using features directly in the JPEG2000 domain.

Chien and Wu [11] used two wavelet decompositions to calculate the approximation band, later to be used in face recognition. Their method performed slightly better than standard PCA. Similarly, in [12] Li and Liu showed that using all the DWT coefficients after decomposition as input to PCA yields superior recognition rates compared with standard PCA.

Two decompositions with Daubechies 8 wavelet were used by Zhang et al. [13] with the resulting approximation band being used as input into a neural network based classifier. By using Daubechies 4 wavelet and PCA and ICA,Ekenel and Sankur [14] tried to find the subbands that are least sensitive to changing facial expressions and illumination conditions. PCA and ICA were combined with L1,

L2 and COS metrics in a standard nearest neighbor scenario.
Earlier studies confirm that the information in low spatial frequency bands plays a dominant role in face recognition. Nastar et al. [15] investigated the relationship between variations in facial appearance and their deformation spectrum. They found that facial expressions and small occlusions affect the intensity manifold locally.

Noise will be inevitably introduced in the image acquisition process and de-noising is an essential step to improve the image quality [2, 4]. The problem in face recognition system is the recognition of noisy face image; the questions that should be answered in this research are: will image de-noising improve the recognition rate for face images? And which image de-noising method is appropriate for face image noise removal? To answer these questions, a wavelet based image de-noising is used for noise removal to all images in the ORL database and the Principal Components Analysis is implemented on both the original ORL database and on the de-noising one, then the recognition rate is measured for both noisy and de-noising databases.

Image de-noising can be used with face recognition methods to improve the face recognition rate. The aim of this research is to compare the recognition rate of noisy faces (default ORL database) and the recognition rate of the de-noising faces (new database after noise removal).

2. A Framework of the de-noising face image for improving recognition rate

Our proposed work is to apply image de-noising process by wavelet transform to the ORL faces database, then implement the PCA on these database to evaluate the recognition rate for a specific face image before and after de-noising. Figure 1 shows that the Discrete Wavelet Transform is applied prior to dimensionality reduction. PCA is then applied with the above technique to find the face recognition accuracy rate and to compare the results of the de-noising method with PCA method.

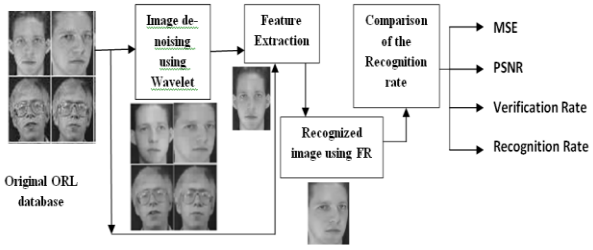


Fig.1 Proposed work block diagram

2.1 Wavelet-based image de-noising

Intuitively, de-noising a noisy face image improves the face recognition performance, provided the right sets of parameters are used [1].
The DWT provides a compact representation of a signal's frequency components with strong spatial support. DWT decomposes a signal into frequency subbands at different scales from which it can be perfectly reconstructed. 2D-signals such as images can be decomposed using many wavelet decomposition filters as shown in Figure 2 and Figure 3 in many different ways. The general wavelet-based procedure for de-noising the image is as follows [16]:
1. Choose a wavelet filter (e.g. Haar, Daubechies, symlet) and number of levels for the decomposition. Then compute the 2D-DWT of the noisy image.
2. Threshold the non-LL subbands.
3. Perform the inverse wavelet transform on the original approximation of LL-subband and the modified non-LL subbands.

2.2 Face Recognition algorithms

All face recognition algorithms consist of two major parts: (1) face detection and normalization and (2) face identification. Algorithms that consist of both parts are referred to as fully automatic algorithms and those that consist of only the second part are called partially automatic algorithms. Partially automatic algorithms are given a facial image and the coordinates of the center of the eyes. Fully automatic algorithms are only given facial images [17]. This subsection consists of description to the general work of PCA [4], LDA [19], KPCA [20] and KFA [21] that will be used in our work.

General work of face recognition algorithms

Step 1: Load images from a database. In our case, from the de-noised ORL databases.
Step 2: Partition data into training and test sets. In our case, the first 3 images of each ORL subject will serve as the training/gallery/target set and the remaining images will serve as test/evaluation/query images.
Step 3: Compute training and test feature vectors using the chosen method. In our case we use different algorithms for feature extraction (PCA, LDA, KPCA, and KFA) and, therefore, first compute the subspace using the training data from the ORL database.
Step 4: Compute matching scores between gallery / training / target feature vectors and test / query

feature vectors. In our case we use the Mahalanobis cosine similarity for computing similarity matrix.

Step 5: Evaluate results: computing of the face recognition rate and the FAR at 0.1%.

3. Experiments and results

In our work three experiments are conducted as trying to verify the effect of wavelet de-noising process on the performance of face recognition system. Figure 2 shows the Graphical User Interface (GUI) of the wavelet de-noising methods used in our experiments. All experiments are applied to the existing face recognition system implemented in the PhD face recognition toolbox. The Toolbox description and user manual are in the referenced section [23, 24].

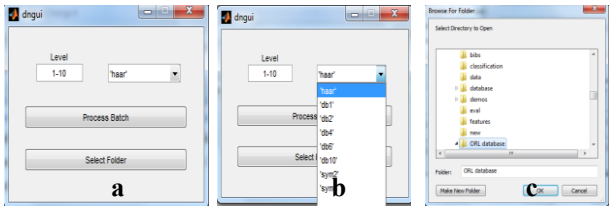


Fig.2 GUI of the wavelet de-noising algorithms

In our work we used ORL database because the demo scripts of the PhD toolbox were written for use with the ORL (AT&T) database. Our Faces Database, (formerly 'The ORL Database of Faces'), contains a set of face images taken between April 1992 and April 1994 in the lab. There are ten different images for each of 40 distinct subjects. Figure 3 illustrates some images of the database. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement).

The files are in PGM format, and the size of each image is 92x112 pixels, with 256 grey levels per pixel. The images are organized in 40 directories (one for each subject), which have names of the form sX, where X indicates the subject number (between 1 and 40). In each of these directories, there are ten different images of that subject, which have names of the formY.pgm, where Y is the image number for that subject (between 1 and 10). The database can be retrieved from http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

In our work we implemented all the de-noising wavelet methods shown in Fig 2-b on this database to generate new de-noised databases (one database for each wavelet de-noising method), then applied the face recognition tool box on them to compare the original results with the de-noising results.

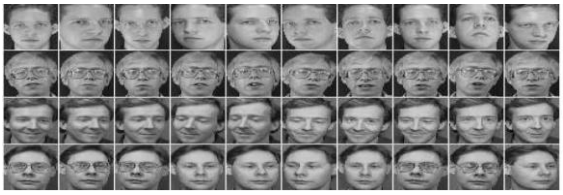


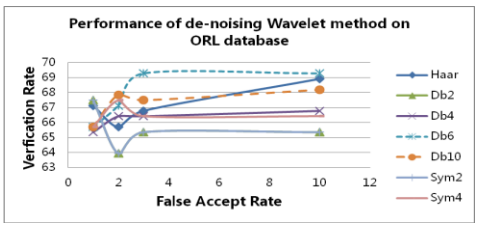
Fig.3: Face samples from the ORL face database.

Experiment 1: Recognized of de-noising image by PCA

Different wavelet de-noising methods represented by (Haar , db2, db4, db6, db10, sym2, sym4) at different levels (1, 2, 3 and 10 decomposition levels) are implemented on PCA face recognition method in this experiment to show the variations in verification rate at 0.1% FAR and results. We find that the performance of PCA on the original ORL database is equal to **66.07%** and we compare this result with the performance of PCA after implementing the de-noising wavelet methods. Table 1 shows the result of the proposed de-noised databases.

Table 1: the verification rate at 0.1% FAR after the de-noising process is implemented on ORL database at level L=1, 2, 3 and 10.

L	Haar	Db2	Db4	Db6	Db10	Sym2	Sym4
1	**67.14%**	**67.50%**	65.36%	65.70%	65.70%	**67.50%**	65.70%
2	65.70%	63.93%	**66.43%**	**67.14%**	**67.86%**	63.93%	**67.50%**
3	**66.79%**	65.36%	**66.43%**	**69.29%**	**67.50%**	65.36%	**66.43%**
10	**68.93%**	65.36%	**66.79%**	**69.29%**	**68.20%**	65.36%	**66.43%**



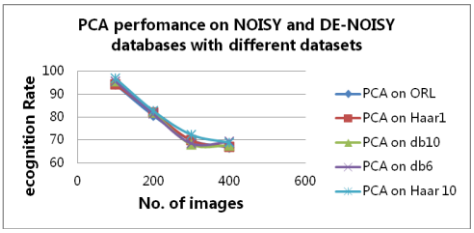
From the result above we find that the verification rate increased in a little variant in level 1 of decomposition using Haar, db2 and Sym2, but the verification rate decreased at level 2 for these methods, in contrast to the other four methods where the verification rate increased. Level 3 has nearly the same result as level 1 with a little decreasing in the percentage rate but it is still higher than the verification rate of origin PCA (66.07). At level 10 only db2 and Sym4 fail to enhance the verification rate in contrast to the other methods that shown increasing reach to more than 2.5% - 3% in both Haar and db6 (at level 3 and level 10) of wavelet.

Experiment 2: Recognizing different de-noising data set by PCA

In this experiment a comparison among different de-noising methods with different number of training image range from 100 to 400 images is done. Table 2 shows the verification rate at 0.1% FAR for the origin PCA implemented on ORL database and PCA implemented after wavelet image de-noising method (1D Haar wavelet, 2D daubchies10, 3D daubchies6, 10D Haar) is applied separately on the ORL database.

Table 2: verification rate at (0.1%) FAR when implementing Haar 10 de-noising wavelet using different dataset on PCA

No. of images	PCA on ORL DB	PCA on Haar1 DB	PCA on db10 DB	PCA on db6 DB	PCA on Haar 10 DB
100	94.29%	94.29%	95.70%	95.70%	**97.14%**
200	80.70%	82.14%	82.14%	81.43%	**82.86%**
300	70%	70%	68.10%	68.57%	**72.38%**
400	66.79%	67.14%	67.68%	**69.29%**	68.93%



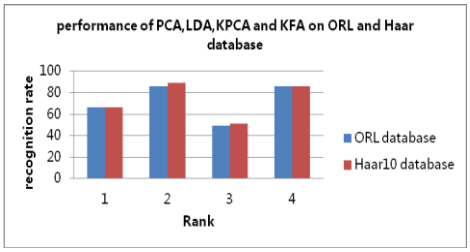
From this experiment we find that wavelet Haar at level 10 has the highest verification rate when the number of images is 100, 200, 300 and the daubchies 6 has the highest rate when 400 images are used for the recognition.

Experiment 3: Evaluate recognition rate using different face recognition method

In this experiment we test the performance of Haar 10 wavelet de-noising on different face recognition method (PCA, LDA, KPCA, FA). We use (PhD-tool) face recognition tool box in matlab to compare the recognition rate for the face recognition methods on original ORL database and the recognition rate for the face recognition methods on the new proposed de-noising database Haar 10 wavelet. The result in Table 3 shows the recognition rate on the original database and the proposed one.

Table 3: recognition rate for PCA, LDA, KPCA, and KFA with noisy and de-noising database

recognition rate	PCA	LDA	KPCA	FA
ORL database (with noise)	66.07%	86.07%	49.29%	85.70%
proposed database (without noise)	66.43%	89.29%	51.07%	86.07%



4. Discussion

In our work we implement three experiments. From the first experiment we find that the verification rate at (0.1%) FAR increases for some methods when we implement PCA with level 1 decomposition wavelet de-noising on ORL database, but there are a big variations in this rate at level 2 of decomposition for all methods that are used, and the methods which increased at level 1 was decreased at level 2 and vice versa, and in level 3 all wavelet methods are improving to the verification rate except db2 and sym2, see Table 1. This is because we do not notice any observable improvement on the rate at level 5 of decomposition. We discard the result of this level and only the result of level 10 is mentioned in this paper since it produces good improvement for both the verification rate at (0.1%) FAR and the recognition rate. For all wavelet de-noising methods from level 1 up to level 10 of decomposition, both the MSE and PSNR are measured to all face images in the original ORL database and the proposed de-noising one, by clicking on any image in the list of face images that have extension .pgm as illustrated in Figure 4 ,where the image on the left side is the face in the original ORL database and the face on the right side is the de-

noising one by Haar10 wavelet with MSE=49.4861 and PSNR=31.186.

Fig.4 An example of implementing Haar 10 wavelet

The result of experiment 1 supports our work to use Haar at level 10 of decomposition in experiment 3 in spite of that it produces 68.93% verification rate at 0.1% FAR which is less than 96.29% that db6 gives at level 10. Since we find db6 has the same verification rate from level 3 up to level 10 decompositions. So we select Haar 10 wavelet because it makes improvement on these rates gradually. In experiment 2, we take the best performance of each wavelet de-noising methods at each (1, 2, 3 and 10) level of decomposition, so that we select Haar 10 wavelet to de-noise the ORL database and produce four new de-noising face image databases, then implement PCA face recognition method on these databases with different data sets (100, 200, 300 and 400) of face images. We do these four tests for each database, and get the results of 20 tests implemented by PCA (4 tests with 4 datasets are done by PCA ORL database, PCA on Haar 1, PCA on db10 at level 2 , PCA on db6 at level 3 and PCA on Haar at level 10), see Table 2.

In spite of db6 at level 3 gives the highest verification rate at 0.1% FAR which is equal to 69.29%, it is not improving the recognition rate at any level of decomposition. For this reason we take Haar 10 in experiment 3, since it is the only wavelet de-noising method that produces improvement for both the verification rate at (0.1%) FAR, and rank one recognition rates, see Table 1 and Table 3.

In experiment 3, we use Haar 10 de-noising wavelet database with different face recognition method represented by (PCA, LDA, KPCA, and FA) and find it produces (2.5%-3%) improvement over the recognition rate when ORL database is used with these face recognition methods. Figure 5 shows some images de-noised by Haar 10 wavelet process.

Fig. 5 Faces after de-noising by Haar 10 wavelet

5. Conclusions

From our work we proved that the image de-noising process has a powerful affect on face recognition rate. We find that the using of wavelet de-noising at higher level for decomposition will discard many metrics that are not important in the face image and focus on the important one like the eyes, nose, mouth, poses, etc., which improves the feature extraction process and leads to good recognition rate. As we see, the using of Haar Wavelet at level 10 produces good improvement on both verification rate at 1% FAR and the recognition rate.

For future work we can use many other wavelet de-noising methods and test them with more levels of decompositions by the same or different face recognition methods.

Acknowledgments

This work is supported by the NSFC-Guangdong Joint Foundation (Key Project) under Grant No.U1135003 and the National Natural Science Foundation of China under Grant No.61070227.

References:

[1] Samarth Bharadwaj, Himanshu Bhatt, Mayank Vatsa, Richa Singh, and Afzel Noore, Quality assessment based de-noising to improve Face recognition performance, Canadian Journal on Image Processing and Computer vision, Vol. 2, No. 4,2011, pp.140-145.

[2] Lei Zhang, Weisheng Dong, David Zhang, Guangming Shi, Two-stage image de-noising by principal component analysis with local pixel grouping, Pattern Recognition Journal, Vol. 43, No. 4, 2010, pp. 1531–1549.

[3] Shan Du and Rabab K.Ward , Adaptive region-based image enhancement method for robust face recognition under variable illumination conditions, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 20, No. 9 , 2010, pp. 1165 – 1175.

[4] D Murugan, S Arumugam, K Rajalakshmi and Manish T I. Performance evaluation of face recognition using gabor filter, log gabor filter and discrete wavelet transform, International Journal of Computer Science and Information Technology (IJCSIT), Vol. 2, No.1, 2010, p.p 125-133 .

[5] Brooks R.R., Grewe L., Iyengar S.S. Recognition in the Wavelet Domain: A Survey, Journal of Electronic Imaging, Vol. 10,No.3, 2001,pp. 757-784.

[6] Delac, K., Grgic, M., Grgic, S.Towards Face Recognition in JPEG2000 Compressed Domain, Proc. of the 14th International Workshop on Systems, Signals and Image Processing (IWSSIP), 2007, pp. 155-159.

[7] Sabharwal C.L., Curtis W., Human face recognition in the wavelet compressed domain, Smart Engineering Systems, 1997,Vol. 7 , pp. 555-560.

[8] Garcia C., Zikos G., Tziritas G. Wavelet packet analysis for face recognition, Image and Vision Computing, Vol. 18, No. 4, 2000, pp. 289-297.
[9] Feng G.C., Yuen P.C., Dai D.Q., Human face recognition using PCA on wavelet subband , Journal of Electronic Imaging, Vol. 9, No. 2, 2000, pp. 226-233.
[10] Xiong Z., Huang T.S., Wavelet-based texture features can be extracted efficiently from compressed-domain for JPEG2000 coded images, Proc. of the 2002 International Conference on Image Processing, ICIP'02, 2002 ,Vol. 1 , pp. 481-484.
[11] Chien J.T., Wu C.C., Discriminate wavelet faces and nearest feature classifiers for face recognition, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24,No. 12 , 2002, pp. 1644-1649.
[12] Li B., Liu Y., "When eigenfaces are combined with wavelets, Knowledge Based Systems, Vol 15, N0. 5, 2002, pp. 343-347.
[13] Zhang B.L., Zhang H., Ge S.S.,.Face recognition by applying wavelet subband representation and kernel associative memory, IEEE Trans. on Neural Networks, Vol.15, No.1, 2004, pp. 166-177.
[14] Ekenel H.K., Sankur B., Multiresolution face recognition, Image and Vision Computing, Vol. 23, No.5, 2005, pp. 469-477.
[15] Nastar C., Ayach N.,Frequency-based nonrigid motion analysis, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.18, No.11, 1996, pp. 1067-1079.
[16] Rafael C. Gonzalez, Richard E. Woods, Steven L. Eddins, Digital image processing using Matlab, China Edition Published by Pearson Education Asia LTD.,2009.
[17] A. S. Tolba, A.H. El-Baz, and A.A. El-Harby, Face recognition: a literature review, International Journal of Information and Communication Engineering, Vol. 2, No. 2, 2006, pp. 88-103.
[18] Daubechies, I. Ten lectures on wavelets, SIAM., SIAM Review, Vol. 35, No. 4, 1993, pp. 115, 132, 194, 242.
[19] Hua Yu*, Jie Yang, A direct LDA algorithm for high-dimensional data-with application to face recognition , The Journal of Pattern Recognition, Vol.34, 2001, pp. 2067-2070.
[20] Bernhard Scholkopf , Alexander Smola, Klaus, Robert Muller, Kernel principal component analysis, Conference of Artificial Neural Networks-ICANN'97, 1997,Vol. 1327,pp. 583-588.
[21] S. Mika, G. Ratsch, J.Weston, B. Scholkopf, and K.-R.Muller , Fisher discriminant analysis with kernels. IEEE Conference on Neural Networks for Signal Processing IX, (1999), pp. 41-48.
[22] Yu, H.; Yang J., A direct LDA algorithm for high-dimensional data — with application to face recognition, Pattern Recognition, Vol.34, No. 10, 2001 pp. 2067–2069.
[23] Struc and N. Pavešić, The complete gabor-fisher classifier for robust face recognition , EURASIP Advances in Signal Processing,2010,page26 , doi =10.1155 / 2010 / 847680.
[24] Struc and N. Pavešić, Gabor-based kernel partial-least-squares discrimination features for face recognition", Informatica (Vilnius), VOL.20,No.1,(2009),pp.115-138.

The Study of High-Speed Passenger Train Operation Plan

Xin Feng[1], Jinbao Luo[2] and Yongsheng Qian[3]

[1] School of Economics and Management, Lanzhou Jiaotong University
Lanzhou, Gansu, 730070, China

[2] School of Traffic and Transportation, Lanzhou Jiaotong University
Lanzhou, Gansu, 730070, China

[3] School of Traffic and Transportation, Lanzhou Jiaotong University
Lanzhou, Gansu, 730070, China

Abstract

The paper sums up the factors with passenger's travel demand, passenger's need of comfort and the wasted train capacity. On the basis of these factors, this paper provides a new idea to solve high-speed passengers train operation plan problem. First, all the possible train operation plans are listed, and then with the aid of intelligent algorithm those train operation plans which can not be sended out are excluded. Finally, the rest of the train operation plans just can meet passenger's travel demand. The paper uses the genetic algorithm to get the best solution and will get train operation quickly.

Keywords: *Passenger Transportation Organization, High-speed Passenger Train Operation Plan, Passenger OD Equilibrium, genetic algorithm.*

1. Introduction

The operation plan of high-speed railway passenger train includes passenger train operation quantity, train type, train operation section and stops along the way, etc. The train operation is the core of the high-speed railway passenger transport organizational management, and it is the basis of establishing the train working diagram and the train schedules. It is a good indicator of the railway passenger transport business strategy and service quality. Excellent passenger train operation plan can improve the business performance of railway passenger transport.

This paper provides a new idea to solve high-speed passenger train operation plan problem. First, all the possible train operation plans are listed, and then with the aid of intelligent algorithm those train operation plans which can not be sended out are excluded. Finally, the rest of the train operation plans just can meet passenger's travel demand. In this way, combining the genetic algorithm the corresponding algorithm is designed in this paper, and by means of an example we can verify the model and the algorithm.

2. Literature Review

In recent years, researchers have done a lot of related researches in the operation plan of high-speed railway passenger train, and have attained some good outcomes.

Some scholars studied the operation plan of high-speed passenger train operation plan considering the influential factors of the passenger train operation plan. Zhang Yongjun, Deng Lianbo, Shi Feng et al. mainly considered stopping at intermediate stations to study the high-speed train operation, established some corresponding models, and put forward the 0-1 programming, multi-objective programming, and intelligent algorithm to the solve the models[1-4]. Cui Bingmou, Luo Yupin et al. studied the problem from the economic perspective, such as the railway transport enterprise cost, the railway transport enterprise profit, passenger traffic fare, waiting cost, transfer cost, congestion tolls. They converted the economic cost into uniform, and established corresponding models [5-9]. Zhou Lixin, Zha Weixiong et al. studied the problem from the passenger OD passenger travel flow and path, and established multi-objective bi-layer models [10-12].

Some scholars studied the operation plan of high-speed passenger train operation plan considering the high speed and convenience of high-speed railway. Ye Huaizhen, Li Qingyun et al. studied the problem from high-speed railway high-speed characteristics, considering minimum train stop time and passenger waiting time, established corresponding model, and designed the solving method [13-14]. Shi Feng, Zheng Li, Qian Yongsheng et al. established some multi-objective bi-layer models considering the passenger convenience degree at different time [15 -17].

In summary, almost no one has studied this problem with the idea of this paper. The paper transforms the high-speed passenger train operation plan problem into passenger's demand of OD equilibrium assignment problem. On the

basis of the the OD equilibrium, we gradually eliminate redundant operation train plan, and this paper is innovative.

3. The Influential Factors of High-speed Passenger Train Operation Plan

Many factors influence the operation plan of high-speed passenger train. For instance, the train's plan, capacity, use ratio of seats, fluctuation coefficient of passenger flow, the relationship of capacity and service frequency, the quantity of passengers, the section density of passengers, the quantity of trains.

The paper sums up the factors with passenger's travel demand, passenger's need of comfort and the wasted train capacity.

4. The Model and Algorithm

4.1 The Model

The symbols are defined as follows:

The stations are defined as S , $S=\{s_i \mid i=1,2,\cdots,N\}$;

a_{ij} --the train travels from i to j;

m_{ij} --the number of passengers travels from i to j;

A_{ij}^{pq} --the number of passengers gets on the train a_{ij} from p to q;

L^{pq} --the distance of p to q;

a_{ij}^0 --the number of passengers gets on the train a_{ij} in the first station i;

a_{ij}^k --the number of passengers gets on the train a_{ij} in station k;

G—passengers' quota of a passenger train;

We defined x_{ij}^k （ $i,j \in S, i<k<j$ ） as stops variable: x_{ij}^k =1 if the train a_{ij} stops in the station k and x_{ij}^k =0 otherwise. Consider a variable x_{ij} such that x_{ij} =1 if exist train from station i to station j and x_{ij} =0 otherwise.

(1)Passenger's travel demand: we assume that if there exists a train a_{kp} starting from station k, which can meets the k station passenger's travel demand, namely $m_{kp} \le G$.

And other trains need not stop in the station k, namely

$$\sum_{i=1}^{k-1} x_{ip}^k = 0 \text{ and } x_{kp} = 1 \quad . \qquad (1)$$

While for a station without a start train, there must exist some trains stop in there, so it meets as follows

$$m_{kp} = \sum_{i=1}^{k-1} x_{ip}^k m_{kp} \qquad (2)$$

Namely $$\sum_{i=1}^{k-1} x_{ip}^k = 1 \quad \text{and} \quad x_{kp} = 0 \qquad (3)$$

In summary, for a station k, it either exists a start train or exists some stations stop in station k, so it meets as follows

$$\sum_{i=1}^{k-1} x_{ip}^k + x_{kp} = 1 \qquad (4)$$

Namely $$x_{kp} = 1 - \sum_{i=1}^{k-1} x_{ip}^k \qquad (5)$$

(2)Passenger's need of comfort: high-speed passenger train cannot overload passengers much. We stipulate a train should not overload passengers more than 10 percent. For train a_{ij} , the number of passengers is m_{ij} at station k-1, and after k stations the number of passengers is $m_{ij} + x_{ij}^k a_{ip}^k$. In order to meet passenger's comfort demand, we define that

$$1.1G < m_{ij} + x_{ij}^k m_{kp} \text{ then } x_{ij}^k = 0 . \qquad (6)$$

If $$1.1G \ge m_{ij} + x_{ij}^k m_{kp} , \text{ then } x_{ij}^k = 1 , x_{kp} = 0 \qquad (7)$$

So the number of train model is as followers:

$$z_1 = Min \sum_{i=1}^{n-1} \sum_{j=2}^{n} x_{ij}, i, j \in S, i < j \qquad (8)$$

(3)The wasted capacity : Because of even the same number of trains, different train operation plan will lead to different number of passengers getting on or getting off in some station, so we should consider the wasted capacity. The wasted capacity is calculated as follows:

$$z_2 = Min \sum_{i=1}^{n} \sum_{j=2}^{n} \sum_{k=2}^{n} (G - A_{ij}^{pq} x_{ij}^k) L_{pq} ,$$
$$i, j, k, p, q \in S, i < j, p < q \qquad (9)$$

We combine the Eq. (7) with the Eq. (8) as follows:

$$y = Min(z_1 w_e + z_2 w_s) \qquad (10)$$

Where w_e is the cost when transport enterprise departures a train, w_s is the cost of the wasted capacity of a seat.

Then the problem is transformed into a 0-1 multi-objective programming model. It is very tedious to use the traditional multi-objective programming solution, and calculation work is added more with the increase of the number of station and will take much time. Therefore it is not practical to use the multi-objective programming method to solve the model. So this paper chooses the genetic algorithm which has advantage of powerful search ability and high efficiency to solve the model.

4.2 Algorithm

In this paper, we first list all the possible solution, and then remove those trains which can not send out with the aid of the genetic algorithm on the base of passenger's travel demand, passenger's comfort demand and trains capacity. We stop our work until the minimum number of trains.

This paper uses genetic algorithm to find the optimal solution. Genetic algorithm for the train operation problem is as follow:

Step 1: Setting initial parameters. Setting population scale M is 200, crossover probability P_c is 0.6, mutation probability P_m is 0.1, and termination generation T is 500.

Step 2: Producing the initial population. On the basis of population scale and chromosomes, chromosome is a string consisted of 0 and 1 in turns which represents the stop or not in a station and start or not a train. 200 chromosomes meet the Eq.(1) to the Eq.(7) are generated randomly,which consists of the initial population, and every chromosome represents a solution.

Step 3: For individual j(j=1, 2, 3, …,M), calculate its fitness $f(X_i)$ by Eq.(8), selected probability p_i, and cumulative probability λ_i, $\lambda_i = \sum_1^i p_i$.

Step 4: Generating a uniform random number ξ on the interval [0, 1]. If $\xi \le \lambda_1$, then choosing first chromosome to copy; if $\lambda_{k-1} < \xi \le \lambda_k$, then choosing the ξth chromosome to copy. Iterate this process 200 times to generate next population.

Step 5: Crossover operation. Generate N uniform random number ξ_1, ξ_2, …, ξ_k, …, ξ_N on the interval [0, 1] in turns. If $\xi_k < 0.6$, choose the kth chromosome as a paternal chromosome to make up paternal population. Then crossover operation happens in third position.

Step 6: Mutation operation. Generate N uniform random number χ_1, χ_2, …, χ_k, …, χ_N on the interval [0, 1] in turns. If $\chi_k < 0.1$, choose the kth chromosome to mutate operation in second position.

Step 7: I=I+1. If I<T+1, go to step 3; Otherwise, go to step 8.

Step 8: Stopping calculation. Showing the solution x.

4.3 Example Test

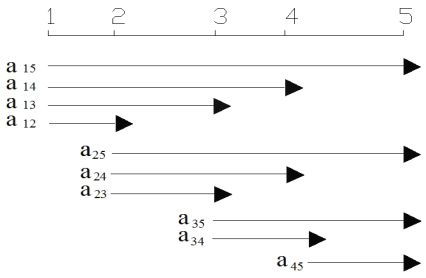


Fig.1 The stations and all train operation.

It shows in the above figure, there are five stations. For convenience, the stations are called as 1, 2, 3, 4, and 5. G=1,000. w_e = 25,000. w_s =15 .The OD traffic flow of network is shown in the following table:

Table 1: The OD traffic flow of network

Origin / Destination	1	2	3	4	5
1	--	--	--	--	--
2	580	--	--	--	--
3	380	410	--	--	--
4	360	350	530	--	--
5	460	450	540	430	--

After using the algorithm, the best train operation is obtained as follows:
Sending out train a_{15} and it stop in station 2; Sending out train a_{14} and it stop in station 3 and 4; Sending out train a_{25} stop and it station 3 and 4.

5. Conclusion

The paper sums up the factors with passenger's travel demand, passenger's need of comfort and the wasted train capacity. On the basis of these factors, this paper provides a new idea to solve high-speed passengers train operation problem and establishes a train operation model. The paper uses the genetic algorithm to get the best solution and will get train operation quickly.

Acknowledgment

This work is partly supported by the Humanities Social Sciences Programming Project of the Ministry of Education of China (no. 10YJA630126), the State Social Science Fund Project (no. 11CJY067), the Natural Science Foundation of Gansu Province (no. 1107RJYA070), the Young Scholars Science Foundation of Lanzhou Jiaotong University (no. 2012056), and the Natural Science

Foundation of Gansu Province (no. 1107RJYA070 and no. 1208RJZA164).

Reference：

[1] Zhang Yongjun, Ren Min, and Du Wen, "Optimization of High Speed Train Operation", Journal of Southwest Jiaotong University, Vol. 33, No. 4, 1998, pp. 400-404.
[2] Deng Lianbo, Shi Feng, and Zhou Wenliang, "Stop schedule plan optimization for passenger train", China Railway Science，Vol. 30, No. 4, 2009, pp. 102-107.
[3] Shi Feng, Deng Lianbo, and HUO Liang, "Bi-Level Programming Model and Algorithm of Passenger Train Operation Plan", China Railway Science, Vol. 28, No. 3, 2007, pp. 110-116.
[4] Xu Ruihua, and Zou Xiaolei, "Study on train plans optimization for passenger traffic special line", Journal of Tongji University: Natural Science, Vol. 33, No. 12, 2005, pp. 1608-1703.
[5] Zhou Li-xin, "Decision Model of Railway Passenger Trains Plan", Journal of Shanghai Jiaotong University, Vol. 34, No. 9, 2000, pp. 11-14.
[6] Shan Wenhao, and Ye huaizhen, "The Study of Confirming Train.s Utilized Rate Based on the Principle of Economic Benefit", Journal of Southwest Jiaotong University, Vol. 35, No. 3, 2000, pp. 235-238.
[7] Cui Bingmou, Fan Jin, "Researching & Developing the Financial Analysis System of Operating Passengers Trains", Railway Transport and Economy, No. 6, 1996, pp. 43- 48.
[8] Qian Yongsheng, Wang Hailong, Kang Hongxia, Zeng Junwei, "Simulation of Unparllel Train Schedule of Double-Track Automatic Blocking Base on Celluar Automaton Medel", International Conference on Information Computing and Automation, 2008, Vol. 13, pp. 464-467.
[9] Luo Yupin, Ye Huaizhen, and Chen Jihong, "The Benefit Principle of Determination of Through Passenger Trains", Journal of Southwest Jiaotong University, Vol. 33, No. 4, 1998, pp. 405-410.
[10] Zhou Lixin, "Decision Model of Railway Passenger Trains Plan", Journal of Shanghai Jiaotong University, Vol. S1, 2002, pp. 11-14.
[11] Zha Wei-xiong, FU Zhuo, "Research on the optimization method of through passenger train plan", Journal of the China Railway Society, 2000, Vol. 22, No. 5, pp. 1-5.
[12] Sang Nguyen, Stefano Pallottino, and Federico Malucelli, "A Modeling Framework for Passenger Assignment on a Transport Network with Timetables", Transportation Science, Vol.35, No. 3, 2001, pp. 238-249.
[13] YIE Huaizhen, YANG Yonglan, and WANG Yan, "The Discussion about the Operation of Long-Short Distance Passenger Trains", Journal of Southwest Jiao tong University, Vol.35, No. 3, 2000, pp. 230-234.
[14] Li Qing-yun, Chen Zhi-ya, "Research on Train Plan of Passenger Train", Journal of Changsha Railway University, Vol. 19, No. 4, 2001, pp. 31-34.
[15] Shi Feng, Deng Lianbo, "Optimal Design of Passenger Transfer Network", Journal of Railway Science and Engineering, Vol. 1, No. 1, 2004, pp. 78-82.
[16] Zheng Li, Song Rui, He Shiwei, et al. , "Optimization model and algorithm of skip-stop strategy for urban rail transit", Journal of The China Railway Society, Vol. 31, No. 6, 2009, pp.1-8.
[17] Qian Yongsheng,Shi Peiji, Zeng Qiong, Ma Changxi, Yin Xiaoting, "Analysis of the influence of occupation rate public trasit vehicles on mixing traffic flow in a two-lane system", Chinese Physics B, Vol.18, No. 9, 2010 , pp. 4037-4041.

Xin Feng received her M.A. in the Southwest University of Finance and Economics, Sichuan, China, in 2006. She is now a lecturer in the School of Economics and Management of Lanzhou Jiaotong University. Her main research interest is traffic economics and accounting.

Jinbao Luo received his B.E. in the School of Traffic and Transportation of Lanzhou Jiaotong University, Gansu, China, in 2010. He is now a postgraduate in the School of Traffic and Transportation of Lanzhou Jiaotong University. His main research interest is simulation of transportation and traffic economics.

Yongsheng Qian received his Ph.D. in the Northwest Normal Uniwersity, Gansu, China, in 2010. He has been a full professor in the School of Traffic and Transportation of Lanzhou Jiaotong University. His main research interest is simulation of transportation and traffic economics.

Design of a Pneumatic Robotic Arm for Solar Cell Tester System by Using PLC Controller

[1]Yousif I. Al Mashhadany*, [2]Nasrudin Abd Rahim

[2] University of Malaya Power Energy Dedicated Advanced Centre (UMPEDAC),
Level 4, Wisma R&D, University of Malaya, Jalan Pantai Baharu,
59990 Kuala Lumpur, Malaysia,
[1]Electrical Engineering Department, College of Engineering, University of Al-Anbar, Al-Anbar, Iraq

Abstract

Solar cell testers sort photovoltaic cells according to their electrical performance, tested under simulated sunlight. A variety of testers exist, but they all face a common challenge of handling cells that are very small and thin, which makes it difficult to transport the cells from the conveyer to the storage box. This paper presents a new design for a handling robot with vacuum end-effectors, which uses a PLC controller to govern the movement of the cells and the testing process. The design applies to solar cell testers for monocrystalline, polycrystalline, cadmium telluride (CdTe), and copper indium diselenide (CIS) cells. Each cell is tested for efficiency and categorized accordingly into four groups (A to D). A Virtual Reality (VR) model was built to simulate the system, keeping in mind real world constraints. Two photoelectric sensors were used to make detections for both the testing process and the robot movement. The PLC controller guides the trajectory of the robot according to the results of the efficiency testing. It was seen that the system worked very well, with the testing process and the robot movement interacting smoothly. The robot trajectory was seen to be highly accurate, and the pick and place operations were done with great precision.

Keywords: *Handling robot, Solar cell tester, Virtual reality, photoelectric sensor.*

1. Introduction

The testing of solar cells both in the laboratory and for commercial uses, solar simulators are used. The simulators employ a single light such as that from a xenon arc lamp to produce a beam of collimated light that matches a reference solar spectrum. Despite some advantageous features for simulating a solar spectrum, the beam from a xenon arc lamp cannot be directly used. This is firstly because a considerable spectral range, especially between 0.8m and 1.2m, is covered in the emission lines of the lamp output. Secondly, the richness of the Ultraviolet (UV) region in the lamp output is an issue. To get around this problem, there are proprietary notch filters that decrease IR and UV components of the output, which is why many types of optical filters are required to obtain a useful spectrum. A water cavity is sometimes used to enhance spectral matching. Still, a number of disadvantages of the xenon lamp remain [1- 6]:

(i) The wavelength of the peak of the solar spectrum and the xenon spectrum do not coincide.
(ii) Intense emission lines remain in the output.
(iii) The liquid water cavity's absorption spectrum and that of the water vapor are not the same.

A cheap alternative to generating solar electric power are Silicon-Film solar cells. The production systems of Silicon-Film use a continuous in-line process to produce polycrystalline silicon sheets. The Apex sheet growth process has continuously progressed and after five design generations, one sheet can give an annual yield of over 15 MW of 200-mm wide polycrystalline silicon sheet, which is used to make large-area APx-8 solar cells. These cells generate over 4 W each and have edge dimensions of 208mm x 208mm. Parallel process systems are often used to increase the volume of solar cells manufactured. However, with modern large-area Apex sheet generation approaches and the development of solar cell process tools, single-thread process systems can be used to design large volume production lines for solar cells [7-11].

The greatest verified efficiencies for a number of photovoltaic cell and module technologies is regularly recorded in many references. The information has helped researchers stay aware of the state of the art and to document their own independent findings in a standardized manner. All results have to be verified by a recognized test center before they are included in the tables [12-16].

There are three significant areas relevant to each photovoltaic device: designed illumination area, aperture area and total area. The efficiencies of the 'active area' are excluded. Depending on the type of device, a particular

minimum area is an important parameter (800 cm^2 for modules, over 0.05 cm^2 for concentrator cells and 1 cm^2 for one-sun cells). Cells and modules can be made from a variety of semiconductors. Also, each semiconductor can be further categorized into different types (such as thin film, crystalline and polycrystalline). By provides spectral information with plotting the external quantum efficiency (EQE) against the wavelength, with normalization done to the peak measured value [16-19].

Recent years have seen a dramatic increase in the solar industry, due to which solar cell and module test solutions are widely sought after. The solar cell modules are generally of two types: comprehensive turnkey solutions and test-system building blocks that must be assembled. The former are easy to set up but expensive. Furthermore, the technology used in the turn-key solution is likely to become outdated quickly, thus requiring an upgrade. With building blocks though, the system is more reasonably priced and is easily modified when required. If there is a need to upgrade to a higher current range or accuracy, only one relevant block would need to be replaced. Also, sets of blocks that are useful for a variety of platforms can be standardized and reused [20,21].

Many automotive facilities are used in the solar cell manufacturing process for mass production. In thin film solar cell production especially, a key task is to handle large size solar cell substrates like the LCD production system. Robots like the serial manipulator, beam type and link type robot can be used to handle the large substrate. A variety of handling robots have been developed over the years for the manufacturing line for solar cells. With an increase in the size of the substrate, vibration control and dynamic analysis assumes greater importance. Since the weight of the solar cell substrate is at least three times more than the LCD glass substrate, the position control needs to be precise, especially considering the vibration of the substrate and forks. The motion analysis of the beam type handling robot using Matlab/ SimMechanics was performed by the authors in a previous research [22 - 24].

In this paper, the design and analysis of a handling robot to transfer solar cells from the surface of the conveyer to four main boxes is proposed, with special consideration of the percentage of efficiency. The paper is divided as follows: Section I is the introduction. Section II explains the problem formulation. Section V presents the design of a prototype by VR. Section IV describes the design of the controller. Section V explains the simulation of design with VR and finally conclusion.

2. Problem formulation

A solar cell tester designed previously at University Malaya's Power Energy Dedicated Advanced Centre (UMPEDAC) was used in this work. The structure of the tester is illustrated in Fig. 1. The tester moves the cells upon the conveyer till it enters the texting box. The entry is detected by a photoelectric sensor, after which the tester box commences work. Each solar cell undergoes a testing process of 2.5 sec duration. Based on the test results, the solar cells are divided into four groups.

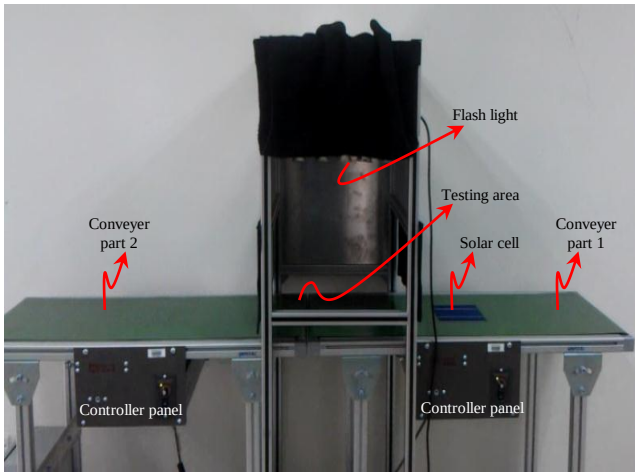


Fig. 1. Real solar cell tester

The tester calculates efficiency values of the solar cells in series, which are used to determine the appropriate box the cells are to be placed in. The robot design thus has two main objectives: to pick the solar cells from the conveyer and place them in the suitable box corresponding to the efficiency value.

3. Design Robot Arm

The robot is designed in two main phases. In the first phase, the robot is designed in a VR environment and the controller performance is tested in simulation. VRLM software was used for the design.The 'classical objects' feature in the software was used to create the preliminary design. The objects were then redrawn using the indexed-face option to obtain the advanced design. Fig. 2 illustrates the design of the robot arm in VR.

To design and simulate the movement of the robot, a trajectory was set for the robot after a number of trials. The trajectory advances in three stages (Fig. 3): first, the cell is handled and moved vertically up to a specified point; second, the cell is moved horizontally to a certain point and finally the cell is lowered vertically to box A, B, C or

D. The robot needs to complete this trajectory in 30 sec. Due to the thin construction of the solar cell and light weight (20 gm), it is quite difficult to pick the cells from the conveyer.

To ensure that each part of the robot was best suited to its job, a number of robot designs were tested in simulation. For picking the cells from the conveyer (which requires accurately capturing them through the conveyer trajectory) and placing the cells in the appropriate box, the vacuum technique was used. Two vacuum grippers were used for this purpose, as is shown in Fig. 3.

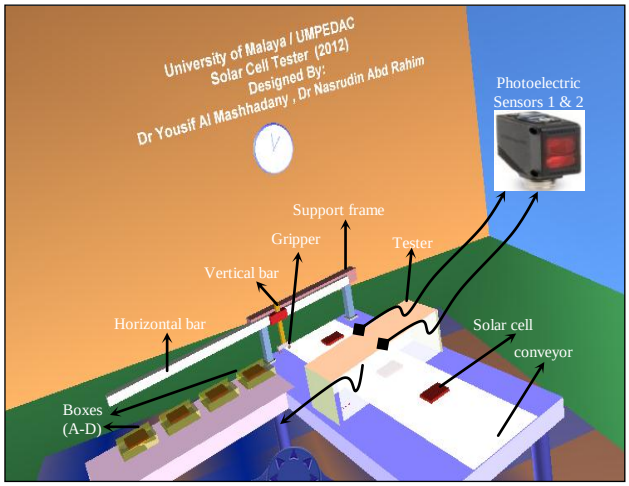


Fig.2. Design robot arm in VR environment

Fig. 3. Suction grippers of robot

For high speed up-down vertical movement, the electric cylinder DNCE with a mechanical linear axis and piston rod was selected (Fig. 4). The cylinder has low weight and fast response. The drive component consists of an electrically driven spindle, which converts the rotary motion of the motor into the linear motion of the piston rod. The mechanical interfaces are largely compatible with the standard cylinder DNC.

The power electronics with positioning controller were designed as an external field device (IP54). Speed, power and position can be set independently of each other and up to 31 travel profiles can be stored. The positioning controller is suitable for stand-alone operation or can be externally controlled via an I/O interface or fieldbus. The movement of the robot in horizontal direction is achieved using the electric toothed belt axis, drive axis (for applications with external guide or for easy handling tasks), plain-bearing guide, toothed belt covered by a steel band and a flexible motor mounting on all 4 sides of the axis. The arrangement provides high feed forces, with speeds up to 5 m/s and acceleration of 50 m/s^2. It also features space-saving position sensing with proximity sensors in the profile slot suitable for electric axis EGC with a recirculating ball bearing guide (Fig.5). All the different parts are assembled virtually and reconstructed in one frame to achieve the complete robot shown in Fig. 6.

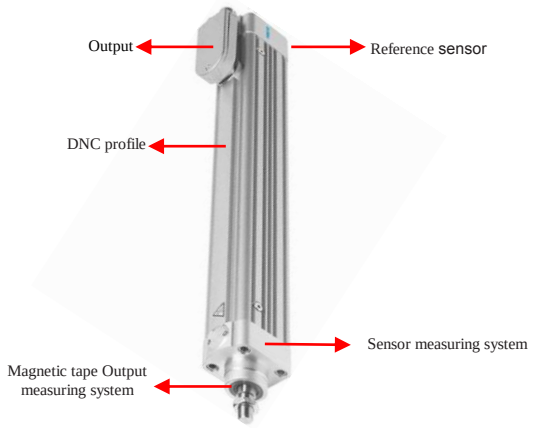


Fig. 4. The electric cylinder DNCE is a mechanical linear axis with piston

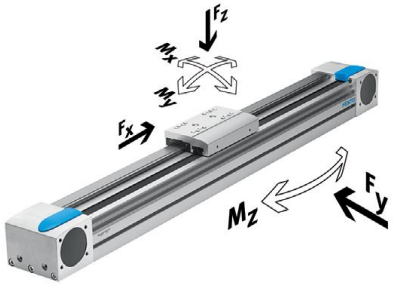


Fig. 5. Electric toothed belt axis

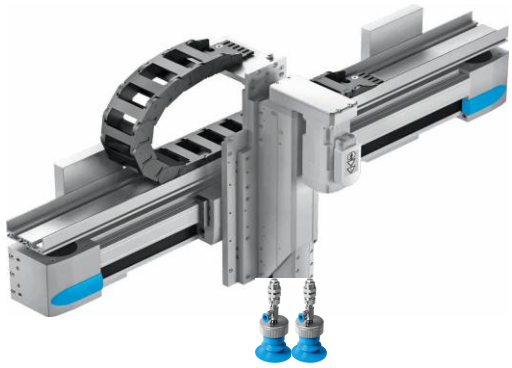


Fig. 6. The overall form of robot design.

Program interfaces made by Festo were used to select the suitable parts of the robot. In the program, application parameters such as mounting position, mass, stroke and precision are input and the required process time is specified. The drive technology can also be preselected. The required solution package is selected, which are sorted by motor and axis technology, component utilization, cycle time or price. The program also provides detailed results such as motor characteristic curve, dynamic characteristic values, system data, product data, and parts list.

4. Controller Design and simulation of solar cell tester

The solar cell tester consists of three main parts: the tester, conveyer and handling robot. Each part requires its own control strategy. The controller has to both determine the working of each part in stand-alone mode and also govern the interaction of the different parts of the tester. Taking the environment of the tester into consideration and the variety of signals that need to handled, a PLC controller was seen as a suitable option to create the controller. Two photoelectric sensors are used; one detects the entry of the solar cell into the testing area (which turns on the tester flash) and the other to detect the movement of the robot to start the handling operation. The duration of robot operation for each cell, from picking the cell and returning to the initial position, is 10 sec. The time for testing each cell and obtaining the efficiency value is 10 sec. The rate of testing for the system is thus 360 cells/hour.

The block diagram of the testing process is shown in Fig.7. The block diagram used for VR simulation is shown in Fig.8. The diagram shows the blocks used for implementing all the steps of the tester: movement of the cells on the conveyer, working of the flash lamp, measurement of efficiency and the handling operation of the robot (vertical and horizontal movements, capture and release of solar cells at the right locations).

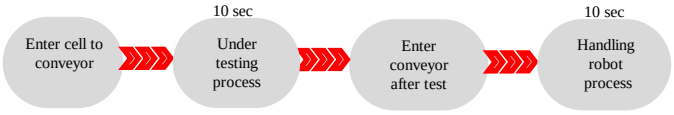


Fig. 7. Block diagram of testing process.

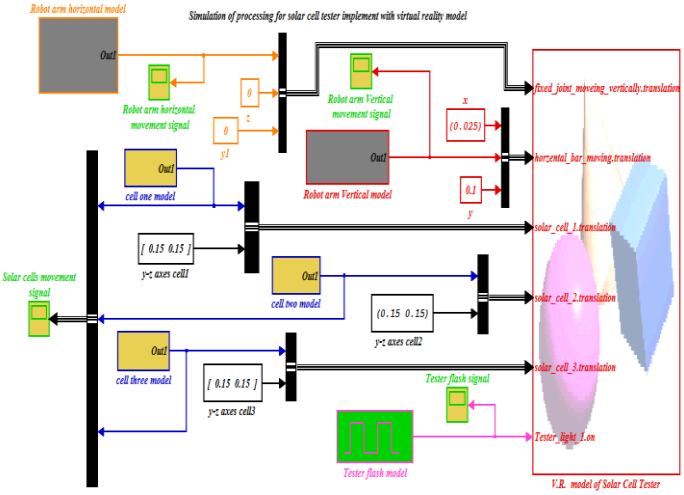


Fig. 8. Simulation of solar cell tester in V. R. environment

5. Simulation Result

The simulation of the design was achieved by using Matlab Ver. 2012a with VR environment to implement the movement of system by interfacing with Matlab/Simulink. Fig.9 presents the simulation results by using oscilloscope display of the control signal for each part of the solar cell tester.

The operation of the cell testing system commences with the movement of the solar cell upon the conveyer. Scope (a) in Fig.9 shows the movement signal for three cells. As is evident from the signal, the sequence repeats after every 50 sec. After 10 sec, the first cell arrives at the tester and the flash tester turns on and off for 5 seconds each. The sequence is then repeated for the second cell (as is clear from the tester flash signal in scope (b) of Fig.9). After 25 sec, the first cell arrives at the capture point to be picked by the gripper of the robot. The down-up signal for the gripper is shown in scope (c) in Fig.9. The horizontal movement of the robot depends on the efficiency of the cell calculated by the tester. Depending on the box number that the cell is to be placed in, the amplitude of the

horizontal movement signal (scope (d) of Fig.9) is limited. The duration of one cycle, from passing the entry point to being placed in a box is 50 sec. The output of the virtual model that simulates the movement is shown in Fig. 10.

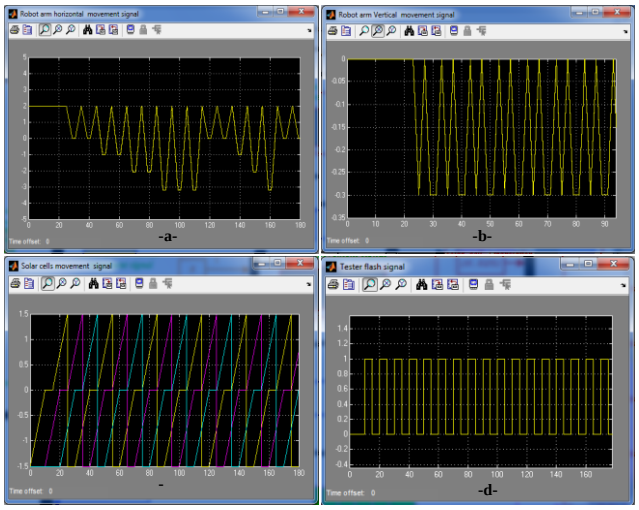


Fig. 9. (a) Robot arm horizontal movement signal. (b) Robot arm vertical movement signal. (c) Solar cells movement signal. (d) Tester flash signal.

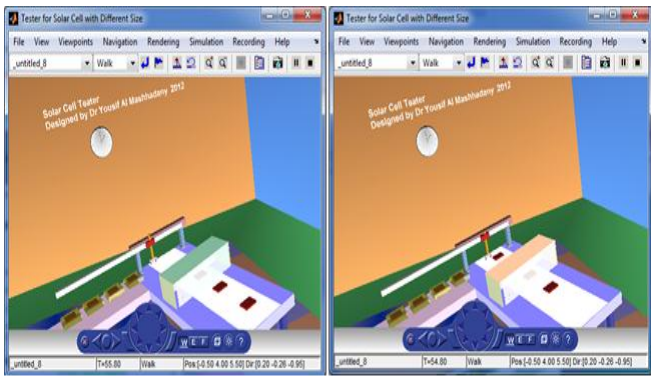


Fig. 10. Implement solar cell tester in V.R. model.

6. Conclusion

In this work, the design of a handling robot with vacuum end-effectors was created to sort solar cells according to their electrical performance. A PLC was used as the controller. The motion of the robot and the testing process were controlled based on detections by photoelectric sensors. The cells were allotted to suitable boxes based on their efficiencies. A highly accurate robot trajectory was obtained, coupled with very accurate position control for placing the cells inside the boxes. Excellent performance is obtained in the simulation of the tester in a V. R. model. The photoelectric sensors were seen to be very effective for detection of cell entry and robot movement. Future works can investigate other designs for the robot and look into different methods of controlling the robot trajectory.

Acknowledgement

Thanks to UMPEDAC for various resources and R&D project 110775 that funded this research, and to University of Malaya for the post-doctoral research fellowship.

References

[1] Bhushan L, Craig M. Design of a fiber optic based solar simulator, IEEE Photovoltaic Specialists Conference, 1991, pp. 783-788

[2] Green M. A., Emery K, Hishikawa Y, Warta W. Solar cell efficiency tables (version 37), progress in photovoltaic: research and applications, 2011, 19, pp:84–92

[3] Glatthaar, M, Rein, S. Separation of series resistance and space charge region recombination in crystalline silicon solar cells from dark and illuminated current–voltage characteristics. IEEE journal of photovoltaic, Vol. 2, issue 3, 2012, pp. 241-246

[4] Sharma K, Peterson D, Jun B, Hanley J. Reliability testing of large area 3J space solar cells. 37th IEEE photovoltaic specialists conference (PVSC), 2011, pp: 003707 – 003712

[5] Alam A, Upadhyay S, Murthy H, Reddy B, Jana C, Mohanta K. Reliability evaluation of solar photovoltaic microgrid. International Conference on Environment and Electrical Engineering (EEEIC), 2012, pp: 490- 495

[6] Nelson J, The physics of solar cells, ISBN: 978-1-86094-340-9, May 2003, pp: 384

[7] Wurfel P, physics of solar cells: From Basic Principles to Advanced Concepts, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, ISBN: 978-3-527-40857-6, 2009, PP:181

[8] Mohammad S, Suria C, Zurita Z, Faridah J, Kasmiran J. Portable Wireless CATV Tester Unit with Solar Panel, IEEE International Conference on Space Science and Communication,2009, PP: 15-18

[9] Hacke P, Terwilliger K, Smith R, Glick S, Pankow J, Kempe M, Kurtz S. System voltage potential-induced degradation mechanisms in PV modules and methods for test. The 37th IEEE Photovoltaic Specialists Conference (PVSC 37), 2011, pp: 1-9

[10] Meiqin1 M, Jianhui S, Chang L, Kai1 P,Guorong Z, Ming D. Research and development of fast field tester for characteristics of solar array. IEEE electrical and computer engineering, 2009, pp: 1055-1060

[11] Osterwald C. Terrestrial photovoltaic module accelerated test-to-failure protocol. Technical report, National Renewable Energy Laboratory 1617 Cole Boulevard, Golden, Colorado 80401-3393 303-275-3000 • www.nrel.gov, March 2008, pp:1-16

[12] Warner H, Messenger S, Lorentz, Summers G. On the need for low energy proton testing of space solar cells. IEEE 4th World Conference on Photovoltaic Energy Conversion, Vol. 2, 2006, pp: 1899 - 1902

[13] Shammas E , Brown B, Choset H, New joint design for three-dimensional hyper redundant robots, IEEE Intl. Conference on Intelligent Robots and Systems, Nevada, 2003, pp:3594-3599

[14] Rigatos G. Control of robotic systems with flexible components using hermite polynomial-based neural networks, Ch-25.book, robot manipulators, new achievements, Unit of Industrial Automation Industrial Systems Institute,26504, Greece, 2012, pp:456-486

[15] Rehiara A. Kinematics of Adept Three Robot Arm, ch. 2, at book, Robot Arms, University of Papua, Indonesia, 2011, pp: 21-38

[16] Potgieter J, Zyzalo J, Diegel O. Reconfigurable Mechatronic Robotic Plug-and-Play Controller. Ch: Cutting Edge Robotics, ISBN 3-86611-038-3, Germany, July 2005, pp: 771-784

[17] Elfasakhany A, Yanez E, Baylon K, Salgado R. Design and Development of a Competitive Low-Cost Robot Arm with Four Degrees of Freedom. Journal of modern mechanical engineering, 2011, 1, pp: 47-55

[18] Lathrop R, Pfluke K, Novel approaches to benchmarking solar cell tabbing solder ability. Proceedings of the 26th European Union Photovoltaic Solar Energy Conference, 2011 , pp: 1-6

[19] Student Guide, Robotics with the Boe-Bot, Ver. 2.2, ISBN 1-928982-03-4, 2011 , pp:1-360

[20] Park C, Park D, Min H. Controller design and motion simulation of solar cell substrate handling robot in vacuum environment, 11th International conference on control, automation and systems, 2011, pp: 1017-1019

[21] Park C, Park D, Min H. Motion simulation model for beam type solar cell substrate transport robot, The 8th IEEE International Conference on Ubiquitous Robots and Ambient Intelligence, 2011, pp:796-799

[22] Wang Z, Gong Z, Wang Y, Wei G. Study on Positioning Control of Transfer Robot with Solar Cell, The ninth IEEE international conference on electronic measurement & instruments, 2009, pp:3,913 3,919

[23] Zhangl X, Panland H, Wul G. Photovoltaic generation and Its applications in DC-motor, IEEE International conference on communications, circuits and systems, 2010, pp: 609-611

[24] Zarafshan P, Ali S, Moosavian A. Manipulation control of a space robot with flexible solar panels, IEEE/ASME International conference on advanced intelligent mechatronics, 2010, pp: 1099-1104

Yousif Ismial Mohammed AL Mashhadany (MIEEE, MIIE) was born in Baghdad 1973. He received the B.Sc. degree in Electrical and Electronic Engineering Department (1995) from the AL-Rasheed College of Engineering and Science / the University of Technology Baghdad / Iraq.M.Sc degree in Control Engineering (1999).and Ph.D degree in Control (2009) from the AL-Rasheed College of Engineering and Science / the University of Technology/ Baghdad / Iraq. Since 2004, he has been working at the University of Anbar – Iraq, as a Lecturer in the Electrical Engineering Department. His research interests include biomedical, robotic and control system. He has more than thirty publishing at journals and International conferences and two books. Now, he has Post-Doctoral research fellows at university of Malaya – UMPEDAC.

Professor Dr. Nasrudin Abd Rahim has been an academician and an active researcher for most of his professional life. Upon graduating with his first degree in 1985, a B.Sc. (Hons.) in Electrical Engineering from the University of Strathclyde in Glasgow, UK, he served briefly in industry, in the capacity of Planning Engineer with a Malaysian telecommunications company. He then entered the world of academia, at the University of Malaya, progressing from Tutor, in 1986, to Lecturer, a year later, and to his doctoral degree in 1988. He was promoted to Associate Professor in 1998, and to Professor in 2003. He was made a Chartered Engineer on 1st January 2000. His professional contribution in imparting and adding to knowledge, and furthering progress in his field, extends beyond the perimeters of the university. At regional and international levels, he has served in various capacities for IEE and IEEE, the most recent has been the Chair of the IEEE Power Engineering Society Motor Sub-Committee Working Group 8 upon his election to the position in 2006. He also has been made the Malaysia SEE Forum Coordinator in 2009. In research, he has led, and co-researched, projects of national and international importance, with funding totalling millions of ringgit. He is a published author of more than 100 refereed journals, 175 articles, and books, and has had more than fifteen PhD candidates graduated. Among many other academic and field-related awards won locally and internationally, is the AUNSEED-Net Research Grant. HE is director for University of Malaya Power Energy Dedicated Advanced Centre (UMPEDAC) from three years.

The study on the spam filtering technology based on Bayesian algorithm

WANG Chunping

Mathematics and Computer Science College, Xinyu University
Jiangxi, Xinyu 338000, China

Abstract

This paper analyzed spam filtering technology, carried out a detailed study of Naive Bayes algorithm, and proposed the improved Naive Bayesian mail filtering technology. Improvement can be seen in text selection as well as feature extraction. The general Bayesian text classification algorithm mostly takes information gain and cross-entropy algorithm in feature selection. Through the principle of Bayesian analysis, it was found that the characteristics distribution is closely related to the ability of the feature representing class, so this paper proposes a new feature selection method based on class conditional distribution algorithm. Finally, the experiments show that the proposed algorithm can effectively filter spam.

Keywords: *Naive Bayes, minimum risk Bayesian, active learning Bayesian, feature selection, email filtering*

1 Introduction

With the rapid development of Internet, interaction between people become more convenient, e-mail, with its quick and low-cost features, gradually become an important tool for interaction. People use it to exchange ideas, transfer files, and express their views, so it has become an indispensable communication tool in daily life. But at the same time it also brings some negative effects and a large part of the mail we receive each day are unsolicited. Some of them are commercials, some political propaganda, some pornographic advertising, there are even viruses. These are what we commonly known as spams.

The economic loss caused by spam to Internet users is quite staggering: According to statistics, only download Internet access fees and phone charges and other expenses a year will cost $ 94 of the world's Internet users. As the sender of the spam, the price is low, usually through mass email in a variety of ways. For e-mail service providers and users, spam gave them a lot of damages and losses, and the losses caused by pornography, computer viruses, and load fraudulent letter are inestimable. Despite some disputes on Bayesians philosophical view, it is undoubted that the ideas and methods are widely used in the social life and production practice. In particular, in recent years, the Bayesian approach, with its uncertain knowledge for its unique form of expression, the probability of rich expressive power and the priori knowledge incremental learning characteristics become the focus of many ways the most compelling data mining.

In 1996, Rvennie set up ifile, a machine learning applications for email filtering system based on Bayesian algorithm, which can use Bayesian algorithm to categorize messages. In the process of establishing ifile system, Rennie noted that each user has a different set of e-mail, and different way to organize messages, thus allowing the user to manually adjust the false positive mail. In 1998, Sahami, in using Bayesian algorithm to filter messages, noted that spam has unique properties different from the legitimate mail: For example, in the class of to get rich quickly spam, in addition to text messages like “free” and “money”, there will be a large number of stressed symbols like "!" and the representative symbol "$". Using Naive Bayes algorithm to filter mail, Sahamiliy hand-joined the domain information for these specific tasks phrase as well as spam features to filter, improving the accuracy of filtering spam; in addition, he is also using a characterization loss rate threshold to reduce false positives of legitimate mail. In 2001, Matthew and others developed a spam filter MEF. MEF can filter out virus e-mail whose attachments have the executable program in UNIX. The mail filter first decodes the binary code of the executable program, compares it with the existing binary code of the virus, uses Naive Bayes algorithm to calculate the probability that it belongs to spam, and makes decisions accordingly.

2. Naive Bayesian spam filtering basic theory

2.1 Naive Bayesian principle

Bayesian approach is an important method of spam filtering, the essence of the method is to identify messages as junk mail or regular mail, which is a classification problem.

Suppose that there are m sample spaces $\{c_1, c_2, ..., c_n\}$, and the mail d has n feature items $(w_1, w_2, ..., w_n)$. The probability of d belonging to the class c_k for a given class c_k (k = 1,2, ..., m) is

$$p(c_k \mid d) = Max\{p(c_1 \mid d), p(c_2 \mid d), ..., p(c_n \mid d)\}$$

By Bayesian probability formula we can get:

$$p(c_k \mid d) = \frac{p(d \mid c_k)p(c_k)}{p(d)} \quad (k=1,2,\ldots,\ m)$$

In which:

$$p(d \mid c_k) = p(w_1, w_2, ..., w_n \mid c_k)$$

The denominator p (d) in formula (3) has nothing to do with the class, so it can be ignored when comparing maximum value in the equation (3). So we only need to calculate the probability $p(c_k)$ and $p(d \mid c_k)$ to categorize mail d.

In equation (4), $p(c_k)$ is a priori probability and easy to calculate, but the calculation of $p(d \mid c_k)$ is more difficult, particularly when the number of feature items is large and the dependence between the feature item is high, so the calculation would take a lot of time. In order to simplify the calculation, we introduced the conditional probability independence assumption, that is assuming that each feature items are independent of each other—the naive Bayes filter, then the formula (2-5) can be converted to:

$$p(d \mid c_k) = p(w_1, w_2, ..., w_n \mid c_k) = \prod_{i=1}^{n} p(w_i \mid c_k)$$

Naive Bayesian filter structure is shown in the following figure:

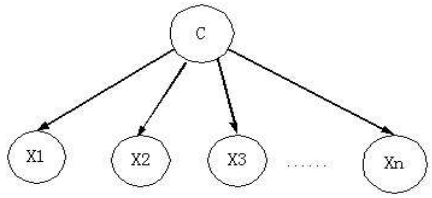


Figure 1 Naive Bayesian filter structure

Naive Bayesian filter uses priori probability to obtain the posterior probability, sets filters according to the training sample, and classifies emails according to the posterior probability of the message text.

2.2 Naive Bayesian mail filtering technology

Literature used Naive Bayes algorithm to design spam filtering system SpamCop [2]. The system is able to identify about 92% of spams with a fault correction rate of 1.16%. SpamCop system made improvements in keyword selection principle, including ignoring spaces, continuous sequence of letters and numbers, as well as getting rid continuous sequence of characters less than three characters apart from above mentioned characters. And the system used m in the calculation of the probability of the feature classes to estimate, and used the following formula:

$$P(Token / Ham) = \frac{N(Token, Ham) + \frac{1}{K}}{N(Ham) + 1}$$

$$P(Token \mid Spam) = \frac{N(Token, Spam) + \frac{1}{K}}{N(Spam) + 1}$$

$$P(Spam \mid Token) = \frac{P(Token \mid Spam)}{P(Token \mid Han) + P(Token \mid Spam)}$$

In these formulas, $P(Spam \mid Token)$ represents the posterior probability of feature Token belonging to the garbage category. $N(Token, Spam)$ represents the number of occurrences of the keyword in the spam, $N(Token, ham)$ represents occurrences keywords in the normal circumstances. $N(Ham)$ is the number of normal mail, and $N(Spam)$ the number of spam. K stands for the number of different keywords in the mail, solving the problem of zero possibility.

Literature provides an effective Bayesian spam filtering method [3]. The filter captures 99.5% of spam with a fault correction rate of less than 0.03%. The filter sets up two hash tables for spam and normal mail to calculate the occurrence of keywords of corresponding Corpus. To calculate the probability of each keyword, we use the following formula:

$$p(W \mid C) = \frac{b / nbad}{2 * g / ngood + b / nbad}$$

In the formula, b represents the number of occurrences of keywords in the spam, g represents the the number of occurrences of the keywords in the regular mail, nbad the total number of spam, ngood the total number of normal mail. A factor 2 in the denominator is a recommended empirical value, used to reduce the probability of normal mail as spam. In the calculation of the joint probability filter uses the following formula:

$$p(d_x \mid c_{spam}) = \frac{p_1 p_2 \cdots p_n}{p_1 p_2 ... p_n + (1-p_1)(1-p_2)...(1-p_n)}$$

Wherein p_i (i = 1,2, ..., N) represents the probability of the I keyword being calculated.

Data sparseness problem is often encountered when

using the formula (1-5) a mail to calculate spam probability, that is, if the message contains a new feature, no matter how high the possibility of this message containing other features items, it will cause zero conditional probability. This is not to ignore the problem. In literature [3], it is given a zero probability smoothing formula, a better solution of zero probability problems. The formula is as follows:

$$f(w) = \frac{a * x + (n * p(w))}{a + n}$$

In this formula, a is an adjustable constant, and n is spam and normal mail aggregate that contains the characteristics of w, x is the initial probability, and when n = 0, f (w) equals to the initial probability, as n increases , f (w) becomes more and more close to the p (w). Based on experience, the initial probability is generally set as 0.52, a is 0.0178. Literature [4] improved formula 1-5 and provided new method in calculating Keywords joint probability.

$$P = 1 - \sqrt[n]{(1-p_1) * (1-p_2) * \ldots * (1-p_n)}$$

$$Q = 1 - \sqrt[n]{p_1 * p_2 * \ldots * p_n}$$

$$S = \frac{P - Q}{P + Q}$$

S value is between -1 and 1. The high value means spam and low means regular mail. 0 means between the two.

Most of the filter built using naive Bayes got improved on the basis of Bayes formula, and did not take in to account the difference between mail filtering and ordinary text classification. On the other hand, these traditional Bayesian methods are based on the minimum error rate of the decision-making methods, not taking into account different characteristics between the legitimate mail and spam, which is that legitimate e-mail misidentified as spam may give users a greater loss. In addition, traditional Bayesian learning algorithm used given the training sample to learn classification parameters. The training samples it dealt with must be with a category label, and be randomly selected with passive acceptance of these samples. In this paper, the Bayesian spam filtering process is studied, the improved method of feature selection process is proposed with two naive Bayes extension models: minimum risk Bayes and active learning Bayes.

4 Improved Naive Bayesian filter design

To better use Bayesian algorithm in spam filtering, this paper improved Bayesian algorithm in the following aspects:

(1) Text representation
In ordinary text classification Bayes algorithm, the text was represented by a word or phrase. Words and phrases are the smallest unit that can represent semantics. In spam, in order to avoid being filtered the spammers use variants of junk words instead of junk words.

(2) Feature selection
The ordinary Bayesian text classification algorithm feature selection mostly take the information gain, and expected cross entropy algorithm. Through analysis of Bayesian principle, it was found that the characteristics distribution is closely related to the ability of the feature representing class, therefore a new feature selection method based on class conditional distribution algorithm is proposed.

4.1 Text representation

In text classification, usually we usually use the vector model (VSM) to represent text, which can be represented as an n-dimensional vector $(t_1, t_2, t_3, ..., t_n)$, in which $t_i (i = 1, \ldots, n)$ represents the weight of the i-th feature items.

The feature item is usually defined as a series of consecutive character string separated by spacebar, tabs or various punctuation marks and accents in the English text. Under normal circumstances, the feature item is a meaningful word or phrase. In character handling, all uppercase letters are converted to lowercase. All spacebar, tabs, line breaks, and various punctuation marks and accents are removed.

In Chinese, text feature item is a character, word, phrase, or some kind of concept. In the Chinese text, they mainly refer to vocabulary after word processing. But in comparison of several spam messages similarity, we found that block phrases appear more often in similar spam. And the spammers now in order to avoid being filtered, often use vocabulary variants to prevent being filtered, so, in the ever-changing spam variants, the simple word characteristics can no longer meet the requirements.

Fingerprint is applied in the comparison of similar mails. When we compare two mails, two mails can be divided into a number of blocks of text (actually a sub-string), if two mails are similar, they must contain a lot of same text blocks. And the comparison operation between these texts blocks is accurate comparison, therefore can be to be optimized with hashing method.

4.2 Feature Selection

Feature selection is an important area of research in text categorization, and its purpose is to select several important features representative of the text, and the text category in a training text. The most important issue in feature selection is the relationship between the characteristics and the class, that is, the features selected are truly representative.

The common feature selection methods are expected cross entropy, information gain method, mutual information method, chi-square test method, principal component analysis method and so on. These methods, from the information theory and from statistical analysis, find out the salient features containing the largest amount of information or influence, while ignoring the rest of the features, to achieve the purpose of feature reduction.

4.2.1 Information gain

Information gain is often used as a method to select the best node in the decision tree technology. It uses the concept of entropy in information theory. In information theory, entropy is a measure of the kind of things that is uncertain. It is based on the individual characteristics values to designate the learning sample spaces, depending on how much of the information gain to select effective feature. The information gain of feature t_k is as follows:

$$IG(t) = -\sum_{i=1}^{n} p(c_i) \log p(c_i) + p(t)\sum_{i=1}^{n} p(c_i \mid t)$$

$$\log p(c_i \mid t) + p(\bar{t})\sum_{i=1}^{n} p(c_i \mid \bar{t}) \log p(c_i \mid \bar{t})$$

$p(c_i)$ is the probability of category c_i in the text; p (t) is probability of features in the text; $p(c_i \mid t)$ represents when t appears in the documentation set, the possibility document belonging to c_i ; $p(c_i \mid \bar{t})$ represents when t does not appear in the documentation set, the probability of the document belonging to c_i .

Whether features appear in the text, they will provide text classification information, to calculate the size of the conditional probability in the different cases the amount of information provided. Information gain use the characteristic values to divide the training sample space, and select features according to the amount of information. During feature selection, we select those characteristics with large information gain. The feature selection method has a problem, that is if a feature appears in the class C_1, but does not appear in the class C_2, this feature is very important in itself, but after summing the values of each log phase offset, the result is 0, and certain words cannot be distinguished. There are two ways to solve this problem: First, take the absolute value of the log value, second, omit the log value that is less than 0. In addition, the method is more complicated.

4.2.2 Expect cross entropy

The only difference with information gain is that, cross entropy does not consider the conditions when characteristics are not happen. The expected cross-entropy of characteristic t is as follows:

$$ETC(t) = p(t)\sum_{i=1}^{n} p(c_i \mid t) * \log_2 \frac{p(c_i \mid t)}{p(c_i)}$$

Expect cross-entropy reflects the probability distribution of the categories of text, as well as the distance between the text class probability distribution in the case of certain characteristic words. During feature selection, we select the characteristics with high cross entropy.

4.2.3 Mutual Information

The mutual information is a feature correlation criterion often used in the field of machine learning, which represents the correlation between the two vectors. Mutual information in characterized t and class c is defined as follows:

$$MI(t,c) = \log_2 p(t \mid c) - \log_2 p(t) = \log_2 \frac{p(t \mid c)}{p(t)}$$

4.2.4 Choose based on the characteristics of the class conditional distribution

This paper is from the essence of the Bayesian classification proposed feature reduction method based on class conditional distribution. Main idea of Bayesian classification is to calculate joint probability of the characteristic class probability, so we base on class possibility in feature reduction in algorithm for Bayesian probabilities. For A_i with number of categories l, and the number the value v, can be expressed with the following matrix of its class conditional probability distribution:

$$(p_{kj})_{l*_v} = (p(a_{ij} \mid c_k))_{l*_v} = \begin{bmatrix} p(a_{i1} \mid c_1), & p(a_{i2} \mid c_1) & \ldots & p(a_{iv} \mid c_1) \\ p(a_{i1} \mid c_2), & p(a_{i2} \mid c_2) & \ldots & p(a_{iv} \mid c_2) \\ \ldots & \ldots & \ldots & \ldots \\ p(a_{i1} \mid c_l) & p(a_{i2} \mid c_l & \ldots & p(a_{iv} \mid c_l) \end{bmatrix}$$

For this two types of classification problems of spam, you can use the following matrix to represent the class

conditional probability distribution characteristics of t_i :

$$p(t_i \mid c_k)_{2*1} = \begin{bmatrix} p(t_i \mid c_{ham}) \\ p(t_i \mid c_{spam}) \end{bmatrix}$$

If distribution of the characteristics of the ham and spam is approximately uniform, then it has not so large influence on calculation, and can be ignored. Such a uniform distribution can make data distribution entropy analysis large. Based on the analysis of IrinaRish on the Naive Bayesian performance data characteristics [8], Naive Bayes can obtain better accuracy low entropy distribution data. Therefore, get rid of these characteristics that enlarge data distribution entropy, we can improve the performance of the Naive Bayes classification. We have taken the following formula as the evaluation of characteristics of the class distribution:

$$CCD(t_i) = \frac{1}{2}\sum_{k=1}^{2}(p_k - p_l)^2$$

$$\text{In this } \quad p_l = \frac{1}{2}\sum_{k=1}^{2} p_k$$

We can see from the formula p_i , in which p_i represents the arithmetic mean of t_i in regular mail and spam. $CCD(t_i)$ represents the distance of representative feature t_i away from the arithmetic average probability. So the larger the value, the farther away t_i from the average probability, indicating that the more uneven it is in the category distribution, the smaller the distribution entropy is. Conversely, the smaller the value of $CCD(t_i)$, the smaller the probability indicated from the average probability distance is, and it indicates that the more uniform distribution of the characteristic in each category. And the higher the distribution entropy is. The purpose of the feature selection is to select a low-entropy distribution of the data, i.e. CCD larger values of characteristics.

5 experiments and analysis

In order to compare the three feature selection methods' influence on classification accuracy, we use three feature selection filters in both offline and online filtering mode to filter. Online filtering mode process on trec07 p mail and the online takes immediate feedback mode and offline mode on the sewm 2008.

In feature selection criteria, the number of key features extracted also has a certain impact on filtering accuracy, so the number of features is 8, 10, and 15, and compare difference of the three feature selection method in mail filtering accuracy. First compare trec07 p online timely feedback experiments, Spam tools of TREC evaluation,experimental results are as follows:

Table 1 Online filtering accuracy of three algorithms when features selected number is 8

Evaluation parameters \ feature selection methods	information gain	expected cross entropy	class conditional distribution (ccd value)
Ham%	2.39(1.76-3.16)	1.18 (1.05-1.32)	1.42 (0.95-2.05)
Spam%	1.10 (0.88-1.35)	0.87 (0.79-0.96)	0.16 (0.09-0.28)
Lam%	1.62 (1.37 - 1.91)	1.01 (0.95 - 1.08)	0.48 (0.34 - 0.69)
1-ROCA%	0.3509 (0.1166 - 1.0512)	0.3728 (0.1699 – 1.8844)	0.2963 (0.1453 – 1. 4129)

Table 2 Online filtering accuracy of three algorithms when features selected number is 10

Evaluation parameters \ feature selection methods	information gain	expected cross entropy	class conditional distribution (ccd value)
Ham%	1.16 (1.13-1.42)	1.20 (1.03-1.29)	1.32(1.05-1.00)

Spam%	0.86 (0.75-0.90)	0.86 (0.79-0.95)	0.14(0.05-0.26)
Lam%	1.02 (0.94 - 1.11)	1.00 (0.93 -1.07)	0.39 (0.35 - 0.44)
1-ROCA%	0.2986(0.2673 -0.3444)	0.2739 (0.2563 – 0.3346)	0.1363 (0.1053 – 0. 3129)

Table 3 Online filtering accuracy of three algorithms when features selected number is 15

Evaluation parameters \ feature selection methods	information gain	expected cross entropy	class conditional distribution (ccd value)
Ham%	2.15(1.56-3.11)	1.17(1.04-1.33)	1.40 (0.97-2.00)
Spam%	0.96(0.78-1.24)	0.88(0.78-0.95)	0.17 (0.11-0.27)
Lam%	1.32(1.27 - 1.85)	1.00 (0.96 - 1.07)	0.47(0.33 - 0.67)
1-ROCA%	0.3423(0.1166- 1.0512)	0.3546 (0.1574 – 1.7633)	0.2567(0.1343 – 1. 3879)

Make Offline filtering on the publicly available data sets in sewm 2008 and take the first 30,000 as training and the last 40,000 to test. The experimental results are in following table:

Table 4 Offline filtering accuracy of three algorithms when features selected number is 8

Evaluation parameters \ feature selection methods	information gain	expected cross entropy	class conditional distribution (ccd value)
Ham%	0.88 (0.24-1.87)	0.85 (0.20-1.83)	0.58 (0.10-1.43)
Spam%	8.89 (7.24-10.61)	8.90 (7.35-10.78)	5.45 (4.53-7.46)
Lam%	6.93 (5.23 - 9.76)	6.89 (5.12 - 9.38)	3.20(2.13-4.42)
1-ROCA%	1.2646 (1.1898 - 1.3688)	1.1978 (1.1832 - 1.2957)	0.7853(0.5346-0.9843)

Table 5 Offline filtering accuracy of three algorithms when features selected number is 10

Table 5 Offline filtering accuracy of three algorithms when features selected number is 10 Evaluation parameters \ feature selection methods	information gain	expected cross entropy	class conditional distribution (ccd value)
Ham%	0.68(0.13-0.87)	0.64(0.10-0.83)	0.50(0.09-1.35)
Spam%	6.89(5.35-8.87)	6.48(5.12-8.53)	5.21(4.41-7.12)
Lam%	4.79(3.34-5.89)	4.45(3.23-5.76)	2.43(1.54-4.78)

1-ROCA%	1.0420(0.6735-1.0241)	0.9879(0.5345-1.0001)	0.5474(0.3214-0.8634)

Table 6 Offline filtering accuracy of three algorithms when features selected number is 10

Evaluation parameters \ feature selection methods	information gain	expected cross entropy	class conditional distribution (ccd value)
Ham%	0.83 (0.21-1.85)	0.85 (0.20-1.83)	0.56(0.08-1.33)
Spam%	8.49 (6.98-9.61)	8.90(7.35-10.78)	5.24(4.32-7.23)
Lam%	6.53 (5.13 - 9.51)	6.89 (5.12 - 9.38)	3.01(2.34-4.32)
1-ROCA%	1.1646 (0.9856-1.5423)	1.1978(1.1832-1.2957)	0.7633(0.5157-0.9621)

The experimental results show that, whether in immediate feedback online or offline mode, with the same mail filtering algorithm, if the number of feature selected is different, spam filtering accuracy is different. When the number of feature selection is 10, whether in information gain, expected cross entropy, or class conditional distribution of feature selection algorithm, mail filtering accuracy is better in the feature selection number 8 and 15 of the algorithm. This indicates that the larger number of feature selection is not the better; More feature selection not only increase the difficulty of the calculation, but also bring some features with low class representation and not clear category. Text content between each word was not completely independent, and the premise of the Naive Bayesian method is assuming features are independent of each other, so when the \the number of eigenvalue extracted increase, the opportunity of interdependence between eigenvalues increases. But with too few selected features, making the classification only consider one-sided to the characteristics of the part, we ignored many on the classification of impact characteristics, resulting in classification accuracy decreased.

It can also be seen from Table 1 to Table 6, our feature selection method selection methods based on the same number of features, classification accuracy of conditional distribution was significantly higher than that based on information gain and expected cross entropy-based feature selection method. Specifically, in legitimate messages missing rate and spam missing rate, on the ROC curve above the area of these three parameters, cross entropy of information gain and expectations are higher than the class of conditional distributions. This also shows that in the Naive Bayes classification, the characteristics with a high amount of information may not contribute to the classification of the characteristics, and the characteristics with uniform class conditional distribution are more significant factors in classification accuracy.

6 Conclusions

This chapter first introduces the classification process of Naive Bayes algorithm, and proposes for its classification process improvements in the text representation fingerprint features in four aspects. It also proposes new joint probability formula and solves probability problems in probability calculations. In feature selection, it creates new feature selection method: feature selection based on class conditional distribution and in classification stage raises the weight integrated classification model. And based on that learning process is deepening, this paper also proposes adaptive algorithm to adjust the threshold.

7 References

[1] Jian Zhu, Hongbing Cao, Haitao Liu, "Parking Space Detection Based on Information from Images and Magnetic Sensors", AISS, Vol. 4, No. 5, pp. 208 ~ 216, 2012
[2] hijuan Deng, Shaojun Zhong, "A Kind of Text Classification Design on the Basis of Natural Language Processing", IJACT, Vol. 5, No. 1, pp. 668 ~ 677, 2013
[3] JianJiao Chen, Anping Song, Wu Zhang, "Hybrid K-harmonic Clustering Approach for High Dimensional Gene Expression Data", JCIT, Vol. 7, No. 3, pp. 39 ~ 49, 2012
[4] Yanhua Tan, Changsheng Zhang, Jing Ruan, "The Comparative Study of Different Models for Feature Selection in Rough Set Theory ", IJACT, Vol. 4, No. 4, pp. 124 ~ 130, 2012
[5] Hao-dong Zhu, Hong-chan Li, Jin-Chao Zhao, "Feature Selection Based on Feature Distinguish Ability And Meta-Information", IJACT, Vol. 4, No. 11, pp. 344 ~ 351, 2012

[6] Jinchao Zhao, Fan Zhang, "Feature Selection Based on Parallel Collaborative Evolutionary Genetic Algorithm", AISS, Vol. 4, No. 6, pp. 296 ~ 304, 2012
[7] Ammar ALmomani, Tat-Chee Wan, Ahmad Manasrah, Altyeb Altaher, Eman Almomani, , "A survey of Learning Based Techniques of Phishing Email Filtering", JDCTA, Vol. 6, No. 18, pp. 119 ~ 129, 2012
[8] ZHANG Qiu-yu, YANG Hui-juan, WANG Peng, MA Wei, "Fuzzy Clustering based on Semantic Body and its Application in Chinese Spam Filtering", JDCTA, Vol. 5, No. 4, pp. 1 ~ 11, 2011
[9] Liu Pei-yu, Zhao Jing, Zhu Zhen-fang, "Email Representation using Noncharacteristic Information and its Application", JCIT, Vol. 5, No. 8, pp. 180 ~ 185, 2010
[10] Anjan Kumar Pau, "Robust Object Classification and Recognition for Video Surveillance Applications", IJIIP, Vol. 3, No. 1, pp. 79 ~ 89, 2012

Building an Automatic Thesaurus to Enhance Information Retrieval

Essam Hanandeh[1]

[1] **Computer Information System,**
Zarqa University, Zarqa, Jordan

Abstract

One of the major problems of modern Information Retrieval (IR) systems is the vocabulary Problem that concerns with the discrepancies between terms used for describing documents and the terms used by the researcher to describe their information need. We have implemented an automatic thesurs, the system was built using Vector Space Model (VSM). In this model, we used Cosine measure similarity. In this paper we use selected 242 Arabic abstract documents. All these abstracts involve computer science and information system. The main goal of this paper is to design and build automatic Arabic thesauri using term-term similarity that can be used in any special field or domain to improve the expansion process and to get more relevance documents for the user's query**.** The study concluded that the similarl thesaurus improved the recall and precision more than traditional information retrieval system in terms of recall and precision level**.**

Keywords: *Information* ***Retrieval,*** *similarity thesaurus****,*** *Vector Space Model****,*** *Query Expansion****.***

1.Introduction

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 (Tengku, 1989). IR is also concerned with text representation, text storage, text organization, and the retrieval of stored information items that are similar in some sense to information requests received from users. The term IR covers a wide range of disciplines, and have some similarities with many other areas of information processing, e.g., management information systems, database management systems, decision support systems, question-answering systems, natural language processing, as well as document retrieval systems.

A thesaurus (plural: thesauri) is a valuable tool in IR, both in the indexing process and in the searching process, used as a controlled vocabulary and as a means for expanding or altering queries (query expansion)[10]. Domain experts and/or experts at document description manually construct most thesauri that users encounter. Manual thesaurus construction is a time-consuming and quite expensive process, and the results are bound to be more or less subjective since the person creating the thesaurus make choices that can affect the structure of the thesaurus. There is a need for methods of automatically construct thesauri, which can in addition to the improvements in time and cost aspects can result in more objective thesauri that are easier to update.

2. Vector Space Model

The vector space model uses non-binary weights that are assigned for the documents and queries index terms[13]. This will suggest a partial matching retrieval instead of the relevant / non-relevant matching. The non-binary weights assigned for both the queries and documents are ultimately used to measure the degree of similarity(equation 1) between each of the documents in store in the system and the user query. Hence, the vector model will also take into consideration documents which match the query terms partially.

The vector model uses the t-dimensional vectors to represent both document and query. For a document **dj** (where j is the document number) and a query **q**, their t-dimensional representations are **dj** and **q** as follows:

The query q representations is :

$$\overrightarrow{q} = (w_{1,q}, w_{2.q}, \ldots, w_{t,q})$$

and the document dj representation is :

$$\overrightarrow{dj} = (w_{1,j}, w_{2.j}, \ldots, w_{t,j})$$

where $\mathbf{w_{i,q} \geq 0}$ and t is a total number of index terms in the system.

The vector model proposes to evaluate the degree of similarity of the document **dj** with regard to the query **q** as the correlation between the vectors **dj** and **q**. This correlation can be quantified, for instance, by the cosine of the angle between these two vector [13], That is,

$$sim(d_j, q) = \frac{\sum_{i-1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i.j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}} \quad (1)$$

Vector model can uses different similarity measures other than cosine similarity as shown in Table 1 [13]:

Table 1: Similarity Measures

Similarity Measure	Evaluation for Binary Term Vector	Evaluation for Weighted Term Vector
Cosine	$sim(d,q) = 2\frac{\lvert d \cap q \rvert}{\lvert d \rvert^{1/2} \bullet \lvert q \rvert^{1/}}$	$sim\ (d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i.j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$
Dice	$sim(d,q) = 2\frac{\lvert d \cap q \rvert}{\lvert d \rvert + \lvert q \rvert}$	$sim(d_j, q) = \frac{2\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sum_{i=1}^{t} w_{i,j}^2 + \sum_{i=1}^{t} w_{i,q}^2}$
Jaccard	$sim(d,q) = \frac{\lvert d \cap q \rvert}{\lvert d \rvert + \lvert q \rvert - \lvert d \cap q}$	$sim(d_j,q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sum_{i=1} w_{i,j}^2 + \sum_{i=1} w_{i,q}^2 - \sum_{i=1}^{t} w_{i,j} \times w_{i,}}$
Inner	$\lvert d_i \cap q_k \rvert$	$Sim = \sum_{k=1}^{t} (d_{ik} \bullet q_k)$

Note $|d|$ is the number of term in document d.

3. Query Expansion:

Information retrieval deals with the representation, storage, organization , and access to information items. It is important that this representation provides users with easy access to the information in which they are interested [15] [18].

Short queries submitted to search engines are behind the missing of important words or terms from the user's queries. To solve this problem, the researcher of information retrieval have been investigated the Query expansion as a method to help the user in formulating a better queries [22].

Many users find it difficult to formulate queries that are well-designed for effective retrieval, and they often use a great variety of words to refer to what they want. Expansion or modification of the user's query can lead to a considerable improvement in the retrieval results [19].

Information retrieval refers to the processing of user requests, commonly referred to as queries to obtain relevant information from a collection of documents [19].

An obvious approach to solve this problem is Query Expansion.

Ricardo Baeza-Yates & Berthier Ribeiro-neto point out that [18] that they examine a variety of approaches for improving the initial query formulation through query expansion and term

reweighing. These approaches are grouped in three categories: (a) approaches based on feedback information from the user; (b) approaches based on information derived from the set of documents initially retrieved (called the local set of documents); and (c) approaches based on global information derived from the document collection, which is the objective of this paper.

4. Previous work

Many techniques and algorithms for Information Retrieval Systems (IRS), building thesauri and query expansion have been devised and proposed in the literature. There have been different approaches for building a thesaurus, some of them based on finding the similar terms for the query term in the documents, while others based on mapping both the query and the documents to some kind of a thesaurus, while others used to expand their queries finding synonyms or build a hierarchy relation between terms. Below are some of the results of using or building thesaurus in information retrieval and query expansion.

David Walker, (2001) in his paper talks about the different types of query expansion, he divides them into three types as (1) human and computer generated thesauri, (2) relevance feedback, and (3) automatic query expansion with taking into account the strengths and weaknesses of each. In the conclusion, he has shown that automatically expanded queries via pseudo-relevance feedback and computationally generated thesauri address the needs of users, but have not improved effectiveness of search engines beyond that encountered in relevance feedback [24].

Abu Salem, (1992) studies the IR in Arabic Language. His study based on 120 documents that he received from the Saudi Arabian National Computer Conference and on 32 queries. In his research, he studies indexing by using full words and by using the roots only. He finds out that using the roots are superior to other ways. He also built a manual thesaurus using the relation between expressions to test the possibility of supporting an IRS through this thesaurus. He finds out that the thesaurus makes IR much better [9].

Al-Shalabi et al, (2004) suggests an algorithm for determining and deleting stop words in Arabic texts. This method depends on Finite State Machine. They tested the system using the 242 documents that were presented in the Saudi Arabian National Computer Conference in addition to some verses taken from the Holy Quran. They reached to an accuracy of 98% [11].

Kanaan et al, (2006) Construct an Automatic Thesaurus to enhance Arabic Information Retrieval System. This study was based on 242 documents taken from Saudi Arabian National Computer Conference and they used 24 queries. Their study find out that using Similar thesaurus will make the efficiency of the Arabic IRS better when using roots of the words **[37]**.

5. Experiments Procedure

We do the following steps:

1 Use vector space model to put text of documents and query in vectors.

2 Normalization.

- Removing stop words those were collected by Al-Shalabi, et al [11], and they gained 98% success in distinguishing in addition to deleting some signs appeared. (stop words are the words that occur so frequently in documents in the collection that it is useless for purposes of retrieval [54]), Elimination of stop words reduces the size of the indexing structure and thus increases the performance of the system and enables it to retrieve more relevant documents.
- Deleting punctuation marks, commas, follow stops, especial signs, numbers (contents that has no meanings).

3. stemming : the following stemming algorithm as in [67] with a little bet modification :

Let T denote the set of characters of the Arabic surface Full word

Let Li denote the position of letter i in term T

Let Stem denote the term after stemming in each step

Let D denote the set of definite articles (ال)

Let S denote the set of suffixes

S ={ ت ا، ن، ي، و، ك، ه،ة،
ير، ار، لي، ري، تك، تا، يا، ما، يه، ته، تن، ني، تم، وا، نا، كن، كم، ها، هن، هم، ات، ون، ين، ان،
ية يل، تي،
{لها ينا، رها، رين، مان، رات، يون، يتش، يان، لين،

Let P denote the set of prefixes

P={ ل ب، ن، ، ت ي،
اس فن، ، في فت، لن، ، لت لي، ، با ، فا ، كا سن، ست، سا، سي، لل، ال،
{ الف الك، للم، الع، الس، الا، الم، لال، مال، الح،

Let n is the total number of characters in the Arabic word

Step 1: Remove any diacritic in T

Step2: If the length of T is > 3 characters then,

Remove the prefix Waw " و" in position L1

Step 3: Normalize آ ,إ ,أ of T to ا (plain alif)

Step 4: Normalize ى in Ln of T to ي

Replace the sequence of ى in Ln-1 and ء in Ln to ئ

Replace the sequence of ي in Ln-1 and ء in Ln to ئ

Normalize ه in Ln of T to ة

Step 5: For all variations of D (ال) do,

Locate the definite article Di in T

If Di in T matches Di = Di + Characters in T ahead of Di

Stem = T – Di

Step 6: If the length of Stem is > 3 characters then,

For all variations of S, obtain the most frequent suffix,
Match the region of Si to longest suffix in Stem
If length of (Stem -Si) >= to 3 char then,
Stem = Stem – Si

Step 7: If the length of Stem is > 3 characters then,
For all variations of P do
Match the region of Pi in Stem
If the length of (Stem -Pi) > 3 characters then,
Stem = Stem – Pi

Step 8: Return the Stem

4 Selection of index terms from the collection of filtered terms. Ricardo Baeza-Yates and Berthier Ribeiro-Neto in [54], show that the inverted file is a word oriented mechanism for indexing a text collection in order to speed up the searching task, Index terms can be Individual words, group of words, or phrases, but most of them are single words [54] for this reason we choose a single words (i.e., single term) as index terms in this work.

5 Building Similar Thesaurus

In this step, the process of building the thesauri includes two important decisions:

1 What is the law used in finding "Term Similarity"/"Term relationship" between the different terms to build a thesauri? Which formula should be used in similar thesaurus? Inner product, Cosine, Jaccard or Dice? And which is better? Will they give the same results?

2 what is the degree of threshold similarity/relationship used between the expressions in the thesauri to be used as a synonym?

We use here Cosine equation(equation 2) , as it is the most common equation in building the similarity thesaurus, and the threshold similarity was a variable to be entered while the system working.

$$\text{Cosine similarity } S_{j,k} = \frac{\sum_{i=1}^{n}(w_{i,j} * w_{i,k})}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2 * \sum_{i=1}^{n} w_{i,k}^2}} \quad (2)$$

All the results were between 0 and 1 as ($0<=W_{i,k}<=1$) & ($0<=W_{i,j}<=1$)

6. Results

This study aims to reinforcing IRS depending on 242 Arabic abstract documents that used by (Hmeidi & Kanaan, 1997) in [36], also to realize the importance of using stemmed words in these systems instead of full words. All these abstracts involve computer science and information system.

To achieve this objective, the researcher designed and built an automatic information retrieval system from scratch to handle Arabic text. Working on these results that we got after applying 59 queries from the Relevance Judgments documents began and results were analyzed using the Recall and Precision criteria. After that, Average of Recall and Precision were calculated.

Researcher has constructed an automatic stemmed words using inverted file technique. Depending on these indexing words, researcher has built two information retrieval systems; in the first system, researcher has used a Traditional Information Retrieval system using a term frequency-inverse document frequency (tf-idf) for index term weights. In the second one, researcher used Similar Thesaurus by using Vector Space Model with four similarity measurements (Cosine, Dice, Inner product and jaccard) using a term frequency-inverse document frequency (tf-idf) for index term weights, and compare between the similarity measurements to find out the best that will be use in building the Similarity thesaurus.

The results of the retrieval systems can browse using the words after stemming in the Traditional retrieval of information and in using thesaurus:

First：

Table (1): Effect of using thesaurus with Stemming than Traditional retrieval

Table(1)

	0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
Traditional with Stemming	0.92	0.88	0.79	0.7	0.63	0.52	0.44	0.29	0.2	0.08	0.05
thesaurus with Stemming	0.874	0.874	0.81	0.71	0.66	0.54	0.44	0.33	0.25	0.14	0.06

Figure (1): Comparison values of the Average Recall Precision when use thesaurus than traditional.

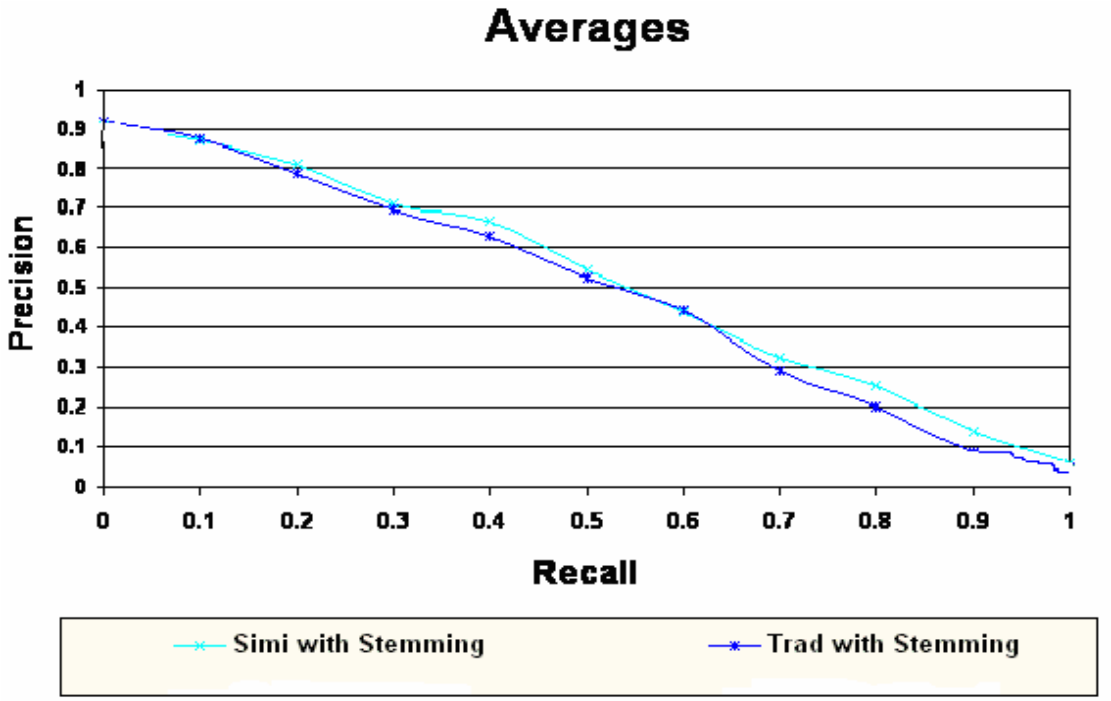


Table (2) shows the number of retrieved documents and how many of them are Relevant and Irrelevant, using thesaurus and with Traditional case.

Table (2)

	Retrieved	Relevant	Irrelevant
Traditional-Stemmed words	2399	1022	1377
thesaurus -Stemmed words	2029	991	1038

Table (3) shows the percentage of the relevant retrieved documents from all the relevant documents in the collection, using thesaurus and with Traditional case

Table (3)

	% of Relevant Docs that Retrieved
Traditional-Stemmed words	61.71497585
thesaurus -Stemmed words	62.681159

Table (4) shows percentage of all the cases together

Table (4)

	Traditional	Similarity
Stemmed words	61.71497585	62.68115942

Second：

Table (5) Effect of using Similarity thesaurus over traditional retrieving (with out using thesauri) by using stemmed words.

Table (5)

Average Recall Precision

Recall	Roots with using Similarity Thesaurus	Roots with using Traditional retrieving	% of Improvement for using Association Thesaurus over Traditional retrieving
0	0.908	0.917966102	-1.00%
0.1	0.87	0.875762712	-0.58%
0.2	0.810178571	0.785762712	2.44%
0.3	0.709464286	0.695254237	1.42%
0.4	0.664821429	0.626237288	3.86%
0.5	0.541428571	0.523389831	1.80%
0.6	0.438571429	0.442542373	-0.40%
0.7	0.325357143	0.290847458	3.45%
0.8	0.251428571	0.198305085	5.31%
0.9	0.13875	0.084745763	5.40%
1	0.056428571	0.047288136	0.91%

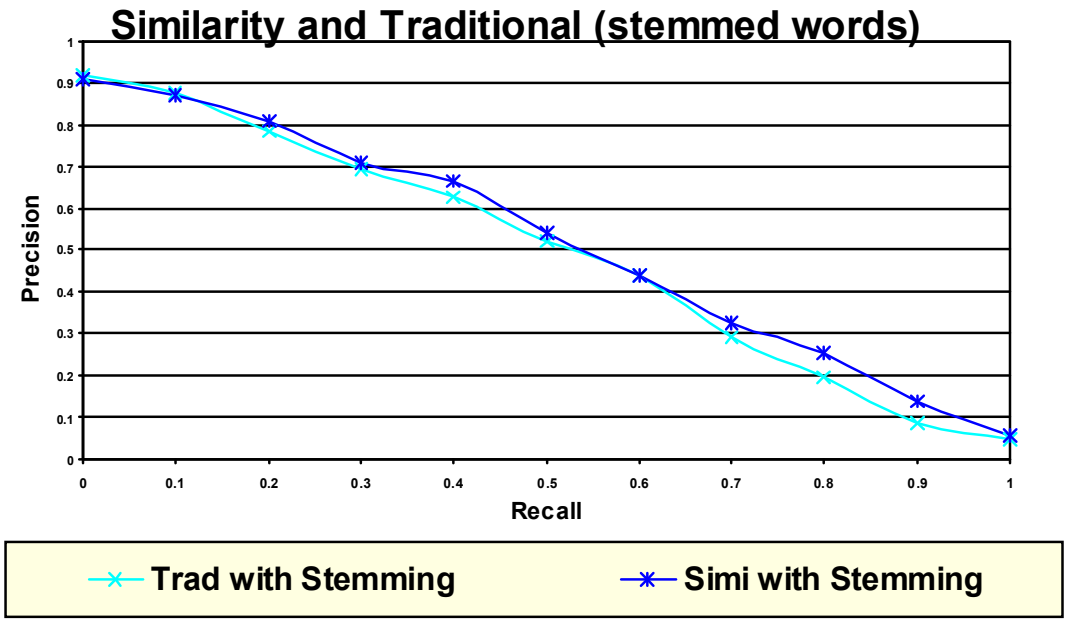


Figure (2) Comparison between the values by using Similarity thesaurus and Traditional retrieving when using stemmed words

Third:

Table (6) shows the effect of using thesauri is much better than using Traditional information retrieval.

Table (6): Average of all the Relative work

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
thesaurus with Stemming	0.91	0.87	0.81	0.71	0.66	0.54	0.44	0.33	0.25	0.14
Traditional with Stemming	0.92	0.88	0.79	0.7	0.63	0.52	0.44	0.29	0.2	0.08

Conclusion

Using stemmed words with similar thesaurus in Arabic language retrieving system is much better than using stemmed words with traditional. [9]. There is a possibility of applying Automatic Indexing and its equations in the Arabic language. Using thesauri enforces IRS. Most researcher agreed on this point. Using the stemming of Arabic words reinforces and supports IRS. This is also true for other languages.

Future work

In this paper, we use stemmed word mechanism. In future, we plan to use all mechanism with stemmed word and full word in traditional retrieval and use thesaurus.

References

[1] Abu Salem, H., A Microcomputer Based Arabic Bibliographic Information Retrieval system With Relation Thesaurus, Ph.D. thesis, University of Illinois, Chicago, USA, 1992.

[2] Adriani, M. and Croft, W. "Retrieval Effectiveness of Various Indexing Techniques on Indonesian News Articles", 1997.

[3] Al-Shalabi, R. Kannan, G., Al-Jaam, J., Hasnah A., and Helat, E., "Stop-word Removal Algorithm for Arabic Language", processing of the 1st International Conference on Information & Communication Technologies: from theory to Applications-ICTTA, Damascus, 2004.

[4] baeza-yates R.,and Rierio-neto B., "Modern Information Retrieval" , Addison-Wesley,New-York,1999.

[5] C. J. van Rijsbergen "information retrieval, Butterworth",1979

[6] Chengfeng Han, Hideo Fujii, and W. Bruce Croft, "Automatic Query Expansion For Japanese Text Retrieval", Umass Technical Report, 1994.

[7] Cui H. Wen J. Nie. J., Ma W., "Query Expansion by Mining User Logs," IEEE Transaction on Knowledge and Data Engineering, 2003; 15(4); 829-839.

[8] David Walker, "Query Expansion using Thesauri: Previous Approaches and Possible New Directions", 2001.

[9] Kanaan, G. "Comparing Automatic Statistical and Syntactic Phrase Indexing for Arabic Information Retrieval", Ph.D.Thesis, University of Illinois, Chicago, USA, 1997.

[10] Kanaan, G. Ghassan and Wedyan, M. (2006). Constructing an Automatic Thesaurus to Enhance Arabic Information Retrieval System. The 2nd Jordanian International Conference on Computer Science and Engineering, JICCSE 2006, Salt, Jordan. 89-97.

[11] Smeaton, A.F., Van Rijsbergen, C.J., The Retrieval Effects of Query Expansion on Feedback Document Retrieval System, The Computer Journal, 26(3), p239-46, 1983.

[12] Aljlayl, M, and Frieder, O, "on Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach" ACM Conference on Information and Knowledge Management, Mcelean, VA , November, 2002

[13] Salton,G., and McGill,M., Introduction to Modern Information Retrieval,McGraw-Hill,New-York, 1983.

A Comparative Approach for Localization Techniques in Wireless Sensor Networks

Mohd Asadullah[1] , Mohd Junedul Haque[2], Mohd Muntjir[3]
College of Computers and Information Technology
Taif University
Saudi Arabia

Abstract: Localization of sensor nodes in wireless sensor nodes plays an important role in many applications. It is important to monitor the location of the data source and event occurrences to track the target and phenomena. This paper provides different kind of localization techniques and their properties. We have also done a comparative study to filter out the better algorithms. Each algorithm's advantages and drawbacks have been highlighted.

Keywords- *Localization, Sensor Nodes, One-hop localization, Hybrid Localization, Centralized Algorithms.*

1. Introduction

A Wireless Sensor Network is a collection of many tiny sensing and wireless communication device called sensor nodes. Each node consists of a processor, a battery and a transceiver for communication [4]. Nodes are connected to each other via transceiver. Wireless Sensor Network consists of one node, called base station which collects sensory information from other nodes in the network and transfers the information to the Computer. They perform specific tasks of sensing some physical phenomena. They are smart, cheaper and deployed in large numbers help in controlling and monitoring the surroundings [1].

There is a wide range of WSNs applications in large number of civil and military needs [2]. There are some civilian applications like environmental habitats, community areas, and smart homes. WSNs are used for surveillance of armed troops and for their tracking, detection of targets [2]. This has also been widely used in smart disaster & relief, search and rescue [5]. The performance of WSNs is quite dependent on how the sensor nodes are located within the network [3].

Fig. 1 A Wireless Sensor

There are different kinds of sensors which can monitor different ambient condition like lightening, pressure, vehicular movement, sound levels, humidity, and availability of certain object [7][8].

2. Localization

The Localization has been a fundamental problem in Wireless sensor networks as the nodes should have the knowledge of the positions of sensors.

Prior information of location enables nodes in a WSN to annotate data with location information. So, the knowledge of location can help to implement feasible message-routing protocols in WSN and Wireless ad-hoc networks [6]. Localization has been categorised into different techniques as shown in fig. 2:

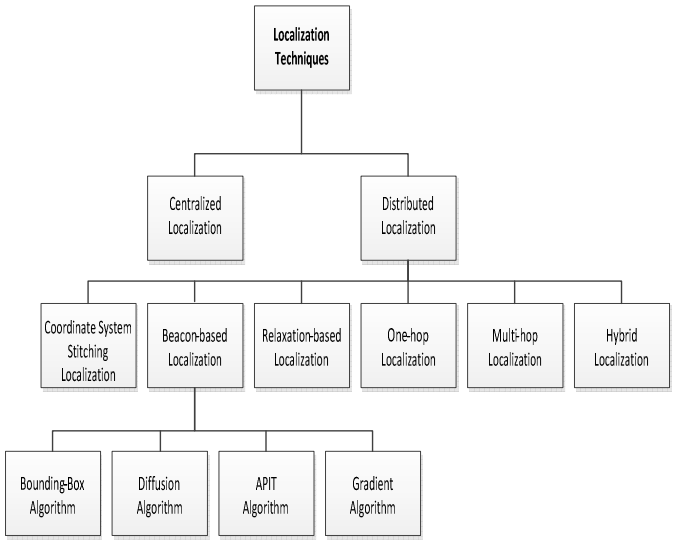


Fig. 2 Different Localization techniques proposed

2.1 Centralized Localization Techniques

In centralized algorithms, there is a central processor which collects the information from each sensor node. This approach tries to obtain maximum accuracy [1]. Centralization requires the migration of the related node connectivity data and ranging

to a centralised base station and then the migration of respective locations return to the related nodes [17]. Centralized algorithms are quite complex with respect to computation [1]. The advantage of centralized algorithms are that it reduces the problem of computation in each node, at the same time the there are certain limitations in the cost of communication of getting data back to the base station [18]. Transmission of data from the sensor nodes to a central base station is very expensive because of limited power supply for each node. Eventually, transmitting time series data within the sensor network results in latency and which also uses energy and bandwidth [19].

Semidefinite programming (SDP) localization algorithm was proposed by Doherty et al [16] in which geometric constraints between sensor nodes are shown in the form of linear matrix inequalities (LMI). In a network, when all the geometric constraints are represented in same manner, the LMIs are added to develop one semidefinite program, which is used to produce a bounding region for every node. This is also called as bounding box. The advantage of this algorithm is that It finds the intersection of the geometric constraints, but some drawback include the inability to access range data in better way and insufficient scaling [17].

In [20], an approach was given based on Simulated Annealing which helps in localizing the sensor nodes in a centralized way. It gets all the information about.

In MDS-MAP centralization algorithm which was developed by Shang et al [15], multidimensional scaling is used for this approach. MDS is an $O(n^3)$ algorithm which reform the relative positions of the sensor node's points using Law of Cosine and linear algebra [17]. This algorithm works on three steps which are as follows:-

Step 1: First of all, we collect data from the network and construct distance matrix by implementing the shortest path computed with the help of Dijkstra's algorithm.

Step 2: We can run classical MDS to compute estimated location for each node.

Step 3: Now transform the relative position map into absolute position map and which help in reducing the error between the correct position and absolute position of each node. MDS-MAP location estimates produce better with the ranges get better [17]. Some disadvantages lies with MDS-MAP is that it all the information about the network and centralized conditions.

the neighbour sensor nodes and accesses the computed locations. It has been described with two steps. In the first step, simulated annealing is used to achieve the location estimate of the sensor nodes with the help of distance constraints. In the second step, some errors are removed with the help of flip ambiguity.

This algorithm provides better accuracy compared to semi-definite programming localization technique.

2.2 Distributed Localization Techniques

This In Distributed localization technique, they do not require a large centralized computer and this technique gives better scalability [17]. In Decentralized or distributed localization techniques, each sensor node gives limited communication with the closer sensor nodes to get the location information [19]. All required computations take place in the sensor nodes themselves and the sensor nodes communicate between one another to get their exact position within the network [18].

A. COORDINATE SYSTEM STITCHING

It is a type of distributed localization technique which is based on Cluster based approach proposed by [21], is used in locating the sensor nodes within the network wherein the node can compute the distance to closer nodes. This algorithm has two stages. Stage 1 is called cluster localization where every node is treated as a centre of the cluster and it calculates the relative position of its neighbour nodes which could be localized in an ad-hoc manner. In the second stage, which is called cluster transformation, the each node's position is overlapped and shared on the local coordinate. It has some advantage in the form of node mobility and insertion of node dynamically [18].

B. BEACON-BASED DISTRIBUTED ALGORITHMS

In this algorithm, an unknown node positions can be estimated using beacon positions. All the required computation can be completed on the relevant sensor nodes themselves in these algorithms. The nodes can be localized into the beacons area [17]. Beacon-based algorithms can be categorized into four approaches: Bounding box, diffusion, APIT and gradient.

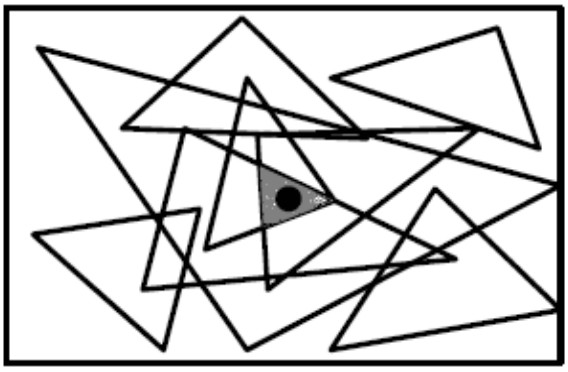


Fig. 3 APIT Technique

1. *Gradient algorithm:* In the Gradient algorithm [24], sensor nodes are randomly scattered in a two dimensional area where every sensor node communicate with the nearest node within some

distance which should not be more than dimension of the plane. The gradient based distance estimates from a beacon are often longer than or equal to the linear distances [17]. This algorithm is divided into two parts: Gradient algorithm and Multilateration algorithm. The advantage of this algorithm that it is scalable whenever needs to add extra sensor or removal of sensors. But, there should be sufficient number of nodes to get the better accuracy [18].

2. *The Bounding Box algorithm:* This algorithm is a method [22] where nodes can be localized within the range of many beacons. It creates a bounding region for each node and then it starts filter their right positions. The collaborative multilateration helps nodes get their location estimate appropriately through identified beacon positions which are hops away [18]. When the node's position is closer to the centre of the beacon nodes, Bounding box algorithm gives accurate results.
3. *Diffusion algorithm:* In diffusion the most likely position of the node is at the centroid of its neighbouring known nodes. Bulusu et al [23] propose the localization of the unknown nodes by getting the average positions of all beacon nodes with which the node is having radio connectivity. The advantage of diffusion algorithm is in the networks where nodes need to do less computation [17].
4. *APIT algorithm:* The APIT algorithm uses a novel area-based approach where nodes can hear huge number of beacons. An unknown node forms a triangle by connecting three beacons. This test is repeated until it gets the required accuracy. At the same moment, APIT computes the centre of gravity of all the triangles to estimate the location of the unknown node. The advantage of APIT is that it is very simple in computation and easy to implement. But, it needs large number of beacons to get accurate result [18].

C. RELAXATION-BASED DISTRIBUTED LOCALIZATION ALGORITHM

It is used to roughly estimate the location within the network. The initial position of the sensor nodes is refined against their neighbouring node's estimate positions. Then, each node changes its position to get the approximate result [17]. In [25], the author has proposed an algorithm which is also called s spring Model where the edges between nodes are called springs and the resting lengths are the actual calculated distance. The nodes adjust their position towards the direction of forces. If the nodes are having zero spring forces applying on them, then the optimization stops. The advantage of Relaxation algorithms are they are fully distributed and can be operated without use of beacons. But, it could not perform well in case of more scalability [17].

D. ONE-HOP LOCALIZATION TECHNIQUES

One-hop localization is a technique where in the non-anchor node which is supposed to be the one-hop nearby from the certain numbers of anchors [9]. In one-hop localization techniques, every blind node should be within the range of its reference node [1]. One-hop localization can be achieved by different techniques which are as follows:- One-hop localization technique can also be achieved by Lighthouse approach in which a base station contains three optical beams which are mutually perpendicular and parallel to one another and which helps in locating optical receivers which comes under the range and line of sight of the optical beams [9]. Ni *et al.* presented strong points on the RSS profiling approach on localization technique which gives a better estimation of the location. In Ni et al.'s weighted version of the RSS-profiling based localization algorithm, the estimation of location for the non-anchor point is shown by:

$$\widehat{X_t} = \sum_{i=1}^{N} \frac{\frac{1}{||\gamma-\beta_i||^2}}{\sum_{i=1}^{N}\frac{1}{||\gamma-\beta_i||^2}} X_i \qquad (1)$$

Where γ is the signal strength vector of the non-anchor node and X_i and β_i are the location vector and signal strength vector for the i^{th} point. $||\gamma - \beta_i||^2$ Denotes nodes [13]. Range-free or Connectivity-based localization algorithms are useful in the situations where needs to get approximate estimate of the location. Niculescu *et al.* [14] have designed the DV-hop approach where in the all anchor nodes cover with other sensor nodes within the network. The signals are propagated hop by hop. Hop-count can be stored in the signal message. It also contains the information about the no. of hops it is away from the respective anchor. There is another approach developed by Shang *et al.* [15] which uses multi-dimensional scaling where in the closest path is calculated with the help of distance matrix and then approximate value is calculated for the each node's relative coordinates. In the another multi-hop localization algorithm which was proposed by Doherty et al. [16] which says that Semidefinite programs are a general form of the linear programs and given as in this form:

$$\text{Minimize} \quad c^T x$$

$$\text{Subject to:} \quad F(x) = F_0 + x_1F_1 + \cdots + x_nF_n$$

$$Ax < b$$

$$F_i = F_i^T \qquad (2)$$

Where $x = [x_1; x_2; :::; x_n]^T$ and x_i are the coordinate vector of node i.

E. MULTI-HOP LOCALIZATION TECHNIQUES

In a multi-hop localization, each node gets the information their anchor nodes via neighbouring stage, some sensor nodes are taken as secondary sensors which are positioned by MDS. The normal sensors which are neither primary nor secondary are positioned by applying PDM. Every primary sensor transmits a packet having unique ID to their neighbours, and then it gets forwarded to the next neighbour until the last value. It also sends proximity which contains hop-count of the packet. In the same way, all the anchors distribute their proximities with another anchor and they can calculate the location using the classical MDS [18].

There is another approach for Hybrid localization which was proposed by A. A. Ahmed et al. [27], which compose two localization techniques: multidimensional scaling (MDS) and Ad-hoc Positioning System (APS). It works in three steps: In the first step, collections of reference sensor nodes are selected in a random manner within the network. In the second step, multidimensional scaling is applied on the set of nodes and then it calculates the shortest-path and then multidimensional scaling is used for mapping. In third step, Ad-hoc Positioning System (APS) is applied where the reference nodes are considered as anchors and the remaining nodes are localized using shortest-path to localize from their anchor nodes. Finally, multilateration process is applied for location estimate [18].

F. HYBRID LOCALIZATION

Hybrid localization can be described with the composition of two or more localization techniques. One approach proposed by King-Yip Cheng et al. in [26], which is combination of two localization methods: multidimensional scaling (MDS) and proximity based map (PDM). Some sensors are placed as primary anchors. There are two stages: in the first the Euclidean distance between the two vectors γ and β_i, and N is the total no. of sample points. Ni et al.'s approach succeeds to get median localization error of 1*m* and a highest localization error of 2*m* [10]. In AOA based localization techniques, optical beams which comes out from the receivers intersect at the particular location which gives the estimation of the location of the transmitter, in the presence of noise. In the absence of noise, some lines do not intersect at one point so a triangulation technique is applied to achieve the estimation of the transmitter location [11]. Stanfield has done a tremendous work in estimate location which provides biased location estimates for the many bearing measurements [12]. ML technique is unbiased at many measurements but contains more errors in terms of root mean square compared to Stanfield approach [11].

Table 1: Centralized Techniques vs Distributed Techniques

Comparison Parameters	**Centralized Localization Techniques**	**Distributed Localization Techniques**
Accuracy	It has approx. 75-80% accuracy.	It gives approx. 75-90% accuracy
Dependency on Specific hardware	No need of specific hardware	Requires specific hardware
Power Usage	It consumes more power	Consumes less power. So, energy efficient.
Deployability and Maintainability	Difficult to deploy and maintain	Easy to deploy and maintain
Communication Cost	High communication cost	It is cost saving.
Robustness	Weak	Robust

3. Conclusion

A lot of significant research work has been done in this area, even though still future works needs to be done. Any technique which is efficient to produce better results should have less communication cost, energy-saving, accuracy, robustness and scalability. The technique should have scalability with energy-saving. In terms of energy saving, distributed techniques are more efficient. In Distributed algorithms, Relaxation algorithms are good because they can be operated without use of beacons but, lags behind in terms of scalability. Gradient algorithm is scalable but needs more nodes for better accuracy. If we hybrid these algorithms and produce a new concept, which could be effective to resolve the no. of beacons and getting better accuracy. We have found many localization techniques which have some advantage but unable to resolve all issues. So, it could be combined few of them and try to overcome major needs of the localizations drawbacks and underperformance.

References

[1] Introduction to Algorithms, 2nd Ed. pp.1027-1033,2001 Peter De Cauwer, Tim Van Overtveldt, Jeroen Doggen, Maarten Weyn and Jerry Bracke, "Localization in Wireless Sensor Networks", PAPERS OF THE E-LAB MASTER THESES' 2008-2009.

[2] O.P. Sahu and Tarun Dubey, "A new approach for self localization of wireless sensor network, Indian Journal of Science and Technology", Vol.2 No. 11 (Nov. 2009) ISSN: 0974- 6846

[3] Bin Xiao, Hekang Chen, Shuigeng Zhou, "A Walking Beacon-Assisted Localization in Wireless Sensor Networks", This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2007 proceedings.

[4] Masoomeh Rudafshani and Suprakash Datta, Localization in wireless sensor networks, In Proceedings of the 6th international conference on Information processing in sensor networks (IPSN '07). ACM, New York, NY, USA, 51-60.

[5] Tian He, Chengdu Huang, Brian M. Blum, John A. Stankovic, and Tarek Abdelzaher. 2003. Range-free localization schemes for large scale sensor networks. In Proceedings of the 9th annual international conference on Mobile computing and networking (MobiCom '03). ACM, New York, NY, USA, 81-95.

[6] Nissanka B. Priyantha, Hari Balakrishnan, Erik D. Demaine, Seth Teller, "Mobile-Assisted Localization in Wireless Sensor Networks" INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies.

[7] D. Estrin, R. Govindan, J. Heidemann, S. Kumar, Next century challenges: scalable coordination in sensor networks, ACM MobiCom'99, Washingtion, USA, 1999, pp. 263–270.

[8] I.F. Akyildiz, W. Su*, Y. Sankarasubramaniam, E. Cayirci , "Wireless sensor networks: a survey", Elsevier Science B.V. PII: S13 8 9-1 2 86 (0 1)0 03 0 2- 4, 1389-1286/02.

[9] Guoqiang Mao, Barıs¸ Fidan and Brian D.O. Anderson, "Wireless Sensor Network Localization Techniques", Comput. Netw. 51, 10 (July 2007), 2529-2553.

[10] L. M. Ni, L. Yunhao, L. Yiu Cho, and A. P. Patil, "LANDMARC: indoor location sensing using active rfid," in Proceedings of the First IEEE International Conference on Pervasive Computing and Communications (PerCom 2003)., 2003, pp. 407–415.

[11] M. Gavish and A. J. Weiss, "Performance analysis of bearing-only target location algorithms," IEEE Transactions on Aerospace and Electronic Systems, vol. 28, no. 3, pp. 817–828, 1992.

[12] R. G. Stanfield, "Statistical theory of DF finding," Journal of IEE, vol. 94, no. 5, pp. 762 – 770, 1947.

[13] Kamin Whitehouse, Fred Jiang, Chris Karlof, Alec Woo, David Culler, "Sensor Field Localization: A Deployment and Empirical Analysis" UC Berkeley Technical Report UCB//CSD-04-1349 April , 2004.

[14] D. Niculescu and B. Nath, "Ad hoc positioning system (APS)," in IEEE GLOBECOM, vol. 5, 2001, pp. 2926–2931.

[15] Y. Shang, W. Ruml, Y. Zhang, and M. Fromherz, "Localization from connectivity in sensor networks," IEEE Transactions on Parallel and Distributed Systems, vol. 15, no. 11, pp. 961–974, 2004.

[16] L. Doherty, K. pister, and L. El Ghaoui, "Convex position estimation in wireless sensor networks," in IEEE INFOCOM, vol. 3, 2001, pp. 1655–1663.

[17] Ivan Stojmenović, Jonathan Bachrach, Christopher Taylor, "Localization in Sensor Networks", Published Online: 23 SEP 2005, DOI: 10.1002/047174414X.ch9.

[18] Amitangshu Pal, "Localization Algorithms in Wireless Sensor Networks: Current Approaches and Future Challenges", Network Protocols and Algorithms ISSN 1943-3581 2010, Vol. 2, No.1.

[19] Tareq Ali Alhmiedat & Prof. Shuang-Hua Yang, "A Survey: Localization and Tracking Mobile Targets through Wireless Sensors Network", ISBN: 1-9025-6016-7 © 2007 PGNet.

[20] Anushiya A Kannan, Guoqiang Mao and Branka Vucetic, "Simulated Annealing based Wireless Sensor Network Localization", Journal of Computers, Vol. 1, No. 2, pp 15-22, May 2006.

[21] David Moore, John Leonard, Daniela Rus, and Seth Teller, "Robust distributed network localization with noisy range measurements", in Proceedings of the Second ACM Conference on Embedded Networked Sensor Systems (SenSys'04), November 2004, Baltimore, MD, pp. 50-61.

[22] S. Simic and S. Sastry. Distributed localization in wireless ad hoc networks, 2002.

[23] Nirupama Bulusu, Vladimir Bychkovskiy, Deborah Estrin, and John Heidemann. "Scalable, ad hoc deployable rf-based localization". In Grace Hopper Celebration of Women in Computing Conference 2002, Vancouver, British Columbia, Canada., October 2002.

[24] J. Bachrach, R. Nagpal, M. Salib and H. Shrobe, "Experimental Results for and Theoritical Analysis of a Self-Organizing a Global Coordinate System from Ad Hoc Sensor Networks", Telecommunications System Journal, Vol. 26, No. 2-4, pp. 213-233, June 2004.

[25] N. Priyantha, H. Balakrishnan, E. Demaine, and S. Teller, "Anchor-free distributed localization in sensor networks", MIT Laboratory for Computer Science, Technical Report TR-892, April 2003, Available HTTP: http://citeseer.ist.psu.edu/681068.html.

[26] King-Yip Cheng, King-Shan Lui and Vincent Tam, "Localization in Sensor Networks with Limited Number of Anchors and Clustered Placement", in Proceedings of Wireless Communications and Networking Conference, 2007 (IEEE WCNC 2007), March 2007, pp. 4425 – 4429.

[27] A. A. Ahmed, H. Shi, and Y. Shang, "Sharp: A new approach to relative localization in wireless sensor networks," in Proceedings of IEEE ICDCS, 2005.

Mohd Asadullah is a lecturer at the College of Computers and Information Technology, Taif University, Saudi Arabia. He holds master's degrees in computer applications. His areas of interest are Database, Data Mining and Wireless Networking.

Mohd Junedul Haque is a lecturer at the College of Computers and Information Technology, Taif University, Saudi Arab. He holds master's degrees in Computer Science. His areas of interest are Data Mining, Image Processing and Networking.

Mohd Muntjir is a lecturer at the College of Computers and Information Technology, Taif University, Saudi Arab. He holds master's degrees in Computer Applications . His areas of interest are Database, and Networking.

Pragmatic Peer Review Project Contextual Software Cost Estimation – A Novel Approach

Manoj Kumar Panda

HEAD OF THE DEPT,CE,IT & MCA

NUVA COLLEGE OF ENGINEERING & TECH NAGPUR , MAHARASHTRA,INDIA

Abstract

Software cost estimation is the process of predicting the effort required to develop a software system. Many estimation models have been proposed over the last 30 years. This Chapter provides a general overview of software cost estimation methods including the recent advances in the field. As a number of these models rely on a software size estimate as input, we first provide an overview of common size metrics. We then highlight the cost estimation models that have been proposed and used successfully. Models may be classified into 2 major categories: algorithmic and non-algorithmic. Each has its own strengths and weaknesses. A key factor in selecting a cost estimation model is the accuracy of its estimates. Unfortunately, despite the large body of experience with estimation models, the accuracy of these models is not satisfactory. The Chapter includes comment on the performance of the estimation models and description of several newer approaches to cost estimation. Keywords: project estimation, effort estimation, cost models. It can be used to determine what resources to commit to the project and how well these resources will be used. It can be used to assess the impact of changes and support re planning. Projects can be easier to manage and control when resources are better matched to real needs. Customers expect actual development costs to be in line with estimated costs. Software cost estimation involves the determination of one or more of the following

Key words :- Work break down structure (WBS),WA, Adjusted Function Points (AFP), Project Delivery Rate (PDRU), Project Elapsed Time (PET), Resource Level (RL) and Average Team Size (ATS)

1.1 Introduction

In recent years, software has become the most expensive component of computer system projects. The bulk of the cost of software development is due to the human effort, and most cost estimation methods focus on this aspect and give estimates in terms of person-months. Accurate software cost estimates are critical to both developers and customers. They can be used for generating request for proposals, contract negotiations, scheduling, monitoring and control. Underestimating the costs may result in management approving proposed systems that then exceed their budgets, with underdeveloped functions and poor quality, and failure to complete on time. Overestimating may result in too many resources committed to the project, or, during contract bidding, result in not winning the contract, which can lead to loss of jobs. Accurate cost estimation is important because:

It can help to classify and prioritize development projects with respect to an overall business plan.

Estimates:

effort (usually in person-months) project duration (in calendar time)

Cost (in Rupees)

Most cost estimation models attempt to generate an effort estimate, which can then be converted into the project duration and cost. Although effort and cost are closely related, they are not necessarily related by a simple transformation function. Effort is often measured in person months of the programmers, analysts and project managers. This effort estimate can be converted into a dollar cost figure by calculating an average salary per unit time of the staff involved, and then multiplying this by the estimated effort required. Practitioners have struggled with three fundamental issues:

Software cost estimation model to use

Software size measurement to use – lines of code (LOC), function points (FP), or feature point.

1.1 A Good Estimation

The widely practiced cost estimation method is expert judgment. For many years, project managers have relied on experience and the prevailing industry norms as a basis to develop cost estimate. However, basing estimates on expert judgment is problematic:

Project Delivery Rate (PDRU):-

This approach is not repeatable and the means of deriving an estimate are not explicit. It is difficult to find highly experienced estimators for every new project. The relationship between cost and system size is not linear. Cost tends to increase exponentially with size. The expert judgment method is appropriate only when the sizes of the current project and past projects are similar. Budget manipulations by management aimed at avoiding overrun make experience and data from previous projects questionable.

Resource level(RL) :- we must find the average resources available and the optimum use of the resources is must including the manpower e.g. the total available computer system add the uninterruptible energy supply to it and the manpower must be in buffer .

Work Breakdown Structure (WBS):- in project management and systems engineering, is a deliverable oriented decomposition of a project into smaller components. It defines and groups a project's discrete work elements in a way that helps organize and define the total work scope of the project.

Work Breakdown Structure(WBS):- element may be a product, data, a service, or any combination. A WBS also provides the necessary framework for detailed cost estimating and control along with providing guidance for schedule development and control

In the last three decades, many quantitative software cost estimation

models have been developed. They range from empirical models such as Boehm's COCOMO models to *analytical* models such as those in. An *empirical model* uses data from previous projects to evaluate the current project and derives the basic formulae from analysis of the particular database available. An *analytical model*, on the other hand, uses formulae based on global assumptions, such as the rate at which developer solve problems and the number of problems available. Most cost models are based on the size measure, such as LOC and FP, obtained from size estimation. The accuracy of size estimation directly impacts the accuracy of cost estimation.

Although common size measurements have their own drawbacks, an organization can make good use of any one, as long as a consistent counting method is used. A good software cost estimate should have the following attributes. It is conceived and supported by the project manager and the development team. It is accepted by all stakeholders as realizable.

It is based on a well-defined software cost model with a credible basis.

Project Elapsed Time (PET):-The project elapsed time is the duration of total time we must finish the project or the stipulated time

during that time at any cost we must deliver the product to the client

It is based on a database of relevant project experience (similar processes, similar technologies, similar environments, similar people and similar requirements). It is defined in enough detail so that its key risk areas are understood and the probability of success is objectively assessed. Software cost estimation historically has been a major difficulty in software development. Several reasons for the difficulty have been identified: Lack of a historical database of cost measurement Software development involving many interrelated factors, which affect development effort and productivity, and whose relationships are not well understood Lack of trained estimators and estimators with the necessary expertise Little penalty is often associated with a poor estimate.

1.2. Process Of Estimation

Estimation is an important part of the planning process. For example, in the top-down planning approach, the cost estimate is used to derive the project plan:

1.2.1.The project manager develops a characterization of the overall functionality, size, process, environment, people, and quality required for the project.

1.2.2 A macro-level estimate of the total effort and schedule is developed using a software cost

Estimation Model.

1.2.3 The project manager partitions the effort estimate into a top-level work breakdown structure. He also partitions the schedule into major milestone dates and determines a staffing profile, which together forms a project plan.

1.2.4The actual cost estimation process involves seven steps:

1.2.5 Establish cost-estimating objectives

1.2.6 Generate a project plan for required data and resources

1.2.7. Pin down software requirements

1.2.8. Work out as much detail about the software system as feasible

1.2.9. Use several independent cost estimation techniques to capitalize on their combined strengths

1.2.10. Compare different estimates and iterate the estimation process

1.2.11. After the project has started, monitor its actual cost and progress, and feedback results to project management.

Average Team Size(ATS):- the total team size dependent on the total project manhours if the average project duration is two years then we must take the sufficient number of human resource in to our project if again the dead line is nearing then the manpower must be enhanced as per the real time situation .

No matter which estimation model is selected, users must pay attention to the following to get best results: coverage of the estimate (some models generate effort for the full life-cycle, while others do not include effort for the requirement stage) calibration and assumptions of the model sensitivity of the estimates to the different model parameters deviation of the estimate with respect to the actual cost.

2 The Outputs Of This Step Are As Follows:

2.1 Assumptions made to revise estimates

2.2 Methods used to revise estimates

2.3 Revised size, effort, schedule, and cost estimates

2.4 Revised functionality and procurements

2.5 Updated WBS

2.6 Revised risk assessment

Review and Approve the Estimates

The purpose of this step is to review the software estimates and to obtain project and line management approval.

3. Conduct A Peer Review With The Following Objectives:

3.1 Confirm the WBS and the software architecture.

3.2 Verify the methods used for deriving the size, effort, schedule, and cost. Signed work agreements may be necessary.

3.3 Ensure the assumptions and input data used to develop the estimates are correct.

3.4 Ensure that the estimates are reasonable and accurate, given the input data.

3.5 Formally confirm and record the approved software estimates and underlying

Assumptions for the project:

4. The software manager, software estimators, line management, and project management approve the software estimates after the review is complete and problems have been resolved. Remember that costs cannot be reduced without reducing functionality.

The outputs of this step are as follows:

1 Problems found with the estimates
2 Reviewed, revised, and approved size, effort, schedule, cost estimates, and assumptions
3 Work Agreement(s), if necessary Track, Report, and Maintain the Estimates
The purpose of this step is to check the accuracy of the software estimates over time, and provide the estimates to save for use in future software project estimates.

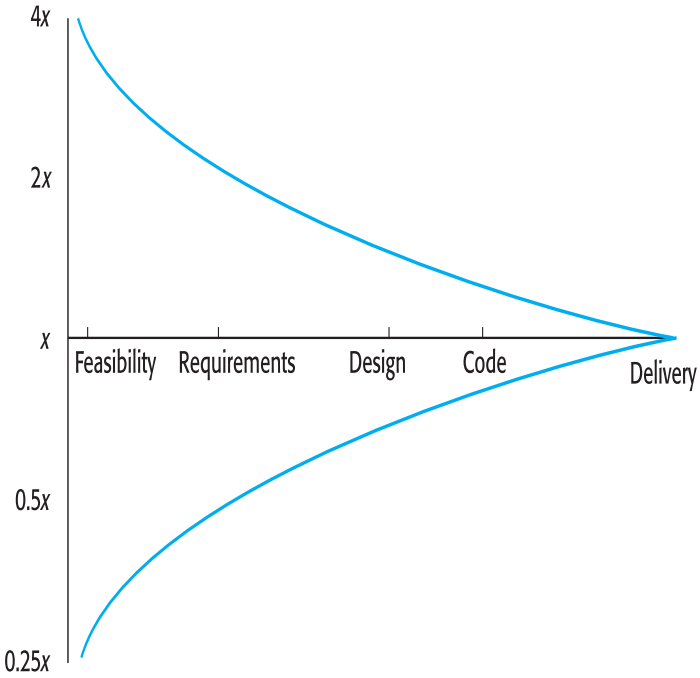
1. Track the estimates to identify when, how much, and why the project may be overrunning or under-running the estimates. Compare current estimates, and ultimately actual data, with past estimates and budgets to determine the variation of the estimates over time. This allows estimators to see how well they are estimating and how the software project is changing over time.

2. Document changes between the current and past estimates and budgets.

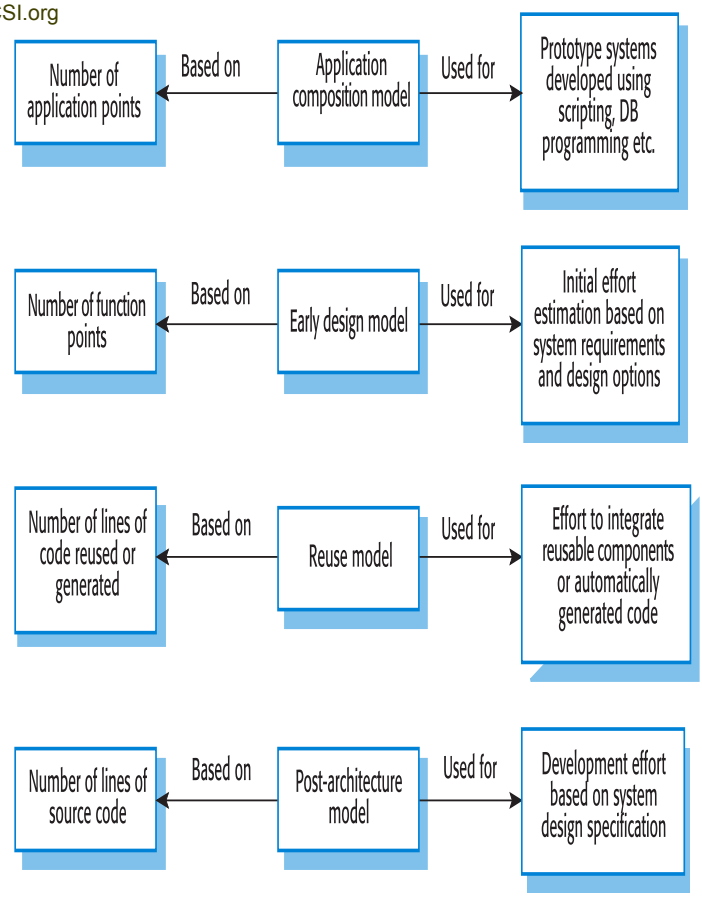
3. In order to improve estimation and planning, archive software estimation and actual data each time an estimate is updated and approved, usually at each major milestone. It is recommended that the following data be archived:

4 Project contextual and supporting information

– Project name
– Software organization
– Platform
– Language
– Estimation method(s) and assumptions
– Date(s) of approved estimate(s)
4.1 Estimated and actual size, effort, cost, and cost of procurements by WBS work element
4.2 Planned and actual schedule dates of major milestones and reviews
4.3 Identified risks and their estimated and actual impacts
The outputs of this step are as follows:
4.4 Updated tracking comparisons of actual and estimated data
4.5 Evaluation of the comparisons
4.6 Updated size, effort, schedule, cost estimates, and risk assessment
4.7 Archived software data, including estimates and actuals



(Figure .1)



(Figure 2)

CONCLUSION

In this approach we aimed at addressing the problem of large variances found in available historical data that are used in software cost estimation. Project data is expensive to collect, manage and maintain. Therefore, if we wish to lower the dependence of the estimation to Computational Intelligence in Software Cost Estimation: Evolving Conditional Sets of Effort Value Ranges 17 the need of gathering accurate and homogenous data, we might consider simulating or generating data ranges instead of real crisp values.

The theory of conditional sets was applied in the present work with Genetic Algorithms (GAs) on empirical software cost estimation data. GAs are ideal for providing efficient and effective solutions in complex problems; there are, however, several trade-offs. One of the major difficulties in adopting such an approach is that it requires a thorough calibration of the algorithm's parameters. We have tried to investigate the relationship between software attributes and effort, by evolving attribute value ranges and evaluating estimated efforts. The algorithm promotes the best individuals in the reproduced generations through a probabilistic manner. Our methodology attempted to reduce the variations in performance of the model and achieve some stability in the results. To do so we approached the problem from the perspective of minimizing the differences in the ranges and the actual and estimated effort values to decisively determine which attributes are the most important in software cost estimates.

We used the ISBSG repository containing a relatively large quantity of data; nevertheless, this data suffers from heterogeneity thus presents low quality level from the perspective of level of values. We formed three different subsets selecting specific cost attributes from the ISBSG repository and filtering out outliers using box-plots on these attributes. Even though the results are of average performance when using the first two datasets, they indicated some importance ranking for the attributes investigated. According to this ranking, the attributes Added Count (AC) and File Count (FC) were found to lay among the most significant cost drivers for the ISBSG dataset. The third dataset included Adjusted Function Points (AFP), Project Delivery Rate (PDRU), Project Elapsed Time (PET), Resource Level (RL) and Average Team Size (ATS). These attributes may be measured early in the software life-cycle, thus this dataset may be regarded more significant than the previous two from a practical perspective. A careful and stricter filtering of this dataset provided prediction improvements, with the yielded results suggesting small value ranges and fair estimates for the mean effort of a new project and its deviation. There was also an indication that within different areas of the data, significantly different results may be produced. This is highly related to the scarcity of the dataset itself and supports the hypoChapter that if we perform some sort of clustering in the dataset we may further minimize the deviation differences in the results and obtain better effort estimates.

Although the results of this work are at a preliminary stage it became evident that the approach is promising. Therefore, future research steps will concentrate on ways to improve performance, examples of which may be: (i) Pre-processing of the ISBSG dataset and appropriate clustering into groups of projects that will share similar value characteristics. (ii) Investigation of the possibility of reducing the attributes in the dataset by utilizing a significance ranking mechanism that will promote only the dominant cost drivers. (iii) Better tuning of the GA's parameters and modification/enhancement of the fitness functions to yield better convergence. (iv) Optimization of the trial and error weight factor assignment used in the present.

BIBLIOGRAPHIES & REFERENCES

1. A. J. Albrecht, and J. E. Gaffney, "Software function, source lines of codes, and development effort prediction: a software science validation", *IEEE Trans Software Eng.* SE-9, 1983, pp.639-648.
2. U. S. Army, *Working Schedule Handbook, Pamphlet No.* 5-4-6, Jan 1974.
3. J. D. Aron, *Estimating Resource for Large Programming Systems*, NATO Science Committee, Rome, Italy, October 1969.
4. R.K.D. Black, R. P. Curnow, R. Katz and M. D. Gray, *BCS Software Production Data*, Final Technical Report, RADC-TR-77-116, Boeing Computer Services, Inc., March 1977.
5. B. W. Boehm, *Software engineering economics*, Englewood Cliffs, NJ: Prentice-Hall, 1981.
6. B.W. Boehm et al "The COCOMO 2.0 Software Cost Estimation Model", *American Programmer*, July 1996, pp.2-17.
7. L. C. Briand, K. El Eman, F. Bomarius, "COBRA: A hybrid method for software cost estimation, benchmarking, and risk assessment", *International conference on software engineering*, 1998, pp. 390-399.
8. G. Cantone, A. Cimitile and U. De Carlini, "A comparison of models for software cost estimation and management of software projects", in *Computer Systems*: *Performance and Simulation*, Elisevier Science Publishers B.V., 1986.
9. W. S. Donelson, "Project planning and control", *Datamation*, June 1976, pp. 73-80.
10. N. E. Fenton and S. L. Pfleeger, *Software Metrics*: *A Rigorous and Practical Approach*, PWS Publishing Company, 1997.
11. D. V. Ferens, and R. B. Gumer, "An evaluation of three function point models of estimation of software effort", *IEEE National Aerospace and Electronics Conference*, vol. 2, 1992, pp. 625-642.
12. G. R. Finnie, G. E. Wittig, AI tools for software development effort estimation*, Software Engineering and Education and Practice Conference*, IEEE Computer Society Press, pp. 346-353, 1996.
13. M. H. Halstead, *Elements of software science*, Elsevier, New York, 1977.
14. P. G. Hamer, G. D. Frewin, "M.H. Halstead's Software Science – a critical examination", *Proceedings of the 6th International Conference on Software Engineering*, Sept. 13-16, 1982, pp. 197-206.
15. F. J. Heemstra, "Software cost estimation", *Information and Software Technology*, vol. 34, no. 10, 1992, pp. 627-639.

The Application of Fuzzy Neural Network to Boiler Steam Pressure Control

Lei Wang

**Department of Information Engineering, Tangshan College,
Tangshan, Hebei 063009, PR China
wanglei122_2000@126.com**

Abstract

The control effect of steam pressure is one of the most important factors influencing stability in chain type boiler. Aimed at the problem that the control of steam pressure is restricted by many factors, a kind of fuzzy neural network (FNN) is presented in this paper. The controller has the advantage of self-adapting, self-learning and tuning on-line. In simulation, this system exerts stably, with a less effect of uncertain factors, so it has a perfect control effect to main steam pressure systems of boiler.

Keywords: *Chain Furnace, Fuzzy Neural Network, Steam Pressure Controller.*

1. Introduction

During the normal working of the boiler in thermal power plant, steam pressure is one of the main operating parameters that needs close watching and controlling. Too high steam pressure will affect the life span of pressure containing components, cause explosion fault, bring serious harm to the equipment; steam pressure dropping will consume more steam and coal, influencing the efficiency of electricity generation and heat supply; too fast pressure change will make the boiler water circulation worse. Therefore, in the operation of the boiler, steam pressure should be controlled at a given value. The traditional PID control is computationally simple, intuitive in nature and strong robustness etc, but it needs to gain control of the mathematics model. While boiler is a nonlinear, strong coupled, multi-variable and complex system with large delay, it is hard to set an accurate mathematical model. And there are many internal and external factors influencing the boiler steam pressure, making it hard for the traditional PID control to achieve ideal controlling effects.

Fuzzy control and neural network control are controlling methods that develop very fast in recent years, the two methods do not depend on mathematical model of the controlled objects, and has good control effect and anti-jamming. But the fuzzy control rules and membership functions of the fuzzy design parameter can only rely on experience to choose, it is very difficult for it to design and adjust automatically, thus it lacks self-learning and self-adaptability. Although the neural network control has strong self-learning and self-adaptability, it has no the function of accessing uncertain information. The fuzzy neural network control combines the advantages of fuzzy control and neural network control, building a fuzzy controlling system with neural network. That is, with the learning method of neural network, it fulfills the automatic update of fuzzy rules' online modification and membership functions, bringing self-study and self-adaptive ability to fuzzy control. [1].

2. Fuzzy Neural Network Steam Pressure Control System

According to the structure and characteristics of boiler combustion in thermal power plant, a kind of fuzzy neural network controller is presented and it may control the system on line. Its structure is shown in Figure 1.

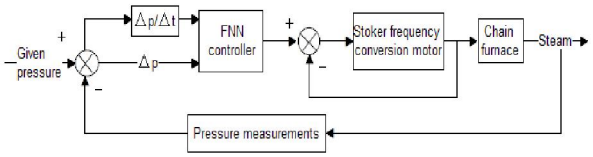


Fig. 1 Boiler steam pressure control loop.

Boiler steam pressure can be controlled by adjustment of coal supply, and the quantity of coal given is in proportion with stoker speed, so we control steam pressure through controlling the stoker speed. If there is a great positive deviation between the actual steam pressure and the given value, the steam pressure value would be high, then it needs reducing stoker speed to control combustion in order to reduce steam pressure. If there is a great negative deviation between the actual steam pressure and the given value, the steam pressure would be low, then it needs increasing stoker speed to help combustion in order to increase steam pressure.[2]

The fuzzy neural network controller fulfills fuzzy control algorithm through neural network structure. Theoretically speaking, the higher the dimension of neural network, the more accurate the control. But if the dimension is too large, fuzzy control rules becomes overly complicated, it is quite difficult to realize the control algorithm. In order to achieve the purpose of preciseness and simplicity, neural network use a five-layer network of two inputs and one output. We take the bias of steam pressure's real duration, given value and error rate as the input of fuzzy neutral network, and the output is the incremental of the stoker frequency conversion motor frequency. Its structure is shown in Figure 2.

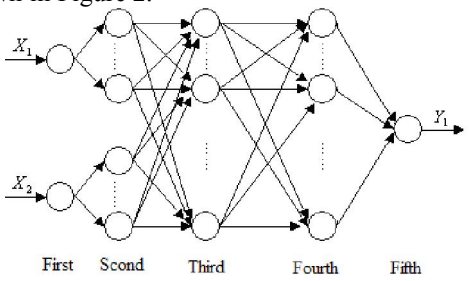


Fig. 2 Fuzzy neural networks controller structure.

The first layer: input layer. Contains two input nodes, they are directly under the signals to the next layer.

$$Net_i^{(1)} = X_i \tag{1}$$

$$O_i^{(1)} = X_i, (i = 1,2) \tag{2}$$

In the equation: $Net_i^{(1)}$—net input of the i-th neurons of the first layer; $O_i^{(1)}$—output of the i-th neurons of the first layer; X_1 is pressure deviation, $X_1 = \Delta P = P_{Real} - P_{Given}$; X_2 is the rate of pressure deviation, $X_2 = \Delta P / \Delta t$. The connecting weight of the first layer is 1.

The second layer: fuzzification layer. The first fuzzy set of the input includes seven linguistic variables. The second fuzzy set of the input also includes seven linguistic variables, and both fuzzy divisions of two input are seven, node is used to realize the value of language input variables of the membership function.

$$Net_i^{(2)} = O_j^{(1)}, i = 1,2,\cdots\cdots,14 \tag{3}$$

Including: $j = 1$， $1 \le i \le 7$； $j = 2$， $8 \le i \le 14$

$$O_i^2 = \exp\{-(\frac{Net_i^{(2)} - m_i^{(2)}}{\sigma_i^{(2)}})^2\}, (i = 1,2,\cdots\cdots,14) \tag{4}$$

m_i and σ_i represent the center and width of the Gaussian membership function of the j-th linguistic value of the i-th input linguistic variable respectively. They are adjustable parameters. The connecting weight of the second layer is 1.

The third layer: Contains 49 nodes, and each node represents a fuzzy rule. Control rules which are shown in table 1. The connection between the third layer and the second layer which is used to match the fuzzy rules. Its output determines each rules of excitation intensity. Function:

$$Net_i^{(3)} = (O_j^{(2)} \times O_k^{(2)}), i = 7(j-1) + (k-7) \tag{5}$$

Including: $j = 1,2,\cdots\cdots 7$； $k = 8,9,\cdots\cdots 14$

$$O_i^{(3)} = Net_i^{(3)}, \quad i = 1,2,\cdots\cdots,49 \tag{6}$$

The connecting weight of the third layer is 1.

The fourth layer: Contains 49 nodes. Each node of this layer executes fuzzy “or” operation so as to form the rules in line with the requirement. Function

$$Net_i^{(4)} = \sum_{j=1}^{49} \omega_{ij} O_j^{(3)} \tag{7}$$

$$O_i^{(4)} = \min(1, Net_i^{(4)}), i = 1,2,\cdots\cdots,49 \tag{8}$$

ω_{ij} represents the i-th output lingual variables and the j-th rule connection strength, which only equals 0 or 1.

The fifth layer: solving fuzzification layer. m_j and σ_j represent the center and width of membership functions of the fuzzification layer, respectively. Now,solving fuzzification layer.

$$Net_1^{(5)} = \sum_{j=1}^{49} (m_j^{(4)} \sigma_j^{(4)}) O_j^{(4)} \tag{9}$$

$$O_1^{(5)} = \frac{Net_1^{(5)}}{\sum_{j=1}^{49} \sigma_j^{(4)} O_j^{(4)}} \tag{10}$$

The connecting weight of the first layer is $m_j^{(4)} \sigma_j^{(4)}$.

3. Simulation of the Control System

The fis structure is generated automatically by using anfisedit in Matlab toolbox and Sugeno system, each of the two inputs is the pressure deviation e and the deviation variance ratio ec, measures are defined at 7.The primary function is Gaussian Function, the error is limited to take default 0. Based on the gradient descent method, the neural network was established and trained by the real experiment data.

The membership function which needs input variables is acquired after training of adaptive learning. The membership functions with input variables before and after learning are shown in Figure.3 and Figure.4.

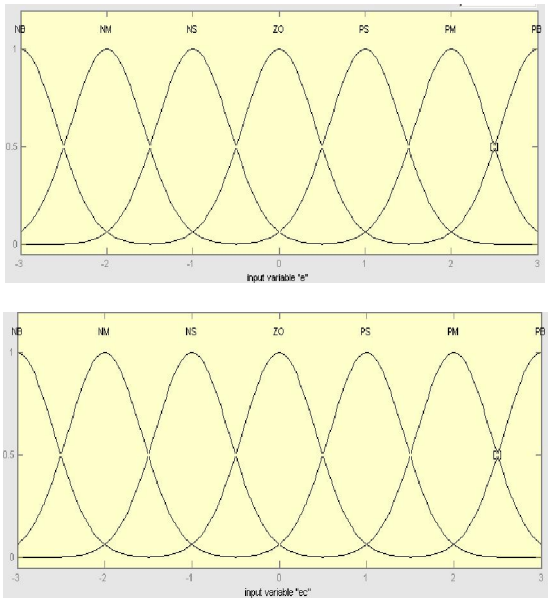


Fig. 3 Formerly membership function diagram of e and ec.

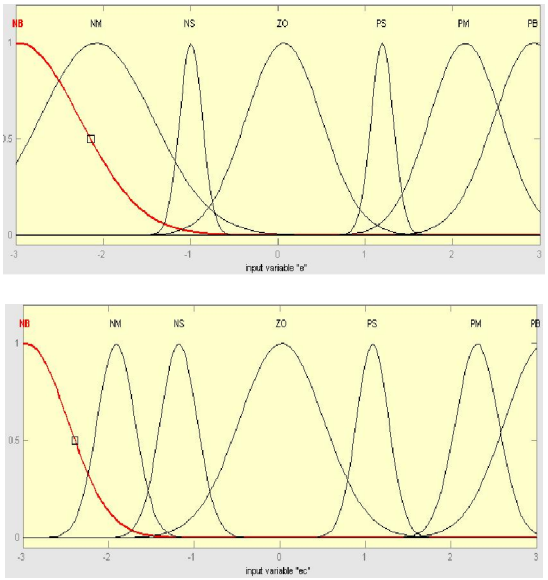


Fig. 4 Membership function diagram of e and ec after learning by fuzzy-neural network.

Through the comparison of the above two diagrams, the correcting changes of membership functions after learning can be observed, the membership functions after calibration can reflect the distribution characteristics of sample data more accurately. Apply the new fuzzy neural network controller after training to the system, the simulation as shown in Figure 5.

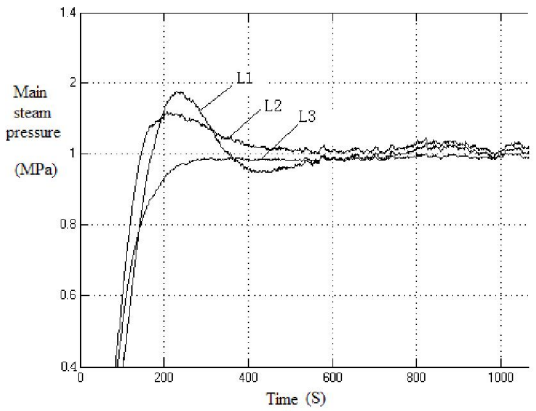


Fig. 5 Proposed beam former.

Including: L1 are traditional PID control, L2 is the fuzzy control, L3 using fuzzy neural network control. Simulations show that the system reached steady-state value at about 220s and the time decreased significantly by using the fuzzy neural network controller. Rising time and overshoot has obvious improvement compared with traditional PID method, steady-state error is less than 2%.The control effect of various control methods are shown in Table 1.

Table 1: Effect compare of three control methods

	Rise time(s)	Overshoot%	Response time s)
PID	180	18	580
Fuzzy control	170	5	400
FNN control	200	2	220

4. Conclusions

This paper focuses on the fuzzy neural network steam pressure controller which has many virtues. For example, the system responds very quickly, it can adjust the control parameters online, the steam pressure fluctuation is relatively small. This controller has not only the neural network self-learning ability and self-adapting ability, but also the advantage of accessing fuzzy information in fuzzy logic way, thus having good anti-disturbance ability and adaptive self-adapting ability. And this controller has simple structure which is easy to fulfill, with high precision in result, so it has wide application fields and great application value.

References

[1] Gao Shan,SHAN Yuan-da.A new neural network short-term load forecasting algorithm using radial basis function network.Automation of Electric Power Systems,1999,23(5):31-34.

[2] W.Yu,State-space recurrent fuzzy neural networks for nonlinear system identification,Neural Process.Lett.22(3) (2005)391-404.
[3] Zhang You-wang. Identification of dynamic system based on dynamic fuzzy neural network[J].Journal of Center South University Technology,2003,34(3):277-280..
[4] Lee Ching-hung,Teng Ching-cheng. Identification and control of dynamic systems using recurrent fuzzy neural networks[J]. IEEE Trans on Fuzzy Systems, 2004,8(4)： 349-366.
[5] JIANG Yong.Fuzzy neural network for short-term load forecasting[J].Relay,2002m3(7):11-13.
[6] ZhaoDeng-fu,ZHANG Tao, YANG Zeng-hui, et al.Short-term load forecasting using radial basis function(RBF) neural networks based on GN-BFGS algorithm.Automation of Electric Power Systems.2003.27(4):23-27.

Author Lei Wang received the B.S. degrees from Hebei institute of technology of department automation, Tangshan, China, in 2003. Currently she is an Assistant Professor at the Tangshan college, Tangshan, China. Her current research interests include nonlinear control systems, control systems design over network.

Two-Dimension Chaotic-Multivariate Signature System

Xiaoyan Sun[1], Maosheng Zhang[2*] , Huanguo Zhang[3], Xiaoshu Zhu[1]

[1] School of computer science and engineering, Yulin normal University, Yulin 537000, China

[2*] School of Mathematics and Information Science, Yulin normal University, Yulin, 537000, China

[3]School of Computer, Wuhan University, Wuhan 430079, China

Abstract

A novel hybrid cryptosystem which is resistant to quantum algorithm is developed. The system is combined with multivariate cryptosystem and chaotic system, two systems which are both secure under quantum attacks. The plaintexts are displaced by an affine transformation and encrypted by central map in multivariable cryptosystem. And then, the outputs of central map act as initial values in a two-dimension chaos system and are transformed by another affine transformation. Finally, the cipher texts are derived by adding the outputs of chaos system and the second affine transformation. Due to the chaos system, the shortcomings of traditional multivariate cryptosystems are offset and therefore the security is enhanced. The analysis shows that the proposed signature system is able to resist common attacks.

Keywords: *quantum computer; cryptosystem; chaotic; security*

1. Introduction

Public Key Cryptology was developed by Diffe and Hellman in 1976[1], which can provide various security services such as confidentiality, credibility (identification), integrity, non-repudiation, usability, access control[2], etc. There are some excellent Public Key Cryptosystems, such as RSA Public Key Cryptosystem which was based on big integer factoring and developed by Rivest in 1978[3]; ElGamal Public Key Cryptosystem which was based on discrete logarithm and developed by ElGamal in 1985[4]; Elliptic Curve Public Key Cryptosystem(ECC) which was developed by Koblitz and Miller in 1987[5]. These systems are widely used in all trades and professions. In 1994, a scientist named Peter Shor working in Bell laboratory proposed a famous algorithm to attack all cryptosystems which can be converted to discrete Fourier transform, including RSA, Elgma and ECC [6]. Shor algorithm and its extended algorithm are efficient with polynomial time in quantum computers. Meanwhile, 28 bit quantum computer was successfully created by D-Wave Company in 2007. NASA bought the first 128 bit quantum computer in 2011. Consequently, the widely used public cryptosystems are not secure under quantum computers. In recent years, Anti Quantum Computation Public Key Cryptology or Post Quantum Computation Public Key Cryptosystem has been received widespread attention and intensive studied all over the world [7, 8]. Multivariate Public Key Cryptology (MPKC) and chaotic cryptosystem are two cryptosystems which can resist quantum computers.

The security of MPKC is based on the difficulty of solving multivariate quadratic equations on finite fields. There is no evidence that multivariate quadratic equations can be solved efficiently on quantum computers [7, 9]. And the calculation resource consuming is much less than traditional public cryptosystems. In recent decades, scholars designed a few famous quadratic multivariate public algorithms, such as MI Cryptosystem developed by Matsumoto and Imai in 1988[10], Hidden Field Equation (HFE) system proposed by Patarin in 1995[11], Unbalanced Oil and Vinegar (UOV) Schemes designed by Patarin in 1997[12], Tame Transformation system originated by T.T.Moh in 1999[13]. Professor Boyin Yang and Minjun Chen proposed a well-known signature system named tame transformation signatures (TTS) in the first International Workshop for Applied PKI (IWAP2002). And then, there were many modification versions based on this signature system for different situations. In 2004, multivariate signature scheme SFlash was accepted as European safety standard for low consumption smart card by European New European Schemes for Signatures, Integrity and Encryption (NESSIE) [14]. Wuhan University developed a kind of new noise factor and noise-operation perturbed mode in order to strength the safety of Sflash in 2011[15]. Though MPKC is well researched by scholars and the great importance is attached by governments, the proposed MPKC above are proved to be unsafe one after another [16-19]. The ways of combining the Multivariate Public Key Cryptosystem with various modification modes can improve the security of cryptosystems, for example, minus mode can strength anti-attack property of Multivariate Public Key Cryptosystem obviously, branch mode can improve computing efficiency [13].

This paper designs a high security hybrid public key cryptosystem, in which two affine transformations, a central map and a chaotic system are combined. Plain texts are first transformed through an affine transformation and then encrypted with a central map. The first two elements of outputs of central map act as initial values in a two-dimension chaotic system. After applying another affine transformation, the outputs are added to the outputs of chaotic system and the cipher texts are derived.

The next section introduces some fundamental theory about MPKC and chaotic algorithm. Section 3 develops our proposed system and section 4 analyzes the security of our crypto scheme. Finally, conclusions are presented in section 5.

2. Multivariate public key cryptosystem and chaos theory

2.1 Multivariate Public Key Cryptosystem

Multivariate Public Key Cryptology System (MPKC) is established on a finite domain in a polynomial ring [20]. Mathematical structure of Multivariate Public Key Cryptology System is shown in Eq. (1).

$$Y=F(X)=T \circ P \circ S,\ \ F_q^n \to F_q^m \tag{1}$$

Among this formula, q is a prime number and F_q^k means the k-dimension vector space on finite domain F_q. T and S are invertible affine transformations on F_q^m and F_q^n respectively. P denotes polynomial equations with n variables and m equations, which is known as central map. The maximum degree of each equation is 2 and the coefficients belong to F_q. The central map is from F_q^n to F_q^m. Non-linear central map P is the core of Multivariate Public Key Cryptology System. The main function of T and S is to hide central map. The calculated result Y is public key. Expression form of its equivalent equation set is shown in Eq. (2).

$$y_i = \sum_{1\le j\le k\le n} c_{ijk}x_jx_k + \sum_{1\le j\le n} b_{ij}x_j + a_i,\ i=1,2,...,n \tag{2}$$

where $c_{ijk}, b_{ij}, a_i \in F_q$.

(S, P, T) are private keys of MPKC. MQ problem is usually expressed as given public key Y=F(X) but x needs to be worked out. IP problem is that Y is divided into S, P and T. Patarin proved that MQ problem on arbitrary domain is a NP complete problem and IP problem is a NP difficult problem on Eurocrypt conference. Multivariate Public Key Cryptology System is designed based on these two mathematic problems [8, 20].

2.2 Chaos Theory

Lorenz, Father of Chaos, defined chaos as "random behavior of deterministic system". Professor Shuisheng Qiu in South China University of Technology considered that if one movement includes the following three characteristics, namely, being sensitive to initial values, having random-like property and unpredictability, then it can be called chaos [21]. According to the number of variables in an equation, chaos equations can be categorized into one-dimension, two-dimension and three-dimension equations which are shown in Eq. (3), (4), (5) respectively.

$$x_{n+1} = \lambda x_n(1-x_n) \tag{3}$$

$$\begin{cases} x_{n+1} = 1 + y_n - C_n^2 \\ y_{n+1} = Bx_n \end{cases} \tag{4}$$

$$\begin{cases} \dfrac{dx}{dt} = \sigma(y-x) \\ \dfrac{dy}{dt} = \rho x - y - xz \\ \dfrac{dz}{dt} = xy - \beta z \end{cases} \tag{5}$$

Where (x, y, z) represent some meaningful variables and act as system trajectory. And also system parameters δ, ρ, β are participating in calculation.

There are many technologies to generate chaos sequence, such as Logistic Mapping, Kent Mapping, Chebyshev Mapping, etc. The expression form of one-dimension Logistic Mapping is shown in Eq.(6).

$$x_{n+1} = 1 - \mu x_n^2, x \in (-1,1), \mu \in [0,2] \tag{6}$$

Whereby x_n are called state variables. μx_n is named driving factor which drives state variable to change from x_n to x_n+1. μ is a bifurcation parameter. When the value of μ satisfies $3.5699456<\mu\le 4$, Logistic Mapping conducts chaotic state [22].

The subtle changes of parameters and initial condition in chaos system will cause avalanche phenomenon to the output of chaos sequence. Meanwhile, the output of chaos system has strong non-linear property as well as strong anti-attack ability against regular cryptanalysis method. So due to its usability, uniqueness, reliability, random-like property and unpredictability, chaos sequence is extraordinary suitable for structuring secure cryptology systems [23].

3. Multivariate-chaos crypto-algorithm

The structure of center map has great influences security performance and so plays a key role in MPKC system. Though it has property to resist quantum calculation, it doesn't receive wide application at present because the mathematical structural properties of center map cannot resist bilinear attack, rank attack, etc. Chaos system is of good random-like property and unpredictability, but many chaos system make use of same initial values and control parameters to start chaos system, meanwhile, the iterative process of chaos system isn't influenced by other factors (i.e. initial state is not given) and doesn't accord with safety criterion of "One-way Pad". As a result, it has potential security flaws. In order to resist quantum algorithm, we can combine the two systems together to make the two systems complement each other. The key idea of two-dimension chaos-multivariate is to add the outputs of multivariate cryptosystem and two-dimension system to demolish the potential mathematical properties of MPKCs. Thus, the hybrid system is secure under common attacks by using the center map of multivariate to change the initial state of chaos system and utilizing chaos system to generate the cipher texts.

3.1 Framework of proposed algorithm

Let X=(x_1, x_2,… x_n) denotes plaintexts and Y= (y_1, y_2,…, y_m) denotes the output of the cryptosystem. S and T are affine transformations. The diagram of proposed multivariate-chaos cryptosystem is depicted in Fig.1.

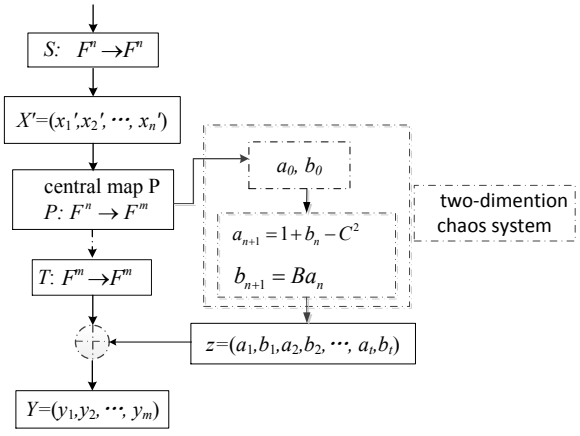


Fig.1. Diagram of multivariate branch chaos signature system

We first define an invertible affine transformation $S: F^n \to F^n$. Affine transformations cannot be directly used to develop cryptosystems because of the linearization relationship between inputs and outputs. We denote X'=(x_1', x_2',…, x_n') which are calculated using S-affine transformation with input parameter X.

Next, we construct a central map $P: F^n \to F^m$(m, n are integers and m is even). The central map consists of m equations with n variables in each equation. There are some existed central maps which can be utilized here, such as the central map of HFE, MI, UOV, TTS, etc. Though they are not secure, the disadvantages will be demolished in the following steps. The mathematic expression of above processes can be shown as Eq. (7).

$$Y' = (y_1', y_2', \ldots, y_m') = P \circ S(X) \quad (7)$$

Taken y'_1, y'_2 as initial stimulus in chaos system, the two-dimension chaotic algorithm is then calculated. After specifying parameters B and C, a two-dimension chaotic system is estimated as shown in Eq. (8).

$$z = (z_1, z_2, ..., z_m) = H(a_0, b_0, B, C) = (a_1, b_1, a_2, b_2, ..., a_t, b_t)$$

$$\begin{cases} a_0 = y'_1 \\ b_0 = y'_2 \\ a_{i+1} = 1 + b_i - C^2 \\ b_{i+1} = Ba_i \\ i = 0,1,...,t,\ t = m/2 \end{cases} \quad (8)$$

Another invertible affine transformation T is applied with input Y'. Adding the outputs of T and chaotic system (i.e. z), the cipher texts are finally generated. The generic construction of the new public key signature system is shown in Eq. (9).

$$Y=F(X)=z \oplus (T \circ P \circ S(x)) \quad (9)$$

The public key, i.e. polynomials of degree 2 over finite fields, is shown as Eq. (10).

$$Y = (y_1, ..., y_m) = F(x_1, ..., x_n)$$

$$= \begin{cases} y_1 = \sum_{j,k=1}^{n} \gamma_{1,j,k} x_j x_k + \sum_{j=1}^{n} \beta_{1,j} x_j + \alpha_1 \\ y_2 = \sum_{j,k=1}^{n} \gamma_{2,j,k} x_j x_k + \sum_{j=1}^{n} \beta_{2,j} x_j + \alpha_2 \\ \quad \ldots \quad \ldots \\ y_m = \sum_{j,k=1}^{n} \gamma_{m,j,k} x_j x_k + \sum_{j=1}^{n} \beta_{m,j} x_j + \alpha_m \end{cases} \quad (10)$$

And the private key consists of maps S,T, p and z.

3.2 Signature process

To sign a document, which is an element Y=(y_1,…,y_m), we need to solve the equation

$$Y=F(X)=T \circ P \circ S(x) \quad (11)$$

We must apply the inverse of S, P and T. First, we have Y':

$$Y'=T^{-1}(Y) = P \circ S(x_1,\ldots,x_n) \quad (12)$$

Next, we need to calculate the inversion of central map P. In this case, we must solve the equation

$$P(x_1,\ldots,x_n)= Y' \qquad (13)$$

It should be pointed out that the central map P is not an invertible map. To overcome this problem, we can calculate the parity-check of $(y'_1,\ldots,y'_m)$, which can be presented as $(o_1,\ldots, o_{n-m})$, and plug them into Y'. As a result, the new equation is shown in Eq. (14).

$$P(x_1,...,x_n) = \overline{Y} = (\overline{y}_1,....,\overline{y}_n)$$
$$where \begin{cases} \overline{y}_i = y'_i, i = 1,...,m \\ \overline{y}_i = o_{i-m}, i = m+1,...,n \end{cases} \qquad (14)$$

Thus, we can solve these invertible equations shown above and have all values of $(x'_1,\ldots,x'_n)$. Then we apply the inverse of S and estimate $X=(x_1,\ldots, x_n)=T^{-1}(X')$. In the next step, we will calculate chaos sequence. By taking y'_1, y'_2 as initial stimulus and specify parameters B and C, we can generate chaos sequence, which we denote by $z=(z_1,z_2,\ldots, z_n)$. We add the plain message Y and z and a new value $\overline{\overline{Y}}$ is calculated.

$$\overline{\overline{Y}} = Y \oplus z \qquad (15)$$

Again, taking $\overline{\overline{Y}}$ as plain messages and applying the inverse of affine transformation T, central map p and affine transformation S, we obtain a totally different values $X=(x_1,\ldots, x_n)$. Finally, the signature $s= (s_1, \ldots, s_{2n})$ is derived.

$$s = X \,||\, z \qquad (16)$$

3.3 Verifying the signature

To verify the signature, one needs to decompose s to X and z first. And then, one must calculate

$$\overline{\overline{Y}} = F(X) \qquad (17)$$

Finally, one checks if indeed

$$Y = \overline{\overline{Y}} \oplus z \qquad (18)$$

4. Security Analysis

There is still no convincing, strict and scientific demonstration for the security of chaotic cryptography system [24-27]. At the same time, any effective solution isn't found too. The subtle changes of initial condition and non-linear property can cause great differences to results, which makes huge amounts of analytical methods for classical cryptography doesn't work on Chaos system. Common attacks which aim at Multivariate Public Key Cryptology System include bilinear attack, rank attack and differential attack, etc. The fundamental principle of bilinear attack is to establish bilinear relationship of plaintext/cipher-text pairs by using the characteristic that both sides of center map equations are linear transformations. Rank attack uses the smallest rank of linear combinations equations in center map and the number of occurrences of the variable which appears least to attack. Differential attack works by using the feature that differential function of center map is a linear relationship of differences. Algorithm proposed in this paper adds center map and chaos sequence together. Hence, it makes full use of the nonlinearity of chaos to prevent the above-mentioned various attacks based on linearity. As a result, signature generated by proposed system cannot be forged.

5. Conclusion

This paper proposes a hybrid signature algorithm combining multivariate public key cryptology system and chaos theory. Taking advantages of nonlinearity and unpredictability of chaos theory, the system is able to offset the matrix relationship between plain-texts and cipher-texts and demolish its potential mathematical structural weaknesses of multivariate public key cryptology system, and improve the security of signature algorithm. But the computing efficiency of proposed algorithm is probably slightly less than traditional MPKC. In conclusion, the proposed system can be applied to any kind of mobile devices whose computing power is not very high.

Acknowledgments

This work is supported by the established Project of Educational Office of Guangxi(Grant Nos:201106LX513;201106LX516) and Key Project of Yulin normal university(Grant No: 2012YJZD17)

References

[1] W. Diffie, M. Hellman, "New directions in cryptography", Information Theory, IEEE Transactions on, vol. 22, no. 6, 1976, pp.644-654.
[2] S. CX, Z. HG, F. DG, C. ZF, H. JW, "REVIEW ON INFORMATION SECURITY", SCIENCE CHINA, vol. 37, no. 2, 2007,pp.129-150.
[3] Gupta, Kamlesh, Silakari, Sanjay. ECC over RSA for Asymmetric Encryption: A review. International Journal of Computer Science Issues,vol.8, no.3-2, 2012, pp. 370-375.
[4] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms", IEEE Transactions on Information Theory - TIT, vol. 31, no. 4, 1985,pp.469-472.
[5] N. Koblitz, "Elliptic curve cryptosystems", Mathematics of computation, vol. 48, no. 177, 1987,pp.203-209.
[6] P. W. Shor, "Algorithms for quantum computation: discrete logarithms and factoring", Foundations of Computer Science, 1994 Proceedings., 35th Annual Symposium on, 1994, pp.124-134.

[7] W. W. Cao, L. Hu, Cryptanalysis of a Multivariate Public Key Encryption Scheme with Internal Perturbation Structure, Berlin: Springer-Verlag , 2009.
[8] H. Z. Wang, H. G. Zhang, Z. Y. Wang, M. Tang, "Extended multivariate public key cryptosystems with secure encryption function", Science China-Information Sciences, vol. 54, no. 6, 2011,pp.1161-1171.
[9] Y. Hashimoto, T. Takagi, K. Sakurai, General Fault Attacks on Multivariate Public Key Cryptosystems, Berlin: Springer-Verlag, 2011.
[10] T. Matsumoto, H. Imai, "Public quadratic polynomial-tuples for efficient signature-verification and message-encryption", In Advances Cryptology -EUROCRYPT, 1988, pp.419-453.
[11] J. Patarin, "Hidden Fields Equations (HFE) and Isomorphisms of Polynomials (IP): Two New Families of Asymmetric Algorithms", Advances in Cryptology-EUROCRYPT '96, 1996, pp.33-48.
[12] A. Kipnis, J. Patarin, L. Goubin, "Unbalanced Oil and Vinegar signature schemes", Theory and Application of Cryptographic Techniques-EUROCRYPT'99, 1999, pp.206-222.
[13] J. Ding, J. E. Gower, D. Schmidt, Multivariate public key cryptosystems, New York: Springer, 2006.
[14] J. Patarin, N. Courtois, L. Goubin, "FLASH, a Fast Multivariate Signature Algorithm Topics in Cryptology—CT-RSA 2001", Cryptographers' Track at RSA, 2001, pp.298-307.
[15] H. Z. Wang, H. G. Zhang, H. M. Guan, H. Q. Han, "A new perturbation algorithm and enhancing security of SFLASH signature scheme", Science China-Information Sciences, vol. 53, no. 4, 2010, pp.760-768.
[16] V. Dubois, P. A. Fouque, A. Shamir, J. Stern, "Practical cryptanalysis of SFLASH", International Crytology Conference-CRYPTO 2007, 2007, pp.1-12.
[17] X. Y. Nie, Z. H. Xu, L. Lu, Y. J. Liao, Security Analysis of an Improved MFE Public Key Cryptosystem, Berlin: Springer-Verlag, 2011.
[18] J. C. Faugere, L. Perret, High Order Derivatives and Decomposition of Multivariate Polynomials, New York: Assoc Computing Machinery, 2009.
[19] D. Smith-Tone, On the Differential Security of Multivariate Public Key Cryptosystems, Berlin: Springer , 2011.
[20] C. Wolf, Multivariate quadratic polynomials in public key cryptography, Mierlo: Leuven, 2005.
[21] S.S. Qiu, Y.F. Chen, M. Wu, Z. Ma, "Discussion on Chaotic Secure Communication and New Schemes of Chaotic Encryption", Journal of South China University of Technology(Natural Science Edition), vol. 30, no. 11, 2002, pp.75-80.
[22] Y. Jing, G.J. shall, S.B. Yu, "An Improved Approach of Logistic Chaotic Series Encryption", Journal of Automatic Technology and Application, vol. 23, no. 2, 2004, pp.58-61.
[23]Ahadpour, Sodeif; Sadra, Yaser; ArastehFard, Zahra. "A Novel Chaotic Encryption Scheme based on Pseudorandom Bit Padding", International Journal of Computer Science Issues,vol.9, no.11-2, 2012, pp. 449-456.
[24] L. Kocarev, "Chaos-based cryptography: a brief overview", Circuits and Systems Magazine, IEEE, vol. 1, no. 3, 2001, pp.6-21.
[25] Akhavan, A, Samsudin, A, Akhshani. "On the speed of 'Image encryption with chaotically coupled chaotic maps' ", International Journal of Computer Science Issues,vol.9, no.3-3, 2012, pp. 452-454
[26] J. Amigó, Chaos-Based Cryptography Intelligent Computing Based on Chaos, Berlin : Springer /Heidelberg, 2009
[27] C. Pellicer-Lostao, R. Lopez-Ruiz, Notions of Chaotic Cryptography: Sketch of a Chaos based Cryptosystem, USA: arXiv, ,2012.

First Author Biographies

Xiaoyan Sun received the bachelor degree from Guangxi Normal University in 2004 and master degree in computer science from the school of mathematics & computing science in Guilin University of Electronic Technology. Currently, she is a lecturer at Yulin normal University, China. Her research interests include software protection and watermarking. She has published over 10 papers and 2 books on journals and/or international conferences. Her research has been supported by 8 provincial-level research projects.

Second Author Biographies

Maosheng Zhang received the bachelor degree from Hubei University in 2004 and master degree in mathematics from Dalian university of technology in 2009. He is a Ph.D. candidate of Wuhan University. Currently he is a lecturer at Yulin normal university. His research interests are in cryptography, watermarking and multimedia coding. He has published over 10 papers and 2 books on journals and/or international conferences and proposed 1 national standard and 2 national patent. He is an ACM member and his research has been supported by 6 provincial-level research projects.

Visual Saliency Based on Local and Global Features in the Spatial Domain

Chao Jia[1], Fang Hou[2] and Liangliang Duan[3]

[1] College of Information Science and Engineering, Yanshan University
Qinhuangdao, HeBei，066004，China

[2] College of Information Science and Engineering, Yanshan University
Qinhuangdao, HeBei，066004，China

[3] College of Information Science and Engineering, Yanshan University,
Qinhuangdao, HeBei，066004，China

Abstract

The human visual system can quickly and efficiently capture the salient objects in a scene. Based on the biological mechanism, a new multi-scale saliency analysis method is proposed in this paper, in which the differences of region colors and spaces are calculated in different scale and their saliency map are fused together. First, we calculate the image saliency by using the color and space information of both local and global in single scale. Then by applying the multi-scale fusion, we can effectively inhibit outstanding but not salient region in each single scale, and different scale can also reflect salient region of the images from different aspects. The experiment results show that this algorithm can effectively predict the salient region attracting human attention. Our method has the state-of-the-art performance and achieves excellent results for salient objects of different sizes and salient region with complicated background in an image.

Keywords: *local and global features , salient object detection, patch saliency, color and space difference, multi-scale fusing*

1. Introduction

Human visual system on the analysis of complicated scene taking a serial calculation strategy, quickly turn attention to stay in a few salient target object, which is processed prior and this reaction process is called visual attention. Imitating the biological mechanism of the computer vision of salient region detection algorithm in image segmentation[1][4], image classification[2][5], target recognition [3], relocation [24], etc takes the high value [6] [7] [8][23]. Generally speaking, there are two different processes that influence visual saliency, one is top-down visual attention model, it uses high-level semantic features and knowledge-driven to compute visual saliency. The other is a bottom-up visual attention model, which is data-driven, automatic processing and it relies on image features. This paper focuses on bottom-up algorithm in salient region detection, which now roughly is divided into two kinds based on local information and global information.

According to cognitive psychology [17] and neurobiology [18][22], we can know that how visual saliency and human treat and deal with salient object of image which is closely related to visual stimuli. When observing an image, the observer's cortical responses to the salient region first. By analyzing the human visual system and summary of the current popular algorithm [9] [10] [11] [12] [13] [14] [15][16], we can know the characteristic of a good visual saliency detection is as follow:

1. Saliency detection algorithm based on the region, the entire region can be separated from the surrounding environment. This method is superior to only highlight salient contour of the object than saliency detection method based on a single pixel.
2. The method based on global color and space differences tends to assign similar saliency value to the similar regions in an image and uniformly highlight the whole salient object.
3. Salient detection algorithm of single scale is able to highlight area having specific characteristics in the image, but they are not all salient region. So using multi-scale salient detection algorithm can inhibit the region obtained at a signal scale which is outstand but not salient.

Based on the analysis of the human visual system and the current problems resolved existing in the algorithm, this paper propose a new detection algorithm by calculating difference between global region colors in different scales and fusing the saliency map in each scale. This algorithm has very good effects to process natural scene, and it can highlight the salient object in the complicated background or salient object repeatedly disturbed by background. The method in a new data set is tested, through the experiments, the proposed detection algorithm have good

results and have achieved better accuracy with the current popular algorithms. It can effectively predict the region human pay attention to.

The remained of the paper is organized as follows: section 2 gives the description and review of related work. In section 3, we elaborated the framework of our salient detection algorithm in details by calculating the image saliency values of single-scale and fusing saliency map of different scales. In Section 4, we demonstrated our experiment results by comparing Ground truths and the results with another three currently popular algorithms. The conclusions are given in Section 5.

2. Related work

Many bottom-up models have appeared in the literature. In [9], by imitating human visual bottom-up attention mechanism, Itti proposed IT model based on the local color, brightness, and other information, through the multi-scale image characteristics of the center-around deviation. The defects of this model are that the effect is poor in the dynamic space. This theory became the base of salient region detection model. Walther [19] expanded the Itti model, through the hierarchical feedback connection made fixed size round of salient region in IT model extend into the shape of salient region in image, which can better guide target recognition, but whose efficiency decline. Liu [8] propose the multi-scale contrast to calculate image saliency based on Gaussian image pyramid contrast linear combination. The advantage of this method is to be able to use the existing mature region segmentation algorithm to carry on the calculation, but it relies on region segmentation algorithm too much. The above methods are based on local information, and these methods tend to give the edge of salient region higher saliency value, not fully highlight the whole object.

Recently, some researchers begin to focus on the global information of the image. R. Achanta et al. proposed a frequency-tuned salient region detection algorithm (FT) in [13]. The FT first converts image to Lab color space then defines the saliency at each location as the difference between the Lab pixel value and the mean Lab value of the entire image. This method proposes a new direction of detecting salient region. In [14], Hou et al. propose a simple and fast algorithm based on spectrum residual (SR) in frequency domain, but the saliency map of SR only use the information contained in the amplitude Spectrum. And in [15], the phase spectrum of the Fourier transform (PFT) was introduced and achieved nearly the same performance as the SR. But SR and PFT tend to find small salient region. When detecting a large salient object, they highlight the contour of the object while ignoring the internal details of the object. Based on SR and PFT, Li proposed HFT [16] detection algorithm based on frequency domain. HFT make full use of the information of amplitude spectrum and phase spectrum in frequency domain after Fourier transform to calculate the saliency value of image. Goferman in [12] propose context-Aware salient region detection algorithm (CA) according to differences of global feature, and the algorithm takes account of the information local and global properties of image, visual organization principle and surface characteristics to achieve salient region calculation.

The above method is based on the global information, whose efficiency has been improved in the calculation, and it considers the global relationship, while most of them ignore the existing of the spatial relationship in the image and can't even uniformly highlight objects for the larger salient objects.

3. Visual saliency model

According to the visual organization principle, each pixel of the image is not independent, and there is usually some relationship between the pixels. In the FT, only the individual pixels and global average color difference are used, and its defect is that only considers a single isolated pixels, and ignores the contact between the pixels and the pixels around, and fails to consider spatial information. According to visual organization principle, the salient region of an image is formed by one or several very important piece of composition, that is to say salient pixel in the image is concentrated. Therefore, we calculate the single scale image saliency based on the region. Here, we will first divide images into several parts according to the different scale, and the integral image is introduced to separate and describe them. Then based on the patch already divided according to each scale, we calculate the global color and space difference of each patch between and considered as image saliency value, thus obtains the saliency map of different scales.

3.1 Region division and representation

To count color features of region of the input image, we need to divide the image into multiple areas. Firstly, images of different sizes uniform converted into the image whose size is 256 * 256, So that we can divide each image at the same scale. We divide image into the multiple patch whose size is k*k, in this paper, k=4,8,16,32. The number of patch in the image is 256 / k * 256 / k.

The size of image can be random. In order to obtain the statistical input image area color characteristics the image is divided in the same scale. First of all there is need to

unify different size images into the same size of 256 * 256. Then the images are divided into N non-overlapping patches with the pixel number for k * k, and the k is the scale adopted in this paper of which the four dimensions, namely k = 4,8,16,32. The number of patch diagram N is 256 / k * 256 / k. In this algorithm, each patch is represent with its region color means value, so the concept of integral image is introduction. Integral image can fast calculate the sum of all pixels on a random area of image. A random point (i, j) in a integral image refers to sum of pixels of all the points in the rectangular area from the top left hand corner of the image to the point., The formula is as follows:

$$s(i, j) = \sum_{m=0}^{i} \sum_{n=0}^{j} f(i, j) \quad (1)$$

The value of each position in the integral image can also be calculated using the following formula:

$$s(i, j) = s(i-1, j) + s(i, j-1) - s(i-1, j-1) + f(i, j) \quad (2)$$

Where, $s(i, j)$, $s(i-1, j)$, $s(i-1, j-1)$ represent the value of the position (i,j), (i-1,j), (i-1,j-1) in integral image. $f(i, j)$ is the value of the position (i,j) in original image. According to the integral image, we can use Eq(3) to calculate the sum of all pixels within each region of original image.

$$F(i_p...i_q, j_p...j_q) = \sum_{i=i_p}^{i_q} \sum_{j=j_p}^{j_q} (s(i_q, j_q) - s(i_q, j_{p-1}) - s(i_{q-1}, j_p) + s(i_{q-1}, j_{p-1})) \quad (3)$$

Where, $F(i_p...i_q, j_p...j_q)$ represent the sum of the pixels within region among four pixels point (p,p),(p,q),(q,p),(q,q) in original image. $s(i_q, j_q)$, $s(i_q, j_{p-1})$, $s(i_{q-1}, j_p)$, $s(i_{q-1}, j_{p-1})$ represent the value of position (i_q, j_q), $(i_q, j_{p-1}), (i_{q-1}, j_p), (i_{q-1}, j_{p-1})$ in integral image. Figure 1 (1) shows the pixels of a simplified image, and Figure 1 (2) is the integral image corresponding to Figure 1 (1). If we want to calculate the sum of pixels in gray region in figure 1(1), we just need to calculate the pixels of yellow position in (2) according to the Eq (3).

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

(1)

3	5	12	14	17
4	11	19	24	31
9	17	28	38	46
13	24	37	48	62
15	30	44	59	81

(2)

Fig 1. (1) Original image. (2) The integral image corresponding to (1).

Therefore, the mean value of pixels of the region can be calculated according to the Eq (4):

$$\overline{F}(i_p...i_q, j_p...j_q) = F(i_p...i_q, j_p...j_q) / ((q-p)*(q-p)) \quad (4)$$

Where, $\overline{F}(i_p...i_q, j_p...j_q)$ represent the mean value of the pixels within region among four pixels point (p,p),(p,q),(q,p),(q,q) in original image. (q-p)*(q-p) is the size of region.

Using integral diagram can effectively calculate the mean pixel of arbitrary area and is very suitable for calculation of the region characteristics, thus it can improve the efficiency of the proposed algorithm.

3.2 Local and global different-scale image saliency

Given an image using 2.1 to do image division, the image is divided into patches and each patch can be expressed. This section will be effective to calculate the saliency value of each piece, make saliency value differences between patches clear, highlight salient region and inhibit the non-salient area. In this process, there are two factors will be considered: one is the difference of the color in the image between any two blocks; the second is distance between them, which should be considered combined with the global information.

3.2.1 The color difference between patches

Color as an important bottom feature, it can well describe the differences between the region and highlight salient region. So, in this paper, we use the mean Lab value of region to describe a region. According to the principles of visual organization, colors of salient patches are similar, and the color difference is large between non-salient patches and salient patches. So, we define the distance between a patch and others in an image as Eq (5).

$$d_c[i][j] = sqrt((\overline{F}_L(i) - \overline{F}_L(j))^2 + (\overline{F}_a(i) - \overline{F}_a(j)) + (\overline{F}_b(i) - \overline{F}_b(j))^2) \quad (5)$$

Where, $d_c[i][j]$ is the color distance between the i-th patch and the j-th patch, $\overline{F}_L(i)$, $\overline{F}_a(i)$, $\overline{F}_b(i)$, $\overline{F}_L(j)$, $\overline{F}_a(j)$, $\overline{F}_b(j)$ respectively represent the mean L,a,b value of the i-th patch and the j-th patch in CIE $L*a*b$ color space.

3.2.2 Combining with the spatial distance

In an image, the number of salient patch is less, If the sum of color difference between a patch and all the other patches in image is larger, then it can be identified as a salient patch. Therefore, we use the sum of color difference between each patch and the other patches in the image to describe the patch saliency.

The spatial distance between the patches is a very important factor when calculating image saliency. Because spatial distribution of most salient patches is concentrated on the adjacent areas of the center of the image, however, non-salient patches can be distributed within the whole image. According to the spatial features of the salient region, if a region is salient, the possibility that its surrounding region is salient is larger and the possibility that regions far away from it are non-salient is larger. For regions in image, with the increase of the space distance between them, the influence between them will be smaller. By integrating the information of region color differences and space distance, image salient formula based on the single-scale is given. The formula as follows:

$$S[i]=\sum_{j=0}^{n}(1/(1+d_p(i,j))*sqrt((\overline{F}_L(i)-\overline{F}_L(j))^2 +(\overline{F}_a(i)-\overline{F}_a(j))^2+(\overline{F}_b(i)-\overline{F}_b(j))^2) \quad (6)$$

$$d_p(i,j)=\sqrt{(i_x-j_x)^2+(i_y-j_y)^2} \quad (7)$$

Where, $d_p(i,j)$ is the Space Euclidean distance between the i-th patch and the j-th patch. S[i] is the image saliency value. By add in space distance, we can control the interaction among all regions in the image. Saliency map of different scale are shown in figure 3.

(1) (2) (3) (4) (5)

Fig 2. (1) The input image. (2) The saliency map that size is 4*4. (3) The saliency map that size is 8*8. (4) The saliency map that size is 16*16. (5) The saliency map that size is 32*32.

3.3 Analysis of saliency map of each scale

When a scene is very far away from observers, the human visual attention mechanism will more focus on the whole salient regions in the image. When a scene is close to us, human visual attention mechanism can pay more attention on more salient or detail part in the salient region. Our method adopted multi-scale to mimic the biological mechanisms. In this section, we analyze the characteristic of single-scale saliency map.

When image was divided by small scale, we can clearly highlight the whole salient region as well as details of the region. If image was divided by larger scale, we can very accurately locate the position of the salient region (Figure 2). Original image is shown in figure 2(1). Figure 2(2) is saliency map whose original image is divided at a scale of 4*4 and we clearly see the contour and details of salient region. The figure 2(5) shows the saliency map at a scale of 32*32, and we can accurately locate the position of the flower in image which is the most significantly bright region. The figure 2(3) and 2(4) is saliency map at scale 4*4 and 8*8, their contour gradually blurred but the positions of salient region increasingly clear. In addition, when image was divided by small scale, we can clearly highlight the more salient part in salient region. However image was divided by larger scale, we can highlight the position of salient region. In figure 2(2), we also can find the stamen portion is brighter and it is more salient. The stamen portion is longer highlight in figure 2(5) and the whole flower is salient.

We can highlight a patch having a specific characteristic according to the color and spatial characteristics of the image by using single-scale, but these patches are not all salient. When image was divided by small scale, saliency map tend to highlight contour and details of the salient object. However, when larger scale is adapted, we can very accurately locate the position of the salient region in image.

3.4 The integration of the salient region in the multi-scale

Natural images contain wide content and degree of complexity is diversified. The salient detection effect of a single scale is not very ideal. Recently, it has been noticed and the multi-scale detection algorithm was introduced. In [6], multi-scale model based on frequency domain has been proposed and achieves good results. By using the multi-scale salient detection algorithm, we can analyze the region of interest and calculate the image saliency at different scales , thus we can do a comprehensive analysis of the salient region of the image and highlight salient region of different sizes in images.

According to gestalt law, the visual organization form is due to one or a few region that is highlight and causes the attention of the visual system. It implies that the region attracting attention of visual system is obviously salient. The region that the algorithm of different scales all can highlight most likely is salient, then we should highlight

this region and suppress other regions of the image in final saliency map. So we superimposed saliency map obtained at different scale by pixel to get our final saliency map. The experiments show saliency maps of four scales all contain important information. The final saliency is calculated as Eq (8).

$$S = \sum_{i=1}^{k} r_i * S_i \quad (8)$$

Where, S is saliency value of final saliency map, S_i is saliency value of saliency map obtained at different scale. r_i is corresponding weight of each saliency map. r_i =1/4 in our experiment. Our final saliency map is shown in figure 3(6), the whole salient region is highlighted and the contour is very clear. At the same time, the dandelion stems which is non-salient but highlighted in figure 3(2) and 3(3) and redundant parts figure 3(5) in are inhibited.

Multi-scale can make better use of the local and global feature information in image. The single-scale saliency map can highlight the region which contains special features and reflect salient objects from different aspects. After fusing the saliency map of different scale, we can highlight unusual region which is belong to salient object and the final saliency map inherit the advantage of most saliency map. By using multi-scale, the accuracy that predict salient region is improved, the efficiency of the algorithm is also high and can achieve real-time detection.

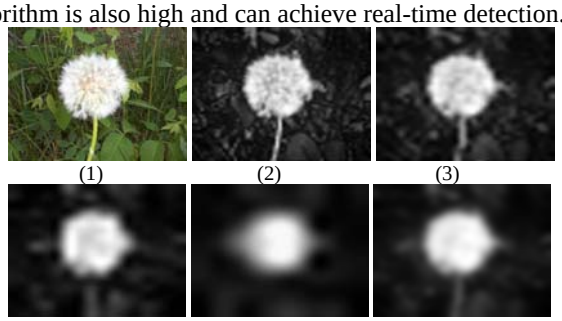


Fig 3. (1) The input image. (2) The saliency map that size is 4*4. (3) The saliency map that size is 8*8. (4) The saliency map that size is 16*16. (5) The saliency map that size is 32*32. (6) The final saliency map

4. Experiment

In this section, we introduce a new database set containing 126 images was collected using Achanta[13] and Li's database [16] as well as the recent literature. Images of the database include salient objects of different sizes, background of some images is complex or repeated interferes with salient object and other images contain more than a salient object. At the same time, the database also provides salient region maps labeled by humans as ground truth. We evaluate our algorithm in the database by comparing with ground truth and three classic and novel algorithms of SR, CA and Itti. In compute vision, the human fixation most concentrated in salient region. So the salient detection algorithm should accurately be able to predict the region human pay attention to [14][20]. Our algorithm use multi-scale fusion and it make the advantage of the saliency map of each scale converge to the final saliency maps. It can clearly highlight the position as well as details of salient object. Therefore, we will analyze the feasibility of our salient region detection algorithm based on region.

We will qualitative evaluate the implementation of our algorithm by using the ground truth of region and compare with SR, Itti and CA. The figure 5 shows the comparison among them. We can clearly find when detecting lager region, SR places extra emphasis on edge detection (figure 4(4)), and also Itti have same characteristic (figure 4(5)). Our algorithm can highlight the edge of salient region as well as details of the region (figure 4(3)).The third column of Figure 4 are the saliency map by using our method. The experimental results show our method can highlight salient object in image, whether complex background or foreground disturb background, besides, shape and detail of salient object is much the same as the ground (figure 4).

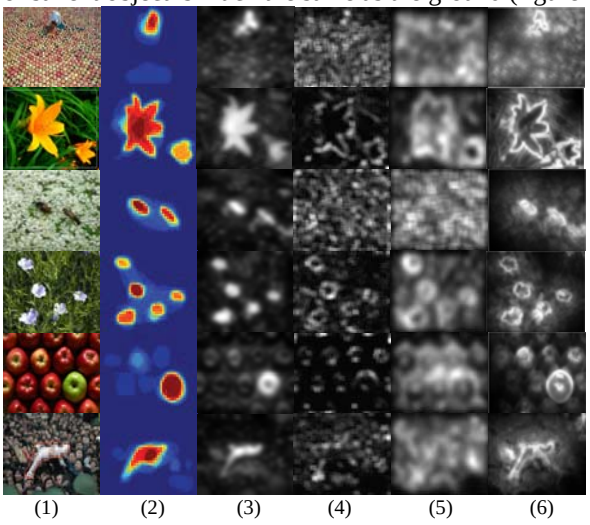


Fig 4. (1) Original image.(2)Ground Truth.(3)Our method.(4)SR.(5)Itti.(6)CA.

To evaluate our algorithm, we introduce Receiver operating characteristic curve (ROC) and the area under ROC (AUC). ROC curve is that the true positive rate (TPR) between saliency map and region human relabeled is as the horizontal axis and the false positive rate (FPR) is as the vertical axis.

$$TPR = TP/(TP + FN) \quad (9)$$

$$FPR = FP/(FP+TN) \quad (10)$$

TP is true positive， FP is false positive ， TN is true negative，FN is false negative.

In this experiment, besides ROC, we also use the DSC (Dice Similarity Coefficient) as a measure to evaluate the overlap between the threshold saliency map and the ground truth. The peak value of the DSC curve (PoDSC) is an important index of performance, as it corresponds to the optimal threshold and the best possible algorithm performance [21].

Table 1 and Figure 5 show the experimental results of our method and other algorithms in the entire database. We can see the value of AUC and PoDSC of SR, CA, Itti and our algorithm and they are calculated under the same conditions. Figure 5 shows corresponding ROC curve. By comparing two values in table 1 and ROC curve of figure 5, we find our algorithm can get better detection effect on the entire database, our algorithm have higher AUC and PoDSC. It indicates that our algorithm can more accurately predict region that human pay attention to and highlights the position and details of the region.

Table 1. Performance of each method.

	The value of AUC	***The value of PoDSC***
Our method	0.9353	0.6363
SR	0.8074	0.4720
Itti	0.9120	0.5602
CA	0.9252	0.6196

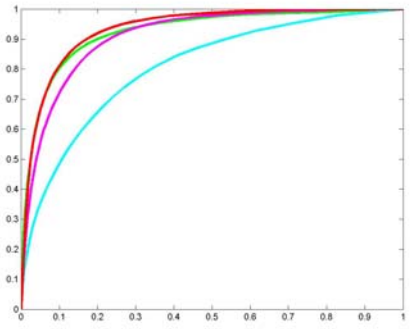


Fig 5.The ROC curves of our model and the other three approaches on dataset of this paper.

5. Conclusions

In this paper, we propose a new saliency detection method. To measure the saliency, we actually combine the color and spatial distance information of image in different scale space, then fusing saliency map of each scale. Experiment results show that our method outperforms some state-of-the-art saliency detection approaches on predicting the region that human pay attention to. One of the limitations of our method is that we focus on regions possessing distinct low-level features. Therefore, next, we will introduce high-level semantic information or increase texture, shape and other information to improve our saliency detection approach. The goal of our research is develop a system for detection of signpost and license plate.

Acknowledgments

This work is supported by the Natural Science Foundation of Hebei Province (No. E2011203212).

References

[1]J. Han, K. Ngan, M. Li, and H. Zhang. "Unsupervised extraction of visual attention objects in color images", IEEE TCSV, 16(1), 2006, pp.141–145.

[2] E. Nowak, F. Jurie, and B. Triggs. "Sampling strategies for bag-of-features image classification", Lecture Notes in Computer Science, 3954:490, 2006.

[3] U. Rutishauser, D. Walther, C. Koch, and P. Perona. "Is bottom-up attention useful for object recognition? ", In CVPR, 2004 ,pp. 37–44.

[4] KK Singh, A Singh, "A Study Of Image Segmentation Algorithms For Different Types Of Images", IJCSI,vol.7, September 2010, pp 414-417.

[5] B Shinde, D Mhaske, AR Dani ,"Study of Image Processing, Enhancement and Restoration", IJCSI, vol. 8, no. 3, November , 2011,pp.262 - 264.

[6] X. Hou, J. Harel, and C. Koch,"Image Signature: Highlighting Sparse Salient Regions", IEEE Trans. Pattern Analysis and Machine Intelligence, 2012, pp. 194 - 201.

[7] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition", IEEE Trans. Pattern Analysis and Machine Intelligence, 2009, pp. 989 - 1005.

[8] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 33, no. 2, 2011,pp. 353 - 367.

[9]L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 11, Nov 1998, pp. 1254 - 1259.

[10]J. Harel, C. Koch, and P. Perona. "Graph-based visual saliency" 2006, In NIPS, pp. 545–552.

[11]T. Kadir and M. Brady, "Saliency, scale and image description", International Journal of Computer Vision, vol. 45, no. 2, 2001 ,pp. 83–105.
[12] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection", IEEE Trans. Pattern Analysis and Machine Intelligence ,2012, pp.1915 – 1926.
[13]R. Achanta, S. Hemami, F. Estrada, and S. Ssstrunk, "Frequency-tuned Salient Region Detection", in IEEE Conf. Computer Vision and Pattern Recognition, 2009.
[14]X. Hou and L. Zhang, "Saliency detection: A spectral residual approach", in IEEE Conf. Computer Vision and Pattern Recognition, 2007.
[15]C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform", in IEEE Conf. Computer Vision and Pattern Recognition, 2008.
[16] Jian Li, Martin D. Levine, Xiangjing An, Xin Xu, Hangen He, "Visual Saliency Based on Scale-Space Analysis in the Frequency Domain", IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 2012.
[17]S. K. Mannan, C. Kennard, and M. Husain. "The role of visual salience in directing eye movements in visual object agnosia", Current biology, 19(6) , 2009,pp.247–248.
[18]J. M. Wolfe and T. S. Horowitz. "What attributes guide the deployment of visual attention and how do they do it?" Nature Reviews Neuroscience, 2004, pp. 5:1–7.
[19]D.Walther, C.Koch. "Modeling Attention to Salient Proto-Objects[J] ". Neural Networks. 19(9), 2006, pp.1394-1407.
[20]L. Elazary and L. Itti, "Interesting Objects Are Visually Salient",Journal of Vision, vol. 8, no. 3, 2008, pp. 1 – 15.
[21]T. Veit, J. Tarel, P. Nicolle, and P. Charbonnier, "Evaluation of Road Marking Feature Extraction", IEEE Conf. Intelligent Transportation Systems, 2008.
[22] L Duan, C Wu, F Fang, J Miao, Y Qiao, J Li,"Visual Attention Shift based on Image Segmentation Using Neurodynamic System",IJCSI, vol. 8, no.3, January , 2011,pp.81 – 86.
[23] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu, "Global Contrast based Salient Region Detection", in IEEE Conf. Computer Vision and Pattern Recognition, 2011.
[24] M. Rubinstein, A. Shamir, and S. Avidan. "Improved seam carving for video retargeting", ACM Trans. on Graphics, 27(3), 2008.

Chao Jia Professor, PhD supervisor, born in 1967. In 1991, he received the B.E. degree in computer profession, Northeast Heavy Machinery Institute. In 1998, he received the M.Sc degree and in 2005 he received the Doctor degree in College of Information Science and Engineering, YanShan University, China. Since 2011 he has been with College of Information Science and Engineering as a Professor. Prof. He has published numerous papers, completed more than one large-scale project. Such as his main research interests include computer graphics, computer vision and image processing, virtual reality and virtual simulation.

Fang Hou was born in 1987. She received the B.S. degree in College of Computer Science and Technology in Harbin Science and Technology University in 2011. Currently, she is a graduate student in College of Information Science and Engineering, Yanshan University, China. Her main research interests include computer vision and image processing.

Liangliang Duan was born in 1985. He received the B.S. degree in College of Information Science and Engineering in Shandong Agricultural University in 2009. He received the M.Sc degree and in 2011. Currently, he is a doctor student in College of Information Science and Engineering, Yanshan University, China. His main research interests include computer vision and image processing.

Study of Online Bayesian Networks Learning in a Multi-Agent System

Yonghui CAO[1, 2]
1, School of Economics & Management, Henan Institute of Science and Technology, Xin Xiang, 453003 ,China
2, School of Management, Zhejiang University, Hang Zhou,310058 ,China

Abstract
This paper introduces online Bayesian network learning in detail. The structural and parametric learning abilities of the online Bayesian network learning are explored. The paper starts with revisiting the multi-agent self-organization problem and the proposed solution. Then, we explain the proposed Bayesian network learning, three scoring functions, namely Log-Likelihood, Minimum description length, and Bayesian scores.

Keywords:*Bayesian Network, Search Algorithms, Heuristic Search, Exhaustive Search*

1. Introduction

We attempt to find how a common task can be performed by a multi-agent self-organizing system. The agents are independent in terms of their model of environment and their actions. Each agent explores the environment and decides its actions by itself. Agents will have no information about the environment at the beginning of their exploration of the environment. They will explore the environment, model the environment and take actions to change the environment according to the common task. We attempt to solve these problems by utilizing Bayesian networks and influence diagrams.

Bayesian networks are employed to model the environment. Because the agents have no or limited information about the environment at the beginning of their exploration, an online Bayesian network learning method will be used. Influence diagrams will be employed to obtain the agents' actions. Bayesian networks and influence diagrams are combined to produce a decision-theoretic agent in a multi-agent system.

Bayesian network learning is examined broadly. There are four cases of Bayesian network learning depending on the availability of the network and the data. The unknown structure and incomplete data case is the nearest case to our problem. Our network structure is not defined in advance and the sensor data may not be complete. On the other hand, for simplicity we will assume the data is complete during the simulations. The agents do not have significant amounts of prior knowledge about the environment. Therefore, the BN will be formed during the agents' exploration of the environment. Each new data case will affect the structure of the network.

Online Bayesian network learning consists of two parts, namely parameter learning and structural learning. Parameter learning is the calculation of the conditional probability table elements of each node in a given Bayesian network. In this research, we use a modified version of Maximum Likelihood Expectation method to calculate the network parameters. Maximum likelihood estimation method is modified so that it has a closed form when the probabilities need to be updated.

Structural learning is the problem of finding the network that represents the data the best. This involves two parameters, complexity of the network and fitness of the network to the data. The structural learning process tries to find the optimal network that provides optimal complexity and fitness. The main building block in structural learning is the search algorithm that generates the network with the highest score.

2 .The Parameter Learning

There are two types of parameter learning techniques used in the literature, MLE and Bayesian estimation. It is stated that with a database having a large number of data cases, these two methods converge to each other. The latter can take prior knowledge if it is available. Also, it is shown that the latter has a closed form. In this section, we have redefined the Maximum Likelihood calculation to have a closed form calculation. Because MLE is computationally simpler than Bayesian estimation, it is employed in our parameter learning. The following paragraphs explain how the parameter learning is performed by modified MLE method.

Let $X = \{X_1, X_2, \cdots X_m\}$ be the discrete variables (nodes) in a Bayesian network, B. Assume that we know that the node X_j is the child of the node X_i, which means $X_i \rightarrow X_j$ .In this case, the parameter learning has to calculate the values in the conditional probability table in the node X_j. The conditional probability can be calculated by utilizing using the fundamental formula for probability calculus as in Equation (1)

$$P(X_i \mid X_j) = \frac{P(X_i, X_j)}{P(X_i)} \quad (1)$$

Since MLE is employed in parameter learning, the probabilities can be calculated by utilizing the natural frequencies of the data cases. A natural frequency of a data case is calculated by counting the number of occurrences of the data case in the database. For individual probabilities, we count the number of occurrences of a state of a variable in the database. Let n_{ij} be the number of occurrences of the state j of the ith variable in the database and n is the total number of data cases in the database. Using these frequency values, we can calculate the probabilities in the following way:

$$P(X_i = x_j) = \frac{n(X_i = x_j)}{n} = \frac{n_{ij}}{n} \quad (2)$$

Thus, the conditional probabilities can be calculated by using the individual probabilities in Equation (1). The conditional probability $P(X_i \rightarrow X_j)$ can be obtained as in the following equations.

$$P(X_i \mid X_j) = \frac{P(X_i, X_j)}{P(X_j)} \quad (3)$$

$$P(X_i, X_j) = \frac{n(X_i, X_j)}{n} \quad (4)$$

$$P(X_i) = \frac{n(X_j)}{n} \quad (4.5)$$

As can be seen in Equations (4) and (5), the denominators are the same in the both terms. When we put these two terms into Equation (3), the denominators cancel each other as shown in the following equation.

$$P(X_i \mid X_j) = \frac{n(X_i, X_j) \,/\, n}{n(X_j) \,/\, n} = \frac{n(X_i, X_j)}{n(X_j)} \quad (6)$$

In the resulting equation, there are only two natural frequencies. There is no need to involve the number of elements in the database for conditional probability calculations. This technique simplifies the computations in the parameter learning. Equation (6) has a closed form because if a new data case is encountered, we can easily update the corresponding natural frequencies accordingly to update the conditional probabilities. The following example provides practical results to the conditional probability calculation technique. For the cases that have not seen yet, the uniform probability distribution is used to fill the conditional probability tables in the nodes. For online Bayesian network learning, the parameter learning is not enough because the agents do not know the system dynamics in advance. Thus, the structural learning part is also necessary to discover the system dynamics.

3. The structural learning

Structural learning is finding the best network that fits the available data and is optimally complex. This can be accomplished by utilizing a search algorithm over the possible network structures. In this research, a greater importance is given to the search algorithm because we have assumed that the data will be complete. That is, each element of the database is a valid state of a variable. If there are non-applicable entries in the database, then the database is said to be incomplete.

The greedy search algorithm is employed to accomplish the structural learning in the online Bayesian network learning. The search algorithm is a score based searching algorithm. The search algorithm is evaluated in terms of the score function used and the technique used to create the candidate networks, such as adding an edge and removing an edge. The greedy search algorithm is also upgraded to have some online properties such as updating the network parameters and its structure adaptively.

The algorithm is a generic greedy search algorithm. How the arc addition is done and which scoring method is used are not specified in the above algorithm. We explore the search algorithms used in this research. In the algorithms, the arcs are added heuristically and exhaustively.

3.1 Search Algorithms

A Bayesian network is not allowed to have a cycle because of the computational difficulties. A cycle in a Bayesian network leads to a "circular reasoning" between the variables. For example, if the dependencies in above

network are: $X_1 \rightarrow X_2$, $X_2 \rightarrow X_3$, and $X_3 \rightarrow X_1$, a cycle will be formed. If evidence is entered into the variable X_1, the Bayesian network will run the evidence to X_2, then to X_3. Then, the evidence will travel to X_1 because X_1 depends on X_3. The evidence may run in the network forever because all the variables depend on each other in a circular way.

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

(1) Heuristic Search

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

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

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

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

If no structural update is necessary go to step 10.

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

another searching algorithm that eliminates the arranging the variables, namely exhaustive search.

(2) Exhaustive Search

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

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

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

The search algorithms are explained in detail. There is a need to analyze the complexity of the search algorithm before there are implemented. The following section gives the complexity analysis of both search algorithms.

(3) Complexity Analysis for Search Algorithms

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

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

Loop 1

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

Loop 2

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

$\vdots$

$\vdots$

Loop (n-1)

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

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

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

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

Then, we can further modify the equation as follows:

$$C_n = (n-1)^2+(n-3)^2+\cdots+2(n-(n-2))^2 \quad (7)$$

Since each element in C_n is less than n^2. We can state that

$$C_n < n^2(n-3) < n^3 \quad (8)$$

Equation (8) illustrates the complexity of the heuristic search. The following paragraphs will explore the complexity of the exhaustive search algorithm.

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

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

Loop 1 $n(n-1)$

Loop 2 $n(n-1)-1$

⋮

⋮

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

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

$$C_e = n(n-1)N-(1+2+\cdots+(N-1)) \quad (9)$$

$$C_e = n(n-1)N-\frac{N(N-1)}{2} \quad (10)$$

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

$$C_e = n^2(n-1)-\frac{n(n-1)}{2} = \frac{2n^2(n-1)-n(n-1)}{2} \quad (11)$$

$$C_e = \frac{(n-1)n(2n-1)}{2} \quad \text{for } n=N \quad (12)$$

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

$$C_e \cong \frac{n\cdot n\cdot 2n}{2} = \frac{2n^3}{2} \Rightarrow C_e \cong n^3 \quad (13)$$

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

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

$$C_e = n(n-1)N-\frac{N(N-1)}{2} = n(n-1)n(n-1)-\frac{n(n-1)(n(n-1)-1)}{2} \quad (14)$$

$$C_e = \frac{2n^2(n-1)^2-n^2(n-1)^2-n(n-1)}{2} = \frac{n^2(n-1)^2-n(n-1)}{2} \quad (15)$$

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

$$C_e \cong \frac{n^2\cdot n^2-n^2}{2} = \frac{n^2(n-1)^2}{2} \Rightarrow C_e \cong \frac{n^4}{2} \quad (16)$$

Two search algorithms are introduced to learn the structure of a Bayesian network in the previous sections. The heuristic search algorithm is simple and explores a limited number of network structures. On the other hand, the exhaustive search algorithm is complex and explores many possible network structures. The complexity of the exhaustive algorithm is approximately n-fold larger than the complexity of the heuristic search algorithm.

3.2 Network scoring functions

Three scoring functions are employed in this research, namely Log-Likelihood, Minimum description length (MDL), and Bayesian (BDE) scores. The Log-Likelihood method measures the likelihood of the network given the available data. The MDL also uses likelihood of the network but it includes the measure of the network's complexity. The Bayesian score involves the calculation of the probability of a network given the data. Bayesian scoring method also penalizes complex networks as the MDL scoring. If the length of the database is large enough these two methods converge to each other. The following sections provide the details of the scoring methods used in the research.

(1) Log-Likelihood Scoring

The Log-Likelihood score of a network, B, is obtained by calculating the likelihood of the data, D, given the network, B, and the network parameters, q_B. After calculating the likelihood of the data, a natural logarithm is applied to get the Log-Likelihood of the data. The following formulas explain the details of the Log-Likelihood calculation.

$$Score_L(B:D) = L(D|B,q_B) \quad (17)$$

$$L(D|B,q_B) = \prod_m P(d[m]|B,q_B) \quad (18)$$

In the above formula, $d[m]$ represents the mth data case in the database. Let us take the logarithm of the likelihood. The logarithm converts the multiplication in to a summation.

$$l(D|B,q_B) = \log L(D|B,q_B) \quad (19)$$

$$l(D|B,q_B) = \sum_m \log P(d[m]|B,q_B) \quad (20)$$

This is basically equal to calculating the probability of each data case in the database, taking their logarithms and adding them together. For example, assume that the network given in the previous section has the relations $X_1 \to X_3$ and $X_3 \to X_2$. Then, we can calculate the log-likelihood of the data with the following equation.

$$\begin{aligned} l(D|B,q_B) &= \sum_m \log P(X_1[m]|q_{X_1}) \\ &+ \sum_m \log P(X_3[m]|x_{10},q_{X_2}) + \sum_m \log P(X_3[m]|x_{11},q_{X_3}) \\ &+ \sum_m \log P(X_2[m]|x_{30},q_{X_2|x_{30}}) + \sum_m \log P(X_2[m]|x_{31},q_{X_2|x_{31}}) \end{aligned} \quad (21)$$

In the log-likelihood approach, the score of the network increases as long as the length of the database and the number of arc in the network increase. Therefore, the search algorithm tries to add as many arcs as possible to the network to get the highest scoring network. At the end of the search, the algorithm ends up with almost a complete network. For the networks with a large number of nodes, this might cause a great increase in complexity of the network. To overcome the complexity problem, we need to find out a way to include the complexity of the network to the scoring function. If the network gets complex, the scoring function should decrease accordingly. The following scoring method handles the complexity problem by introducing the complexity parameter in the scoring function.

(2) Minimum Description Length Scoring

The MDL method combines the likelihood of the data and the complexity of the network to find optimally complex and accurate networks. The MDL method penalizes networks with complex structures. The MDL has two parts, the complexity of the network, $L_{NETWORK}$, and the likelihood of the data, L_{DATA}. Then, the MDL score can be calculated by the following.

$$Score_{MDL} = L_{DATA} - L_{NETWORK} \quad (22)$$

The complexity part involves the dimension of the network, $Dim(B)$, and structural complexity of the network, $DL(B)$. The dimension of the network can be calculated using the number of states in each node, S_i. The following equation illustrates the dimension of the network.

$$Dim(B) = \sum_{i=1}^{N} (S_i - 1) \prod_{j \in pa(x_i)} S_j \quad (23)$$

Where N is the number of nodes in the network. Let M be the number of data cases in the database. Using the central limit theorem, each parameter has a variance of $\sqrt{M}$. Thus, for each parameter in the network, the number of bits required is given by the following.

$$d = \log \sqrt{M} \Rightarrow d = \frac{\log M}{2} \quad (24)$$

The structural complexity of the network depends on the number of parents of the nodes. The following formula calculates the structural complexity.

$$DL(B) = \sum_{i=1}^{N} k_i \log_2(N) \quad (25)$$

Where k_i is the number of parents the node X_i has. Finally, the following formula presents the complexity part of the MDL score by combining the dimension of the network and the structural complexity.

$$L_{NETWORK} = \frac{\log M}{2} Dim(B) + DL(B) \quad (26)$$

$$L_{NETWORK} = \frac{\log M}{2}\left[\sum_{i=1}^{N}(S_i - 1)\prod_{j \in pa(x_i)} S_j\right] + \sum_{i=1}^{N} k_i \log_2(N) \quad (27)$$

The likelihood of the data needs to be defined after presenting the network complexity part of the MDL score. The likelihood of the data given a network can be calculated by using cross-entropy. The difference between the distribution of the data (P) and the estimated distribution (Q) is from the network. Kullback-Leiber and Euclidean distance are the commonly used cross-entropy methods. Therefore, the likelihood of a data can be calculated by measuring the distance between two distributions. If we use the Kullback-Leiber cross-entropy, the likelihood of the data can be calculated by the following.

$$l(D|B,q_B) = \sum_{i=1}^{M} p_i \log \frac{p_i}{q_i} \quad (28)$$

$$L_{DATA} = \sum_{i=1}^{M} p_i \log \frac{p_i}{q_i} \quad (29)$$

Where p_i is the probability of data case i using the database and q_i is the estimate of the probability of data case i from the network parameters. If Euclidean distance measure is employed to calculate the distance between the distributions, the likelihood of the data is calculated by the following.

$$l(D|B,\hat{q}_B) = \sum_{i=1}^{M} (p_i - q_i)^2 \quad (30)$$

$$L_{DATA} = \sum_{i=1}^{M} (p_i - q_i)^2 \quad (31)$$

After defining the likelihood and complexity parts, the MDL score can be given as

$$Score_{MDL}(B:D) = l(D|B,q_B) - \frac{\log M}{2} Dim(B) - DL(B) \quad (32)$$

(3) Bayesian Scoring

Another commonly used scoring method is Bayesian score. Now, we will provide the details of the Bayesian scoring technique. Bayesian scoring is calculated by utilizing the Dirichlet parameters of the network.

Bayesian statistics tells us that we should rank a prior probability over anything we are uncertain about. In this case, we put a prior probability both over our parameters and over our structure. The Bayesian score can be evaluated as the probability of the structure given the data:

$$Score_{BDE}(B:D) = P(B|D) = \frac{P(D|B)P(B)}{P(D)} \quad (33)$$

The probability $P(D)$ is constant. Therefore, it can be ignored when comparing different structures. Thus, we can choose the model that maximizes $P(D|B)P(B)$ .Let us assume that we do not have prior over the network structures. Assume that we have uniform prior over the structures. One might ask whether we get back to the maximum likelihood score. The answer is 'no' because the maximum likelihood score for B was $P(D|B,q_B)$, i.e. the probability of the data in the most likely parameter instantiation. In Bayesian scoring, we have not given the parameters. Therefore, we have to integrate over all possible parameter vectors:

$$P(D|B) = \int P(D|q_B, B)P(q_B|B)dq_B \quad (34)$$

This is, of course, different from the maximum likelihood score. To understand the Bayesian scoring better, consider two possible structures for a two-node network, where $B_1 = [A\ B]$ and $B_2 = [A \rightarrow B]$ .Then, the probability of the data given the network structures can be calculated by the following equations.

$$P(D|B_1) = \int_0^1 P(q_A, q_B)P(D|[q_A, q_B])d[q_A, q_B] \quad (35)$$

$$P(D|B_2) = \int_0^1 P(q_A, q_{B|q_0}, q_{B|q_1})P(D|[q_A, q_{B|q_0}, q_{B|q_1}])d[q_A, q_{B|q_0}, q_{B|q_1}] \quad (36)$$

The latter is a higher dimensional integral, and its value is therefore likely to be somewhat lower. This is because there are more numbers less than 1 in the multiplication. Multiplying the numbers less than 1 results in a number smaller than any of the number in the multiplication. For example, multiplying three small numbers (less than 1)is likely to be smaller than the number obtained by multiplying two small numbers (less than 1). Since the probabilities in the integrals are less than 1, the above argument applies to the integrals. Therefore, it can be said that the higher dimensional integral is likely to have lower value that the lower dimensional integral. This idea presents preference to the networks with fewer parameters. This is an automatic control in the complexity of the network.

Let us analyze $P(D|B)$ a little more closely to understand the Bayesian score calculations. It is helpful to first consider the single parameter case even though there is no structure learning to learn there. In that case, there is a

simple closed form solution for the probability of the data given by the following.

$$P(D) = \frac{\Gamma(a)}{\Gamma(a_0 + a_1)} \cdot \frac{\Gamma(a_0 + n_0) \cdot \Gamma(a_1 + n_1)}{\Gamma(a + n)} \quad (37)$$

Where $\Gamma(m)$ is equal to $(m-1)!$ for an integer m, n is the number of data cases in the database, n_0 and n_1 are the number of zeros and ones, respectively, and $a = a_0 + a_1$. Let us assume we have 40 zeros and 60 ones in the database. Assuming that we have uniform priors, $a_0 = a_1 = 3$, the probability of data is

$$P(D) = \frac{\Gamma(6)}{\Gamma(3)\Gamma(3)} \cdot \frac{\Gamma(3+40) \cdot \Gamma(3+60)}{\Gamma(6+100)} \quad (38)$$

The probability for a structure with several parameters is simply the product of the probabilities for the individual parameters. For example, in our two-node network, if the same priors are used for all three parameters, and we have 45 zeros and 55 ones for the variable B, then, the probability of the data for the network B_1 can be calculated as

$$P(D|B_1) = \frac{\Gamma(6)}{\Gamma(3)\Gamma(3)} \cdot \frac{\Gamma(43) \cdot \Gamma(43)}{\Gamma(106)} \cdot \frac{\Gamma(6)}{\Gamma(3)\Gamma(3)} \frac{\Gamma(48) \cdot \Gamma(58)}{\Gamma(106)} \quad (39)$$

For the second network, let us assume that $a_{00} = 23$, $a_{01} = 22$, $a_{10} = 29$ and $a_{11} = 26$, where $a_{ij} = n(a_i, b_j)$ is the number of cases with $A = a_i$ and $B = b_j$. Then, we can compute the probability of the data for the network B_2 using following equation.

$$P(D|B_2) = \frac{\Gamma(6)}{\Gamma(3)\Gamma(3)} \cdot \frac{\Gamma(43) \cdot \Gamma(43)}{\Gamma(106)} \cdot \frac{\Gamma(6)}{\Gamma(3)\Gamma(3)} \frac{\Gamma(23+3) \cdot \Gamma(22+3)}{\Gamma(45+3)} \cdot \frac{\Gamma(6)}{\Gamma(3)\Gamma(3)} \frac{\Gamma(29+3) \cdot \Gamma(26+3)}{\Gamma(55+3)} \quad (40)$$

The intuition is clearer. The analysis shows that we get a higher score by multiplying a smaller number of bigger factorials rather than a larger number of small ones.

It turns out that if we approximate the log posterior probability, and ignore all terms that do not grow with M, we can obtain

$$\log P(D|B) = l(D|q_B, B) - \frac{\log M}{2} Dim(B) \quad (41)$$

i.e, as M grows large, the Bayesian score and the MDL score converge to each other using Dirichlet priors. In fact, if we use a good approximation to the Bayesian score, and eliminate all terms that do not grow with M, then we are left exactly with MDL score. Therefore, it can be concluded that the Bayesian score gives us, automatically, a tradeoff between network complexity and fit to the data. The Bayesian score is also decomposable like the MDL score since it can be expressed as a summation of terms that corresponds to individual nodes. In this research, we have decomposed the Bayesian score to make efficient calculations and a uniform distribution is employed for Dirichlet priors. The simulation results will show that the Bayesian score provides optimally complex and accurate network structures.

4 .Conclusions

Structural learning is finding the best network that fits the available data and is optimally complex. This can be accomplished by utilizing a search algorithm over the possible network structures. A greater importance is given to the search algorithm because we have assumed that the data will be complete. That is, each element of the database is a valid state of a variable. If there are non-applicable entries in the database then the database is said to be incomplete. We explore the search algorithms used in this research. In the algorithms, the arcs are added heuristically and exhaustively.We calculate the quality (score) of the networks to find the best network. In this paper, three scoring functions are employed, namely Log-Likelihood, Minimum description length (MDL), and Bayesian (BDE) scores. The Log-Likelihood method measures the likelihood of the network given the available data. The MDL also uses likelihood of the network but it includes the measure of the network's complexity. The Bayesian score involves the calculation of the probability of a network given the data.

Acknowledgments

This work is financially supported by the National Natural Science Foundation of China (Project No. 90718038). Thanks for the help.

References

[1] S. Russell and P. Norvig, Artificial Intelligence: A modern Approach, New Jersey: Prentice Hall, 1995.
[2] F. V. Jensen, an Introduction to Bayesian Networks. London, UK: University College London Press, 1996.
[3] D. Heckerman, "A tutorial on learning Bayesian networks," Technical Report MSR-TR-95-06, Microsoft Research, 1995.
[4] Y. Shoham, "Agent-oriented programming," Artificial intelligence, vol. 60(1), pp. 51-92, 1993.

[5] J. Pearl, "Bayesian networks", in M. Arbib (Ed.), Handbook of Brain Theory and Neural Networks, MTT Press, pp. 149-153, 1995
[6] J. Pearl, "Bayesian networks," Technical Report R-246, MTT Encyclopedia of the Cognitive Science, October 1997.
[7] F.V. Jensen, "Bayesian network basics," AISB Quarterly, vol. 94, pp. 9-22, 1996.
[8]W. Lam and F. Bacchus, "Learning Bayesian belief networks: an approach based on the MDL principle," Computational Intelligence, vol. 10, pp. 269-293, 1994.
[9] N. Friedman, M. Goldszmidt, D. Heckerman, and S. Russell, "Challenge: Where is the impact of the Bayesian networks in learning?" In Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI), pp.10-15, 1997.
[10] N. Friedman, K. Murphy, and S. Russell, "Learning the structure of dynamic probabilistic networks," in G.F. Cooper and S. Moral (Eds.), Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98), San Francisco, CA: Morgan Kaufmann, 1998.
[11]G. F. Cooper and E. Herskovits, "A Bayesian method for constructing Bayesian belief networks from databases," in Proceedings the Conference on Uncertainty in AI, pp.88-94, 1990.
[12] B. Theisson, C. Meek, and D. M. Chickering, and D. Heckerman, "Learning mixtures of Bayesian networks," in G.F. Cooper and S. Moral (Eds.), Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98), San Francisco, CA: Morgan Kaufmann, 1998.
[13] N. Friedman, "The Bayesian structural EM algorithm," in G.F. Cooper and S. Moral (Eds.), Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98), San Francisco, CA: Morgan Kaufmann, 1998.
[14] D. Spiegelhalter, P. Dawid, S. L. Lauritzen, and R. Cowell, "Bayesian analysis in expert systems," Statistical Science, vol. 8, pp. 219-282, 1993.
[15] D. Heckerman, D. Gieger, and M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," Technical Report MSR-TR-94-09, Microsoft Research, Redmond, WA, 1994.
[16] C. Claus, "Dynamics of multi-agent reinforcement learning in Cooperative multi-agent systems," Ph.D. Dissertation, Univ. of British Colombia, Canada, 1997.
[17] S. Sen and M. Sekaran, "Multi-agent coordination with learning classifier systems," in Proceedings of the IJCAI Workshop on Adaptation and Learning in Multi-agent Systems, Montreal, pp. 84-89, 1995.
[18] C. Boutilier, "Planning, learning and coordination in multi-agent decision processes," in Sixth conference on Theoretical Aspects of Rationality and Knowledge (TARK'96), The Netherlands, 1996

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

The suggested system for health insurance Application based on Smart Cards

M. El-Sayed Wahed[1] and Esam M. El Gohary[2]

[1] Faculty of Computers and information , Suez Canal University, Egypt

[2] Faculty of Computer and Information Systems, Mansoura University Mansoura, Egypt

Abstract

This paper concentrates on designing a system for Health insurance using smart card technology .The system is called HISS (Health insurance system using smart card).As we will see the system is web based application based on central database ,uses smart card for two reasons first, as a data carrier for patient and professionals. Second reason is for authentication purposes. There are some figures that describe system architecture and processes then each component well be explained.

1. Healthcare & Medical Service in Egypt [URL 1]

The Egyptian Government, through the Ministry of Health and Population, gives high priority to the provision of public health services. It has made substantive progress in improving the health status of its constantly increasing population (more than 70 million & is estimated to reach 100 million by year 2020). In 2001/2 the government expenditure on healthcare was LE 3.64 billion & investment was LE 2.12 billion.

Past and ongoing Public Health sector improvement projects either financed locally or by donors, have generally focused either on specific narrow sector issues and/or on targeted population groups. These have not been sufficient to face increased financial pressures resulting from the growing disease burden, population growth, and new and costly medical technologies. They also failed to sustain initial improvement in health conditions and needed to be complemented by fundamental changes in policies. It is generally considered that centralized and fragmented management and combined financing and provision have also led to inefficient use of resources. Consequently, the need for reform existed to increase accountability, transparency and efficiency by clarifying roles, separating finance from provision, and developing needed technical capacities. Public Health Sector being restructured to improve the health service and to provide all Egyptians with effective 'health security'. A pilot reform programmed has been launched.

1.1. Problems in HIO (Health insurance Organization):

HIO (Health insurance Organization) is the biggest organization that provide healthcare in Egypt .HIO serves about 25 million insured persons, it has many branches all over Egypt. Although HIO has a hierarchal structure and considered organized more than any healthcare authority, it has some problems affects service quality to insured persons and wastes lot resources, problems such as:

- Poor IT infrastructure

- Wasted time and money for professionals and patients
- Bad inventory management
- Limited primary health care services .
- Absence of real recognition of 'infection control' procedures and protocols .
- Most of the equipment in Public hospitals is obsolete or under-maintained .
- Centralized management of the Public healthcare sector .
- Public healthcare sector management is fragmented as public hospitals belong to the Ministry of Health, Ministry of Higher Education or military, as well as by some public companies to serve their employees.
- State-employed doctors and nurses suffer from difficult working conditions, low pay and lack of oversight to enforce performance standards.

The private sector has developed during the last two decades & achieved some regional recognition. It continues to grow, and is targeting high worth Egyptians and attracting international visitors.

2. Proposed System architecture

Figure 1 illustrates an overview of a typical integrated information system to manage health insurance over multiple hospitals and medical centers.

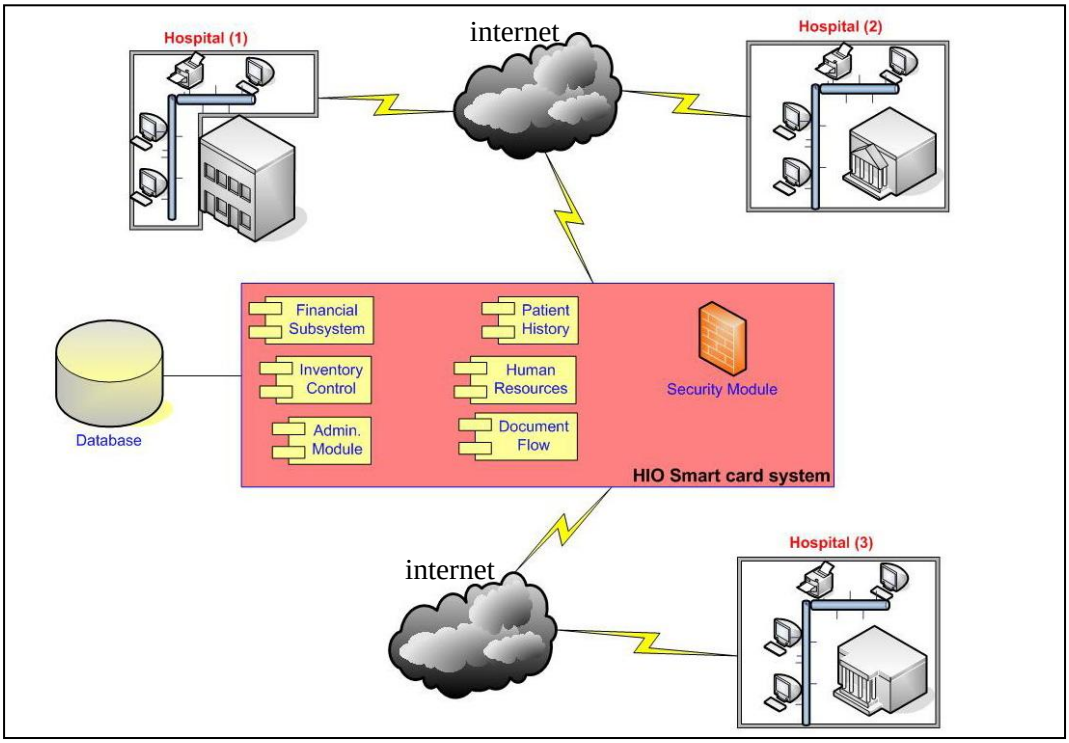


Figure 1: A common model of interoperable architecture of a web-smart-card-based information system

The proposed system is typically a web-based system as it connects all HIO locations with internet connection .The proposed system enable users to access the system from internet browsers. Each Location has a local database which replicates to the central database. At this point we have to notice that all subsystems work on the same central database located on the server farm. Also, all subsystems are integrated with each other to perform the functionality of the whole system. It is important to point that using smartcard , web

based connection which considered a client server architecture and database replication increase the proposed system reliability and availability. The system consists of the following subsystems:

- Patient Information and Case History Subsystem.
- Human Resources Subsystem.
- Document Flow Subsystem.
- Inventory Control Subsystem.
- Administration subsystem.
- Security subsystem.

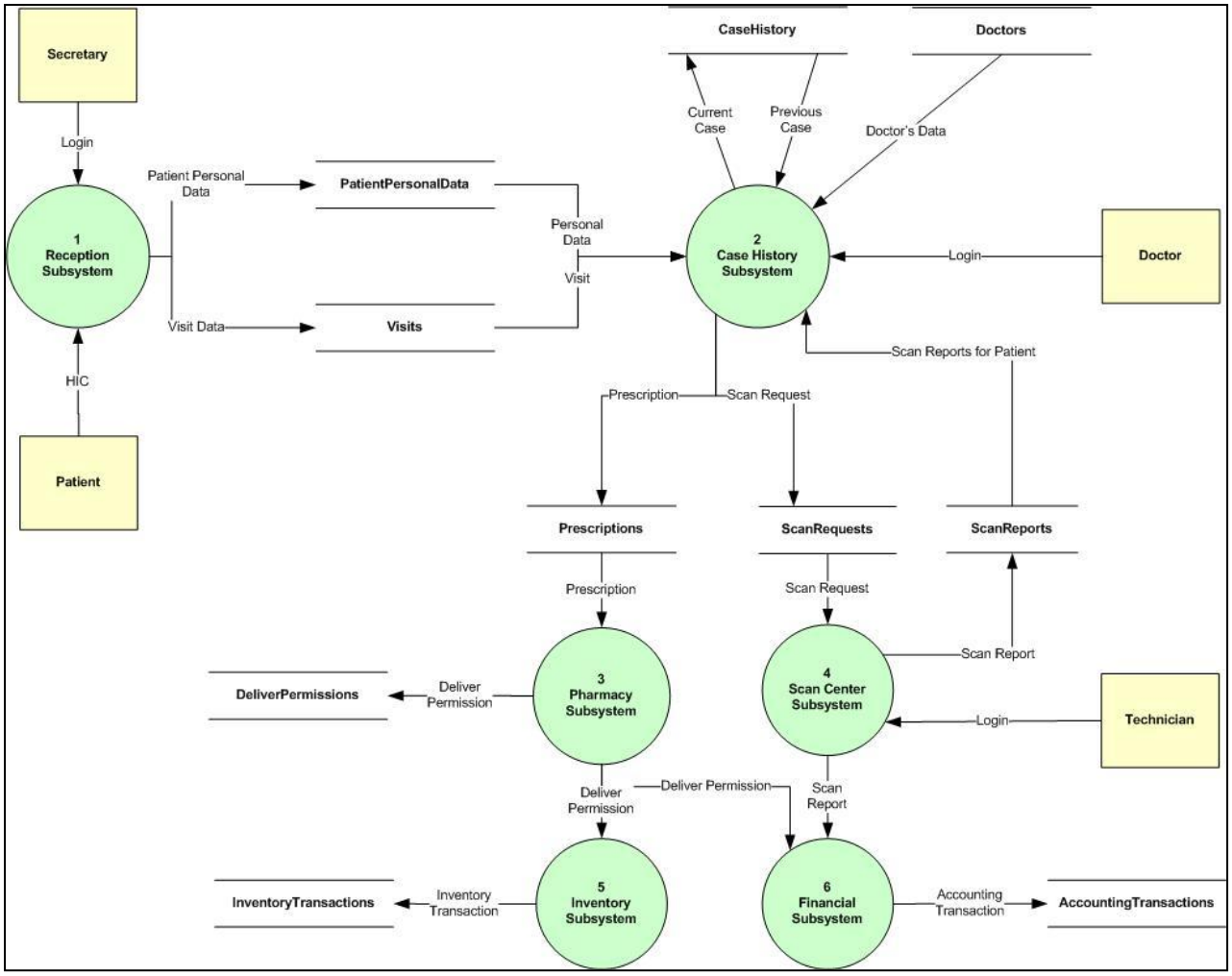
2.1. Proposed System Objectives

The proposed System has several objectives such as:

- Increase the quality of medical services to the insured and simplify the procedures involved.
- Improve communication between the HIO, healthcare users and healthcare providers.
- Improve data security and confidentiality.
- Reduce administrative tasks for the insured, healthcare providers, health insurance companies, employers, and the HIO.
- Increase operating efficiency within the HIO and the healthcare providers.
- Make the investment at a national level as cost-beneficial as possible.

2.2. Level zero data flow diagram

Figure 2 shows the major subsystems that are considered the components of the system .The figure demonstrates the main data flows between subsystems. Each subsystem details will be explained in the following subsections:



A. Patient Information and Case History Subsystem.

The responsibility of this subsystem is to manage and handle all information related to patient either personal or medical data. It creates a whole isolated profile for each patient includes all related data. As proposed; patient profile consists of patient personal data and the history of all visits and medical reports of the patient occurred in all hospitals of health insurance system. .This subsystem uses HIC (*health insured person* card) and using some features they can increase service quality, independence and autonomy in confirming administrative data. Patient can register electronically at any HIO health care facility in a manner that is fast, user friendly and reliable. Detailed features are:

1- Reading and writing personal data can be done based on HIC see example for HIC figure (3).
2- Adding, browsing and search personal data of patient.
3- Adding new patient personal data creates and initiating a new patient profile.
4- Entering basic medical information of patient like weight, height, allergies any previous surgeries.
5- Adding new page for each visit includes symptoms, diagnosis, treatment and any required investigations or scans.
6- All fields in the case history module are auto-complete fields which assists doctor for fast data entry.
7- Every recorded visit for patient preserves the user name of its doctor and can't be edited by anyone rather than the doctor who created it and system admin according to security rules.
8- Attaching files (Electronic files or scanned reports) to patient's medical profile so HIC can save only URL to these attached files to save storage area on HIC smart card.

   The card will store pointers (URL addresses) linking the card to the storage of x-rays, lab results and health records located in the databases of organizations where the services were performed.
9- Fast print of prescription details including medicines, dosage, times, required investigations or scans and also logo of the center and doctor name which all are configurable to print or not.
10- Ability to print a report of the whole visit.
11- Ability to print a report of the whole case history of the patient which may include also the case history of his/her couple.
12- View Case History of couples into coupled view, which make it very easier to review progress over both at once.
13- Getting over the problems of managing same case between many doctors as the system gives abilities to all doctors to see notes of other doctors on the case.
14- This subsystem can Decrease the amount of time taken to perform patient registration and Decrease the number of recording errors in medical records.

Patient data cards

A patient data card is a mobile data vehicle for the patient, storing up-to-date specialist health information on the patient and enabling differentiated read and write access authorization.

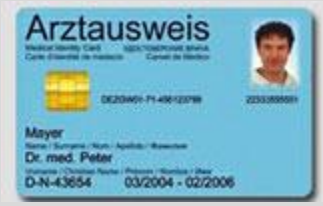
Figure 3: Example for Patient card

B. Human Resource subsystem

HR subsystem is responsible for managing employees' data, attendance and payrolls integrating with Financials subsystem to generate financial transactions of payrolls. It perform the following tasks:

1- Managing HPC (Health Professional Card) or (Health Physicians Card) allows to Enter, browse, read and write employee personal and work data. See example for HPC figure (4).
2- Supporting entering attendance times of employees using HPC smart cards.
3- Supporting loans and scheduling loans payments back.
4- Supporting insurance and other salary items.
5- Handling different weekend days for each employee and different work shifts.
6- Generating monthly salary report according to gathered data.
7- Reporting wizard for employee attendance.
8- Integration with Financial module to generate financial transactions.

Health professional cards



Health professional cards are for doctors at hospitals, in practices and in ambulances as well as for pharmacists. Authorization cards granting the health professionals individual read and/or write access rights for defined data fields on the patient data card. Only the emergency data may be read without the patient's additional permission.

Figure.4: Example for professional card

C. Document Flow Subsystem

This module is responsible for storing management documents and managing moving it forward and backward through the system. It keeps a history for each document about creation date, all processes applied to it and user who applied each process. This module work as EDI technique (Electronic Data Interchange) it helps in the following processes:

1- Adding and browsing documents.
2- Attaching document either in an electronic format (Word document, excel file …) or by scanning paper document.
3- Applying customized security levels by the creator of the document.
4- Attaching notes and related data (ex. Document type, Department, Source, Distinction …) to document.

5- Sending single document or group of related documents to another user on the system.
6- Simple and advanced search for reports and documents.

This subsystem increase in productivity among staff due to less time dealing with paperwork.

E-prescription offers high savings potential

To keep patient documents two alternatives for transfer of the electronic prescription data to pharmacies and insurers are currently under discussion. Both are equally feasible from the technical point of view. In the card-based solution, the electronic prescription is saved directly to the patient's card after the doctor has used his health professional card to authenticate himself and has signed the prescription with his digital signature. At the pharmacy, the electronic prescription is read with the aid of a card reader and the pharmacist's HPC. When the medicine has been dispensed, the prescription is deleted from the health card and stored on the pharmacist's computer. In the server-based data transfer system, the signed and encrypted electronic prescription is transferred to a server via a protected connection. Using his HPC, the doctor writes a "ticket", generated by the server, to the patient's health card. This ticket and the pharmacist's HPC allow the prescription to be retrieved from the server, after which the drug can be dispensed. The e-prescription is then deleted from the server and stored locally at the pharmacy.

At regular intervals, all e-prescriptions stored at the pharmacy are sent to the central clearing points serving the pharmacy sector. As under the present system, these clearing centers settle accounts with the health insurance organizations. According to a study carried out by Debold&Lux, introduction of the e-prescription alone will allow the German health service to save up to EUR 250 million a year .

Besides the cost savings achievable through fast data transfer, the e-prescription also offers a range of benefits for patients. The medication record stored optionally – and only with the patient's consent - on the card will let the pharmacist check for possible interaction between the prescribed drugs and any nonprescription medicines the customer may also be taking, or wish to take. The quality of the advice given in the pharmacy thus improves – especially for older people, who are not always able to remember what drugs they may be taking. Also, the registration of co-payment exemptions on the card will optimize the dispensing process. And the e-prescription, not being hand-written, is always legible – no more need to refer back to the doctor.

D. Financial Subsystem

Financial subsystem integrates with other modules in the system in order to manage all financial transactions. That’s beside managing purchases and journals.

1- Enter, edit and search purchasing bills which automatically generates its financial transactions.
2- Handling suppliers’ accounts and payments out operations to suppliers.
3- Generating supplier’s balance reports.
4- Integrates with Appointments Module to generate financial transactions of patient payments.

5- Integrates with Scan Center Module to generate financial transactions of patient payments.
6- Ability to enter petty expenses and generating its financial transactions automatically.
7- Ability to enter manual journals.
8- Report wizard of purchases in which purchases can be reported by period, supplier, item or category of items and also grouped by date, supplier or item category.
9- Generating final financial reports such as Income Statement report and balance sheet anytime of the year.

E. Inventory Subsystem

Inventory subsystem is responsible of all inventory transactions and reporting inventory amounts and transactions.

1- Enter, browse and search items of inventory.
2- Supporting multi-inventory system.
3- Ability to enter items into hierarchical tree of categories with unlimited levels.
4- Entering documents of importing items automatically generates its related inventory transactions.
5- Entering documents of exporting items to use automatically generates its related inventory transactions.
6- Alerting user when critical items amounts get to reorder level.
7- Reporting wizard of inventory transactions and inventory amounts within period of time grouped by item, date or category of items.
8- Report of items which reached reorder level.

F. Administration & Security subsystem

This subsystem gives the main tool of system administrator to manage the system and central database. It also gives him the ability to create different security levels and assign them to system users. This subsystem can do the following:

1- Creating main schedule of database backup creation.
2- Restoring database backups on any problem.
3- Creating defined security levels of the system.
4- Create, suspend system users and assign them security levels.

This subsystem also is responsible for insuring that data transfer is secure .It supports system in interacting with SSL protocol that make proposed system save.

2.3. Inside one Location

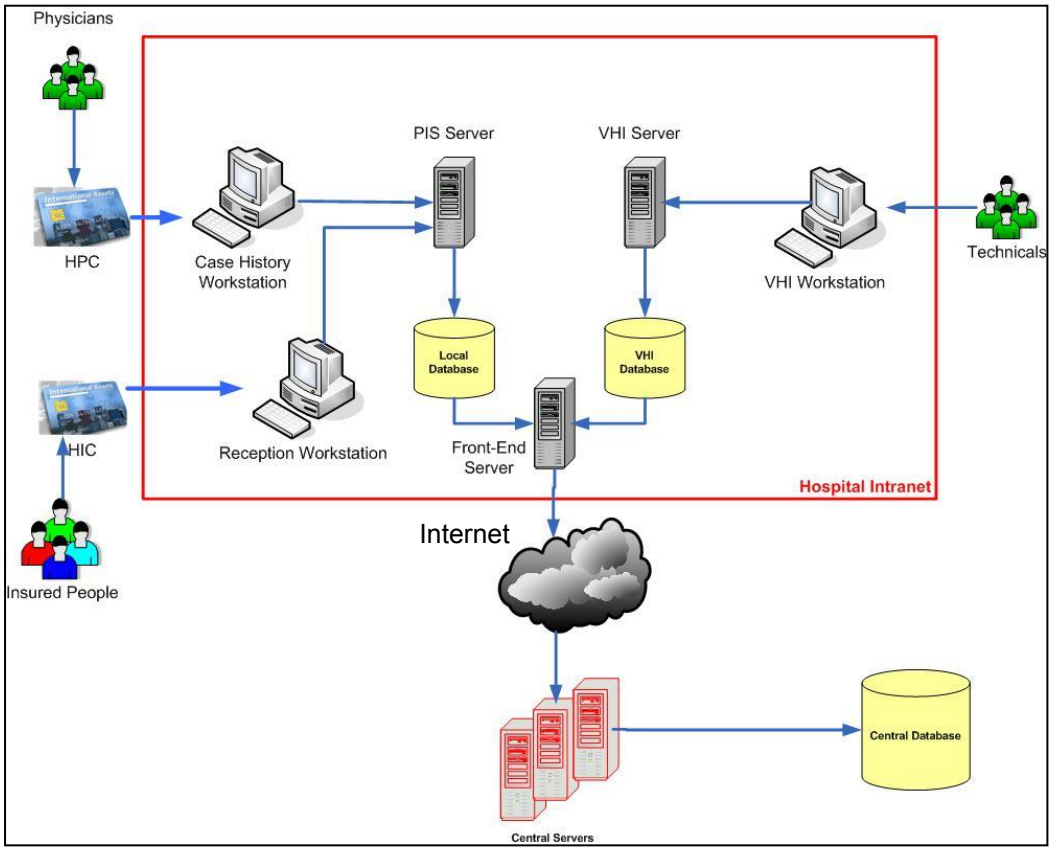


Figure 5: Subsystem in one location (single site)

Having a deeper look inside one location (hospital) to illustrate system deployment and business work flow, each location has three servers

A. PIS Server

The Patient information server (PIS) is where a local database is located. It is a database dedicated to store patient profiles locally before batching them to the main server.

B. VHI Server

The visual health information server (VHI) is where a local VHI database is located. It is a database dedicated to store high resolution health images.

C. Front-End Server

The server is responsible for sending/querying data to and from the central server. It represents the interface between local world and central servers. Front-End server is also responsible for synchronizing local database with the central database especially in cases of internet connection failure.

In such case, the Front-End server start to compare time stamp of rows in local database with the time of last batch sent to the central server. Rows which have time stamp later than last batch time are queued for batching to the central server as soon as internet connection is back.

2.4. DFD for case history subsystem

It may also called (patient information subsystem), Figure (6) illustrate Level 1 data flow diagram for patient information subsystem .the figure demonstrates the main data flows between processes.

2.5. ERD for case history subsystem

Figure(7) show main entities in case history subsystem and relations between them.

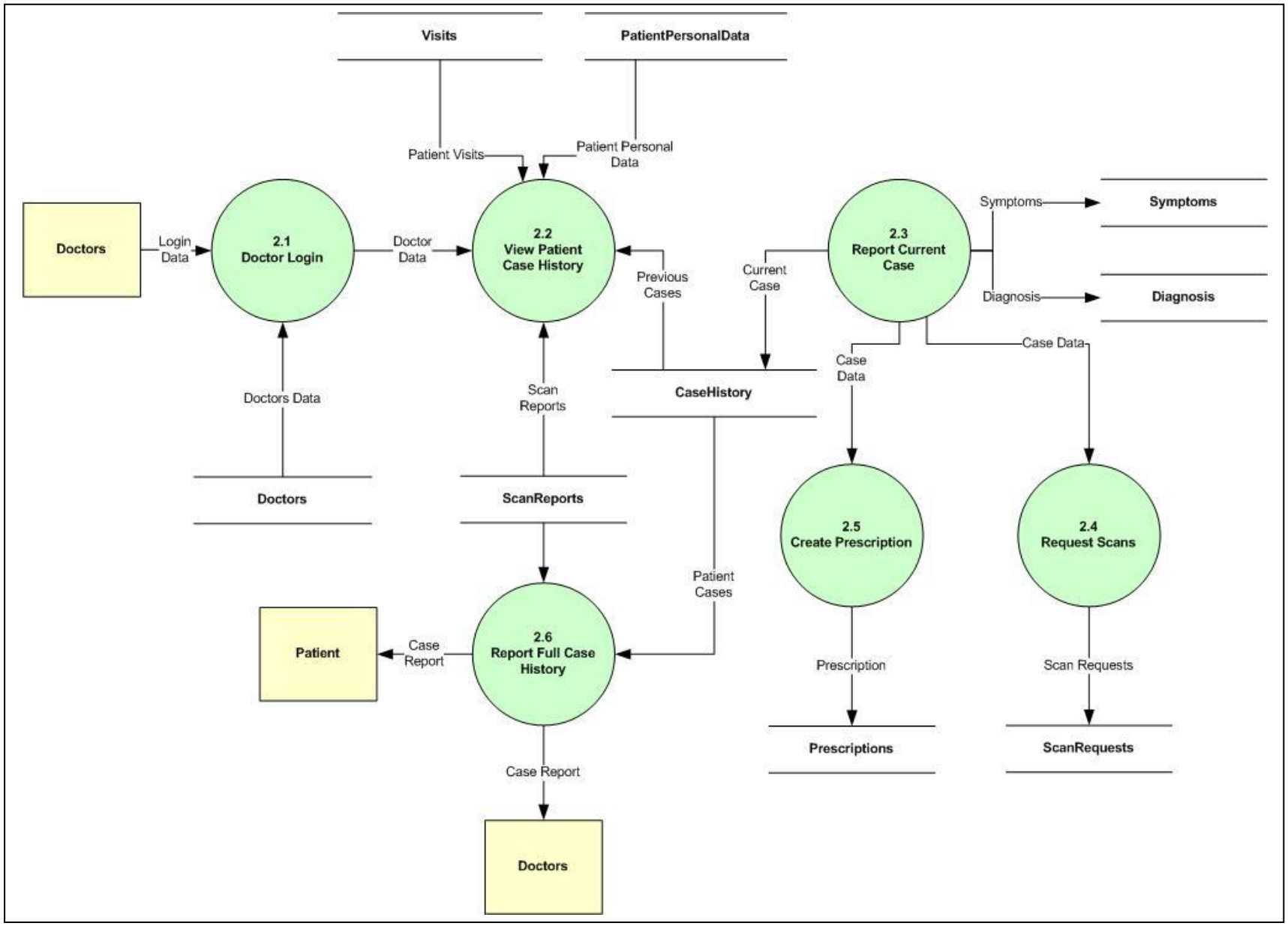


Figure.6: DFD level 1 for process 2 PIS

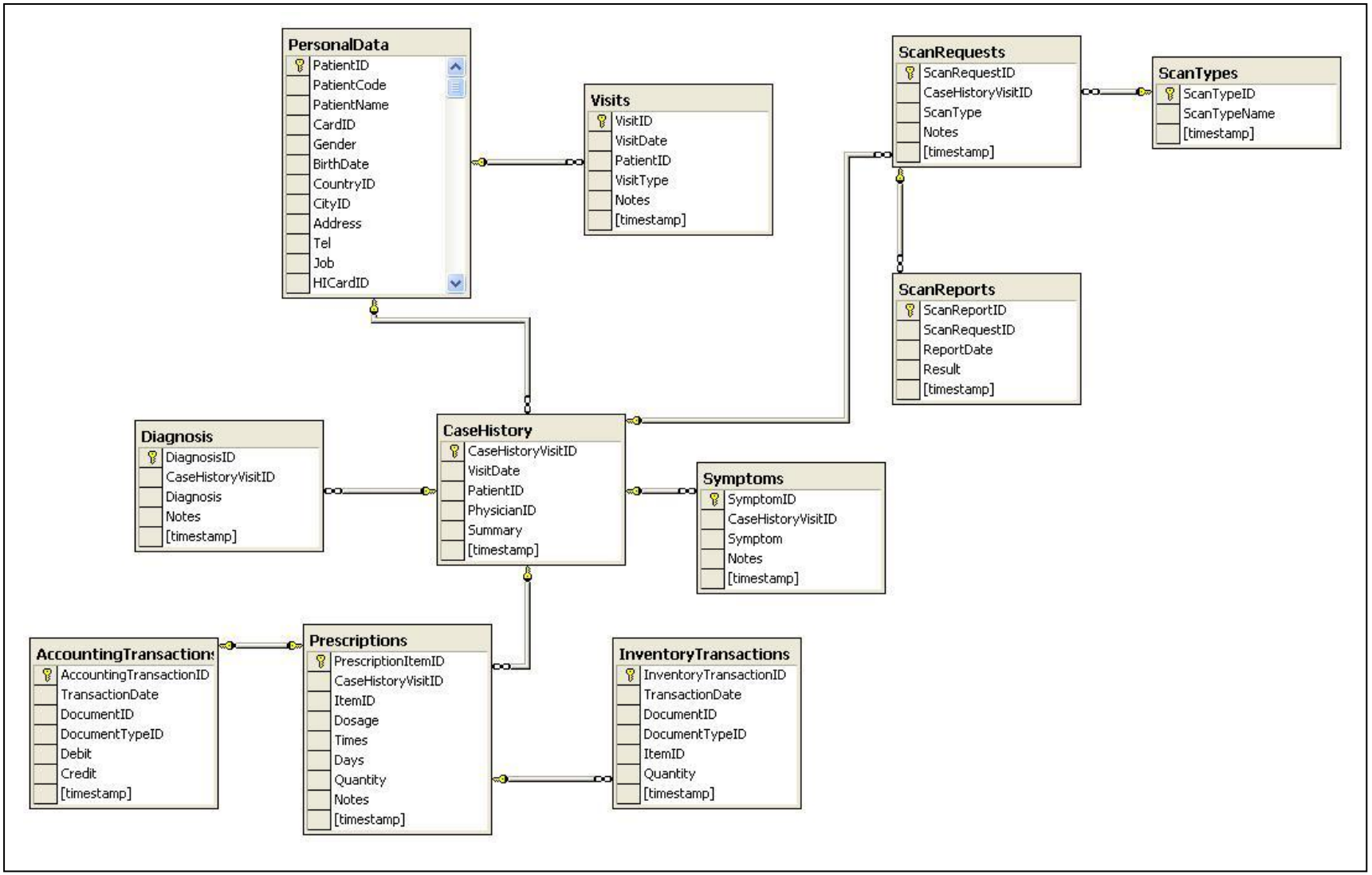


Figure. 7: ERD for PIS

2.6. Data Flow and synchronization of card data

The confirmation of health insurance and updating of other data is carried out through the network, which links the self-service terminals to the central compulsory and voluntary health insurance databases.

The current data are downloaded to the card memory. When presenting the card at the doctor's, the insured person thus transfers current data to the health care and health insurance officers, holders of the health professional

The health care worker inserts his/her health professional card into a special purpose device, a card reader, and logs in with his/her personal password. When the patient enters the doctor's office, his card is inserted in the second slot of the same reader. The card reader enables the health professional card holder to read and modify card data in his/her competencies. The data are displayed on the local computer screen and, as required, saved in the local databases.

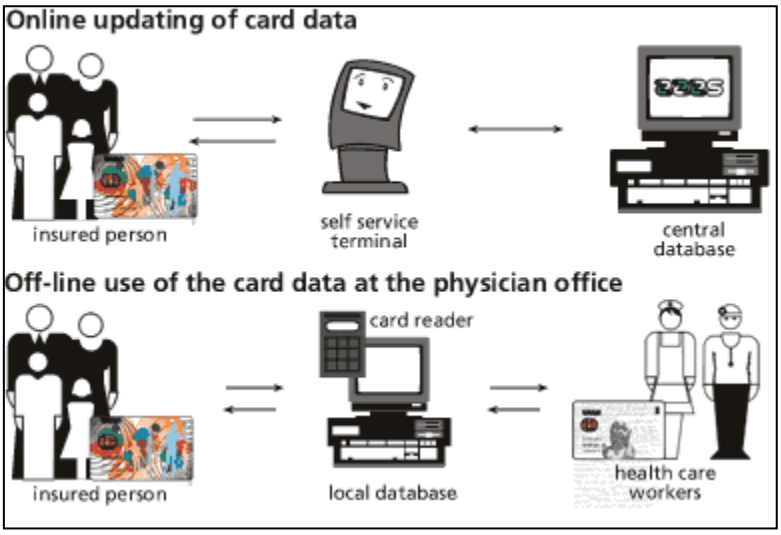


Figure 8: Online and offline data updating

The system grantee that if there is no internet connection between local site and central database, data will be saved in local site and when connection established automatic update will be made.

3. Components of a Smart Card System [1]

The configuration of the smart card platform will vary substantially from project to project depending upon the card management approach, card personalization and issuance procedures, card capabilities and applications, and technical environment selected by the project. However, the following components will typically comprise a smart identification card platform see fig. (9)

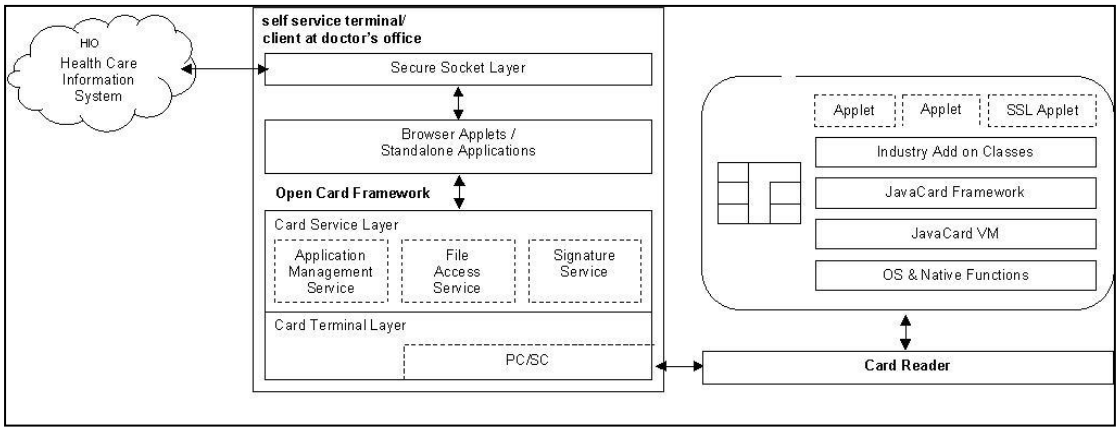


Figure 9: Card platform

3.1. Data Types on the Patient Data Card

The administrative part of the information is stored on the card by the health-insurance funds. The medical data is voluntary and is inserted on the card by health professionals or the patient .

- **Constant Data :** Constant data comprises administrative data such as insurance and address data as well as emergency information. It may, for example, also include information on chronic illnesses or organ-donor information .
- **Transfer Data :** Transfer data refers to information such as electronic prescriptions, referrals and admission requests .
- **Dynamic Data :** Dynamic data is the data which is modified in the course of the patient's life, e.g. the medication, treatment and health history. Depending on the scope of the data or the system architecture preferred, the card can be used as data carrier or as a pointer to data and as an access key.
- **Cards :** Smart cards contain an ICC that provides computational power similar to that of a PC. Smart cards have the capability to implement multiple authentication technologies such as PKI and biometrics. They also have a certain amount of storage capability. Smart cards are generally used for both physical and logical access, and are available with both contact and contactless interfaces.
- **Central Card Management System**. The central card management system should function as the core of the smart card system, and as such, requires connectivity and interfaces with all other system components. It houses the central cardholder database that supports the capture, storage, retrieval, retention, integrity, and management of data necessary for the Life Cycle Management (LCM) of smart cards. LCM includes: pre-issuance, issuance, status, replacement, renewal, post-issuance capabilities and audit of smart cards for each agency.
- **Smart Card Equipment and Software.** Smart card equipment and software includes the computers, peripherals, and software needed to capture the information used to enroll a cardholder, personalize the card, load the card with

any necessary PKI certificates, issue the card to the cardholder, and perform post-issuance capabilities such as PIN reset and certificate updates on the card. Card issuance equipment typically includes:

- **Enrollment Workstation.** The enrollment workstation is used to capture enrollment information and route it to the central card management system and to the equipment actually personalizing and issuing the cards (if not the enrollment workstation itself). At agency discretion, attachments to the enrollment workstation may include a digital video camera to capture the cardholder's digitized photo, a digitized signature capture device, a biometric capture device (most commonly a fingerprint capture device but could include a wide variety of biometric capture devices), and a key pad used for generating a user's PIN. Depending on the procedures for capturing demographic data (e.g., through manual entry or legacy system upload), the enrollment workstation may be used to collect demographic data for card personalization. In some implementations, the biometric data and/or public keys captured through the enrollment workstation could be directly routed to the certificate/attribute authority workstation as part of a certificate request.
- **Key Generation Workstation**. Although key pairs generally will be generated by a cryptoprocessor on the smart card, some agencies may choose to use a separate workstation to generate keys (i.e., using software-generated keys rather than token-generated keys). Once keys have been generated, they are securely transmitted (using mutual authentication protocols and encryption (symmetric or asymmetric)) and loaded onto the card at the point of card personalization and issuance.
- **Card Personalization System.** The card personalization system is used to personalize the card with data, photos, key pairs (if not generated on the card itself), and digital or attribute (i.e., biometric) certificates. Attached to the card personalization workstation is a card reader that is used to load information to the chip on the card and a card printer that is used to print information and photos on the face of the card. In some scenarios, the card personalization workstation and enrollment workstation may be the same device, depending on whether a centralized (i.e., bulk personalization) or decentralized (i.e., on-site issuance) process is used for card personalization and issuance.
- **Registration Authority System.** In some scenarios, if an agency has a designated registration authority, a separate workstation may be used to read public keys from the card (or verify biometric data), document identity proofing, and generate a digital certificate (or attribute certificate) request. In turn, the registration authority system may receive signed certificates from the certificate authority (or attribute authority) and place them on the card. The registration authority workstation could be the same as the enrollment workstation and the card personalization system in an on-site card issuance location.
- **Certificate/Attribute Authority System.** The certificate and/or attribute authority

system is a trusted computer system that receives certificate requests (that would contain public keys and data or a biometric template) from the entity acting as a registration authority, and, in turn, signs and issues certificates that are returned to the registration authority (or enrollment workstation/card personalization system) for loading onto cards. The certificate or attribute authorities typically will maintain their own repositories (i.e., Lightweight Directory Access Protocol (LDAP) servers) that are used to publish certificates.

- **Card Reader.** A card reader is used to communicate with the smart card during a transaction. It is the interface between the card and the host system. Card readers provide power and timing to the ICC and can operate with either contact or contactless interfaces.
- **Applications**. Smart cards used to implement physical and logical access control applications, as well as other applications that are components of an agency's card system. Depending on the card management approach, these applications may communicate with the central card management platform to upload back-up transactions and/or to download hot lists.
- **Interfaces to Legacy Databases**. Many agencies will choose to personalize their smart cards with data from existing legacy systems. Thus, important components of the platform architecture are the interfaces from legacy systems to the central cardholder database or to the card issuance workstation.

3.2. Smart card Operating System [URL2]

Every smart card has an operating system. It is the hardware-specific firmware that provides basic functionality as secure access to on-card storage, authentication and encryption. Only a few cards allow writing programs that are loaded onto the smart card - just like programs on a computer. This is a great way to extend the basic functionality of the smart card OS. Some of popular operating system:

COS OS

The smart card's Chip Operating System (frequently referred to simply as COS; and sometimes referred to as the Mask) is a sequence of instructions, permanently embedded in the ROM of the smart card. Like the familiar PC DOS or Windows Operating System, COS instructions are not dependent on any particular application, but are frequently used by most applications. Chip Operating Systems are divided into two families:

- The general purpose COS which features a generic command set in which the various sequences cover most applications, and
- The dedicated COS with commands designed for specific applications and which can even contain the application itself. An example of a dedicated COS would be a card designed to specifically support an electronic purse application.

The baseline functions of the COS which are common across all smart card products include:

- Management of interchanges between the card and the outside world, primarily in terms of the interchange protocol.
- Management of the files and data held in memory.

- Access control to information and functions (for example, select file, read, write, and update data).
- Management of card security and the cryptographic algorithm procedures.
- Maintaining reliability, particularly in terms of data consistency, sequence interrupts, and recovering from an error.
- Management of various phases of the card's life cycle (that is, microchip fabrication, personalization, active life, and end of life).

In most cases the issuer has to commit to a specific application developer, operating system and chip for each service the issuer wished to provide to its customer base. This leaves almost no flexibility to change any of these components without having to invest funds into a new software and/or hardware implementation. As a result early smart cards were costly and inflexible. But today we can clearly see a development towards open operating systems that support multiple applications.

For on-card application development of programs that run inside the secure environment of the smart card chip, we highly recommend operating systems that have bigger market exposure such as JavaCard OS, MultOS and lately Windows for smart cards.

Multi Application Card Operating Systems (MACOS)

Until the emergence of multi-application smart cards, each software application representing a product or service on a card was written for a specific operating system, which in turn was specific to a particular hardware (chip) or silicon platform supplier.

Multi-application operating systems allow the development of multiple applications that run on one card. Ideally the on-card applications can't interfere with each other and are protected by a firewall. Currently there are three major operating systems on the market

Java Card is a multi-application

JavaCard is a multi-application operating system for smart cards. JavaCard is an open, multi-application operating system for smart cards. Diverse parties can develop applications using their Java programming skills. The resulting applets run on the same card and they all co-reside independently. This way applications from various vendors can be combined, all separated from each other.

Until the emergence of multi-application smart cards, each software application representing a product or service on a card was written for a card specific operating system, which in turn was particular to a hardware (chip) or silicon platform supplier. In most cases there wasn't even an operating system between the hardware layer and the card edge.

From a card issuer perspective, an issuer had to commit to a specific application developer, operating system and chip for each service the issuer wished to provide to its customer base. The issuer had almost no flexibility to change any of these components without having to invest funds into a new software and/or hardware implementation. Early smart cards were therefore costly and inflexible.

From a consumer perspective, cardholders were forced to carry a different card for each service or function they wished to benefit from. If the product or service they benefited from changed in any way, they would receive a replacement card.

JavaCard has changed the smart card proposition for both issuers and cardholders. Java cards provide increased convenience and flexibility for users while delivering savings and a wealth of opportunities for issuers across all business sectors.

Application Load & Unload in JavaCard OS.

JavaCard allows applications to be loaded on-the-fly. This means that a card with the JavaCard operating system on it can change features during its lifetime. For example a student who has been issued a smart card with JavaCard on it, can load applications (java applets) over the Internet. Of course this would require the correct authorization. But the interesting part is that this can happen securely over insecure networks. This way the student can change the set of available applications over the smartcard's lifetime. One day it could contain an electronic purse and a metro travel application. The next day the student will add an electronic key to get logical access the university network. This is extremely beneficial for both, the cardholder and the card issuer

MULTOS

MULTOS is a multi-application operating system for smart cards for highest security needs. MULTOS is the first, open, high security, multi-application operating system for smart cards (hence 'MULT-OS'). The beauty of this system is that diverse parties can develop applications that are running on the same card and they all co-reside both independently and securely. This way applications from various vendors can be combined, all securely separated from each other.

The open nature of the MULTOS platform allows anyone to issue cards, write applications, implement the operating system on a specific chip, manufacture smart cards or provide value added products which support MULTOS.

From a card issuer perspective, an issuer had to commit to a specific application developer, operating system and chip for each service the issuer wished to provide to its customer base. The issuer had almost no flexibility to change any of these components without having to invest funds into a new software and/or hardware implementation. Early smart cards were therefore costly and inflexible.

From a consumer perspective, cardholders were forced to carry a different card for each service or function they wished to benefit from. If the product or service they benefited from changed in any way, they would receive a replacement card.

As the leading high security, multi-application operating system, MULTOS has changed the smart card proposition for both issuers and cardholders. MULTOS provides increased convenience and flexibility for users while delivering savings and a wealth of opportunities for issuers across all business sectors.

3.3. System benefits

Major system benefits may be listed as follows:

Insured persons: benefits in terms of service quality, independence and autonomy in confirming administrative data;

health professionals: reduced administrative and paper work, more time available for quality professional tasks, electronic transfer of data onto the existing forms, electronic

linking to different expert information systems;

Employers: total elimination of the issuing and confirmation of the health care identification booklets, reduced administrative tasks;

Health insurance providers: improved currency and accuracy of data, rationalized data flows, reduced administrative tasks, improved quality of services, support to analyses and timely and appropriate implementation of measures of proper fund allocation.

Society: simpler and better transfer of data between the partners in the health care system, improved personal data security, transparency in the field of financial liabilities among different cooperating subjects, positive long-term national scale economic effects due to optimization of operation of the health care sector.

4. Conclusion

The movement to e-government, at its heart, is changing the way people and businesses interact with government. E-Government offers a huge potential in seeking innovative way to reach the ideal of government of people, by people and for people.

Although there are a number of applications that can benefit from smart card technology, the main driving force for using smart cards in medical science has been identifying the patient and critical information about the patient such as allergies, the medication that the patient is on, blood type and other information that can help doctors and other medical personnel apply proper care to the patient without suffering long delays. The portability and security provided by smart cards make them appealing to health organizations. When coupled with secure web sites holding detailed data about the history of a patient, medical smart cards seem to be the smart choice for solving our hard-rooted problems.

Now after designing smartcard-web-based system it is recommended to apply it for HIO in order to: first, to overcome its problems, second to make this sector as a part of Egyptian e-government, third for the following:

- Decrease in the amount of time taken to perform patient registration.
- Decrease in the amount of time taken to access a patient's medical records.
- Decrease in the number of recording errors in medical records.
- Increase in productivity among staff due to less time dealing with paperwork.
- Decrease in the amount of time that patients have to wait for service.
- Increase in successful life saving medical encounters due to the availability of emergency data.
- Using web-based system and smartcard technology achieve reliability and availability

The new system discussed in the paper is cost efficient improvement in patient care because it:

a. Provides pertinent patient information for emergency use that:
   - Gives patient demographics.
   - Describes the patient's current medical condition(s).
   - Lists the patient's current medications.
   - Alerts health care professionals to allergies and other need to know conditions.
   - Enables effective contact with the patient's primary physician.

b. Allows the patient to quickly and securely access medical records in the patient's primary database, using a web site on the Internet.

c. Information being updated to HIIS, can also be updated to the patient's smart card concurrently, or added to a patient's smart card at another time.

Also Security architectures proposed using SSL protocols and user authentication within the system allow for increased patient medical data confidentiality by providing access only to entitled professionals. Providing better, faster, and secure access to patient clinical information, smart cards sit at the heart of the qualitative evolution of medicine.

References

1- Bill Holcombe ,GOVERNMENT SMART CARD HANDBOOK ,Office of Governmentwide Policy,February 2004

2- [URL 1] http://www.bi-me.com/main.php?id=29&t=1 last seen june-2012

3- [URL 2] http://www.cardwerk.com/smartcards/smartcard_operatingsystems.aspx last seen jan-2012

Synchronization Criteria of Chaos Systems with Time-delay Feedback Control

Chao Ge[1], Hong Wang[2]

[1]College of Information Engineering, Hebei United University, Tangshan, Hebei 063009, PR China

[2]College of Qing Gong, Hebei United University, Tangshan, Hebei 063009, PR China

Abstract

For the time-delay feedback control of chaos synchronization problem, an idea of Lyapunov functional with time-delay decomposition is presented. Some delay-dependent synchronization criteria are formulated in the form of matrix inequalities. The controller gain with maximum allowed time-delay can be achieved by solving a set of linear matrix inequalities (LMIs). A simulation example is given to illustrate the effectiveness of the design method.

Keywords: synchronization, chaos system, time-delay feedback, delay decomposition, linear matrix inequality (LMI).

1. Introduction

During the last two decades, chaotic synchronization has received considerable interests, e.g., [1–3] and references cited therein. It is found to be useful or has great potential in a variety of fields including physical, chemical and ecological systems, human heartbeat regulation, secure communications, and so on.

Recently, the effect of delay on synchronization between two chaotic systems has been reported in many literatures due to the propagation delay frequently encountered in remote master–slave synchronization scheme. In particular, some delay-independent[4] and delay-dependent synchronization criteria was derived in [5,6]. Liao and Chen in [7] improved some results in [6] and gave some simple algebraic conditions which are easy to be verified. In [8] Cao et al. further generalized and improved the results in [6,7]. However, when deriving delay-dependent sufficient conditions for master-slave synchronization, Yalcin et al. [6] and Cao et al. [8] employed model transformation, which led to some conservative synchronization criteria for inducing additional terms. In order to avoid using model transformation, some new approaches had been employed to derive much less conservative synchronization conditions. Xiang et al. [9] and He et al. [10] used integral inequality and free weighting matrix approach in the derivative of Lyapunov functional respectively. It is interesting and valuable issue to proposed new method to obtain a larger delay threshold below which synchronization can be ensured theoretically.

In this paper, we employ a delay decomposition approach recently proposed in [11,12] and fully use information from the nonlinear term of the error system to derive the synchronization criteria. Based on the synchronization criteria, we will give some sufficient conditions on the existence of a state error feedback controller. These sufficient conditions will be formulated in the form of matrix inequalities. Moreover, we will design the controller by solving a set of LMIs. We will use one simulation example to illustrate the effectiveness of synchronization criteria and the design method.

2. Problem statement

Consider a general master–slave synchronization scheme using time-delay feedback control.

$$M:\begin{cases}\dot{x}(t)=Ax(t)+B\varphi(Cx(t))\\ p(t)=x(t)\end{cases} \tag{1}$$

$$S:\begin{cases}\dot{y}(t)=Ay(t)+B\varphi(Cy(t))+u(t)\\ q(t)=y(t)\end{cases} \tag{2}$$

$$C:u(t)=K(p(t-\tau)-q(t-\tau)) \tag{3}$$

With master system M, slave system S and controller C, where the time-delay $\tau>0$. The master and slave system are chaos systems with state vectors $x, y\in R^n$, and the output vector $p, q\in R^l$, respectively. The matrices $A\in R^{n\times n}$, $B\in R^{n\times m}$, $C\in R^{m\times n}$, $H\in R^{l\times n}$ are known constant matrices. The nonlinearity $\varphi(\cdot)$ is time-invariant, decoupled, and satisfies a sector condition with $\varphi_i(\sigma)(i=1,2,....,m)$ belonging to a sector $[0,k]$, i.e.,

$$\varphi_i(\sigma)(\varphi_i(\sigma)-k\sigma)\le 0 \quad \forall t\ge 0 \quad \forall \sigma\in R \quad (4)$$

Now, define the synchronization error as $e(t)=x(t)-y(t)$. Then, an error dynamical system is obtained in the form:

$$\dot{e}(t)=Ae(t)+B\eta(Ce(t),y(t))-Ke(t-\tau) \quad (5)$$

where $\eta(Ce,y)=\varphi(Ce+Cy)-\varphi(Cy)$.

Let $C=[c_1,c_2,...c_m]^T$ with $c_i\in R^n, i=1,2,...m$ .The nonlinearity $\eta(Ce,y)$ is assumed to belong to the sector $[0,k]$, i.e., for $\forall t\ge 0$, $\forall e,y$

$$\eta_i(c_ie,y)(\eta_i(c_ie,y)-kc_ie)\le 0 \quad (6)$$

The purpose of this paper is to study the master-slave synchronization of chaos systems and design the controller (3), i.e., to find the controller gain K, such that the system described by (5)-(6) is globally asymptotically stable, which means that the master system and the slave system synchronizes.

The following lemma is useful in deriving synchronization criteria.

Lemma1.(Ding [3]) For any constant matrix $R>0$, $R=R^T\in R^{n\times n}$,scalar $\tau>0$,and vector function e and $\dot{e}:[0,\tau]\to R^n$ such that the following inequality is well defined, then

$$-\tau\int_0^\tau \dot{e}^T(s)R\dot{e}(s)ds\le\begin{pmatrix}e(t)\\e(t-\tau)\end{pmatrix}^T\begin{pmatrix}-R & R\\ R & -R\end{pmatrix}\begin{pmatrix}e(t)\\e(t-\tau)\end{pmatrix}$$

3. Main results

Then we are in the position to give the main result.

Theorem 3.1. For a given scalar $\tau>0$,the error system (5) is globally asymptotically stable if there exist matrices $P=P^T>0$, $Q_i=Q_i^T>0$, $R_i=R_i^T>0 (i=1,2,...N)$, and positive diagonal matrices $\Lambda=diag(\lambda_1,\lambda_2,...,\lambda_m)>0$, and any matrices G_1,G_2 such that

$$\Xi=\begin{pmatrix}
\Xi_{11} & R_1 & 0 & \cdots & 0 & -G_1K & \Xi_{1N+2} & G_1B+kC^T\Lambda & 0\\
* & \Xi_{22} & R_2 & \cdots & 0 & 0 & 0 & 0 & 0\\
* & * & \Xi_{33} & \cdots & 0 & 0 & 0 & 0 & 0\\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots\\
* & * & * & \cdots & \Xi_{NN} & R_N & 0 & 0 & 0\\
* & * & * & \cdots & * & \Xi_{N+1N+1} & -K^TG_2^T & 0 & 0\\
* & * & * & \cdots & * & * & -G_2-G_2^T & G_2B & h\left(\sum_{i=1}^N R_i\right)\\
* & * & * & \cdots & * & * & * & -2\Lambda & 0\\
* & * & * & \cdots & * & * & * & * & -\sum_{i=1}^N R_i
\end{pmatrix}<0 \quad (7)$$

where

$\Xi_{11}=A^TG_1+G_1^TA+Q_1-R_1$, $\Xi_{ii}=-R_{i-1}-R_i+Q_i-Q_{i-1}$,

$\Xi_{N+1N+1}=-Q_N-R_N$, $\Xi_{1N+2}=-G_1^T+A^TG_2+P$

Proof. Consider the following Lyapunov-Krasovskii function candidate for system (5) as :

$$V(e(t))=V_1(e(t))+V_2(e(t))+V_3(e(t)) \quad (8)$$

with

$$V_1(e(t))=e^TPe+2\sum_{i=1}^m\int_0^{c_i^Te}\lambda_i\varphi_i(s)ds$$

$$V_2(e(t))=\sum_{i=1}^N\int_{t-ih}^{t-(i-1)h}e^T(s)Q_ie(s)ds$$

$$V_3(e(t))=\sum_{i=1}^N\int_{-ih}^{-(i-1)h}\int_{t+\theta}^t\dot{e}^T(s)hR_i\dot{e}(s)ds\,d\theta$$

where $h=\tau/N$, N is the positive integer of division on the interval $[-\tau,0]$ and h is the length of each division.

According to (5), for any appropriately dimensioned matrices G_1,G_2, the following equations are true:

$$\begin{aligned}&2[e^T(t)\quad \dot{e}^T(t)\ ][G_1\quad G_2]^T\times[-\dot{e}(t)+Ae(t)\\&+B\eta(Ce(t),y(t))-Ke(t-\tau)\ ]=0\end{aligned} \quad (9)$$

From (4), (5) and Ti=diag $(t_{i1},t_{i2},...,t_{im})>0$(i=1,2), we have

$$-2\eta^T\Lambda\eta+2ke^TC^T\Lambda\eta\ge 0 \quad (10)$$

Taking the derivative of $V(e(t))$ with respect to t along the trajectory of (5) yields

$$\dot{V}_1(e(t))=2e^TP\dot{e} \quad (11)$$

$$\begin{aligned}\dot{V}_2(e(t))&=\sum_{i=1}^N e^T(t-(i-1)h)Q_ie(t-(i-1)h)\\&-\sum_{i=1}^N e^T(t-ih)Q_ie(t-ih)\end{aligned} \quad (12)$$

$$\begin{aligned}\dot{V}_3(e(t))&=\sum_{i=1}^N h^2\dot{e}^T(t)R_i\dot{e}(t)\\&-\sum_{i=1}^N h\int_{t-ih}^{t-(i-1)h}\dot{e}^T(s)R_i\dot{e}(s)ds\end{aligned} \quad (13)$$

Adding the left side of (9)-(10) to the right side of $\dot{V}(e(t))$, and using Lemma 1we have

$\dot{V}(e(t))\le q^T(t)\hat{\Xi}q(t)$

Where

$q(t)=[e^T(t)\ \ e^T(t-h)\ \ e^T(t-2h)\ \ .....$

$e^T(t-(N-1)h)\ \ e^T(t-Nh)\ \ \dot{e}^T(t)\ \ \eta^T(t)]$

$$\hat{\Xi}=\begin{pmatrix}\Xi_{11} & R_1 & 0 & \cdots & 0 & -G_1K & \Xi_{1N+2} & G_1B+kC^T\Lambda\\ * & \Xi_{22} & R_2 & \cdots & 0 & 0 & 0 & 0\\ * & * & \Xi_{33} & \cdots & 0 & 0 & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots\\ * & * & * & \cdots & \Xi_{NN} & R_N & 0 & 0\\ * & * & * & \cdots & * & \Xi_{N+1N+1} & -K^TG_2^T & 0\\ * & * & * & \cdots & * & * & -G_2-G_2^T+h^2\left(\sum_{i=1}^{N}R_i\right) & G_2B\\ * & * & * & \cdots & * & * & * & -2\Lambda\end{pmatrix}<0$$

It follows from Schur complement that $\hat{\Xi}<0$ is equivalent to (7), then $\dot{V}(e(t))\le q^T(t)\Xi q(t)<0$ for $q(t)\neq 0$, which means that the system described by (5)-(6) is globally asymptotically stable. This completes the proof.

Remark 3.2: In order to reduce the conservative, the delay-decomposition is proposed in the Lyapunov functional. Therefore, the delay-dependent stability criterion is expected to be less conservative than the existing ones, which will be illustrated through an example in the next section.

In order to get the controller gain we let $G_2=\mu G_1, Y=-G_1^TK$,then we can establish the following synchronization criterion.

Corollary 3.3. For a given scalar $h>0$,the error system (5) is globally asymptotically stable if there exist matrices $P=P^T>0$, $Q_i=Q_i^T>0$, $R_i=R_i^T>0(i=1,2,...N)$, and positive diagonal matrices $\Lambda=diag(\lambda_1,\lambda_2,...,\lambda_m)>0$, and any matrices Y,G_1 such that

$$\Xi=\begin{pmatrix}\Xi_{11} & R_1 & 0 & \cdots & 0 & Y & \hat{\Xi}_{1N+2} & G_1B+kC^T\Lambda & 0\\ * & \Xi_{22} & R_2 & \cdots & 0 & 0 & 0 & 0 & 0\\ * & * & \Xi_{33} & \cdots & 0 & 0 & 0 & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots\\ * & * & * & \cdots & \Xi_{NN} & R_N & 0 & 0 & 0\\ * & * & * & \cdots & * & \Xi_{N+1N+1} & Y^T & 0 & 0\\ * & * & * & \cdots & * & * & -uG_1-uG_1^T & G_1B & h\left(\sum_{i=1}^{N}R_i\right)\\ * & * & * & \cdots & * & * & * & -2\Lambda & 0\\ * & * & * & \cdots & * & * & * & * & -\sum_{i=1}^{N}R_i\end{pmatrix}<0 \quad (14)$$

Where $\hat{\Xi}_{1N+2}=-G_1+uA^TG_1^T+P$

Moreover, the delay feedback controller gain matrix is given by $K=-G_1^{-T}Y$. This completes the proof.

4. An example

Consider the following Chua's circuit

$$\begin{cases}\dot{x}=\alpha(y-h(x)),\\ \dot{y}=x-y+z,\\ \dot{z}=-\beta y,\end{cases}$$

with nonlinear characteristic:

$h(x)=m_1x+0.5(m_0-m_1)(|x+c|-|x-c|)$,

and parameters $m_0=-1/7$, $m_1=2/7$, $\alpha=9$, $\beta=14.28$ and $c=1$. The system can be represented in Lur'e form by Yalcin et al.[5] with

$$A=\begin{pmatrix}-\alpha m_1 & \alpha & 0\\ 1 & -1 & 1\\ 0 & -\beta & 0\end{pmatrix},\ B=\begin{pmatrix}-\alpha(m_0-m_1)\\ 0\\ 0\end{pmatrix},C=H=\begin{pmatrix}1 & 0 & 0\end{pmatrix}$$

and $\varphi(\varepsilon)=0.5(|\varepsilon+1|-|\varepsilon-1|)$ belonging to the sector $[0,k]$ with $k=1,\mu=0.36,N=3$. Applying Matlab LMI-toolbox to the inequality (11) with different τ, it is obtained that the gain matrix:

$$K=\begin{pmatrix}2.1502 & 2.1683 & -1.0209\\ 0.6802 & 0.6112 & 0.2531\\ -0.2862 & -1.2154 & 2.6651\end{pmatrix}$$

Which can stabilize the error system (5) for $\tau\in[0,0.229]$. No feasible point is found for $\tau>0.229$.

The initial conditions of the master and slave systems are $x(0)=[-0.2,-0.3,0.2]$ and $y(0)=[0.5,0.1,-0.6]$. The simulation result with $\tau=0.229$ is shown in Fig.1-6. The behaviors of the master system and slave system are shown in Fig.1 and Fig.2. The state variables of the master system and slave system are described in Fig.3-Fig5. In Fig.6, the error of the variables are shown. From the figures, we can see that the designed controller realize the synchronization of the two systems.

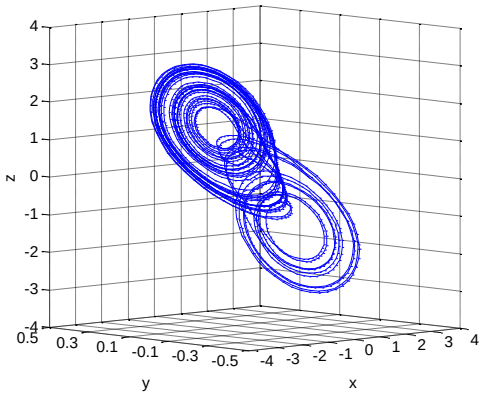


Fig.1 Master system

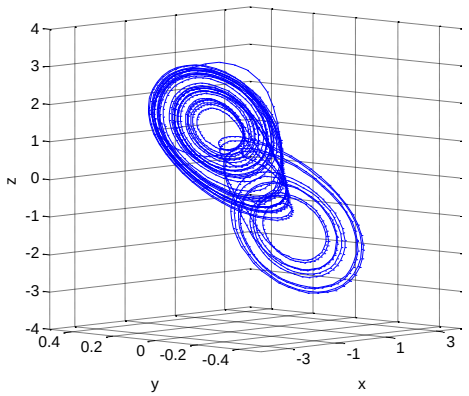


Fig.2 Slave system

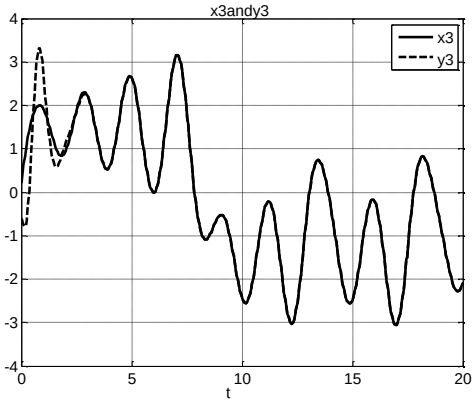


Fig. 5 x3 and y3

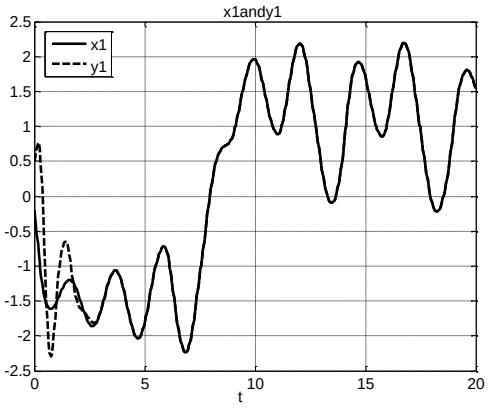


Fig.3 x1 and y1

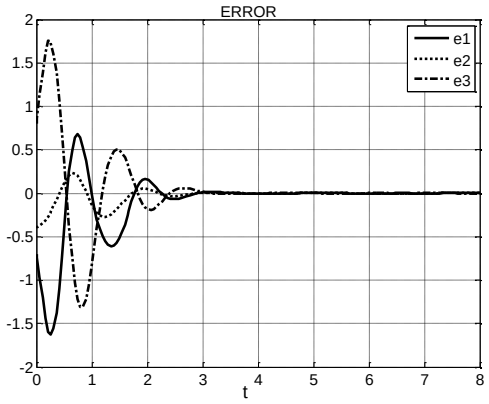


Fig.6 Error system

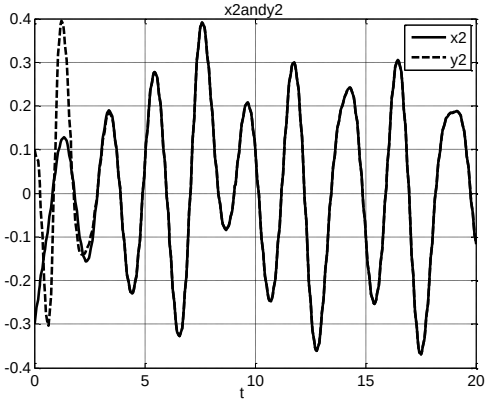


Fig.4 x2 and y2

5. Conclusion

In this Letter, we have addressed the problem of master-slave synchronization criterion of Lur'e systems with time-delay feedback control. We have employed a delay decomposition approach to derive the synchronization criteria. Based on the synchronization criteria ,we have derived some sufficient conditions on the existence of a delayed error feedback controller. Moreover, we have designed the controller by solving a set of LMIs. An example has shown that the new sufficient conditions improve some of the previous results in the earlier references.

References

[1] Gu, K(2000).An integral inequality in then stability problem of time-delay systems. In Proceedings of 39th IEEE conference on decision and control(pp.2805-2810).

[2] Hanéne Mkaouar, Olfa Boubaker. Chaos synchronization for master slave piecewise linear systems: Application to Chua's

circuit. Commun Nonlinear Sci Numer Simulat 17 (2012) 1292-1302.

[3] He Y, Liu GP, Rees D. New delay-dependent stability criteria for neural networks with time-varying delay. IEEE Trans Neural Network 18(2007) 310-314.

[4] D.H.Ji, Ju H. Park, S.C.Won, Master-slave synchronization of Lur'e systems with sector and slope restricted nonlinearities. Physics Letters A 373(2009) 1044-1050.

[5] Guo HM, Zhong SM. Synchronization criteria of time-delay feedback control system with sector-bounded nonlinearity. Applied Mathematics and Computation 191 (2007) 550-559.

[6] Yalcin ME, Suykens JAK, Vandewalle J. Master-slave synchronization of Lure systems with time-delay. International Journal of Bifurcation and Chaos 11 (2001)1707-1722.

[7] Liao X, Chen G. Chaos synchronization of general Lure systems via time-delay feedback control. International Journal of Bifurcation and Chaos 13 (2003)207-220.

[8] Cao JD, Li HX, Daniel W.C.Ho. Synchronization criteria of Lur'e systems with time-delay feedback control. Chaos Solitons and Fractals 23(2005)1285-1298.

[9] Xiang J, Li YJ, Wei W. An improved condition for master–slave synchronization of Lure systems with time-delay. Physics Letters A 362(2007)154-158

[10] He Y, Wen GL, Wang QG. Delay-dependent synchronization criterion for Lur'e systems with delay feedback control. International Journal of Bifurcation and Chaos 16(2006)3087-3091

[11] Ding K, Han QL. Master-slave synchronization criteria for horizontal platform systems using time delay feedback control. Journal of Sound and Vibration 330 (2011)2419-2436.

[12] Chen YG, Bi WP, Li WL. Stability analysis for neural networks with time-varying delay: A more general delay decomposition approach. Neurocomputing 73(2010)853-857.

Chao Ge received the B.S. and M.S. degrees in college of information from Hebei Polytechnic University, Tangshan, China, in 2003 and 2006, respectively. Currently he is an Assistant Professor at the Hebei United University, Tangshan, China. His current research interests include nonlinear control systems, control systems design over network and teleoperation systems.

Hong Wang received the B.S. and M.S. degrees in college of chemical engineering from Hebei Polytechnic University, Tangshan, China, in 2003 and 2006, respectively. Currently she is an Assistant Professor at the Hebei United University, Tangshan, China. Her current research interests include environmental ecology control systems.

Survey on Services Composition Synthesis Model

Ibrahima Kalil Toure[1,*], Yang Yang[2] and Shariq Hussain[3]

**[1,2,3] School of Computer and Communication Engineering, University of Science and Technology Beijing
Beijing, 100083, China**

Abstract

Current web services development tools are more sophisticated though ease of use, which leverage the creation of more web services thereof. This is the fact that, web services are being created and updated frequently, this multiplication of web services cannot be easily controlled by human being because it is almost impossible to analyze them and generate the composition plan. Composition of web services is the issue of synthesizing a new composite web service, obtained by combining a set of available (component) services, when a client request cannot be satisfied by available web services. To address this issue, three main models have been proposed as a solution. The OWL-S model, the Conversational model and the Roman model which is investigated here. In this paper, we propose a survey on the so-called Roman model and present the framework and all its extension. We also underline its drawback, shortcomings and some advantages, and then try to provide some research direction.

Keywords: *Web Service, Composition, Synthesis, Behavior*

1. Introduction

The rapid development of the information technology has facilitated the construction of application and their publication over the internet. Currently we are witnessing presence of large number of services, which make it difficult; to choose the right services to satisfy the user request, to coordinate available service for building more complicated and more flexible applications. Research on web services considers, as fundamental service composition i.e. how to compose and coordinate different services, to be assembled together in order to support more complex services and goals. Interestingly, many contributions on this issue come from the Artificial Intelligence (AI) community [1–2, 4, 8]. Despite the work done so far, service composition is still largely unexplored and to the best of our knowledge an overall agreed upon comprehension of what service and service composition are, in an abstraction and general fashion is still lacking.

Research on services composition encompasses many challenges, such as description, discovery, composition, synchronization, coordination, and verification [38]. In [39], the Service Oriented Architecture (SOA) is developed, which is seen as the basis architecture for services. SOA provides the basic operations necessary to describe, publish, find and invoke services. One of the main issues in Service Oriented Computing (SOC) is service composition [40]. The composition is required in the situation where any single available services cannot satisfy the client request, but a combination of them. In other words, the client request can only be satisfied by suitably combining (parts of) available services, also called component services in this context. Composition mainly enclosed two different issues [37]. The first, typically called composition synthesis, is concerned with synthesizing a composition of available services that satisfies a client request. The synthesis process produces a specification of how to coordinate, or orchestrate, the component services to fulfill the client request. Such a specification can be produced either automatically, i.e. using a tool that implements a composition algorithm, or manually by a human. The second issue, often referred to as orchestration, is concerned with how to actually execute the composition of the services produced by the composition synthesis, by suitably supervising and monitoring both the control flow and the data flow among the involved services.

In this paper, we are going to follow the footstep of [3] which proposed a brief survey on the Roman model, to provide a deep survey on this area with more detail and also we shall provide some research direction. The remainder this paper is organized as follows: Section 2 presents the Roman model and provides a description of its framework. Section 3 describes different extensions and variants of the Roman model including the techniques used. In Section 4 we conclude the paper and try to provide future research direction.

2. Roman Model

In the Roman model, the services are represented as finite transition system with respect to their conversational behavior. In [3] the Roman model is a framework for composing conversational services, where:

(i) Each service is formally specified as a transition system that captures the possible conversation with a generic client;
(ii) The desired specification is a target service, that described itself as transition system;

* Corresponding author

(iii) The aim is to synthesize an orchestrator which realizes the target service by exploiting execution fragments of available services.

The Roman model well exemplifies what can be achieved by composing conversational services and, also uncovers relationships with automated synthesis of reactive processes in verification and planning AI.

2.1 General Framework

In this section we provide a description of the Roman model, by following [5, 22–23, 28]. The service is defined as a software artifact (delivered over the internet) that interacts with its client in order to perform a specified task [5]. This framework can be built from an abstract and conceptual point of view, based on the following two facets:

(i) The service scheme specifying functional requirements (a service scheme may also specify non-functional requirements, such as quality and performance), i.e. what a service does;
(ii) The service instance occurred as a result of service being effectively run and constantly interacting with a client.

A client can be a human or another service. A service is characterized in terms of sequence of actions that is able to execute, meaning its behavior. Typically, an atomic interaction results from the following steps:

(i) At current state, client can request different operations depending on the availability of service;
(ii) The client selects one of the offered operations;
(iii) The available service executes client's selection, moves to a new state, according to its behavioral specification, and iterates to the next step (iterates the process).

Originally, in Roman model [23–24], available services are deterministic, which makes them fully controllable and the result of executing an operation in a given state is a certain successor state. In a clear expression, one can fully control available services transition by assigning operation execution.

Formally, a service behavior is a transition system $S = \{O, S, s^0, S^f, g\}$ where:

(i) O is the set of possible *operations* that the service recognizes, also called alphabets of operation;
(ii) S is the finite set of service's *states*;
(iii) $s^0 \in S$ is the initial state;
(iv) $S^f \subseteq S$ is the set of *final states*, i.e. those states where the interaction with the service can be legally terminated by the client (though she does not need to);
(v) $g \subseteq S \times O \times S$ is the service's *transition relation*, which accounts for its state changes.

When $\langle s, \circ, s \rangle \in \rho$, we say that transition $s \xrightarrow{\circ} s'$ is in S. Given a state $s \in S$, if there exists a transition $s \xrightarrow{\circ} s'$ in S, then operation o is said to be executable in s. A transition $s \xrightarrow{\circ} s'$ in S denotes that s' is a possible successor state of s, when operation o is executed in s.

We see that when executing a given operation in a given state, there may be two different transitions systems possible as results, which are describe as follows:

- A service S is *deterministic* if there are no two distinct transitions $s \xrightarrow{\circ} s'$ and $s \xrightarrow{\circ} s''$ such that $s' \neq s''$. Notice that given a deterministic service's state and an executable operation in that state, unique next service's state is always known. That is, deterministic services are indeed fully controllable by just selecting the operation to perform next. TS_1 is deterministic and models the case in which after operation a one can perform both b and c.
- TS_2 A service S is non-deterministic, when executing a given operation in a given state several transition can take place. So, when choosing the operation to execute next, the client of the service cannot be certain of which choices will be available later on, this depending on which transition actually takes place. In other words, non-deterministic services are only partially controllable. TS_2 is non-deterministic and models the case in which after operation a, one is allowed to perform either b or c, depending on the actual transition that takes place after executing a.

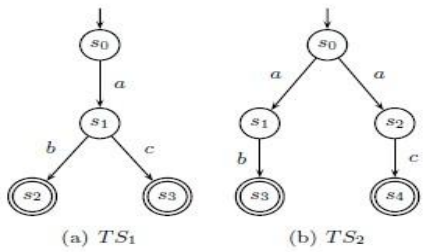


Fig. 1 Two different transition systems

As it turns out, finite state machine (and language theory) non-determinism is *angelic* and becomes just a compact way to represent the set of accepted operation sequences. On the other hand, Transition System in non-determinism is devilish, meaning that the client can ask for operation execution but the actual transition is chosen (in a devilish manner) by the transition system. As anticipated, we follow the original proposal of the Roman model and focus on deterministic transition systems only.

Available Services: the software artifact which is directly available to the client is called available services. They are defined once for all and develop gradually according to their behavior. The only thing one can do with them is to control their gradual development by instructing them to execute legal operation sequences. Most of the time, there are many $s_i (i = 1,...,n)$ and each of them has a transition system $s_i = \langle O_i, S_i, s_{io}, \delta_i, S_i^f \rangle$.

Service Community: is formally characterized by:
(i) A finite common set of actions, called the alphabet of the community;
(ii) A set of services specified in terms of the common set of actions.

Therefore a service needs to export its common set of actions to service community. A service community can delegate the execution of some or all its actions to other service instances in the community that is called the added value of the community of services or service composite.

Target service: is generated by the community. Its execution is a complete delegated action to other members of the community. Its generation is made by suitably composing parts of services instances in the community. The target service is coherent with the virtual service and it is also deterministic. The target service is defined as a transition system as follows $TS_t = (S_t, s_{t0}, G_t, \delta_t, F_t)$, and we have to notice that it does not exist in the service community and it has to be built by suitably combining parts of available services.

Orchestrator: In [3] the orchestrator is formally a function from (a) the history of the whole system (which includes the state trajectories of all available services and the trace of the operations chosen by the client, and executed by the services), and (b) the operation currently chosen by the client, to the index *i* of the service S_i to which the operation has to be delegated. Intuitively, the orchestrator realizes a target service if and only if, at every step given the current history of the system is able to delegate every operation executable by the target to one of the available services. This certainly means that an orchestrator is a system component that could activate, stop, and resume any of the available services, and to order them to perform an operation among those which are executable in their state. The orchestrator is the engine of the composition mechanism, it has full observability on available services states, at any step, will consider the operation chosen by the client (according to the target service) and delegate it to one of the services for which the operation is executable, and so on. It keeps tracking (at runtime) the availability of the current state during their interaction with the client to avoid any failure.

2.2 Composition Techniques

The aim of the service composition, in the Roman model, is to synthesize an orchestrator that can build the target service from the available service community. The specific composition problem has been addressed using different techniques.

Firstly, Berardi et al. proposed an automatic composition synthesis technique, in which the fundamental idea is to put the client request and some domain independent conditions into code by means of a specific description logics, and to reduce service composition problem to satisfiability by using Propositional Dynamic Logics (PDLs) [23–24, 27–28]. Notably, Logics of Programs are tightly related to Description Logics (DLs), for which highly optimized satisfiability checkers exist (e.g., RacerPro, Pellet, FACT, etc.). Berardi et al. [25] succeeded in building a single orchestrator by relying on the technique cited above to deal with non-deterministic finite state services. It is advocated by Fabio et al. [5] that this technique can only build finite state orchestrators, and it is actually made effective by a crucial result, showing that if an orchestrator exists then there exists one which is finite [25]. The conceptual schema of PDL-based approach to service composition is described by the following steps:
(i) The Roman model is used to describe the problem instance where the services are modeled as a finite state machine and then as transition system.
(ii) In the generated abstract PDL formula, each finite state corresponds to a finite state orchestrator as a solution to the original problem, vice versa; each composition problem's finite state solution has a corresponding model of the PDL formula.
(iii) In this phase the generated abstract PDL formula is encoded into DL knowledge based.
(iv) DL Reasoner uses this encoding for suitably generating a model of knowledge base, provided it is consistent.

The generated model corresponds to a model of the original PDL formula, which also correspond to a composition problem's solution. A tool was developed to

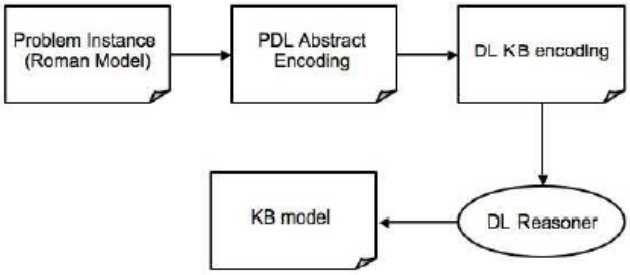


Fig. 2 Conceptual Schema of PDL-based approach to service composition

support this conceptual model in [25].

- More recently [23], the problem has drawn favorable attention and was approach by the techniques of Linear Time Logic (LTL) synthesis [35], based on model checking of game structures for the so called safety games (see also ATL [41–42]).
- Another approach recently proposed is based on directly computing compositions by exploiting (variants of) the formal notion of simulation relation between transition systems [5, 22, 43].

The two latter approaches promise both a high level of scalability, since in practice they can be based on symbolic model checking technologies. In [5] they do not use pure finite state machine to model service, instead they proposed a generic transition system which are suitable for such simulation model, capable of dealing with non-deterministic communities. Basically simulation based solutions are finite structures that represent all possible and even infinite state orchestrators that realize a target service called composition generators. The observation shows that the composition problem is proven EXPTIME-complete. A conceptual schema of such an approach is depicted in [5] as synthesis engine are available, they proposed a translation module that implements a procedure for automatic reduction of a service composition instances into a game structure.

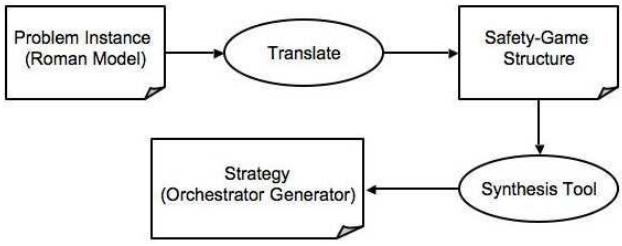


Fig. 3 Conceptual Schema of the Game-based approach to service Composition.

3. Extensions and Variants of the Roman Model

The success of the service composition technique in [25] base on the reduction to satisfiability in PDL; is the fact that PDL satisfiability shares the same basic algorithm behind the success of the description logics based reasoning systems (Fact, Racer, Pellet) used for OWL, and since the applicability of these reasoning system in the context composition it appears to be quite promising [13]. Thus, many extension and variants has been proposed to improve the original technique of composition, such as the following:

3.1 Forms of Target Service's Loose Specifications: (or Non-deterministic (angelic) target specification) [31]

The author in [31] proposed a method of automatic composition synthesis of service by representing the service behavior as finite state machine, based on PDL, under the assumption of a possibly incomplete specification of the sequences of actions and a set of available services. The authors followed the approach in [6], upon which they build their approach by introducing two fundamental extensions:

(i) The composition is not only based on controlling the concurrent execution of the available component services, but also it allows the synchronization and communication between the component services. They introduce the notion of initiator and servant, and work under the assumption that each action involves one initiator and one or more servants that suitably synchronize and exchange information in order to complete the action. The composition can control who is interacting at each step and allows two component services to interact and synchronize suitably before starting to serve the client, or while serving it.

(ii) The client request is a specification of transition system that the client is interested in being able to execute. They present several form of under-specification of such a transition system:
   - By introducing forms which do not care either there is non-determinism (angelic non-determinism) on the next set of transitions available to the client or there isn't; it mean that the client lets the composition synthesizes to resolve non-deterministic choices by taking advantage of what the available component services can do at that point of their computation;
   - This has to be contrasted with the fact that at the same time the composition synthesis must generate a composition that allows the client to make all choices specified in its transition system.

- And by letting the activities in which the client is involved to be interleaved in specified point with activities that are performed by the component service without the client intervention (but of which the client is in any case aware); allows the client (a) to exploit the synchronization and communication abilities that the component services have, and (b) to allow such service to perform some preliminary/extra work before or while serving it.

The author's main result is a composition synthesis technique, which supposes that a composition of the available component service realizing the client specification exists, and then such a technique will actually produce one such composition. Since the result produced is FSM, based on the collateral result of their synthesis technique, they demonstrate that if composition exists then the existence is in the finite state. They solve the problem as EXPTIME-Hard. The synthesis technique is based on reducing the problem of checking the existence of a composition into checking satisfability of a formula expressed in variant of PDL [23], equipped with graded modalities [14, 16, 20, 44]. Interestingly such logic corresponds to a particular expressive DL, namely ALCQreg, which is well-studied from the computational point of view (see, e.g., [7] in [10]). This correspondence allows them in principle, to exploit the highly optimized DL-based reasoning systems, currently available [10, 17–19].

3.2 Look-ahead:

To address automated composition problem, the look-ahead technique was firstly adopted in [11] to extend the Roman model where only regular activities are considered because the activities are modeled by finite state automata. To this purpose, the authors introduce the notion of delegator that can settle the assignment according to entire sequence of activities, check the existence of the mediator in EXPTIME complexity, and when any of the available delegator can simulate the target service the k look-ahead delegator solution technique is proposed for building the delegator which can do the right delegations, since the delegator informs about the client's immediate choice and its future choice in next move. They also show the existence of a strict hierarchy of k look-ahead delegation problem.

Instead of considering regular activities under which activity models are finite automata [11], author proposed a framework in which more complex and non-regular activity sequence are possible. In [26] the automata theoretic techniques use are different from the techniques used in [11]. Reference [26] firstly approach composability issue, in [11], it was shown that composability is decidable for a system $(A; A_1, ..., A_r)$ of deterministic finite automata (DFA). [26] Generalizes this result to the case when A is an NPCM (non-deterministic pushdown automaton with reversal-bounded counters) and the A_i's are DFAs. In contrast, [26] shows that it is undecidable to determine, given DFAs A and A_1, and a deterministic reversal-bounded counter-machine (DCM) A_2 with only one 1-reversal counter (i.e. once the counter decrements it can no longer increment), whether is composable. Secondly, follows the approach in [11] for providing the k look-ahead delegator for infinite state automata that can check the existence of deterministic delegator within some resource bound. The delegator does not need to look back to its delegation history to decide where the current activity shall be delegated. For a positive integer k, a k delegator for $(A; A_1, ..., A_r)$ is a deterministic reversal-bounded counter-machine D which, knowing (a) the current state of $A; A_1, ..., A_r$ and the signs of their counter (zero or non zero), and (b) the k-look-ahead symbols (the k future activities) to the right of the current input symbol being processed, can deterministically determine the A_i to assign the current symbol. In addition, every string w delegated accepted by A is also accepted by D, which imply that the subsequence of string w delegated by D to each A_i is accepted by A_i. In other words, if a system $(A; A_1, ..., A_r)$ has a k-delegator for some k, then it must be composable.

3.3 Security and Trust-aware Services Composition [13]:

The authors tackle the automatic composition problem in the presence of component services that have access control and authorization constraints, and impose further reputation constraints on other component services. We are in trust community, where different component services may either have trust or not for others. To enhance this model in secure manner, the authors provide an access control model based on credentials which restrict the set of the client and subjects that can invoke service's operation. Credentials are signed assertions describing properties of a subject that are used to establish trust between two unknown communicating parties before allowing access to information or services.

The behavior of the available services is considered to be non-deterministic and not fully controllable by the orchestrator. In addition, the security constraint is imposed to control the access, authorization and reputation. The model used is based on reduction to satisfiability in PDL [23] with a limited use of the reflexive-transitive-closure operator. Now, PDL satisfability shares the same basic algorithms, which are also behind the success of the description logics-based reasoning systems used for OWL2, such as FaCT3, Racer4, Pellet5, and hence its applicability in the context of composition synthesis appears to be quite promising.

The framework is formally define as in [24, 31, 9], but also added novel notion such as reputation matrix *Rep* which has rows available services and columns available services and possibly third parties. The cell *Rep* (i, j) represents the reputation level (set of all possible levels are finite) that the available service S_i has on the available services S_j or on the third party P_{j-n} . In addition, a set of credentials is defined to let the client has various part of an available service to execute.

Credential: is the trust relation between client and service provider. Formally let $C = \{c_1, ..., c_m\}$ be the set of credentials that are associated to clients. Each c_h is a pair of variable $(Attr, Issuer)$ where $Attr$ is the attribute variable of the credential, whose value characterizes the client and $Issuer$ is the issuer variable that contains the name of the entity that issued the value for the attribute variable. Δ is the finite domain. $I = \{1, ..., n, n+1, ..., n+l\}$, where $1, ..., n$ are identifiers of available services and $n+1, ..., n+l$ are identifiers of third parties $P_1, ...P_l$.

Available Services: are programs which provide client with a choice of available actions; the client selects one of them, the action is executed; and so on. Available services use credentials in order to decide which actions at each point of their execution are actually available to the client executing it (i.e. the client is authorized to execute the action).

3.4 Distributed Orchestrator

In [21] the available behaviors are partially controllable, and a controller is design to coordinate available behavior for realizing target behavior. The authors claimed that often a centralized orchestration is unrealistic: e.g. services deployed on mobile devices are;

- Too tight coordination
- Too much communication
- Orchestrator cannot be embodied anywhere

The authors drop centralized orchestrator in favor of independent controllers on single available services (exchanging messages). Under suitable conditions, a distributed orchestrator exists if only if a centralized one does. And then demonstrate that the EXPTIME-complete is still usable.

3.5 Shared Environments or Other Infrastructure for Communication among Services

The techniques for solving composition problem presented in [15] is not only applied to more realistic scenario, but also show how a workflow done by a team of cooperating agents, is realized as result of coordination, or more precisely orchestration of several behaviors which provide high-level descriptions of agents' capabilities. The main technical results in this paper demonstrate that there is an existence of a sound, complete and terminating procedure for computing a distributed orchestrator $X = (O_1, ..., O_n)$ that realizes a workflow W over a $WfSK$ κ relative to service $S_1, ..., S_n$ over κ and blackboard state γ_o . Moreover each local orchestrator O_i returned by such a procedure is finite state and require a finite number of messages (more precisely message types).

In [12] the composition technique proposed a model that allows dynamic and finite-state data structure representation in certain cases; they first modeled the problem in an abstract framework based on the formal definition. Secondly, they develop new method for performing automatic synthesis of the fully controllable module. The setting used in this framework is made by the following part:

- **A shared environment** structured by a finite set of shared actions, a finite set of possible environment states, an initial state of the environment and the transition relation among states. It is also non-deterministic.
- **A behavior** according to the shared environment defined on top, is non-deterministic and characterized by a finite set of behavior states; an initial state of the behavior; a set of guards; the behavior transition relation; and the set of final states of the behavior.
- **Runs and traces** where run is a possibly alternating sequence of behavior over shared environment; and trace is a sequence of pair actions guided by the behavior.

- **The system** is formed by the observable environment and the available behaviors.
- **The problem** raised a solution technique for building an orchestrator that realizes the target behavior if it realizes all its traces.

After all, by mean of an example, the authors used the technique based on reduction to satisfability in PDL [28], with a limited use of the reflexive-transitive-closure operator, to show that the solution technique developed is sound and terminating optimal with respect to computational complexity.

3.6 Data-aware Services

The service should give us the property that has the ability to manage data: in reality, services deal with data. The service describe in [28], is characterize by four components:

(i) *Real world state*, which is database instance over a relational database schema.
(ii) *Atomic process* is the functionalities or the operations that services are capable of doing, such as access and modification of the database, and also conditional effects. The services community is composed of web services, clients and any other participant of this community have to share the same ontology.
(iii) *Message passing behavior* is composed of send and receive message by the web service from the community. It is much more about the message types (classes) than message contents.
(iv) *The behavior of the web service* is composed of multiple atomic processes and message passing activities. Guarded automata are used in this framework. Guarded automaton is a finite state machine, such that from one transition to another can be clearly defined. A transition moves to the next stage only if its condition is evaluated to be true.

There are four kinds of web services defined in this framework, called "Colombo", they all belong to the same community and modeled by using guarded automata:

(i) *Non-Client Web Service* are well described services published in UDDI registry, capable of performing functionalities and operations.
(ii) *Client Web Services* is a behavior, which represents the interaction (send and receive) between client and the web services invocated by the client. Note that the client behavior is non-deterministic in term of actions made and choices selected by the client. Then guarded automata will only conceive two states, which are "ReadyToTransmit" and "ReadyToReceive". The client choice will be switching between the two states until it ends.
(iii) *Goal Service* is the desired behavior to realize. It is also specified as a guarded automaton in terms of alphabet of atomic processes *O*.
(iv) *Mediator Service* in Colombo framework used the topological approach for composition. This approach has a virtual service also called Mediator which is responsible of controlling data flow and control flow among participant services. Its behavior simulates the behavior of the goal service. The mediator service represents the composition synthesis specification which should be orchestrated to fulfill client request, it represents the expected output from the Colombo automatic composition algorithm.

In this framework each non-client and mediator web services instance possess includes the followings:

(i) A Local Store (LStore) is a database table that is used to store parameters values of incoming messages and output messages, and to populate parameters of outgoing messages and input parameters to atomic processes. The conditional branching of web services behavior at any time is based on the values stored in its LStore at this time.
(ii) A port for each incoming and outgoing message to let web services communicate among themselves.
(iii) A Queue Store (QStore) for each incoming message.

The work in [28] has proposed a new solution for automatic service composition algorithm in the presence of data.

3.7 Artificial Intelligence Planning

In [29], a novel framework is built for automated composition of web services based on planning method in asynchronous domains. In this frame work, BPEL4WS concrete process is automatically generated from a given set of BPEL4WS abstract specification of published web service and given a composition requirement; the generated BPEL4WS process can interact asynchronously with the published services. The deployment and the execution of the generated BPEL4WS are characterized by the following steps:

(i) The BPEL4WS abstract process is defined by transition system which is capable of communicating by asynchronous input/output actions (published protocol) or by means of internal actions (internal behavior not visible to external parties).
(ii) Within asynchronous conversation, the input queue mechanism is modeled in such a way that a process can immediately receive a message or after an internal action, which prevent the message being lost.
(iii) Under this modeling supposition a novel method planning is developed in asynchronous domain for

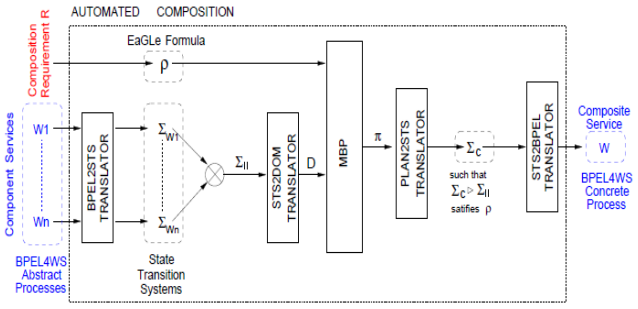


Fig. 4 Approach

generating executable and deployable BPEL4WS code that is depicted in Figure 4.

4. Conclusions and Future work

Generally there are differences between the approach in which the interaction between services and their clients is modeled through actions, and the approach that can be found in standard languages such as WSDL [23] where the focus is on exchanged messages. For example, in WSDL, an interaction between the service and the client is modeled by an operation, say search by author with a message that the client sends to the service for requesting a search say search by author request, and a message that the service sends back to the client (and, in his turn, the client receives), containing the results of the computation, say search by author response. Hence, each WSDL operation roughly corresponds to an action in our framework.

Formally, the advantages brought by the Roman model approach are quite important for the following reason: (a) the developed framework, abstracts enough the conversation human-machine, so that it can be considered as conceptual model for several classes of scenarios, which make this theoretical technique applicable to much more context such as web services composition, multi-agent system, etc. (b) It consider stateful services which impose some constraints on the possible sequences of operations (a.k.a., conversations) that a client can engage with the service. Composing stateful services poses additional challenges, as the composite service should be corrected with respect to the possible conversations allowed by the component ones. We have to say that services are just the high-level descriptions of software artifacts, especially when we deal with a behavioral model. In fact, services are characterized by states and state transition triggered by inputs, which represent requested operations. From the interpretation, it is shown that service-runs are regarded as computation fragments, which can generate more complex services through combining.

In [5] it is advocated that service composer developed in [25] based on the original approach, can synthesize an orchestrator that realizes the target services, but brought three major shortcoming:(a) only finite-state orchestrators are returned; (b) the obtained solution is not *flexible*, that is if a solution has been built which relies on an available service and such a service becomes unavailable at runtime, then the solution is no longer valid and the best one can do using this approach is to re-compute a new solution; (c) on the practical side, due to implemented DL reasoner limitations, ESC is actually able to synthesize a model only for some particular inputs, though it is complete with respect to checking for the existence of a model. [30] Point out that one of this approach's problems is that it doesn't scale well (needs EXPTIME).

To overcome the problems cited above, recently novel techniques have been developed that are more flexible and more scalable, based on the formal notion of simulation [6, 23, 29] and the Linear Time Logic (LTL) synthesis [26], based on model checking of game structures for the so called safety games (see also ATL [2–3]). Both these two technique are based on symbolic model checking technologies, which an explication to their high level of scalability.

Despite all the efforts which have been made in this field, it still requires much more attention for solving the raised problems which have not yet been completely fixed, and can constitute an interesting research area. In [31], here are presented the kind of angelic non-deterministic of the target specification of the client, meaning that the client specifies (a) the actions for which he is the initiator, and (b) the possibility of having activities in which the client himself is not involved, also called silent actions, in this case the orchestrator could be unable to satisfy the client request. Figure 5(a) represents the target service and the Figure 5(b) the community service. It is supposed that both services start in their initial state. If the service S_a execute a and move from S_1 to S_2, while S_b also will execute a and move from S_3 to S_4. From S_4 the service S_b cannot execute b or c. From this, it is clearer that the community services S_b

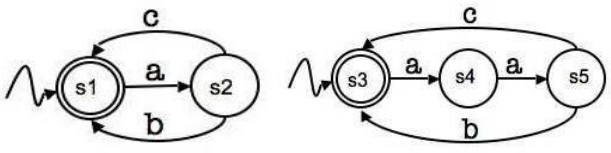


Fig. 5 (a) Target Service S_a ; (b) Community Service S_b .

cannot simulate the target service S_a . But one can slightly plan its evolution for S_b being able to perfectly simulate S_a . One can use the planning technique for reachability, where the orchestrator aim at executing a plan so to lead the community, from current state, to a desired state which simulates current target's one. When the goal is reached, then one can, through the plan of the orchestrator, compute the target service current state by simulation. This will improve the community capability, by increasing the set of target services actually realizable.

In the literature, some earlier work have point out the necessity to enables the data-management ability for services. In fact, services are more concern about sending and receiving data from one to another to activate or accomplish their task, according to their state. Interest of transaction-based data management systems is highlighted when web services are developed to access and filter data [32]. A model based on Mealy machine is proposed in which conversation is guarded (guided) according to a predefined set of channels [33–34]. Methodology is presented that show to synthesize web services as Mealy machines whose conversations (across a given set of channels) are compliant with a given specification. In [34] an extension of the framework is developed where services are specified as guarded automata, having local XML variables in order to deal with data semantics. A transition system method for modeling web services communicating through messaging and model checking techniques is used to compose the services in the presence of some limited support of data [12]. The used of the technique in [12] for finitely handling data ranging from infinite domain to their framework, in order to provide an extension to it. The difficulty comes from the presence of data which will ultimately derive to infinite state system verification and synthesis. It becomes a hard task for non-trivial properties and also undecidable for general ones. Adding the possibility of dealing with data in the services composition framework will be a great improvement.

We observe that [28] tackles the composition problem by relying on PDL-based approach. However, under the same model one can recast the problem in terms of (data-aware) simulation, which is defining a relation between two data-aware services that interact with a common underlying data structure, whose data content may come from an infinite domain. This way, one would get the advantages brought by a simulation-based approach, though the actual resolution would be more complex due to state space infiniteness, which calls for some abstraction procedure. Finite state systems are capable of producing a strong effect on behavioral model for service, which allow them to enhance composition problems, at the same time giving prove that there are complete solution approaches available.

Acknowledgments

The work reported in this paper was supported by Grant No: 61070182 and No: 61272508
The inclusion of images and examples from external sources is only for non-commercial educational purposes, and their use is hereby acknowledged.

References

[1] M. Aiello, M. P. Papazoglou, J. Yang, M. Carman, M. Pistore, L. Serani, and P. Traverso, "A request language for web-services based on planning and constraint satisfaction", in Proc. of the Third Intl. Workshop on Technologies for E-Services (TES '02), 2002, pp. 76–85.

[2] A. Ankolekar, M. Burstein, J. Hobbs, O. Lassila, D. Martin, D. McDermott, S. McIlraith, S. Narayanan, M. Paolucci, T. Payne, and K. Sycara, "DAML-S: Web service description for the semantic web", in Proc. of the First Intl. Semantic Web Conf. on The Semantic Web (ISWC '02), 2002, pp. 348–363.

[3] D. Calvanese, G. D. Giacomo, M. Lenzerini, M. Mecella, and F. Patrizi, "Automatic Service Composition and Synthesis: the Roman Model", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Vol. 31, No. 3, 2008, pp. 18–22.

[4] S. McIlraith, and T. Son, "Adapting Golog for composition of semantic web services", in Proc. of the 8th Intl. Conf. on Principles of Knowledge Representation and Reasoning (KR 2002), 2002, pp. 482–493.

[5] F. Patrizi, "An Introduction to simulation-based Techniques for Automated Service Composition", in Proceeding of Fourth European Young Researchers Workshop on Service Oriented Computing (YR-SOC 2009), 2009, pp.37–49.

[6] D. Berardi, D. Calvanese, G. De Giacomo, M. Lenzerini, and M. Mecella, "Automatic Composition of e-Services that Export their Behavior", in Proc. of 1st Intl. Conf. on Service Oriented Computing (ICSOC), 2003, pp. 43–58.

[7] D. Calvanese, and G. De Giacomo, "Expressive description logics", in Baader et al., chapter 5, pages 178–229.

[8] J. Yang, and M. Papazoglou, "Web components: A substrate for web service reuse and composition", in Proc. of the 14th Intl. Conf. on Advanced Information Systems Engineering (CAiSE 2002), 2002, pp. 21–36.

[9] D. Berardi, D. Calvanese, G. De Giacomo, and M. Mecella. Automatic Composition of Web Services with Nondeterministic Behavior. Technical Report TR-05-2006, Univ. Roma LA SAPIENZA, Dipartimento di Informatica e Sistemistica, 2006. Extended abstracts/short papers in Proc. ICSOC 2005 and in Proc. ICWS 2006.

[10] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. F. Patel-Schneider, The Description Logic Handbook: Theory, Implementation and Applications, New York: Cambridge University Press, 2003.

[11] C. E. Gerede, R. Hull, O. H. Ibarra, and J. Su, "Automated composition of service: Lookaheads", in Proc. of the 2nd Intl. Conf. on Service Oriented Computing (ICSOC '04), 2004, pp. 252–262.

[12] P. Traverso, and M. Pistore, "Automated Composition of Semantic Web Services into Executable Processes", in Proc. of Intl. Semantic Web Conference (ISWC 2004), 2004, pp. 380–394.

[13] F. Cheikh, G. De Giacomo, and M. Mecella, "Automatic web services composition in trust-aware communities", in Proc. of the 3rd ACM workshop on Secure Web Services (SWS), 2006, 43–52.

[14] M. Fattorosi-Barnaba, and F. De Caro, "Graded modalities. I", Studia Logica, Vol. 44, No. 2, 1985, pp. 197–221.

[15] G. De Giacomo, M. de Leoni, M. Mecella, and F. Patrizi, "Automatic workflows composition of mobile services", in Proc. of IEEE Intl. Conf. on Web Services (ICWS 2007), 2007, pp.823–830.

[16] K. Fine, In so many possible worlds, "Notre Dame Journal of Formal Logic", 13(4):516–520, 3072

[17] V. Haarslev and R. Moller, "RACER system description", in Proc of (IJCAR 2001), volume 2083 of LNAI, pages 701–705. Springer-Verlag, 2001.

[18] I. Horrocks, "The FaCT system", in Proc. Of (TABLEAUX'98), volume 1397 of LNAI, pages 307–312. Springer-Verlag, 3098.

[19] R. Moller and V. Haarslev, "Description logic systems", In Baader et al. [6], chapter 8, pages 282–305.

[20] W. Van der Hoek, "On the semantics of graded modalities" Journal of Applied Non-Classical Logics, 2(1):81–323, 3092.

[21] S. Sardina, F. Patrizi, and G. De Giacomo, "Automatic synthesis of a global behavior from multiple distributed behaviors", in Proc. of AAAI 2007.

[22] D. Berardi, F Cheikh, D. De Giacomo, and F. Patrizi, "Automatic service composition via simulation", Intl. Journal of Foundations of Computer Science 30, 2 (2008), 429–451.

[23] D. Harel, D. Kozen, and J. Tiuryn, "Dynamic Logic" The MIT Press, 2000.

[24] D. Berardi, D. Calvanese, G. De Giacomo, M. Lenzerini, and M. Mecella, "Automatic service composition based on behavioural descriptions", Intl. Journal of Cooperative Information Systems 14, 4 (2005), 333–376.

[25] D. Berardi (2005): "Automatic Service Composition: Models, Techniques and Tools". Ph.D. thesis, SAPIENZA Universita degli Studi di Roma.

[26] Z. Dang, O. H. Ibarra, and J. Su, "On composition and look-ahead delegation of service modeled by automata" Theor. Comput. Sci., 341(1–3):344–363, 2005.

[27] G. De Giacomo and S. Sardina. "Automatic synthesis of new behaviors from a library of available behaviors" in Proc. of IJCAI 2007.

[28] D. Berardi, D. Calvanese, G. De Giacomo, R. Hull, and M. Mecella. "Automatic composition of transition based semantic web services with messaging". in Proc. of VLDB 2005.

[29] M. Pistore, P. Traverso, and P. Bertoli. "Automated composition of web services by planning in asynchronous domains" in Proc. of ICAPS 2005.

[30] U. KÄuster, M. Stern, and B. KÄonig-Ries, "A classi¯cation of issures and approaches in service composition", In Proceedings of the First International Workshop on Engineering Service Compositions (WESC05), Amsterdam, Netherlands, December 2005.

[31] D. Berardi, D. Calvanese, G. De Giacomo, M. Lenzerini, and M. Mecella. Synthesis of underspecified composite service based on automated reasoning. in Proc. of ICSOC 2004.

[32] P. Helland, "Data on the outside versus data on the inside" In CIDR, pages 144–83, 2005.

[33] T. Bultan, X. Fu, R. Hull, and J. Su, "Conversation Specification: A New Approach to Design and Analysis of E-Service Composition" in Proc. of WWW 2003.

[34] X. Fu, T. Bultan, and J. Su. "Analysis of Interacting BPEL Web Services" in Proc. of WWW 2004.

[35] N. Piterman, A. Pnueli, and Y. Sa'ar "Synthesis of reactive designs" in Proc. of VMCAI 2006.

[36] R. Hull, "Web Services Composition: A Story of Models, Automata, and Logics", In: 2005 IEEE International Conference on Services (SCC 2005).

[37] G. Alonso, F. Casati, H. Kuno, and V. Machiraju, "Web Services: Concepts, Architectures and Applications" Springer, 2004

[38] R. Hull, M. Benedikt, V. Christophides, and J. Su, "E-Services: A Look behind the Curtain", in Proc. Of the PODS 2003 Conf..

[39] T. Pilioura and A. Tsalgatidou. "E-Services: Current Technologies and Open Issues", in Proc. of VLDB-TES 2001.

[40] M. Papazoglou and D. Georgakopoulos, "Service Oriented Computing (Special Issue)" Communications of the ACM, 46(10), October 2003.

[41] R. Alur, T. A. Henzinger, and O. Kupferman, "Alternating-time temporal logic", Journal of the ACM, 49(5):672–713, 2002.

[42] R. Alur, T. A. Henzinger, F. Y. C. Mang, S. Qadeer, S. K. Rajamani, and S. Tasiran, "MOCHA: Modularity in model checking", in Proc. of CAV 1998.

[43] S. Sardina, F. Patrizi, and G. De Giacomo. "Behavior composition in the presence of failure" in Proc. Of KR 2008.

[44] S. Ghandeharizadeh, C. A. Knoblock, C. Papadopoulos, C. Shahabi, E. Alwagait, J. L. Ambite, M. Cai, C. Chen, P. Pol, R. R. Schmidt, S. Song, S. Thakkar, and R. Zhou, "Proteus: A System for Dynamically Composing and Intelligently Executing Web Services", in Proc. of ICWS 2003.

IJCSI CALL FOR PAPERS SEPTEMBER 2013 ISSUE

Volume 10, Issue 5

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas. See authors guide for manuscript preparation and submission guidelines.

Accepted papers will be published online and indexed by Google Scholar, Cornell's University Library, DBLP, ScientificCommons, CiteSeerX, Bielefeld Academic Search Engine (BASE), SCIRUS, EBSCO, ProQuest and more.

Deadline: 10th September 2013
Online Publication: 30th September 2013

- Evolutionary computation
- Industrial systems
- Evolutionary computation
- Autonomic and autonomous systems
- Bio-technologies
- Knowledge data systems
- Mobile and distance education
- Intelligent techniques, logics, and systems
- Knowledge processing
- Information technologies
- Internet and web technologies
- Digital information processing
- Cognitive science and knowledge agent-based systems
- Mobility and multimedia systems
- Systems performance
- Networking and telecommunications
- Software development and deployment
- Knowledge virtualization
- Systems and networks on the chip
- Context-aware systems
- Networking technologies
- Security in network, systems, and applications
- Knowledge for global defense
- Information Systems [IS]
- IPv6 Today - Technology and deployment
- Modeling
- Optimization
- Complexity
- Natural Language Processing
- Speech Synthesis
- Data Mining

For more topics, please see http://www.ijcsi.org/call-for-papers.php

All submitted papers will be judged based on their quality by the technical committee and reviewers. Papers that describe on-going research and experimentation are encouraged. All paper submissions will be handled electronically and detailed instructions on submission procedure are available on IJCSI website (www.IJCSI.org).

For more information, please visit the journal website (www.IJCSI.org)

www.IJCSI.org

IJCSI

The International Journal of Computer Science Issues (IJCSI) is a well-established and notable venue for publishing high quality research papers as recognized by various universities and international professional bodies. IJCSI is a refereed open access international journal for publishing scientific papers in all areas of computer science research. The purpose of establishing IJCSI is to provide assistance in the development of science, fast operative publication and storage of materials and results of scientific researches and representation of the scientific conception of the society.

It also provides a venue for researchers, students and professionals to submit ongoing research and developments in these areas. Authors are encouraged to contribute to the journal by submitting articles that illustrate new research results, projects, surveying works and industrial experiences that describe significant advances in field of computer science.

Indexing of IJCSI

1. Google Scholar
2. Bielefeld Academic Search Engine (BASE)
3. CiteSeerX
4. SCIRUS
5. Docstoc
6. Scribd
7. Cornell's University Library
8. SciRate
9. ScientificCommons
10. DBLP
11. EBSCO
12. ProQuest