Computer Science & Information Technology 22

Dhinaharan Nagamalai
Sundarapandian Vaidyanathan (Eds)

Computer Science & Information Technology

Fourth International Conference on Advances in Computing and Information Technology (ACITY 2014)
Delhi, India, May 24 ~ 25 - 2014

AIRCC

Volume Editors

Dhinaharan Nagamalai,
Wireilla Net Solutions PTY LTD,
Sydney, Australia
E-mail: dhinthia@yahoo.com

Sundarapandian Vaidyanathan,
R & D Centre,
Vel Tech University, India
E-mail: sundarvtu@gmail.com

ISSN : 2231 - 5403
ISBN : 978-1-921987-22-9
DOI : 10.5121/csit.2014.4501 - 10.5121/csit.2014.4534

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

Fourth International Conference on Advances in Computing and Information Technology (ACITY 2014) was held in Delhi, INDIA, during May 24~25, 2014. Sixth International Conference on Wireless & Mobile Networks (WiMoN 2014), Fourth International Conference on Artificial Intelligence, Soft Computing & Applications (AIAA 2014), Fourth International Conference on Digital Image Processing and Pattern Recognition (DPPR 2014), Fifth International Conference on Internet Engineering & Web services (InWeS 2014), Third International Conference of Networks and Communications (NECO 2014), Fifth International Conference on Communications Security & Information Assurance (CSIA 2014) were collocated with the ACITY-2014. The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The ACITY-2014, WiMoN-2014, AIAA-2014, DPPR-2014, InWeS-2014, NECO-2014, CSIA-2014 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, ACITY-2014, WiMoN-2014, AIAA-2014, DPPR-2014, InWeS-2014, NECO-2014, CSIA-2014 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the ACITY-2014, WiMoN-2014, AIAA-2014, DPPR-2014, InWeS-2014, NECO-2014, CSIA-2014

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Dhinaharan Nagamalai
Sundarapandian Vaidyanathan

Organization

Program Committee Members

A.K.Daniel	M.M.M University of Technology, India
A.S.N. Chakravarthy	J.N.T.U. Kakinada University College of Engineering, India
Aababu Akella	JNTU College of Engineering Anantapur, India
Aasha Savan	Krupanidhi School of Management, India
Aayush	Stony Brook University, USA
Abdel-Aziz Ahmed	Anna University, India
Abirami	Anna University, India
Abraham Varghese	Adi Shankara Institute of Engineering and Technology, India
Adamu Murtala Zungeru	Massachusetts Institute of Technology, Cambridge
Ahmad Azzazi	Applied Science University, Jordan
Ajay Kumar	Mewar University, India
Akanksha	Center For Development of Advanced Computing, India
Akhil jabbar Meerja	Aurora's Engineering college, India
Akhilesh A Waoo	Beijing Genomics Institute, India
Alaa Hamami	Amman Arab University, Jordan
Ali Abid D. Al-Zuky	Mustansiriyah University, Iraq
Ali kartit	Laboratory of Research in Informatics and Telecommunication, Morocco
Ali Mohamed Jaoua	Qatar University, India
Allali Abdelmadjid	University of Oran, Algeria
Amandeep Singh	NIT Jalandhar, India
Amandeep Verma	Punjabi University, India
Amanpreet Kaur	ITM University, India
Amir Khusru Akhtar	Cambridge Institute of Technology, India
Amish Kumar	Chouksey Engineering College, India
Amit Choudhary	Maharaja Surajmal Institute, India
Amit Srivastava	Jaypee University of Engineering & Technology, India
Amoli bel	Yeshwantrao Chavan College of Engineering, India
Amritansh Singh	Lovely Professional University, India
Anandkumar Mani	SRM Easwari Engineering College, India
Andrea Zisman	The Open University, Walton Hall, UK
Anil Kumar Dubey	Govt. Engineering College, India
Ankit Chaudhary	Maharishi University of Management, USA
Anshuman Kalla	Jaipur National University, India
Anuja Arora	Jaypee Institute of Information Technology, India
Anwar Basha.H	S.A. Engineering College, India
Apai	University Malaysia Perlis, Malaysia
Apirajitha Ps	Sree Sastha Institute of Engineering and Technology, India
Artur Arsenio	Lisbon University, Portugal
Arun R	Sree Narayana Gurukulam College of Engineering,India
Aruna Pathak	Govt Engineering college, India
Aruni Singh Kamla	Nehru Institute of Technology, India
Arvind Kumar Sharma	Sine International Institute of Technology, India

Asadollah Shahbahrami	Guilan University, Iran
Ashish Gupta	Jay Pee University of Engg. And Technology (JUET), India
Ashish Kumar Dass	National Institute of Science and Technology, India
Ashok Kumar	Yellama Dasappa Institute of Technology, India
Ashutosh Kumar Dubey	TITR Bhopal, India
Ashwani Kumar	Jaypee University of Information Technology, India
Astha Pareek	ICG The IIS University Jaipur, India
Ayad Ismael	Erbil Technical Eng. College, IRAQ
B.Narendra Kumar Rao	Sree Vidyanikethan Engg. College, India
B.Sridevi	Velammal College of Engineering and Technology, India
Babu Karuppiah	Velammal College of Engg & Tech, India
Baddi Youssef	Youssef BADDI Isert- ENSIAS - UM5S, Morocco
Badri Subudhi	Indian Statistical Institute, India
Barbaros Preveze	Cankaya University, Turkey
Basappa Kodada	Canara Engineering college, India
BDCN Prasad	Prasad V. Potluri Siddhartha Institute of Technology, India
Beulah Christalin	Karunya University, India
Bijoy Kumar Raut	Birla Institute of Technology and Science, India
Bouchaib Falah	Al Akhawayn University, Morocco
Brij	National Institute of Technology - Kurukshetra, India
C Raghavendra Rao	University of Hyderabad, India
C S Lamba	Rajasthan College of Engineering for Women Jaipur, India
Can Zhang	Beijing University of Posts and Telecommunications, China
Chaitali Biswas	Girijananda Chowdhury Institute of Management and Technology, India
Chandra Mohan	Anna University, India
Chandrakala C.B	Manipal Institute of Technology, India
Chandrappa	SJB Institute of Technology, India
Channa Jayanath Kumarage	St Cloud State University, USA
Chhaya Dalela	JSS Academy of Technical Education, India
Chin-Chih Chang	Chung Hua University, Taiwan
Chintan patel	Gujarat Technological University, India
Cleopas	Niger Delta University, Nigeria
Cristian Alejandro	Torres Valencia Universidad Tecnologica de Pereira, Colombia
D.S.Vinod	Sri Jayachamarajendra College of Engineering, India
Dac-Nhuong Le	Vietnam National University, India
Daniel A. K	M.M.M University of Technology, India
Daniel Ajai	Madan Mohan Malaviya Engineering College, India
Debajyoti Pal	Camellia Institute of Technology, India
Deborah Vijayakumar	Anna University, India
Deepak Dembla	JECRC University, India
Deepak Raj S	Anna University, India
Deepali Kamthania	Institute of Computer Applications and Management, India
Demian Antony D'Mello	St. Joseph Engineering College, India
Demian D'Mello	St. Joseph Engineering College, India
Devasena Radhakrishnan	IFHE University, India
Dhanya Pm	Rajagiri School of Engineering and Technology, India
Dilli Ravilla	Manipal Institute of Technology, India

Dinesh B. Bhoyar	Yeshwantrao Chavan College of Engineering, India
Dinesh Baburao Bhoyar	Yeshwantrao Chavan College of Engineering, India
Divan Raimangiya	Institute of Technology and Science Rajkot, India
Doreswamy	Mangalore University, India
Ekbal Rashid	Cambridge Institute of Technology, India
Fadhil A. Ali	Oklahoma State University, USA
Faiyaz Ahmed	Integral University, India
Farhad Soleimanian	Islamic Azad University, Iran
G. Sankara Malliga	Dhanalakshmi College of Engineering,India
G.Radhamani	Dr G R D College of Science, India
G.Sankara Malliga	Dhanalakshmi College of Engineering, India
G.V.Jayaramaiah	Dr Ambedkar Institute of Technology, India
G.Vithya	St.joseph's College of Engineering, India
Gaurav Kumar Tak	Lovely Professional University, India
Gaurav Ojha	Indian Institute of Information Technology and Management, India
Geetha V	National Institute of Technology Karnataka, India
Geetha	Thiagarajar College of Engineering, India
George Prakash	Bharathidasan University, India
Gnana Jayanthi J	J.J. College of Engineering and Technology, India
Gnana Seelan	Centre for Development of Advanced Computing, India
Gnanam Jayanthi	J.J.College of Engineering and Technology, India
Grckaladhar Sarma	IBM India Pvt Ltd, India
Grienggrai Rajchakit	Maejo University, Thailand
Gsthyagaraju	SDMCET, India
Gulista Khan	Teerthanker Mahaveer University, India
Gullanar M Hadi	Salahaddin University, Iraq
Habib Rasi	Shiraz university of technology, Iran
Hamid Tairi	Universite Sidi Mohamed Ben Abdellah, Morocco
Hanumanthu Rajasekhar	Jawaharlal Nehru Technological University, India
Harish Bhaskar	Samsung Electronics India, India
Harish BS	Jayachamarajendra College of Engineering, India
Hassini Noureddine	University of Oran, Algeria
Hazem Al-Najjar	Misurata University, Libya
Himanshu Gupta	Amity University, India
Hitesh Rajput	Bhabha Atomic Research Centre, India
Hossein jadidoleslami	MUT University, Iran
Humera Khanam	S V University College of Engineering, India
Ijaz Ali Shoukat	King Saud University, Saudi Arabia
Inderpal Singh	Punjab Technical University, India
Indrajit Bhattacharya	Kalyani Government Engineering College, India
Indrajit Mandal	Sastra University, India
Isa Maleki	Islamic Azad University, Iran
Iti Mathur	Banasthali University, India
Ittipong Khemapech	University of the Thai Chamber of Commerce, Thailand
J. K. Mandal	Kalyani University, India
J.Usha	R V College of Engineering, India
Jacob W. Buganga	Makerere University, Uganda

Jalel Akaichi	University of Tunis, Tunis
Jan Zizka	Mendel University in Brno, Czech Republic
Jayadev Gyani	Jayamukhi Institute of Technological Sciences, India
Jayakumar C	RMK Engineering College, India
Jayalakshmi Sekar	Sudharsan Engineering College, India
Jayanthi K	Pondicherry Engineering College, India
Jayashree Mallapur	Basaveshwar Engineering College, India
Jeeva susan Jacob	Rajagiri School f Engineering and Technology, India
Jeevaa Katiravan	Anna University, India
Jinu Elizabeth John	Saintgits College of Engineering, India
Jitendra Maan	Tata Consultancy Services, India
John Tenvile	Sunyani Polytechnic, Ghana
Julie M	MES College, India
Jyothi Pillai	Bhilai Institute of Technology, India
Jyoti Gautam	Gautam Buddha University, India
K C Tiwari	Delhi Technological University, India
K. Daniel	Madan Mohan Malaviya Engineering College, India
K. Reddy Madhavi	JNTUA, India
K.Hemant Kumar Reddy	Biju Patnaik University of Technology (Bput), India
K.M.Anandkumar	SRM Easwari Engineering College, India
K.Ramudu	Kakatiya Institute of Technology and Science, India
K.Vanitha	Dr.G.R.Damodaran College of Science, India
Ka Chan,	La Trobe University, Australia
Kailas I Patil	G.H.Raisoni Institute of Engineering, India
Kalavathi Alla	Vasireddy Venkatadri Institute of Technology, India
Kamlesh Dutta	National Institute of Technology, India
Karan Singh	Motilal Nehru National Institute of Technology, India
Karthikeyan B	VIT University, India
Kavita Thakur	Ravishankar Shukla University, India
Kavitha Rajamani	St. Aloysius College,India
Kayhan Zrar Ghafoor	Koya University, Iraq
Keneilwe Zuva	University of Botswana, Botswana
Ketki Khante	Nagpur University, India
Khushbu Tikhe	Mumbai University, India
Kirubakaran	Bharat Heavy Electricals, India
Kishore Bhamidipati	Manipal Institute of Technology, India
Kishorjit Nongmeikapam	Manipur Institute of Technology, India
Kishwar Sadaf	Jamia Millia Islamia, India
Koushik Majumder	West Bengal University of Technology, India
Krishna Prakash	Manipal Institute of Technology, India
Krishna Prasad	University College of Engineering Kakinada, India
Krmadhavi Raju	JNTU College of Engineering, India
Kuljeet Kaur	Technivarana (Centre For Technical Solutions), India
Kunwar Singh Vaisla	BT Kumaon Institute of Technology, India
L. Vallikannu	Hindustan University, India
Lakhan Singh	Vishveshwarya Institute of Engineering & Technology, India
Leela Gopal	Visvesvaraya Technological Univerity, India
Leelavathi G	Visvevaraya Technological University, India

Lin Jin Cheng	Tatung University, Taiwan
Luiz Eduardo	University Federal of Amazonas, Brazil
M Mahbubur Rahman	Military Institute of Sceince and Technology, Bangladesh
M. N Rao	Sarvajanik College of Engineering and Technology, India
M. Nagabhushana Rao	SCET, India
M. Nirmala Devi	Amrita School of Engineering, India
M.K. Sharma	Amrapali Group of Institutes, India
M.Kamaraju	Gudlavalleru Engineering College, India
M.Mohamed Ashik	Salalah College of Technology, Sulthanate of Oman
M.Neelakantappa	AVR & SVR Institute of Tech.& Science, India
M.Seetha	G.Narayanamma Institute of Technology and Science, India
Madasamy Sornam	University of Madras, India
Madhavi Vaidya	University of Mumbai, India
Mahesh Pk	Don Bosco Institute of Technology, India
Mahsa Ghasembaglou	Qazvin Azad University, Iran
Malliga Raguraman	Dhanalakshmi College of Engineering, India
Mamta Rajshree	Shobhit University, India
Manik Gupta	Chitkara University, India
Manju Khari	Ambedkar Institute of Technology, India
Manjula K. Shenoy	Manipal University, India
Manoj Gupta	ITS Engg College, India
Manoj Sharma	Chandigarh Group of Colleges, India
Mansaf Alam	Jamia Millia Islamia, India
Mansouri Ali	Université Claude Bernard Lyon1, Tunisia
Marjan Mahmoodi	Islamic Azad University, Iran
Maryam Rastgarpour	Islamic Azad University, Iran
Masoud Ziabari	Mehr Aeen University, Iran
Md. Ibrahim Chowdhury	City University, Bangladesh
Meenakshi Tripathi	Malaviya National Institute of Technology, Jaipur
Meenakshi	Ambedkar Institutte of Technology Bangalore, India
Meenu	Periyar Maniammai University, India
Melih Kirlidog	Marmara University, Turkey
Met	ZTE University, China
Meyyappan	Alagappa University, India
Mihir Narayan Mohanty	Soa University, Odisha
Milind M. Mushrif	Y. C. College of Engineering, India
MK Sharma	UttarAkhand Technical University, India
Mohamed Abbas	B.S.Abdur Rahman University, India
Mohamed Alajmi	King Saud University, Saudi Arabia
Mohammad Zunnun Khan	Integral University, India
Mohammed H. AL-Jammas	University of Mosul, Iraq
Mohan Mahalakshmi Naidu	International Institute of Information Technology, India
Mohd Dilshad Ansari	Jaypee University of Information Technology, India
Mohd Umar Farooq	Muffakham Jah College of Engineering and Technology, India
Mohd. Amjad	Jamia Millia Islamia , India
Mohit Mathur	Jagan Institue of Managemnt Studies, India
Monica Mehrotra	Jamia Millia Islamia (Central University),India
Monica R Mundada	M.S.Ramaiah Institute of Technology, India

Moses Ekpenyong	University of Uyo, Nigeria
Mounir Gouiouez	laboratory Informatic and Simulation, Morocco
Murugan	Aalim Muhammed Salegh College of Engineering, India
Musheer Ahmad	Jamia Millia Islamia, India
Muthu Senthil	Anna University Chennai, India
N. Mohankumar	Amrita Vishwa Vidyapeetham University, India
N.Doshi	Sardar Vallabhbhai National Institute of Technology, India
N.Kaliammal	RVS College of Engineering & Technology, India
N.Uma Maheswari	Psna College o Engineering & Technology, India
Nag SV	RMK Engineering College, India
Nageswararao K	Mother Terisa Institute of Sci &Tech, India
Namita Tiwari	Maulana Azad National Institute of Technology, India
Namrata Dave	Gujarat Technology University, India
Nandhini	Anna University, India
Naresh Sharma	SRM University, India
Narges Shafieian	Azad University, Iran
Nataraj Urs H D	Reva Intistitute of Technology and Management, India
Natarajan Meghanathan	Jackson State University, USA
Neetesh Saxena	IIT Indore, India
Neetirajsinh Chhasatia	G H Patel College of Engineering & Technology, India
Neha Chaudhary	U.P. Technical University, India
Nilanjan Dey	JIS College of Engineering, India
Nilay Khare	Maulana Azad National Institute of Technology, India
Nirbhay Chaubey	Gujarat Technological University, India
Nirmalya Kar	National Institute of Technology, India
Nishant Doshi	NIT Surat, India
Nisheeth Joshi	Banasthali University, India
Nitin J. Janwe	Rajiv Gandhi College of Engineering, India
Omar Almomani	World Islamic Sciences & Education University, Jordan
Oss Noui	University of Batna, Algeria
Ouarda Barkat	University of Constantine, Algeria
Oussama Ghorbel	National Engineers School of Sfax University, Tunisia
P Mahesha	S.J.College of Engineering, India
P. Balamurugan	Government Arts College, India
P. E. S. N. Krishna Prasad	Prasad V. Potluri Siddhartha Institute of Technology, India
P.Danajayan	Pondicherry Engineering College, India
P.R.S.M.Lakshmi	Vignan University, India
P.Vijayakumar	Christ College of Engineering & Technology, India
Padmaja M	VR Siddhartha Engg College, India
Partha Pratim Bhattacharya	Mody University of Science & Technology, India
Pavan Kumar	PVP Siddhartha Institute of Technology, India
Pavan Pandey	Globallogic, India
Philomina Simon	University of Kerala, India
Pinki Nayak	AMITY, India
Piotr Zwierzykowski	Poznan University of Technology, Poland
PKS Krishna Sankar	KSR Institute for Engineering and Technology, India
Poonam	LNM Institute of Information Technology, India

Prasad Halgaonka	Maharashtra Institute of Technology College of Engineering, India
Prasenjit Chanak	Bengal Engineering and Science University, India
Preetha Manish	Rajagiri School of Engineering & Technology, India
Priti S. Sanjekar	R.C.Patel Institute of Technology, India
Pushpalatha D.V	Gokaraju Rangaraju Institute of Engineering and Technology, India
Pushpendra Kumar Pateriya	Lovely Professional University, India
Pushpendra Singh	Delhi Technological University, India
R Manjula	Vellore Institute of Technology Univeristy, India
R. Ramalakshmi	Kalasalingam University, India
R. Suresh	Vel Tech University, India
R.C. Joshi	Graphic Era University, Dehradun
R.Geetharamani	Anna University, India
R.I.Minu	Jersalem College of Engineering, India
R.S.Ponmagal	Dr.M.G.R.Educational And Research Institute University, India
R.Vadivel	Bharathiar University, India
R.Venkatesh	Psna College of Engg.& Technology, India
Radhika Kavuri	Chaitanya Bharathi Institute of Technology, India
Rafah M. Almuttairi	University of Babylon, Iraq
Rahmath Safeena Abdullah	Taif University, KSA
Rahul Jassal	SSGPURC-Hoshiarpur, India
Rahul Moriwal	Acropolis Institute of Technology & Research, India
Raj kumar Thenua	Anand Engineering College, India
Rajan Vohra	Guru Nanak Dev University, India
Rajesh Bawa	Punjabi University, India
Rajesh Mehra	NITTTR, India
Rajesh P Barnwal	Central Mechanical Engineering Research Institute, India
Rajeshwari Hegde	BMS College of Engineering, India
Rajib Kumar Jha	Indian Institute of Technology Patna, India
Rajiv Pandey	Amity University, UK
Rajput	Bhabha Atomic Research Centre, India
Ram Gopal L	Nokia Solutions and Networks, USA
Ramachandra Rao Kurada	Shri Vishnu Engineering College for Women, India
Ramakrishnan	Dr. M.G.R. Educational And Research Institute, India
Ramesh Babu	VIT University, India
Ramjeet Singh Yadav	Ashoka Institute of Technology and Management, India
Ramon Adeogun	Victoria University of Wellington, New Zealand
Ranjeet Bidwe	Pune Institute of Computer Technology, India
Ranjeet Vasant Bidwe	Pune Institute of Computer Technology, India
Ranjita Swain	Rourkrla Institute of Management Studies, India
Rashid Ali	Taif University, Saudi Arabia
Rashmi Mukhija	YMCA University of Science and Technology, India
Rastgarpour M	Science and Research University, Iran
Ravendra Singh	MJP Rohilkhand University, India
Ravinder Ahuja	Jaypee University of Information Technology Waknaghat Solan, India
Ravindrakumar	Chettinad College of Engineering and Technology, India

Ravitheja Perla	Vellore Institute of Technology, India
Reda Mohamed Hamou	Taher Moulay University of Saida, Algeria
Rekha jain	Banasthali University, India
Rengarajan	Sakunthala Engineering College, India
Renjith Kurup	Rajagiri College of Social Sciences Cochin, India
Revathi Venkat	RM University, India
Reza Ebrahimi Atani	University of Guilan, Iran
Rim Haddad	Laboratory Innov'com Sup'com,Tunisia
Ripal Patel	BVM Engineering College, India
Rizwan Beg	Integral University ,India
Roopali Garg	Panjab University, India
Rupali D	Rajiv Gandhi Proudyogiki Vishwavidyalaya, India
S. Adlin Jeena	Velammal Engineering College, India
S.Appavu alias Balamurugan	K.L.N.College of Information Technology, India
S.Hariharan	TRP Engineering College (SRM Group), India
S.Karunakaran	Kongu Engineering College, India
S.Murugavalli	Panimalar Engineering College, India
S.P.Karthik	Anna University, India
S.Padmavathi	Amrita University, India
S.Rajagopalan	Alagappa University, India
S.Sangeetha	Avinashilingam Deemed University, India
S.Selvaperumal	Syed Ammal Engineering College, India
S.Taruna	Banasthali University, India
Sachidananda	Berhampur University (Odisha), India
Saeed Agbariah	George Mason University, USA
Sai Kumar	CMR Technical Campus, India
Saikat Banerjee	Naraina Group of Institution, India
Samitha Khaiyum	Dayananda Sagar College of Engineering, India
Sandhya Magesh	B.S.Abdur Rahman University, India
Sangeeth	Amrita University, India
Sanjay Dorle	G.H. Raisoni College of Engineering, India
Sanjay K Sharma	Banasthali University, India
Sankara Malliga G	Dhanalakshmi College of Engineering, India
Santosh K. Pandey	The Institute of Chartered Accountants of India, India
Saptarshi Chakraborty	NIT Agartala, India
Sarita Simaiya Lilhore	Vidyapeeth Institute of Science & Technology (VIST), India
Saritha	Maulana Azad National Institute of Technology, India
Saruladha K	Pondicherry Engineering College, India
Sasanko Sekhar Gantayat	GMR Institute of Technology, India
Sasi rekha	Rathinam college of Arts and Science, India
Satria Mandala	Universiti Teknologi Malaysia, Malaysia
Satya Tazi	Govt. Engg. College Ajmer, India
Selvakumar S	GKM College of Engineering & Technology, India
Selvarani Rangasamy	ACS College of Engineering, India
Senthilnath	Indian Institute of Science, India
Sesha Bhargavi Velagaleti	G.Narayanamma Institute of Technology and Sciences, India
Seyed Ziaeddin Alborzi	Nanyang Technological University, Singapore
Seyyed Reza Khaze	Islamic Azad University, Iran

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Networks & Communications Community (NCC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

ACADEMY & INDUSTRY RESEARCH COLLABORATION CENTER (AIRCC)

TABLE OF CONTENTS

Fourth International Conference on Advances in Computing and Information Technology (ACITY 2014)

Sixth International Conference on Wireless & Mobile Networks (WiMoN 2014)

Fifth International Conference on Communications Security & Information Assurance (CSIA 2014)

Fourth International Conference on Artificial Intelligence, Soft Computing & Applications (AIAA 2014)

Fourth International Conference on Digital Image Processing and Pattern Recognition (DPPR 2014)

Third International Conference of Networks and Communications (NECO 2014)

Fifth International Conference on Internet Engineering & Web services (InWeS 2014)

TRANSLATION OF TELUGU-MARATHI AND VICE-VERSA USING RULE BASED MACHINE TRANSLATION

Dr. Siddhartha Ghosh[1], Sujata Thamke[2] and Kalyani U.R.S[3]

[1]Head of the Department of Computer Science & Engineering, KMIT,
Narayanaguda, Hyderabad
siddhartha@kmit.in
[2]R&D Staff of Computer Science & Engineering, KMIT,
Narayanaguda, Hyderabad
sujata.thamke@gmail.com
[3]R&D Staff of Information Technology, KMIT, Narayanaguda, Hyderabad
upadhyayula.kalyani@gmail.com

ABSTRACT

In today's digital world automated Machine Translation of one language to another has covered a long way to achieve different kinds of success stories. Whereas Babel Fish supports a good number of foreign languages and only Hindi from Indian languages, the Google Translator takes care of about 10 Indian languages. Though most of the Automated Machine Translation Systems are doing well but handling Indian languages needs a major care while handling the local proverbs/ idioms. Most of the Machine Translation system follows the direct translation approach while translating one Indian language to other. Our research at KMIT R&D Lab found that handling the local proverbs/idioms is not given enough attention by the earlier research work. This paper focuses on two of the majorly spoken Indian languages Marathi and Telugu, and translation between them. Handling proverbs and idioms of both the languages have been given a special care, and the research outcome shows a significant achievement in this direction.

KEYWORDS

Machine Translation, NLP, Parts Of Speech, Indian Languages.

1. INTRODUCTION

Machine Translation(MT) is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one natural language to another natural language. Machine Translation performs a simple translation of words from one natural language to another language, but that cannot produce a good translation of text i.e. recognition of whole phrases and their equivalent meaning should be present in the target language.

Machine Translation mentions the use of computers to convert some or the entire task of translation between human languages. Development of bilingual Machine Translation system for any two natural languages with electronic resources and tools is a challenging task. Many

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 01–13, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4501

practices are being done to develop MT systems for different languages using rule-based and statistical-based approaches. Machine Translation systems are specially designed for two particular languages, called a bilingual system, and for more than a single pair of languages, known as multilingual system. A bilingual system may be either unidirectional, from one Source Language (SL) to Target Language (TL), or may be bidirectional. Multilingual systems are bidirectional, but most bilingual systems are unidirectional. Machine Translation methodologies are commonly categorized as direct, transfer, and Interlingua. The methodologies differ in the analysis of the SL and extent to reach a language independent representation of meaning between the source and target languages. Barriers in good quality Machine Translation output can be attributed to ambiguity in natural languages. Ambiguities are classified into two types: **structural ambiguity** and **lexical ambiguity**.

India is a linguistically rich area. It has 22 constitutional languages, which are written in 10 different scripts. Hindi is the official language of the Union. Many of the states have their own regional language, which is either Hindi or one of the other constitutional languages. In addition, English is very widely used for media, commerce, science and technology, and education only about 5% of the world's population speaks English as a first language. In such a situation, there is a large market for translation between English and the various Indian languages.

1.1 Telugu Language

Telugu is one of the major languages of India. It is a Dravidian language frequently spoken in the Indian state of Andhra Pradesh. There were 79 million speakers in 2013. In India Telugu language occupies third position which is been spoken by large number of native speakers.

1.2 Marathi Language

Marathi is an Indo-Aryan language .It is mainly spoken in Maharashtra. Marathi is one of the 23 official languages of India. There were 74.8 million speakers in 2013. In India Marathi language occupies fourth position which is been spoken by large number of native speakers.

2. RELATED WORKS

There has been a growing interest in Machine Translation. Machine Translation has been brought a great change in making the Indian language more flexible to learn and understand which has been brought into consideration by various translation techniques addressed for decades in the form of Language Translator. The well-known Parts Of Speech (POS) Tagging has been used to the two Indian Languages i.e. Telugu and Marathi where the dictionary has been created which consist of the text meaning as well as its POS, also proverbs/idioms which are difficult to retrieve the exact meaning in different Indian languages are been represented in a database which consist of the meaning of the proverbs/idioms. Retrieving the exact translated sentence is difficult but these are a small practice using direct translation.

3. PROBLEM DEFINITION

Indian Language Translation is one of the serious problems faced by Natural Language Processing where the proverbs/idioms is also one within them. To get the exact meaning of the word in different languages we should also know the grammatical arrangement of the sentences in every language. So we are working with Parts of Speech tagging for each and every word as we are working for Marathi to Telugu Translation and vice-versa. Getting the exact meaning of the proverbs/ idioms is difficult in Indian Languages because the meaning of the proverbs/idioms changes when the word to word translation takes place. The problem here is to resolve how to convert the source text to the target text without changing its meaning. So we have focused on two Indian Languages which have same grammatical arrangement of sentence.

4. MACHINE TRANSLATION APPROCAH

Generally, Machine Translation is classified into seven categories i.e. Rule-based, Statistical-based, Hybrid-based, Example-based, Knowledge-based, Principle-based, and online interactive based methods. The first three Machine Translation approaches are widely used. Research shows that there are fruitful attempts using all these approaches for the development of English to Indian languages as well as Indian languages to Indian languages. Figure.1, shows the classification of MT in Natural language Processing (NLP).

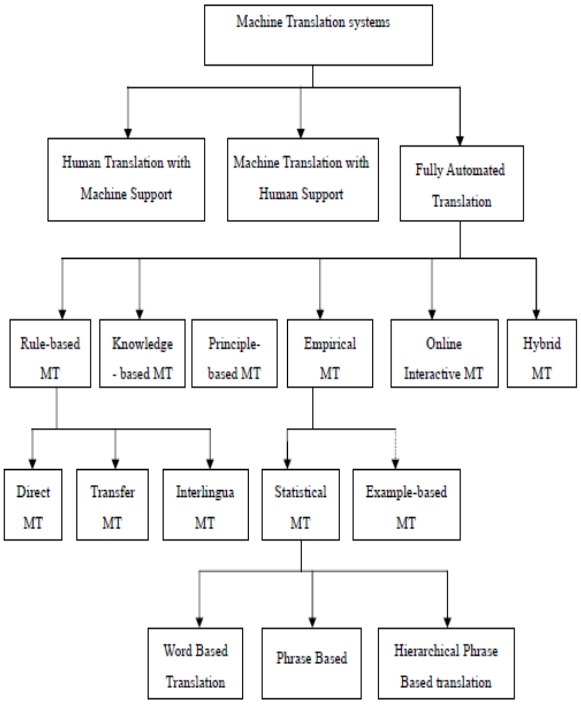


Figure. 1 Classification of Machine Translation

4.1 Rule-based Approach

The rule-based approach is the first strategy in Machine Translation that was developed. A Rule-Based Machine Translation (RBMT) system consists of collection of rules, called grammar rules, a bilingual or multilingual lexicon to process the rules. Rule Based Machine Translation approach requires a large human effort to code all of the linguistic resources, such as source side part-of-speech taggers and syntactic parsers, bilingual dictionaries. RBMT system is always extensible and maintainable. Rules play a major role in various stages of translation, such as syntactic processing, semantic interpretation, and contextual processing of language. Generally, rules are written with linguistic knowledge gathered from linguists. Transfer-based Machine Translation/Direct Translation, Interlingua Machine Translation, and dictionary-based Machine Translation are the three different approaches that come under the RBMT category. In the case of English to Indian languages and Indian language to Indian language MT systems, there have been fruitful attempts with all four approaches.

In this paper we have applied direct translation because both the Indian languages follow same sentence format i.e. SOV, so direct translation is applicable.

4.1.1 Direct Translation

In this method, the Source Language text is structurally analyzed up to the morphological level, and designed for a particular pair of source and target language. The performance of this system depends on the quality and quantity of the source-target language dictionaries, morphological analysis, text processing software, and word-by-word translation with minor grammatical adjustments on word order and morphology.

5. WORK CONTRIBUTED FOR TELUGU TO MARATHI TRANSLATION AND VICE-VERSA

To create a dictionary of Telugu and Marathi fonts in the database we have to go through the following process.

1. Download the appropriate Telugu http://telugu.changathi.com/Fonts.aspx) and Marathi (http://marathi.changathi.com/Fonts.aspx) fonts in your system then copy .ttf file and paste the file in the FONTS folder which is available in the Control Panel\Appearance and Personalization\Fonts.

2**.** There are two ways to type the text in the Indian languages

i. This method is used by the people who know the keyboard formats of Telugu and Marathi fonts. After adding the fonts we need to go to Control Panel-→Clock, Language, and Region→region and languages→change keyboards or other input methods→general→installed services→add→Telugu (India) and Marathi (India)→ok.

ii. The following method is a simple method to type any Indian language.

To convert the words into appropriate Indian language we need to type the text in English which will directly convert the text into selected Indian language .To achieve this process we need to download ***Microsoft Indic Language Input Tool for Telugu*** (http://www.bhashaindia.com/ilit/Telugu.aspx install desktop version) and ***Microsoft Indic Language Input Tool for Marathi***(http://www.bhashaindia.com/ilit/Marathi.aspx install desktop version).

3. Here we have developed software for conversion by using .NET as front end and SQL as back end. The databases which are created are the collection of Telugu and Marathi words with its parts of speech. The following are the databases which are created:-

a) Telugu words and its Parts of Speech (POS).
b) Marathi words and its Parts of Speech (POS).
c) Telugu and its equivalent Marathi words.

d) Marathi and its equivalent Telugu words.
e) Proverbs/ Idioms in Telugu and Marathi.
f) Proverbs/ Idioms in Marathi and Telugu.

This paper deals with two Indian languages (Telugu and Marathi) which follows the same grammar rule i.e. SOV (Subject Object Verb) for both the languages so direct word to word translation can be possible only by providing source language and target language dictionaries.

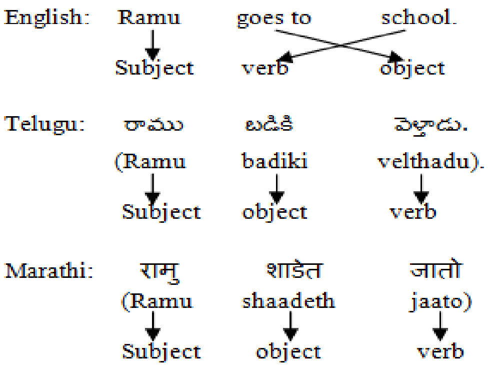
Here we have concentrated on the POS Tagging. Basically rule formations are mainly depending upon the 'Morpho-syntactic' information's. With the use of these rules the Parts Of Speech Tagger helps us to add the appropriate POS Tags to each and every word. Our main goal is to translate the Marathi to Telugu and vice versa. The contribution of our work defines as follows.

6. PARTS OF SPEECH

Tagging means labelling. Parts Of Speech Tagging is the one where we add the Parts of Speech category to the word depending upon the context in a sentence. It is also known as Morpho-syntactic Tagging. Tagging is essential in Machine Translation to understand the Target Language. In NLP, POS Tagging is the major task. When the machine understands the TEXT then it is ready to do any NLP applications. For that the machine should understand each and every word with its meaning and POS. This is the main aim of our research. Particularly in MT when the system understand the POS of source language Text then only it will translate into target language without any errors. So POS Tagging plays an important role in NLP.

7. MACHINE TRANSLATION IN INDIAN LANGUAGES (TELUGU TO MARATHI-MARATHI TO TELUGU)

We have considered the two neighbouring states which are familiar with each other and the grammatical rule of both the languages are same by which the directly translation of the language takes place i.e. Maharashtra using Marathi as its mother tongue and Andhra Pradesh using Telugu as its mother tongue, both having same arrangement of words in the sentences i.e. SOV where the rearrangement of words doesn't takes place so direct translation is been considered. Let us take an example.



In English language the sentence formation follows Subject Verb Object where are in Telugu and Marathi the sentence formation is in the form of Subject Object Verb. As we are translating.

Telugu to Marathi both follows same format so direct translation has been done. Considering the above example where Ramu is the subject goes to is the verb and school is the object where as in Telugu and Marathi language Ramu/రాము/ रामु is the subject Badiki/బడికి/शाडेत is the object Velthadu/వెళ్తాడు/Jato जातो is the verb.

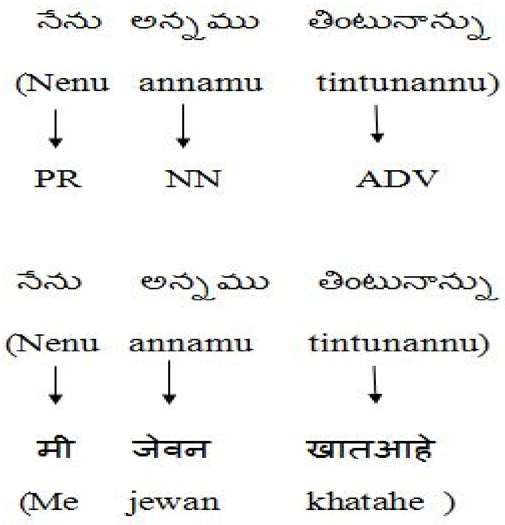
8. PROCEDURE FOR TRANSLATION

i. Select the language for translation then we need to enter the text in Telugu format with the help of Microsoft Indic Language Input Tool (Ref Figure.1).

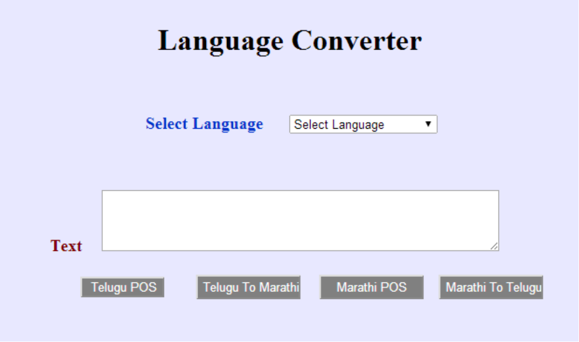
ii. After entering the text, the sentence has been separated into words based on the delimiters (space, commas etc).The separated words are stored in an Array List.

iii. From the Array List each and every word will check with the database and gets its parts of speech for every word (Ref Figure. 2 and 3)

Consider the following example in this we have taken Telugu text which we want to convert it into Marathi. Let us consider an example; it shows how the translation has been done.



Here is the screenshots of the software which has been developed using Microsoft VisualStudio2010.To insert the Telugu and Marathi words in the database we need to keep the data type as **nvarchar** where n is used to support Multilanguage's i.e. Telugu and Marathi text in the database. Figure.2 shows the Home page of the developed software.

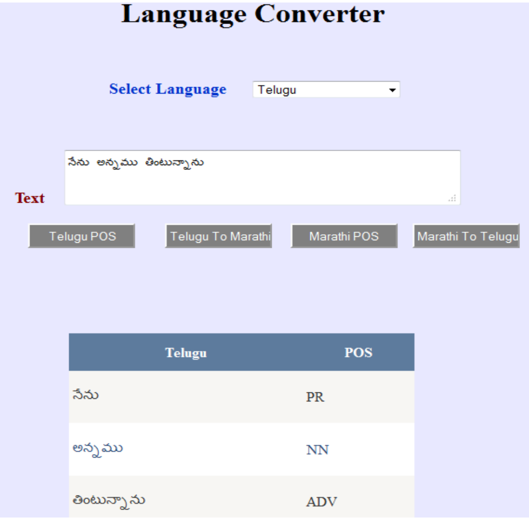


Language Converter

Select Language: Select Language

Text:

Figure. 2 Home Page



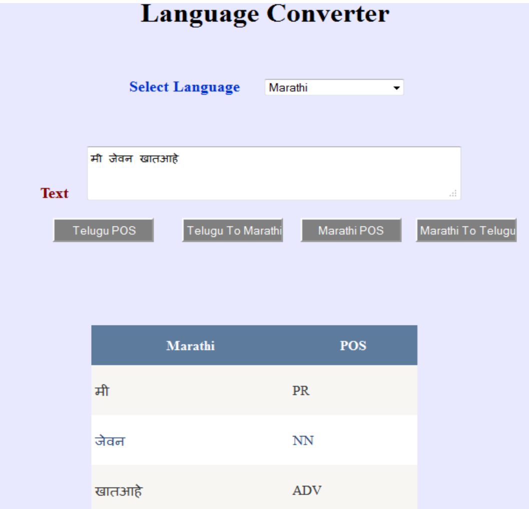
Language Converter

Select Language: Telugu

Text: నేను అన్నము తింటున్నాను

Telugu	POS
నేను	PR
అన్నము	NN
తింటున్నాను	ADV

Figure. 3 shows the Parts of Speech for Telugu Text



Language Converter

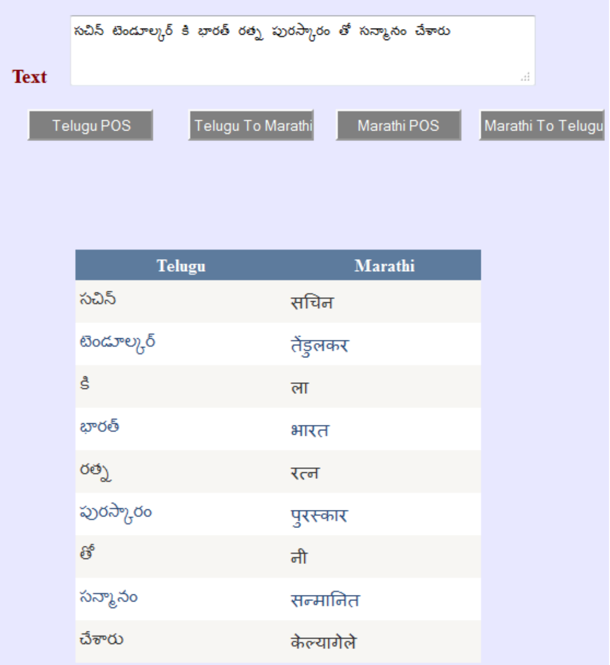
Select Language: Marathi

Text: मी जेवन खातआहे

Marathi	POS
मी	PR
जेवन	NN
खातआहे	ADV

Figure. 4 Shows Parts of Speech for given Marathi text

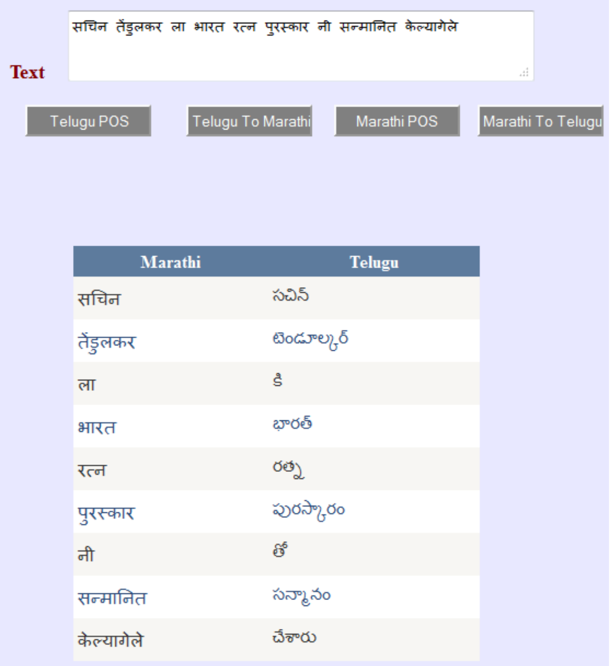
Converting Telugu Language into Marathi Language and Vice-Versa refer to Figure. 5 & 6 which shows how translation takes place with the help of database in the software which has been developed at our lab.



సచిన్ టెండూల్కర్ కి భారత్ రత్న పురస్కారం తో సన్మానం చేశారు

Telugu	Marathi
సచిన్	सचिन
టెండూల్కర్	तेंडुलकर
కి	ला
భారత్	भारत
రత్న	रत्न
పురస్కారం	पुरस्कार
తో	नी
సన్మానం	सन्मानित
చేశారు	केल्यागेले

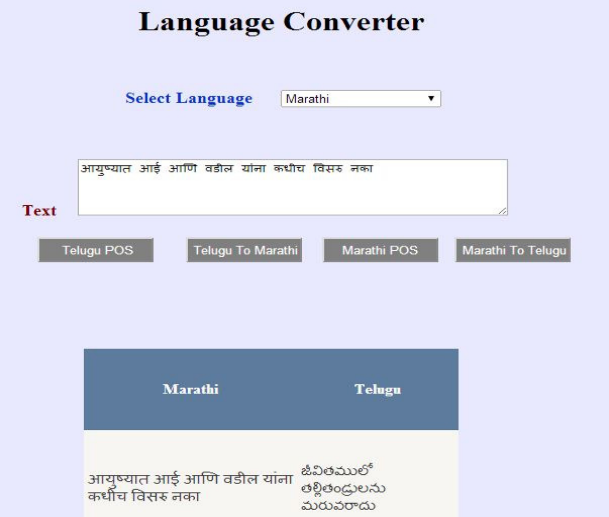
Figure. 5 shows the translation of given Telugu to Marathi Text



सचिन तेंडुलकर ला भारत रत्न पुरस्कार नी सन्मानित केल्यागेले

Marathi	Telugu
सचिन	సచిన్
तेंडुलकर	టెండూల్కర్
ला	కి
भारत	భారత్
रत्न	రత్న
पुरस्कार	పురస్కారం
नी	తో
सन्मानित	సన్మానం
केल्यागेले	చేశారు

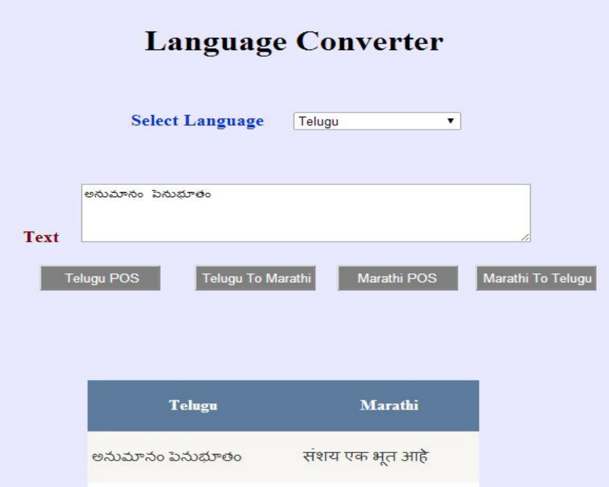
Figure. 6 shows the translation of given Marathi to Telugu Text



आयुष्यात आई आणि वडील यांना कधीच विसरु नका

Marathi	Telugu
आयुष्यात आई आणि वडील यांना कधीच विसरु नका	జీవితములో తల్లితండ్రులను మరువరాదు

Figure. 7 shows the translation of Marathi proverb to Telugu proverb

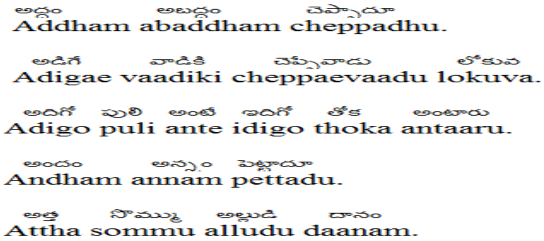


అనుమానం పెనుభూతం

Telugu	Marathi
అనుమానం పెనుభూతం	संशय एक भूत आहे

Figure. 8 shows the translation of Marathi proverb to Telugu proverb

9. PARTIAL PROVERBS LIST OF TELUGU AND MARATHI



అద్ధం అబద్ధం చెప్పాదూ
Addham abaddham cheppadhu.

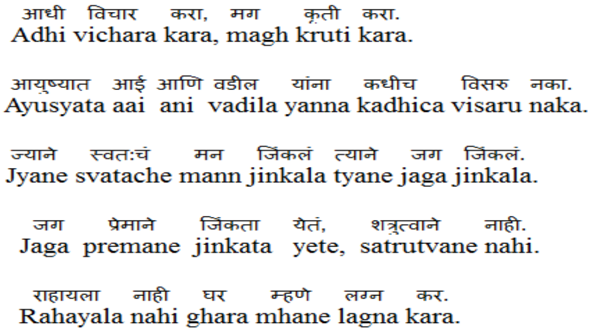
అడిగే వాడికి చెప్పేవాడు లోకువ
Adigae vaadiki cheppaevaadu lokuva.

అదిగో పులి అంటే ఇదిగో తోక అంటారు
Adigo puli ante idigo thoka antaaru.

అందం అన్నం పెట్టాదూ
Andham annam pettadu.

అత్త సొమ్ము అల్లుడి దానం
Attha sommu alludu daanam.

Figure. 9 List of Telugu proverbs



आधी विचार करा, मग कृती करा.
Adhi vichara kara, magh kruti kara.

आयुष्यात आई आणि वडील यांना कधीच विसरु नका.
Ayusyata aai ani vadila yanna kadhica visaru naka.

ज्याने स्वत:चं मन जिंकलं त्याने जग जिंकलं.
Jyane svatache mann jinkala tyane jaga jinkala.

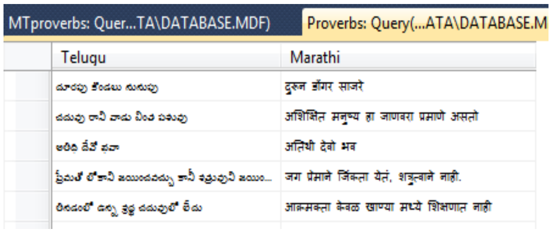
जग प्रेमाने जिंकता येतं, शत्रुत्वाने नाही.
Jaga premane jinkata yete, satrutvane nahi.

राहायला नाही घर म्हणे लग्न कर.
Rahayala nahi ghara mhane lagna kara.

Figure. 10 List of Marathi proverbs

10. PROVERBS HANDLING ENGINE OF MACHINE TRANSLATION SYSTEM

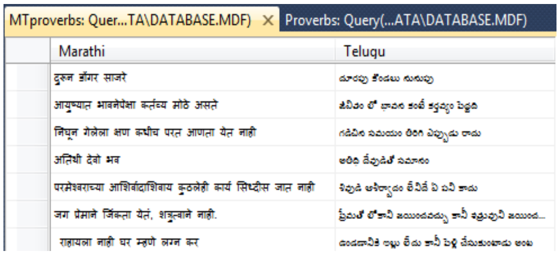
In the case of proverbs, as the languages have its own proverbs which will not match with another language proverb so we have inserted the proverbs meaning in Telugu/Marathi proverb databases.



MTproverbs: Quer...TA\DATABASE.MDF)
Proverbs: Query(...ATA\DATABASE.M

Telugu	Marathi
దూరపు కొండలు నునుపు	दुरून डोंगर साजरे
చదువు రాని వాడు వింత పశువు	अशिक्षित मनुष्य हा जाणवरा प्रमाणे असतो
అతిథి దేవో భవా	अतिथी देवो भव
ప్రేమతో లోకాని జయించవచ్చు కానీ శత్రువుని జయిం...	जग प्रेमाने जिंकता येतं, शत्रुत्वाने नाही.
తినడంలో ఉన్న శ్రద్ధ చదువులో లేదు	आक्रमकता केवळ खाण्या मध्ये शिक्षणात नाही

Figure. 11 Telugu to Marathi Proverbs Translation



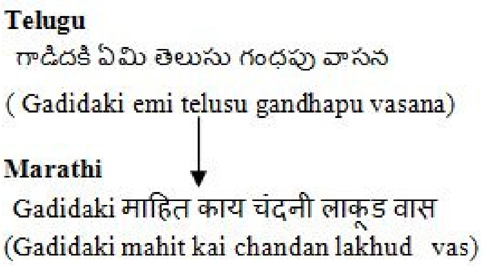
MTproverbs: Quer...TA\DATABASE.MDF) ✕
Proverbs: Query(...ATA\DATABASE.MDF)

Marathi	Telugu
दुरून डोंगर साजरे	దూరపు కొండలు నునుపు
आयुष्यात भावनेपेक्षा कर्तव्य मोठे असते	జీవితం లో భావన కంటే కర్తవ్యం పెద్దది
निघून गेलेला क्षण कधीच परत आणता येत नाही	గడిచిన సమయం తిరిగి ఎప్పుడు రాదు
अतिथी देवो भव	అతిథి దేవుడితో సమానం
परमेश्वराच्या आशिर्वादाशिवाय कुठलेही कार्य सिध्दीस जात नाही	శివుడి ఆశీర్వాదం లేనిదే ఏ పని కాదు
जग प्रेमाने जिंकता येतं, शत्रुत्वाने नाही.	ప్రేమతో లోకాని జయించవచ్చు కానీ శత్రువుని జయించ...
राहायला नाही घर म्हणे लग्न कर	ఉండడానికి ఇల్లు లేదు కానీ పెళ్లి చేసుకుంటాడు అంట

Figure. 12 Marathi to Telugu Proverbs Translation

11. RESULT

We have observed that in Google translator if the word is not present then it is displaying the same word which is written in Telugu and also in the case of proverbs it is not converting the sentence in meaningful way and the conversion takes direct word to word translation. For example consider the below where we have given Telugu proverb as



Telugu

గాడిదకి ఏమి తెలుసు గంధపు వాసన

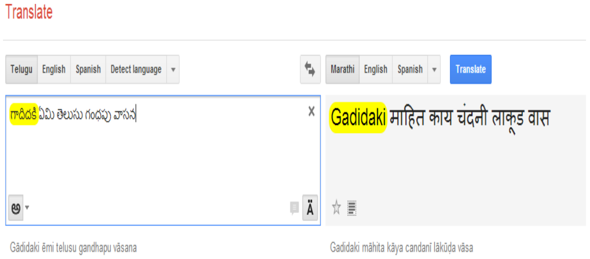
(Gadidaki emi telusu gandhapu vasana)

Marathi

Gadidaki माहित काय चंदनी लाकूड वास

(Gadidaki mahit kai chandan lakhud vas)

In the above example it is not changing "Gadidaki" (means donkey) into Marathi and the meaning of the proverb is also changing when the translation has been done.



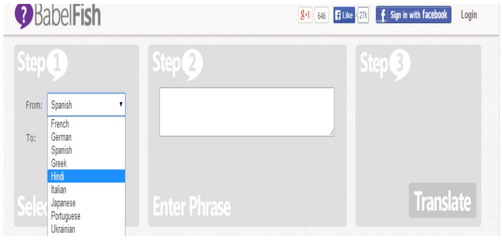
గాడిదకి ఏమి తెలుసు గంధపు వాసన

Gadidaki माहित काय चंदनी लाकूड वास

Gādidaki ēmi telusu gandhapu vāsana

Gadidaki māhita kāya candanī lākūḍa vāsa

In case of BabelFish there are 14 languages present for translation out of that one Indian Language is present i.e. Hindi.



From: Spanish

French
German
Spanish
Greek
Hindi
Italian
Japanese
Portuguese
Ukrainian

To overcome this problem we have added a separate dictionary for proverbs in which direct meaning of the proverbs has been inserted with its equivalent Telugu/Marathi text.

12. CONCLUSION

In this paper the translation of Indian Languages introduced and shown the parts of speech as well as the different Machine Translation approaches. Direct Translation was the technique to translate the Telugu and Marathi Language which have the same grammatical arrangement of sentence i.e. SOV. The Direct Approach can be possible only in some of the translation, but for more complex sentence the words are interchanged to get its proper meaning in the target language. This is especially true for real world problems where translation requires much complex algorithm to solve this problem when considered to the Indian Languages. Despite these already promising results, translation takes place using direct translation as well as and giving POS for each and every word, and proverbs/idioms are been solved by using dictionary which consist of meaning of the proverbs/idioms instead of word to word translation where there is no chance of wrong information. Also proverbs/idioms translation can most probably be improved. Further research might include a rule based approach, statistical approach for better translation. It is of great interest where Google translator as well as Babel Fish also failed to translate the exact meaning.

ACKNOWLEDGEMENT

This work is supported by the Research & Development Department of under Computer Science Department of Keshav Memorial Institute of Technology, Hyderabad, India.

The author thanks to the director of institute to believe us and encourage us for the work. And also the reference papers which helped us a lot for our research.

REFERENCES

[1] IEEE format [Online] available at http://www.coep.org.in/page_assets/491/IEEE_Template_4.pdf

[2] Microsoft Indic language input tool for Telugu [Online] available at http://www.bhashaindia.com/ilit/Telugu.aspx

[3] Microsoft Indic language input tool for Marathi [Online] available at http://www.bhashaindia.com/ilit/Marathi.aspx

[4] Telugu Fonts [Online] available at http://telugu.changathi.com/Fonts.aspx

[5] Marathi Fonts [Online] available at http://marathi.changathi.com/Fonts.aspx

[6] Wren & Martin, High School English Grammar & Composition, S.CHAND Publications

[7] Sneha Triparthi and Juran Krishna Sarkhe, Approaches to Machine Translation [online] available at http://nopr.niscair.res.in/bitstream/123456789/11057/1/ALIS%2057%284%29%20388-393.pdf

[8] Monika Sharma and Vishal goyal, Extracting Proverbs in machine translation from hindi to Punjabi Using relational data approach [online] available at http://www.csjournals.com/IJCSC/PDF2-2/Article%2060.pdf

[9] Sneha Triparthi and Juran Krishna Sarkhe, Hierarchical phrase based Machine Translation: Literature Survey [online] available at http://www.cfilt.iitb.ac.in/resources/surveys/Hierarchical_MT_Bibek.pdf

[10] Antony P.J, Machine Translation Approaches and Survey for Indian Languages [online] available at http://www.aclclp.org.tw/clclp/v18n1/v18n1a3.pdf.

AUTHORS

Dr. Siddhartha Ghosh is the Head of the Department of Computer science & Engineering Branch in Keshav memorial Institute Of Technology, Hyderabad. He received his Ph.D and from the Osmania University, Hyderabad in Computer Science & Engineering. He is the best innovative faculty “smart city contest” winner of IBM in 2010. He is interested in the research areas of NLP, AI, Data Mining, Machine Learning, and Computing in Indian languages. He has about 24 national and international research publications. He has combindly authored two books published from USA.

Sujata M. Thamke is working as a Research Assistant in Keshav memorial Institute Of Technology, Hyderabad. She received her Polytechnic from MSBTE and B.E from Amravati University in Computer Science & Engineering and Pursuing M.Tech in Computer Science & Engineering from Jawaharlal Nehru Technological University.

Kalyani U.R.S is a working as a Research Assistant in Keshav memorial Institute Of Technology, Hyderabad.She received her B.Tech degree in Information Technology From St.Mary’s College of Engineering and Technology affiliated to Jawaharlal Nehru Technological University.

INTENTIONAL BLANK

DISTANCE BASED TRANSFORMATION FOR PRIVACY PRESERVING DATA MINING USING HYBRID TRANSFORMATION

Hanumantha Rao Jalla[1] and P N Girija[2]

[1]Department of Information Technology, CBIT, Hyderabad, A.P, INDIA
hanu_it2007@yahoo.co.in
[2]School of Computer and Information Sciences, UOH, Hyderabad, A.P, INDIA
pn_girija@yahoo.com

ABSTRACT

Data mining techniques are used to retrieve the knowledge from large databases that helps the organizations to establish the business effectively in the competitive world. Sometimes, it violates privacy issues of individual customers. This paper addresses the problem of privacy issues related to the individual customers and also propose a transformation technique based on a Walsh-Hadamard transformation (WHT) and Rotation. The WHT generates an orthogonal matrix, it transfers entire data into new domain but maintain the distance between the data records these records can be reconstructed by applying statistical based techniques i.e. inverse matrix, so this problem is resolved by applying Rotation transformation. In this work, we increase the complexity to unauthorized persons for accessing original data of other organizations by applying Rotation transformation. The experimental results show that, the proposed transformation gives same classification accuracy like original data set. In this paper we compare the results with existing techniques such as Data perturbation like Simple Additive Noise (SAN) and Multiplicative Noise (MN), Discrete Cosine Transformation (DCT), Wavelet and First and Second order sum and Inner product Preservation (FISIP) transformation techniques. Based on privacy measures the paper concludes that proposed transformation technique is better to maintain the privacy of individual customers.

KEYWORDS

Privacy preserving, Walsh-Hadamard transformation, Rotation and classification

1. INTRODUCTION

Explosive growth in data storing and data processing technologies has led to the creation of huge databases that contains fruitful information. Data mining techniques are retrieving hidden patterns from the large databases. Sometimes, the organizations share their own data to third party or data miners to get useful information. So, the original data is exposed to many parties. It violates privacy issues of individual customers. Privacy infringement is an important issue in the Data Mining. People and organizations usually do not tend to provide their private data or locations to the public because of the privacy concern [1]. The researchers are intended to address this problem on the topic of Privacy Preserving Data Mining (PPDM). These methods have been developed for different purposes, such as data hiding, knowledge hiding, distributed PPDM and privacy aware knowledge sharing in different data mining tasks [2].

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 15–23, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4502

The issue of privacy protection in classification has been raised by many researchers [3, 4]. The objective of privacy preserving data classification is to build accurate classifiers without disclosing private information while the data is being mined. The performance of privacy preserving techniques should be analysed and compared in terms of both the privacy protection of individual data and the predictive accuracy of the constructed classifiers.

Recent research in the area of privacy preserving data mining has devoted much effort to determine a trade-off between privacy and the need for knowledge discovery, which is crucial in order to improve decision-making processes and other human activities. Mainly, three approaches are being adopted for privacy preserving data mining namely, heuristic based, cryptographic based and reconstruction based [5]. Heuristic based techniques are mainly adopted in centralized database scenario, whereas cryptographic based technique finds its application in distributed environment. There is a clear tradeoff between accuracy of knowledge and the privacy. That is higher the accuracy-lower the privacy and lower the accuracy-higher the privacy. Hence, privacy preserving data mining remains as an open research issue. Some data perturbation techniques which are maintaining data mining utilities may not satisfy statistical properties. However some perturbation techniques like SAN and MN may satisfy statistical properties which are lagging in privacy issues.

In this paper we suggest a Hybrid transformation technique it maintains data mining utilities and statistical properties like mean and standard deviation of the original data without information loss. Also we preserve the Euclidean distance between the data records before and after the transformation. WHT is an attractive alternative to the Fourier Transforms because it is computationally more efficient, and thus performs fast on digital computer.

This paper is organized as follows: section 2 discuss about the related work. Section 3 focus on Walsh-Hadamard Transformation, section 4 talk about usage of Rotation transformation, section 5 explains the proposed algorithm, section 6 presents experimental results and finally section 7 discuss conclusion and future scope.

2. RELATED WORK

PPDM techniques are mostly divided into two categories such as random perturbation and cryptographic techniques. A number of proposed privacy techniques exist based on perturbation. Agarwal and Srikanth [3], build classifier from the perturbed training data, later in 2001 a distortion-based approach for preserving the privacy was introduced by Agrawal and Aggarwal [6]. Reconstruction-based techniques for binary and categorical data are available in the literatures [7, 8]. M.Z Islam and L.Brankovic [9] proposed an algorithm known as DETECTIVE. In their work they addressed the perturbation is used for the categorical attributes based on clusters.

Cryptographic techniques are applied in distributed environment. Secure Multiparty Computation (SMC) is the well known technique in this category. In SMC two or more parties compute secure sum on their inputs and transfer to the other party without disclosing the original data [10, 11 and 12].

Jie Wang and Jun Zhang [13] addressed a frame work based on matrix factorization in the context of PPDM [13], they have used Singular Value Decomposition (SVD) and Non negative Matrix Factorization (NMF) methods. The framework focuses the accuracy and privacy issues in classification.

Recently, Euclidean distance preserving transformation techniques are used such as Fourier related transforms (DCT), wavelet transforms and linear transforms which are discussed in [14, 15 and 16].

In this paper, we also present a Euclidean distance preserving transformation technique using Walsh-Hadamard (WHT) and Rotation Transformation. WHT generates an orthogonal matrix, it preserves the Euclidean distance after transformation and as well as preserves statistical properties of the original data then we apply Rotation Transformation, it also preserve distance between data points. Hybrid Transformation technique preserves individual privacy of the customers.

3. WALSH-HADAMARD TRANSFORMATION (WHT)

Definition: The Hadmard transform H_n is a $2^n \times 2^n$ matrix, the Hadamard matrix (scaled by normalization factor), that transforms 2^n real numbers X_n into 2^n real numbers X_k.

The Walsh-Hadamard transform of a signal x of size N=2^n, is the matrix vector product $x.H_n$. Where

$$H_N =_{i=1}^{n}\otimes H_2 = \underbrace{H_2 \otimes H_2 \otimes \ldots \otimes H_2}_{n}$$

The matrix $H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ and $\otimes$ denotes the tensor or kronecker product. The tensor product of two matrices is obtained by replacing each entry of first matrix by that element multiplied by the second matrix. For example

$$H_4 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

The Walsh-Hadamard transformation generates an orthogonal matrix(H_n), it preserves Euclidean distance between the data points.

Definition: Matrices A for which, $A^T.A = I$ are called orthogonal matrices. They have the property that the transpose of A is also the inverse: $A^T = A^{-1}$ and$(A^T)^{-1} = A$.

Theorem 1: Suppose that $T: R^n \rightarrow R^n$ is a linear transformation with matrix A, then the linear transformation T preserves scalar products and therefore distance between points/vectors if and only if the associated matrix A is orthogonal.

Proof: Suppose that the scalar product of two vectors $u, v \in R^n$ is preserved by the linear transformation. Recall that a scalar product is the same as the matrix product of one vector as a row matrix by the other vector is a column matrix: $u.v = u^T.v$ then

$$\begin{aligned}(Au).(Av) &= (Au)^T(Av) = u^T A^T A v \\ &= u^T(A^T A)v\end{aligned}$$

Hence, if the scalar product is preserved then $(Au).(Av) = u^T(A^T A)v = u.v$ which shows that the product $A^T A$ must disappear from $u^T(A^T A)v$. This certainly happens if $A^T A = I_n$, where I_n the identity matrix, for then $u^T(A^T A)v = u^T I_n v = u^T v$, as required. It is also intuitively clear, at least, that this happen only if $A^T A = I_n$. Since distance is defined in terms of the scalar product, it follows that distance is also preserved. ■

Theorem 2: Suppose that $T: R^n \rightarrow R^n$ is a linear transformation with matrix A, then the linear transformation T preserves angles between the vectors (may or may not preserve distance) if the associated matrix B is a scalar multiple of an orthogonal matrix. *i.e. B=kA,* where $A^T A = I_n$ and $k \in R$.

Proof: this theorem proof is similar to the above theorem. ■

4. ROTATION TRANSFORMATION

In Cartesian co-ordinate system Translation, rotation and reflection are Isometric Transformations. Using Translation transformation failed to protect privacy of individual customers [17]. In this paper we are using Rotation transformation to hide underlying data values with combination of WHT. The purpose of Rotation Transformation is, increase complexity to unauthorized people while accessing the data for their personal use.

Definition: let Transformation $T: R^n \rightarrow R^n$ be a transformation in the n-dimensional space. T is said to bean isometric transformation if it preserves distances satisfying the constraint

$|T(U) - T(V) = |U - V|$ for U, V ∈ R^n

$$T: \quad \begin{pmatrix} x' \\ y' \end{pmatrix} = T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$

In this work we choose transform angle θ based on Variance of attributes before and after Transformation.

$\mathrm{Var(X)} = \mathrm{Var}(x_1, x_2, x_3, \ldots.., x_M) = \frac{1}{M} \times \sum_{i=1}^{M}(x_i - \overline{x})^2$

Where $\overline{x}$ is arithmetic mean of$x_1, x_2, x_3, \ldots.., x_M$.

We are following guidelines in [17] for choosing the transform angle θ, is calculated as follows $p_0 = min\big(var(A_i - A_i'), var(A_j - A_j')\big)$ Where A_i and A_i' are original and transformed data respectively and $\theta = p_0 * pl$.

5. EXPERIMENTAL WORK

Assume that, we represent a dataset as a matrix format. A row indicates an object and a column indicates an attribute. If the number of columns is less than2^n, here n=0, 1, 2, 3… Then we are adding the columns to its nearest value of 2^n. All the added columns are padding with zeros. Every element is discrete and numerical missing element is not allowed.

Algorithm:

Input: Dataset D, privacy_level Pl;

Output: Modified DatasetD';

1. Pre-Process the data if no. of columns less than N (N=2^n, n=0, 1, 2, 3….)
2. Generate Walsh NxN Matrix, N= number of columns.
3. Obtain the modified dataset by multiply original dataset with Walsh matrix.
4. Divide modified dataset into N/2 pairs.
5. For each pair apply rotation transformation.
6. Based on privacy level we choose optimal transform degree value.
7. Obtain the modified dataset by multiplying with Rotation matrix

6. EXPERIMENTAL RESULTS

We conducted experiments on two real life datasets Iris and Australian Credit dataset obtained from UCI Machine learning Repository [18]. The dataset properties are as follows.

Table 1. Dataset Description

Dataset name	No. of Records	No. of attributes	No. of classes
Iris	150	4	3
Australia credit	690	14	2

The Iris consists of flower dataset. It contains features of three types of flowers (classes) like Iris *Setosa*, Iris *Versicolor* and Iris *Virginica*. The four attributes are Sepal Length (SL), Petal Length (PL), Sepal Width (SW) and Petal Width (PW).

The Australian credit is banking dataset. It consists of two types of classes, *good* and *bad*. It consists of 690 instances with 14 attributes. Out of these 14 attributes, 6 attributes are numerical and 8 attributes are categorical. In this work we consider only numerical attributes. Two extra columns are added to dataset and those columns are appended with zeros.

We are using KNN (K-Nearest Neighbor) as a classifier in WEKA Tool [19]. KNN classifier is well known classifier. That works based on the distance between records. In the experiments parameter k is set with the values 3, 5, and 7, this transformation preserves distance between the records before and after transformation. Original data takes a matrix format, row is treated as an object and column is treated as an attribute.

Table 2. Original Dataset

SL	PL	SW	PW
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2

Table 3. Modified Dataset

SL	PL	SW	PW
-1.69866	8.109623	-2.54931	6.198623
-1.13457	7.34827	-1.98522	5.437271
-1.54761	7.465414	-2.29962	5.653045
-1.36575	7.382185	-2.31502	5.372555

Original and modified values are shown in Table 2 and 3 respectively. Distance between first and remaining records in original dataset are 0.5385, 0.5099 and 0.6480. In modified dataset the distances are 1.0770, 1.0198 and 1.2961. WHT transformation is worked based on Theorem 2. Distance matrix of modified dataset is an integer multiple of original dataset distance matrix. Due to this reason the distance between data records is not modified, next we apply rotation with some angle of degree which is based on privacy level of customer then the modified dataset is obtained. In this work the privacy level is set to 0.6. K-NN classification algorithm works well on modified dataset without information loss.

Accuracy of K-NN classifier on IRIS dataset is compared between existing perturbations methods like SAN, MN etc., and also with our proposed method is Hybrid Transformation , which is comparatively also better than the other distance preserving transformation methods given in Table 4.

Table 4. Accuracy of K-NN Classifier on IRIS

Method	Accuracy (%)		
	K=3	K=5	K=7
Original	95.33	95.33	95.33
SAN	95.33	95.33	95.33
MN	95.33	95.33	95.33
DCT	95.33	93.33	94.00
FISIP	96.00	95.33	96.67
Hybrid	96.67	95.33	95.33

Table 5. Accuracy of K-NN Classifier on Australian Credit

Method	Accuracy (%)		
	K=3	K=5	K=7
Original	73.33	72.60	72.31
SAN	73.33	72.60	72.31
MN	73.33	72.60	72.31
DCT	66.23	66.52	68.98
FISIP	66.56	67.39	69.42
Hybrid	67.82	68.98	67.68

Accuracy of KNN classifier on Australian Credit dataset is compared with existing methods shown in Table 5.

6.1 Privacy Measures

Privacy measures are adopted from [20].

6.1.1 Value difference

After a dataset is modified, the values of its elements are changed. The Value Difference (VD) of the dataset is represented by the relative value difference in Frobenius form. VD is the ratio of the Frobenius norm of the difference of D and $|D^{'}|$ to the Frobenius form of D.

$VD = \left\|D - |D^{'}|\right\|_F \,/\, \|D\|_F.$

6.1.2 Position Difference

After a data modification, the relative order of the value of the attribute changes, too. We use RP represents the average change of order for all the attributes. After data modification, the order of each value changes. Assume dataset D has n data objects and m attributes. Ord_j^i Denotes the ascending order of the j^{th} value in i^{th} Attribute, and $\overline{Ord_j^i}$ denotes the ascending order of the modified valueD_{ij}. Then RP is defined as

$$RP = \left(\sum\nolimits_{i=1}^{m}\sum_{j=1}^{n}\left|Ord_j^i - \overline{Ord_j^i}\right|\right) / \,(m * n)$$

RK denoted as percentage of elements keep their order in modified data.

$$RK = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} RK_j^i\right) / \,(m * n)$$

Where RK_j^i represents whether or not an element keeps its position in the order of values.

$$RK_j^i = \begin{cases} 1, if\ ord_j^i\ = \overline{ord_j^i} \\ 0, \quad\ otherwise \end{cases}$$

The metric CP is used to define the change of order of average value of attribute.

$$CP = \left(\sum_{i=1}^{m} \left|\left(ordAV_i - \overline{ordAV_i}\right)\right|\right) / \ m$$

Where $ordAV_i$ is the ascending order of the average value of attribute i, while $\overline{ordAV_i}$ denotes its ascending order after modification.

CK is to measure the percentage of the attributes that keep their orders of average value after distortion.

$$CK = \left(\sum_{i=1}^{m} CK^i\right)/m$$

Where CK^i is calculated as

$$CK^i = \begin{cases} 1, if\ ordAV_i = \overline{ordAV_i} \\ \quad 0, othrwise \end{cases}$$

The higher the value of RP and CP and the lower the value of RK and CK, the more privacy is preserved [18]. We calculate above data distortion measures on both modified datasets, results are shown in Table 6. Our transformation technique is compared with existing distance preserving transformation techniques such as FISIP and wavelet transformations. Privacy measures of IRIS dataset using Wavelet Transformations are taken from [15]. Based on these values, we say that our proposed transformation preserves distance as well as knowledge without loss.

We adopted the data distortion metrics used in [19] to measure the degree of data perturbed. According to their definitions, we know that a larger RP and CP, and smaller RK and CK value indicates more the original data matrix is distorted. Which implies the data distortion method is better in preserving Privacy. Data distortion measures on Iris dataset are showed

Table 6. Privacy Measures

Data (Method)	VD	RP	RK	CP	CK
IRIS (Hybrid)	-0.4390	47.028	0.50	0	1
AUS (Hybrid)	-0.5428	259.86	9.66e-4	0	1
IRIS (FISIP)	-0.032	42.0883	0.0033	0	1
IRIS (Wavelet)	0.91276	29.6266	0.015	1.0	0.25

7. Conclusion And Future Work

Some data mining algorithms works based on statistical properties based on that we propose a Hybrid transformation for PPDM. It preserves distance between data records so, knowledge should be same. It modifies the data but maintains Accuracy of classifier as original data without information loss. Our proposed transformation is applicable only to numerical attributes. It can be extended to categorical attributes.

REFERENCES

[1] D. Lin, E. Bertino, R. Cheng, and S. Prabhakar,(2008) "Position transformation: a location privacy protection method for moving objects" ,Proc. of Int'l Workshop on Security and Privacyin GIS and LBS, pp. 62–71.

[2] Giannotti, F., and Pedreschi, D.(2006)" Mobility, Data Mining and Privacy", Springer, Germany.

[3] R.Agrawal and R. Srikant,(2000) "Privacy Preserving Data Mining", In Proceeding of SIGMOD Conference on Management of Data, pp 439-450

[4] Y. Lindell and B. Pinkas,(2000) "Privacy-Preserving Data Mining", In Advances in Cryptology-CRYPTO, pp36-54.

[5] Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y., Theodoridis, Y,(2004) "State-of-the-art in Privacy Preserving Data Mining", SIGMOD Record 33 ,pp50—57.

[6] D. Agrawal and C. C. Aggarwal,(2001) "On the Design and Quantification of Privacy Preserving Data Mining Algorithm", In Proceeding of ACM SIGMOD, pp247-255.

[7] A. Evfimieski, R. Srikant, R. Agrawal and J. Gehrke,(2002) "Privacy Preserving Mining of Association Rules", In Proceedings of the 8th ACM SIGKDD, Edmonton, Canada ,pp 217-228

[8] S. Rizi and J.R. Haritsa,(2002) "Maintaining Data Privacy in Association Rule Mining", In the proceedings of the 28th VLDB Conference, Hong kong, China ,pp 682-693.

[9] M. Z. Islam, and L. Brankovic,(2005) "DETECTIVE: A Decision Tree Based Categorical Value Clustering and Perturbation Technique in Privacy Preserving Data Mining", In Proc. of the 3rd International IEEE Conference on Industrial Informatics, Perth,Australia

[10] A. C. Yao,(1986) "How to Generate and Exchange Secrets", In Proceedings 27th IEEE Symposium on Foundations of Computer Science, pp 162-167.

[11] J. Vaidya and C. Clifton,(2002) "Privacy Preserving Association Rule Mining in Vertically Partitioned Data", In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 639-644.

[12] J. Vaidya and C. Clifton,(2003) "Privacy-Preserving k-Means Clustering over Vertically Partitioned Data", In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 206-215.

[13] Jie Wang and Jun Zhang, "Addressing Accuracy Issues in Privacy Preserving Data Mining through Matrix Factorization ".

[14] Shibnath Mukharjee , Zhiyuan Chen, Aryya Gangopadhyay,(2006)"A Privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier-related transforms ",the VLDB Journal,pp (293-315).

[15] Vinod patel and yogendra kumar jain,(2009)"wave let transform based data perturbation method for privacy protection",IEEE.

[16] Jen-Wei Huang,Jun-Wei Su and Ming-Syan Chen,(2011) "FISIP: A Distance and Correlation Preserving Transformation for Privacy Preserving Data Mining"IEEE.

[17] ZHANG guo-rong,(2012)" An Effective Transformation Approach for Privacy Preserving Similarity Measurement", FSKD.

[18] http://kdd.ics.uci.edu/

[19] http://www.wekaito .ac.nz/ml/weka

[20] Shuting Xu, Jun Zhang, Dianwei Han, and Jie Wang,(2005)"Data distortion for privacy protection in a terrorist Analysis system", P. Kantor et al (Eds.): ISI 2005, LNCS 3495, pp. 459-464.

AUTHORS

P.N Girija is presently working as Professor in the School of Computers and Information Sciences, university of Hyderabad, Hyderabad. Her research areas are Speech Recognition, Speech Synthesis and Human Computer Interaction. She has published nearly eighty papers in various national and International journals and conferences. She visited School of Computer Science, Camegie Mellon University, Pittsburgh, U.S.A as a visiting Scholar during Jun-August 2004. She chaired several Sessions like COCOSDA, NTU Singapore etc, She completed sanctioned research projects from DST, AICTE, UPE etc.

HanumanthaRao Jalla completed B.Tech in computer science and Engineering from the VRSEC, Nagarjuna University, Guntur, A.P, in 2003 and M.Tech in information Technology from University of Hyderabad, A.P. Presently working an assistant professor in the Department of Information Technology, Chaitanya Bharathi Institute of Technology, Hyderabad, A.P. His research interests Privacy-Preserving Data Mining

INTENTIONAL BLANK

AN IMPROVISED MODEL FOR IDENTIFYING INFLUENTIAL NODES IN MULTI-PARAMETER SOCIAL NETWORKS

Abhishek Singh[1] and A. K. Agrawal[2]

[1,2]Department of Computer Engineering, IIT(BHU) Varansi
abhishek.singh.cse09@iitbhu.ac.in
akagrawal.cse@iitbhu.ac.in

ABSTRACT

Influence Maximization is one of the major tasks in the field of viral marketing and community detection. Based on the observation that social networks in general are multi-parameter graphs and viral marketing or Influence Maximization is based on few parameters, we propose to convert the general social networks into "interest graphs". We have proposed an improvised model for identifying influential nodes in multi-parameter social networks using these "interest graphs". The experiments conducted on these interest graphs have shown better results than the method proposed in [8].

KEYWORDS

Viral Marketing, Community Detection, Influence Maximization

1. INTRODUCTION

In today's era of the internet, the enormous growth and penetration of social networks into people's daily lives has brought a number of opportunities and challenges. It is not only a way to connect to the rest of the world but has also become an integral part of business, economy, politics and almost all such fields.

Identifying community structure [10] has been a central area of research in identifying groups in the network based on multiple factors such as common interests, friendship, organizations etc among the 'actors'. These communities are important as they can be viewed as a platform for sharing knowledge, data, emotions, sentiments etc. Another application of social network analysis has been viral marketing [] which is a very crucial area of research in business analytics. Viral marketing is an advertisement technique where one identifies a subset of 'actors' of the social network so as to obtain a "word of mouth" effect in promoting the product.

A major task in tackling both these problems is identifying influencers in the network, as whether it is the survival of a community or spreading of an innovation/idea/product in a network, *there's always a need for certain 'actors' who have influence over the rest of the community.* This can be seen in viral marketing as companies trying to identify a seed set of individuals to introduce their product so as to have as much spread of word as possible. In terms of community detection, a

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 25–30, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4503

certain set of individuals that share interests can be seen influencing each other. For example, an individual who shares interests in terms of his movie preferences with another individual would be compelled to watch a new movie if he/she sees positive feedback from the other. When this happens, the community detection algorithms increases the parameter used to represent the common interests among a certain set of individuals which results in detecting the community. So, although both community detection and viral marketing are different areas, the underlying problem in both the cases is the same. This is the problem of maximizing the spread of influence in a network.

Formally, the problem of Influence maximization involves finding few initial users in an online social network to adopt an innovation and spread the information, so that the influence of innovation or product in the network is maximized. Influence maximization is a problem applied not only to tasks related to social networks but can be used for different other applications.'

Finding these 'few initial users' in these large networks is the major challenge of the problem and for which, huge amount of data is needed to be processed. Not only the processing of huge amount of data is required, timeliness of the processes are also important. For this, the time complexity of the process should be small. The most popular approaches in this area are greedy algorithm and/or optimizations to the greedy algorithms.

It was observed that the individuals are connected to many others based on different interests. So a product/idea/event, which is to be 'spread' in the network, belong to a particular community of the individual only. So, instead of targeting the entire network to find 'influential seeds', one can find the community first, where the probability of spreading is high and subsequently finding the 'seed'/'seeds'. In this paper, we propose an improvement of the three methods for choosing the seeds by using community detection methods. The approaches are discussed in the next sections. In our approach, we emulate relationship between community detection and viral marketing.

The rest of the paper is organized as follows: section 2 discusses about the related works, section 3 describes our method, section 4 provides experimental results and section 5 concludes and discusses the future aspect of the work.

2. RELATED WORK

The concept of spreading of an idea, or an innovation or influence for that matter was first studied in the field of economy, giving birth to viral marketing. Several models were proposed to simulate this process. There are two models in particular have gained widespread acceptance. These are the Linear Threshold Model and the Independent Cascade Model. The Linear Threshold model states that every node in a network has some threshold which is needed to be achieved after which it can become active. The Independent Cascade model on the other hand gives a probabilistic methodology to this. It proposes that any active node in a network would get a dingle chance to activate an inactive node, which it can do with certain probability. The problem of Influence Maximization was first studied by Domingos and Richardson [1][2]. Although there attempts at solving this problem were probabilistic. Since Domingos and Richardson studied the problem of Influence Maximization, the other researchers [5][6][7] have proposed greedy approaches rather than the probabilistic approach suggested by the formers. In [3], Kempe et.al, proposed a greedy approach to solve what they viewed as a discrete optimization problem. After this several modifications and improvements have been proposed over this original approach.

In 2010 Tieyun Qian et.al [8] proposed a different approach towards studying the same problem by identifying seed nodes in implicit social networks. The suggested approach uses the Reverse Nearest Neighbours logic presented by [4], and build on that by defining Social Network Potential of an individual in a network. Inspired by this idea, we propose a new model to identify influential nodes and understand the spread of influence in a network where the connections between actors incorporate various factors such as the reasons for these connections and how strong these connections are.

3. PROPOSED METHOD

In this work we propose a model based on the method of [8]. As discussed in section 1 it was observed that the actors in the social networks are connected to each other for various interest/reasons. So a product/idea/event, which is to be 'spread' in the network, may be of interest to a particular set of individuals only. Here we propose a model for the Influence Maximization problem taking into account the specific area which may be of interest to one who wants to influence the network, out of the available areas.

Generally, the social networks are represented as graphs where the nodes are actors and the edges represent the connections between these actors. This representation of the edge is a accumulation of multiple parameters on which the social network is based. These parameters are of diverse nature like location, interests, likes etc. The application of any method traditionally involves considering all the above mentioned parameters instead of focusing on a particular interest. So in our model we converted the traditional social graph into "interest graph", which represents a particular interest(s) of the network. By this the volume of the network and the parameters of the network are reduced subsequently bringing down the time complexity and the computational overhead of applying the method on the original graph. An overview of the proposed model is shown in Figure 1.

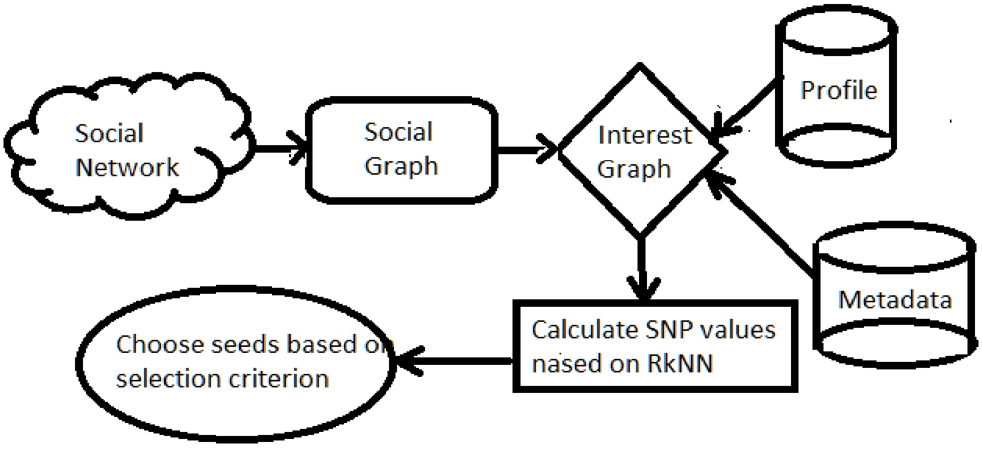

Social Network
Social Graph
Interest Graph
Profile
Metadata
Calculate SNP values nased on RkNN
Choose seeds based on selection criterion

Figure 1. Pictorial description of the proposed model

The model shows a social network, for example a blogger network. Now actors in a blogger network may be connected to each other through links that represent multiple shared interests.

The graph extracted from such network has edges that account for a number of parameters which may or may not be of interest. Hence it is required that the graph be filtered to form an interest graph containing edges with weights specific to the interest. Once we have this graph we can move forward with the basic algorithm as proposed by [8]

4. EXPERIMENT

The data used in [11], forms the basis of our experiments. The data contains information about blogs and bloggers from a particular organization. The author of the blog use different tags for each blog entry they make. These blogs are represented as a graph where the blogs are the nodes and the edges are defined as the relationship between the blogs. This relationship is defined by the common tags giving weightage to particular tag which in this case represents the interests of the authors.

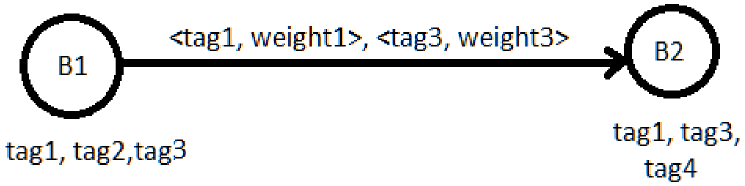


Figure 2. An example edge in the multi-parameter network

Any graph which represents this blog network can be defined by representing the blogs as nodes and the tags can be used to represent the weightage between the nodes. For converting this graph into "interest graph" we have given weightage to a particular tag in which one is interested. For example, let there be two blogs b1 and b2 having the tags{t1,t2,t3} and {t1, t4} respectively as shown in fig 2, then the nodes represent b1 and b2, and the weightage of the edge between b1 and b2, will be calculated as:

The graph so prepared is of 500 nodes and 87436 edges, by taking the first 500 nodes and their connections from the original dataset. The weights of this newly formed graph are then used as similarity measure for the nearest neighbourhood algorithm [4]. Then we use the methods explained in[8], to calculate the SNP values for the actors and find the desired set of seed nodes.

5. RESULTS

Table 1. Top 5 users for k=1

User-Id	R1NN	SNP
#211	8 22 41 45 61 100 107 137 157 169 171 175 186 209 227 247 249 280 295 296 297 298 312 316 337 359 380 391 429 435 437 455 461 481	34
#12	5 8 9 23 34 40 41 108 114 129 162 169 171 209 216 217 229 232 243 245 297 311 337 345 385 391 408 428 455 481	32

#119	2 29 41 142 169 171 209 217 239 243 245 247 263 271 297 300 337 355 368 370 383 391 401 409 417 421 425 454 455 481 485	31
#65	2 26 33 39 43 79 82 115 122 136 203 205 240 243 250 264 271 295 311 333 336 347 377 398 399 472 475 485	28
#327	4 16 23 27 32 47 56 92 94 103 129 130 154 163 181 182 186 197 217 228 283 339 375 385 418 427 457	27

6. CONCLUSION

In this work, we have proposed an improvised model for Influence Maximization in a general social network with edges having weights that are combination of multiple parameters. In such networks, connection between any two actors may be seen due to multiple shared interests. Thus, application of any algorithm on such a graph directly would not be completely realistic. Our work has addressed this problem by isolating these parameters by creating “interest graphs”.

The results so obtained are not very different from those obtained on the original graph, but the isolation of parameters has shown improvements in terms of the computational and time complexity requirements of the algorithm. Although the original algorithm would work perfectly in case of single parameter connections, we have proved that using “interest graphs” improves the accuracy and efficiency of the algorithm in multi parameter connection graphs.

As mentioned in the introduction, the fields of viral marketing and community detection are connected. Hence, as part of our future work we plan to use this hypothesis and work towards improving our model by incorporating community detection in this work.

REFERENCES

[1] P. Domingos & M. Richardson. “Mining the network value of customers”, 2001

[2] M. Richardson & P. Domingos. “Mining Knowledge-Sharing Sites for Viral Marketing”, 2002.

[3] D. Kempe, J. Kleinberg, & E. Tardos. “Maximizing the spread of influence through a social network”, 2003

[4] S. M. Flip Korn & S. Muthukrishnan. “Influence sets based on reverse nearest neighbour queries”, 2000

[5] W. Chen, Y. Wang, & S. Yang. “Efficient influence maximization in social networks”, 2009.

[6] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos,J. VanBriesen, & N. Glance. “Cost-effective outbreak detection in networks”.

[7] W. Chen, Y. Wang, & C. Wang. “Scalable Influence Maximization for Prevalent Viral Marketing in Large-Scale Social Networks”, 2010.

[8] Tieyun Qian & Jiangbo Liu“Influence Maximization through Identifying Seed Nodes from Implicit Social Networks”, ICUIMC’10 proceedings

[9] en.wikipedia.org/wiki/Viral_marketing

[10] http://en.wikipedia.org/wiki/Community_structure

[11] Nitin Agarwal, Huan Liu, Lei Tang & Philip S. Yu "Identifying influential blogger in a community", in proceedings of the 1st International Conference on web search and data mining(WSDM '08), PP 207-218,Feb 11-12 2008, Stanford, California

AUTHORS

Abhishek Singh is a post graduate student at the department of computer engineering, IIT-BHU. His area of interests are social networks, data mining, graph theory.

Prof. A. K. Agrawal is the head of department of Department of Computer Engineering at IIT-BHU. His areas of interest include database systems, theory of computation, compiler design and graph theory.

FINDING PROMINENT FEATURES IN COMMUNITIES IN SOCIAL NETWORKS USING ONTOLOGY

Vijay Nayak[1] and Bhaskar Biswas[2]

Department of Computer Engineering,
Indian Institute of Technology (BHU), Varanasi 221005 India
vijay.nayak.cse09@iitbhu.ac.in
bhaskar.cse@iitbhu.ac.in

ABSTRACT

Community detection is one of the major tasks in social networks. The success of any community depends upon the features that were selected to form the community. So it is important to have the knowledge of the main features that may affect the community. In this work we have proposed a method to find prominent features based on which community can be formed. Ontology has been used for the said purpose.

KEYWORDS

Social networks, community detection, Ontology, feature selection.

1. INTRODUCTION

Community Detection is one of the major tasks in social network analysis. Communities in social networks have a wide range of applications like viral marketing, sharing of information, sentiments, emotions etc. Communities are group of people/actors/ items who share some common topic(s) of interest. Most of the times, while detecting any community in a social network, many parameters are taken into account. For example, if anyone wants to find a community of interest, let say movies, in a social network he/she is connected to, may apply some parameter like actors, name of movie, genre of movie etc. to find such community. In general, more is the number of parameters, better is the community formed. At the same time, increasing the number of parameters increases the time complexity and computational complexity both. So, while adding any parameter or features for community detection, one must ensure that the addition should enhance the results of community detection algorithms in use. The major problem is that many times, it is difficult to find out the prominent features that can be used in community detection. There are two reasons behind this. First one, being the size of data to which the community detection algorithm is applied is huge. Second one is related to the first one, that is, due to this huge data many times the features selection cannot be done properly or some features are hidden or so.

One of the ways to reduce the size of data as stated in [14] is Ontology. In [14], the authors proposed Ontology as a means to represent the conceptual view of the data thereby reducing the size of the data. Ontology is explicit specification of conceptualization of a domain. In [14], the authors proposed a model to evaluate algorithms that can be applied to social networks using

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 31–36, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4504

Ontology. Inspired by this we extend the work of [14] to find out the prominent features effecting the formation of communities in social networks.

Ontology was proposed for the semantic web for W3C.org. This inspired many researchers like [2],[4],[6],etc. for web search engine and web crawlers. In [1],[3],[8],[7],[12] and [11]the authors demonstrated the use of ontology in social networks in different ways. In 2013, Régine Lecocq et.al. [13] proposed a generalised prototype for analysing social networks through Ontology.
The organisation of the rest of the paper is as follows: In section 2, the description of the proposed method/model is done. Section 3 describes the experiments we have performed and section 4 gives the results and conclusion which is followed by references in section 5.

2. METHOD

In the proposed method, the data from the social networks is extracted by web crawler/ apps/ any other tool and is stored in the database. Ontology was created capturing the features and properties of the extracted database. Since the ontology is used to represent different features of the dataset it can be used to represent the data itself for conducting any type of experiments. Any predefined or trivial communities based on the features of the dataset can be used as input for communities. Different community detection algorithms can be used to find communities from the dataset. As the community detection can be done using clustering algorithms, so in this work Markov Chain Clustering* algorithm is used to find communities. Further, communities and clusters are used interchangeably. Two nodes/actors in social networks are connected to each other by a specific relationship. This relationship is based on a set of properties or features of the nodes. These properties are represented in the form of ontology. Such a relationship is used to form communities. These communities are combined with previously stored ontology to form a modified ontology.

This modified ontology is then used to find the degree of overlap for every feature that is included in the ontology. The degree of overlap in this context is defined as intersection between communities. In an ideal community there should be distinctness in the nodes. So, lesser is the overlapping of the nodes, greater is the distinctness between them. The higher the degree of overlap the lesser is the intra–cluster distance between two communities or clusters and vice-versa. This observation resulted in concluding that if a certain feature had lesser degree of overlap than the other features over the specified communities then that feature had more influence in the formation of that specific cluster. The degree of overlap for every feature is calculated and compared. A feature which has minimum degree is observed to be influencing the formation of the community. For example, a person is a friend of some individual, this relationship of friendship can be used to form clusters. The properties/features which influence the friendship like location, movies, music and books, etc. are used in designing the ontology. A modified ontology is then created to find the degree of overlap for each property. These degrees of overlap are then compared to find which feature has minimum value thus influencing the formation of the clusters. The above method is described in the model shown in figure 1.

3. EXPERIMENT

3.1. Data preparation:

The ontology was created from a dataset extracted from Facebook (www.facebook.com). This dataset consisted of properties of each individual like location (city, state and country), likes of music and movies. The communities were formed by using MCL algorithm. The parameter used for clustering in MCL algorithm is friendship between each individual. The ontology and the clusters were combined to form a modified ontology. The degree of overlap for each feature/property was calculated using this modified ontology.

*www.Wikipedia.org

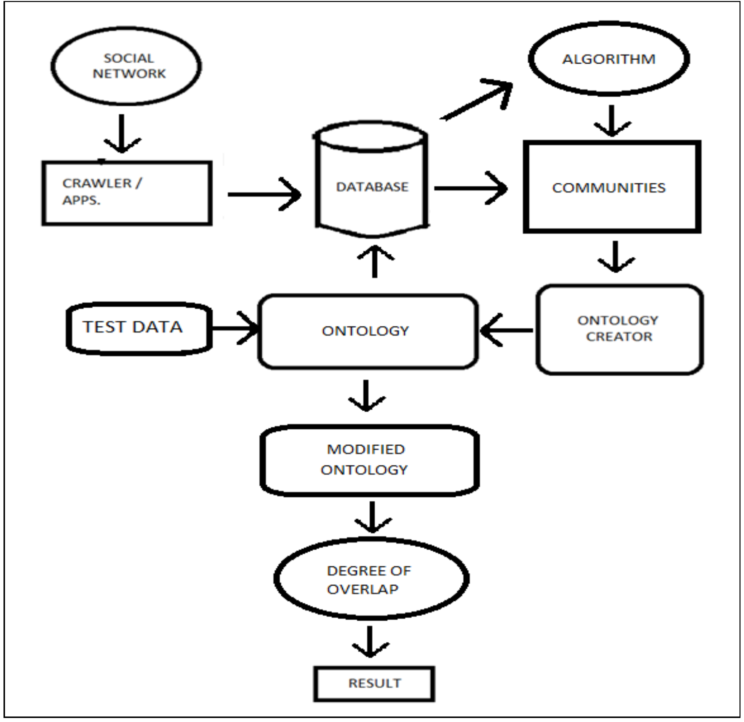


Figure 1. Model

3.2. Ontology Design

The tool Protégé was used for creation of Ontology [16] [18]. Ontology was designed with the features/properties as classes. The object properties in the ontology were defined so that it reflected the relation between the users and their features. This ontology was stored so that it could be reused for any further experiments using different algorithms. Here we used the MCL algorithm for finding clusters. The ontology was then combined with these clusters. The following is the class hierarchy of the modified ontology and its object properties:

Class hierarchy:

- Thing
    - Users
    - City
    - State
    - Country
    - Movies
    - Music
    - Cluster

Object properties:

- hasHomeCity
- hasHomeState
- hasCountry
- hasLikedMovie
- hasLikedMusic

- isHomeCityOf
- isHomeStateof
- isCountryOf
- MovieLikedBy
- MusicLikedBy

The relations between users and its features were defined using the above object properties. For example, if the user U1 liked a movie M1 then the relation was defined as, U1 *hasLikedMovie* M1. The clusters obtained from the MCL algorithm were also mapped to their respective users in the same manner.

3.3. Observations:

Firstly, the degree of overlap was calculated. The degree of overlap for every feature with each cluster was calculated. It was defined as the ratio between the numbers of individuals of that feature belonging to cluster with the total number of individuals of that feature. For example, degree of overlap of feature 'City' for any cluster[i] is as follows:

Degree of overlap (city, cluster[i]) = (number of cities which have users in cluster[i]) / (total number of cities) ...(1)

Then a total degree of overlap was calculated for every feature. It was defined as the ratio between the number of individuals in all the clusters with the product of total number of individuals of that feature and total number of clusters.

Total Degree of overlap (city) = (total number of cities in all the clusters) / (total number of cities) X (number of clusters) ...(2)

The number of individuals in a certain cluster was found out using the DL Query in the tool Protégé. The tool Protégé used its inbuilt 'reasoner', in our case 'Fact++' to help DL Query. For example, the Table 1 below shows calculation of the values of degree of overlap for the feature home 'City' in the ontology using formula (1).

Table 1 Degree of Overlap (City)

	CITY		
	NUMBER OF CITIES	**TOTAL**	**DEGREE OF OVERLAP**
cluster 1	8	70	0.114286
cluster 2	8	70	0.114286
cluster 3	59	70	0.842857
cluster 4	4	70	0.057143
cluster 5	2	70	0.028571
TOTAL	81	350	0.231429

In the same manner, the rest of the values in Table 2 were calculated using formula (2).

Table 2 Total Degree of Overlap

Degree of overlap	**City**	**State**	**Country**	**Movies**	**Music**
cluster 1	0.114286	0.2	1	0.37415	0.617902
cluster 2	0.114286	0.333333	1	0.609524	0.331088
cluster 3	0.842857	0.933333	1	0.462585	0.342637
cluster 4	0.057143	0.2	1	0.160544	0.117421
cluster 5	0.028571	0.066667	1	0	0.00385
TOTAL	0.231429	0.346667	1	0.321361	0.282579

We observed the following:

1. For each cluster, some of the features had their degree of overlap greater than the other features. As shown in Table 2, the degree of overlap for ‘cluster 1’ for the feature ‘city’ i.e. 0.11 is lesser than that for music in the same cluster i.e. 0.61.

2. The total degree of overlap for some features was lesser than that for other features. For example, the total degree of overlap for ‘City’ is the lower than the other features, i.e.0.23 as shown in Table 2. Whereas that of the feature ‘Country’ is ‘1’ in all the cases because every user had the same country. This showed the maximum overlap condition that was possible is all the data in cluster are similar and overlapping.

In the proposed method, the data from the social networks is extracted by web crawler/ apps/ any other tool and is stored in the database. Ontology was created capturing the features and properties of the extracted database. Since the ontology is used to represent different features of the dataset it can be used to represent the data itself for conducting any type of experiments.

4. RESULTS AND CONCLUSION:

The graph (Figure 2) below reflects the observations from Table 2. For each cluster, the bar with lowest height represents the feature which is more prominent than the others for that cluster. As shown in the graph, for cluster 1 the prominent feature is ‘City’ and in cluster 3 the prominent feature is ‘Music’. The total degree of overlap shows the feature which is prominent for formation of all the communities.

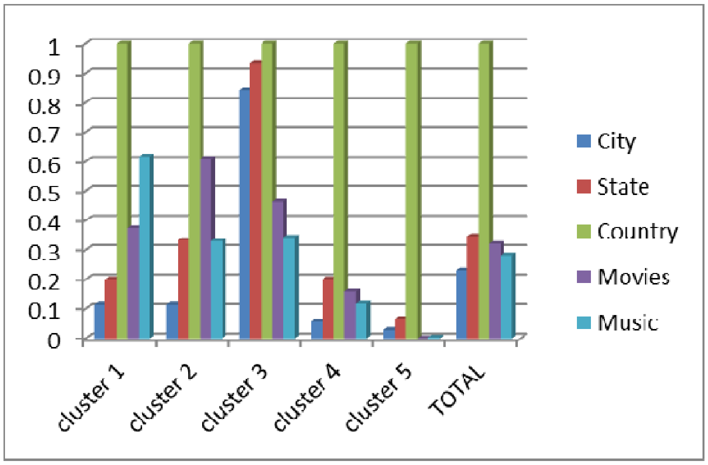


Figure 2. Degree of overlap for different clusters

Thus, the proposed method can be used to find which of the features are prominent in the formation of communities for different algorithms. Moreover, the ontology that was previously stored can be reused for different algorithms and also updated whenever needed for improving the results. Depending upon the prominent features we can use it for viral marketing, strategy planning, feature selection, etc. For future work, we plan to find sub-clusters among the already formed clusters. Then finding which features have affected the formation of these sub-clusters or sub-communities. This may give us knowledge as to how communities and their sub-communities are related with each other and on through which features.

REFERENCES

[1] G.Aghila et.al, “Ontology-based Web Crawler”, Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC’04), 2004.

[2] Peter Mika, “Ontologies are us: A unified model of social networks and semantics”, Lecture Notes in Computer Science (Springer/ The Semantic Web – ISWC 2005) Volume 3729, pp 522-536, 2005.

[3] Debajyoti Mukhopadhyay, Arup Biswas, Sukanta Sinha, “A New Approach to Design Domain Specific Ontology Based Web Crawler”, 10th International Conference on Information Technology (ICIT 2007), 289-291, 2007.

[4] Jason , “J. Jung, Jérôme Euzenat, “Towards Semantic Social Networks”, Lecture Notes in Computer Science (Springer/The Semantic Web: Research and Applications), Volume 4519, pp 267-280, 2007.

[5] Wu Chensheng, Hou Wei, Shi Yanqin, Liu Tong, “A Web Search Contextual Crawler Using Ontology Relation Mining”, International Conference on Computational Intelligence and Software Engineering (CiSE 2009) Page 1-4, 2009.

[6] Wu Peng, Li SiKun, “Social Network Visualization via Domain Ontology”, International Conference on Information Engineering and Computer Science (ICIECS 2009) DOI: 10.1109/ICIECS.2009.5362898, 2009.

[7] Liu Chen, Shan Wei, Zhang Qingpu, “Semantic Description of Social Network Based on Ontology”, Proceedings of the 2010 International Conference on E-Business and E-Government (ICEE ‘10), Pages 1936-1939, 2010.

[8] Morteza Jamalzadeh, Navid Behravan, “Using Semantic Web Ontologies for better inter-operability on social network sites”, IEEE International Conference on Control System, Computing and Engineering, Page 103-108, 2011.

[9] Reena Mishra,Shashwat Shukla,Deepak Arora, Mohit Kumar, “An Effective Comparison of Graph Clustering Algorithms via Random Graphs”, International Journal of Computer Applications (0975 – 8887) Volume 22– No.1, May 2011.

[10] Sam K. M. and Chatwin C.R., "Ontology-Based Text-Mining Model for Social Network Analysis", Proceeding of the Sixth IEEE International Conference on Management of Innovation and Technology, pp. 226 - 231, 2012.

[11] Moharram Challenger, “The Ontology and Architecture for an Academic Social Network”, International Journal of Computer Science Issues (IJCSI), Vol. 9, Issue 2, No 1, PP 22-27, SSN (Online): 1694-0814, 2012.

[12] Régine Lecocq, Étienne Martineau, Maria Fernanda Caropreso, “An Ontology-based Social Network Analysis Prototype”, IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), DOI: 10.1109/CogSIMA.2013.6523839, Page(s): 149 – 154, 2013.

[13] Bhaskar Biswas, Vijay Nayak and Harish Kumar Shakya, “Comparison of Algorithms for Social Networks using Ontology”, International Journal of Computer Applications, 85(13):31-34 January 2014.

[14] http://en.wikipedia.org/wiki/Ontology.

[15] http://en.wikipedia.org/wiki/Web_Ontology_Language.

[16] http://protege.stanford.edu.

AUTHORS

Vijay Nayak is a post graduate student at Department of Computer Science and Engineering, Indian Institute of Technology (BHU), Varanasi, India. His research interests are mainly on social networks and data mining.

Dr. Bhaskar Biswas is Assistant Professor at Department of Computer Engineering, Indian Institute of Technology (BHU), Varanasi, India. Dr. Bhaskar had done his Ph.D from the Same Institute in Web Mining. His research interest includes Data mining, Social Networks Analysis, and Web Mining.

EFFECT OF REFACTORING ON SOFTWARE QUALITY

Noble Kumari[1] and Anju Saha[2]

[1]USICT, Dwarka, Delhi, India
noblevashishta_127@yahoo.com
[2]USICT, Dwarka, Delhi, India
anju_kochhar@yahoo.com

ABSTRACT

Software quality is an important issue in the development of successful software application. Many methods have been applied to improve the software quality. Refactoring is one of those methods. But, the effect of refactoring in general on all the software quality attributes is ambiguous.

The goal of this paper is to find out the effect of various refactoring methods on quality attributes and to classify them based on their measurable effect on particular software quality attribute. The paper focuses on studying the Reusability, Complexity, Maintainability, Testability, Adaptability, Understandability, Fault Proneness, Stability and Completeness attribute of a software .This, in turn, will assist the developer in determining that whether to apply a certain refactoring method to improve a desirable quality attribute.

KEYWORDS

Metrics, Refactoring, Attributes & External software quality attributes.

1. INTRODUCTION

Refactoring is defined as the "process of improving the design of existing code by changing its internal structure without affecting its external behavior" [7, 8].The poorly designed code is harder to maintain, test and implement and hence the quality of software degrades. The basic goal of refactoring is the safe transformation of the program to improve the quality. The benefit of undertaking refactoring includes improvement of external software quality attributes.

In software program the word “smell” means potential problem in the code. In the refactoring cycle as the smell is found, refactoring methods are applied and code is improved. The cycle continues till we find the maximum efficient code [8].

The external software quality attributes like reusability, complexity, maintainability, testability and performance are dependent of the software metrics. Software metrics are used to predict the value of the software quality attributes. A large number of software metrics have been proposed which are quantifiable indicators of external quality attributes[22].The value of internal software quality metrics like Coupling Factor (CF), Lack of Cohesion of Method (LCOM), Depth of Inheritance Tree (DIT), Weighted Method per Class (WMC), Lines of Code (LOC) and Cyclomatic Complexity (Vg) are desired to be lower in a system whereas Attribute Hiding Factor (AHF) and Method Hiding Factor (MHF) are desired to be higher [20].

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 37–46, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4505

Refactoring changes the value of software metrics and hence the software quality attribute. Not all the refactoring methods improve the software quality, so there is need to find out the refactoring methods which improve the quality attributes [6].

The aim of this paper is to find out the effect of refactoring methods on the software metrics. From the relation between the software metrics and external quality attributes direct relation between refactoring methods and software quality attributes is derived.

The paper analyzes the effect of refactoring on the software quality attributes. The classification of refactoring methods is done for particular desired quality attributes and metrics set.

The study also shows that refactoring does not ensure to improve the software quality always. It has to compensate with some attributes to improve the other.

This paper is organized as follows. Section 2 describes the literature review. Section 3 and 4 explains about the research data and refactoring methods respectively. Section 5 explains about the analysis and result of refactoring methods. Section 6 and 7 explains the threats to validity and Conclusion respectively.

2. RELATED WORK

The goal of this paper is to find the effect of refactoring on the external software quality attributes, using software metrics. In this section we review the study of various researchers on the effect of refactoring on software quality attributes.

Cinneide, Boyle and Moghadam [1] studied the effect of automated refactoring on the testability of the software. The aim is to find the refactoring method which improves the cohesion metric and hence the testability of the software. Code-Imp platform is explored for the refactoring purpose and available metrics in the tool are applied. The survey is done with the volunteers where further testing is required to validate that automated refactoring improves the testability of the software.

Sokal, Aniche and Gerosa [2] took data from Apache software and applied refactoring on it. The authors randomly selected the fifty refactoring methods. They classified them in two groups according to their effect on cyclomatic complexity and analyzed the change in code after refactoring. Their studies show that refactoring does not necessarily decrease the cyclomatic complexity but increases the maintainability and readability of the program.

Alshayeb [6] assess the effect of refactoring on the external software quality attributes. The quality attributes taken were Adaptability, Maintainability, Understandability, Reusability and Testability. Code for refactoring is taken from the open source UMLTool,RabtPad and TerpPaint. The author applied different types of refactoring on the code and studied the effect of refactoring on the software metrics. From the relation between the software metrics and external quality attributes, the effect of refactoring is studied. The author found the inconsistent trend in the relationship of refactoring method and external quality attributes.

Elish and Alshayeb [3] studied effect of refactoring on testability of software. They used five refactoring methods: Extract Method, Extract Class, Consolidated Conditional Expression, Encapsulate Field and Hide Method. Chidamber and Kemerer metrics suite [17] is used to find the software metric values. The authors concluded that all the refactoring methods they used increase the testability except the Extract Class method.

Kataoka [5] used coupling metrics to find the effect of refactoring on the maintainability of the software. He proposed a quantitative evaluation method to measure the maintainability enhancement effect of program refactoring and helped us to choose the appropriate refactoring.

Stroggylos [28] analyzed the source code version control system logs of some of the popular open source software system. They found the effect of refactoring on the software metrics to evaluate the impact of refactoring on quality. The results found the increase in metric valued of LCOM, Ca and RFC which degrades the software quality. They concluded that refactoring does not always improve the software quality.

Shrivastava [29] presented a case study to improve the quality of software by refactoring. They took open source and with the Eclipse refactoring tool produced three version of refactored code. The results found that the size and complexity of a software decreases with refactoring and hence maintainability increases.

The study to find effect of refactoring on the software quality attributes has a wide scope. Fowler [7] has given 70 types of refactoring methods and each refactoring method can be linked to the various software quality attributes. So, our focus is to find the effect of fourteen randomly chosen refactoring methods on the various object oriented metrics and hence on the external software quality attributes.

The following quality attributes will be used in the study:

Maintainability: It is defined as the ease with which modification is made on set of attributes. The modification in the attributes may comprise from requirement to design. It may be about correction, prevention and adaptation [6].
Reusability: It is defined as the reusable feature of the software in the other components or in other software system with little adaptation [6].

Testability: It is defined as the degree to which software supports testing process. High testability requires less effort for testing.

Understandability: It is defined as the ease of understanding the meaning of software components to the user [6].

Fault proneness: Fault Proneness in the programs is more prone to the bugs and malfunctioning of the module.

Completeness: Completeness of the program refers for all the necessary components, resources, programs and all the possible pathways for execution of program [9].
Stability: Stability is defined in terms of ability of the program to bear the risk of all the unexpected modification [23].

Complexity: In an interactive system it is defined as the difficulty of performing various task like coding, debugging, implementing and testing the software.

Adaptability: Adaptability of the software is taken in terms of its ability to tolerate the changes in the system without any intervention from any external resource [26].

3. RESEARCH DATA

The classes used for research data in this paper are from an open source code JHotDraw7.0.6 [10]. Erich Gamma and Thomas Eggenschwiler are the authors of JHotDraw [10]. It has been developed as a quite powerful design exercise whose design is based on some well-known design patterns. We took 120 classes of JHotDraw7.0.6 and applied refactoring methods on it.

The aim of making JHotDraw an open-source project is:

- To refactor and hence enhance the existing code.
- To identify new refactoring and design patterns.
- To set it for an example of a well-designed and flexible framework.

4. REFACTORING METHODS

The refactoring methods applied in this paper are taken from the catalog defined by Fowler [7]. The following refactoring methods are applied [12, 18]:

1. Extract Delegate: This refactoring method allows extracting some of the methods and classes from a given class and added them to newly created class. The refactoring resolves the problem of the class which is big in size and performs much functionality. The name of newly created class is given by the user.

2. Encapsulate field: This refactoring allows modifying the access of data from public to private and generating getter and setter method for that field in the inner class.

3. The Replace Inheritance with Delegation: This refactoring allows removing a class from inheritance hierarchy, while maintaining the functionality of the parent class. In this refactoring a private inner class is made, that inherits the former super class. Selected methods of the parent class are invoked through the new inner class.

4. Replace Constructor with Builder method: The Replace Constructor with Builder refactoring helps hide a constructor, replacing it with the references to a newly generated builder class or to an existing builder class.

5. Extract Interface: Extract Interface is a refactoring operation that allows making a new interface with the members from the existing class, struct and interface.

6. Extract Method**: It** is a refactoring operation that allows creating a new method from the existing members of the class.

7. Push Member Down: The Push Members down refactoring allows in relocating the class members into subclass/sub interface for cleaning the class hierarchy.

8. Move Method: This refactoring allows moving a method from one class to another. The need of moving a method comes when the method is used more in other class than the class in which it is defined.

9. Extract Parameter: The Extract parameter refactoring allows selecting a set of parameters to a method or a wrapper class. The need of the refactoring comes when the number of parameter in a method is too large. The process of refactoring is done by delegate via overloading method also.

10. Safe Delete: The Safe Delete refactoring allows you to safely remove the class, method, field, interface and parameter from the code with making the necessary corrections while deleting.

11. Inline: The Inline Method refactoring allows putting the method’s body into the body of its caller method.

12. Static: This refactoring is used to convert a non-static method into a static. This allows the method functionality available to other classes without making the new class instance.

13. Wrap Method Return Value: The Wrap Return Value refactoring allows selecting a method and creating a wrapper class for its return values.

14. Replace Constructor with Factory Method: The Replace Constructor with Factory Method refactoring allows hiding the constructor and replacing it with a static method which returns a new instance of the class.

The tool used for refactoring and studying the values of software metrics is Intellij Idea: IDE for java and a reliable refactoring tool. It knows about code and gives suggestion also as a tip. Refactoring methods referenced from Fowler [7] are available in this tool [12]. All the object oriented metrics can be computed using the tool. The tool gives the module, package, class, project and method level program metrics. It is available and easy to use. Table 1 shows the "wrap method return value" refactoring using the tool.

Table 1. Example of "Wrap Method Return value" Refactoring using the IntelliJ Idea tool.

Before Refactoring	After Refactoring

public newadded getScrollPane()
{ if (desktop.getParent() instanceof JViewport)
{ JViewport viewPort =
(JViewport)desktop.getParent();
if (viewPort.getParent() instanceof
JScrollPane) return new
newadded((JScrollPane)
viewPort.getParent());
}
return new newadded(null);
}
	public noble getScrollPane()
	{ if (desktop.getParent() instanceof
	JViewport)
	{JViewport viewPort =
	(JViewport)desktop.getParent();
	if (viewPort.getParent() instanceof
	JScrollPane) return new noble(new
	newadded((JScrollPane)
	viewPort.getParent()));
	return new noble(new newadded(null));
	}

In inner class name "noble" is made and then refactoring is performed in the tool.

5. ANALYSIS AND RESULTS

The focus of this paper is to find the effect of refactoring methods on the software quality attributes and hence categorized the refactoring methods according to particular quality attributes and software metric domain. The values of object oriented software metrics is found before and after refactoring. The result is analyzed according to the value of the software metrics.

To focus our study on the category of refactoring methods, we set up the following hypothesis. For each hypothesis, H0 represents null hypothesis and H1 represents the alternative hypothesis of H0.

Hypothesis 1
H0: Refactoring does not improve software adaptability.
H1: Refactoring improves the software adaptability.

Hypothesis 2
H0: Refactoring does not improve software maintainability.
H1: Refactoring improves the software maintainability.

Hypothesis 3
H0: Refactoring does not improve software Understandability.
H1: Refactoring improves the software Understandability.

Hypothesis 4
H0: Refactoring does not improve software Reusability.
H1: Refactoring improves the software Reusability.

Hypothesis 5
H0: Refactoring does not improve software Testability.
H1: Refactoring improves the software Testability.

Hypothesis 6
H0: Refactoring does not decrease software Complexity.
H1: Refactoring decreases the software Complexity.

Hypothesis 7
H0: Refactoring does not make software less Fault Proneness.
H1: Refactoring makes the software less Fault Proneness.

Hypothesis 8
H0: Refactoring does not improve software Stability.
H1: Refactoring improves the software Stability.

Hypothesis 9
H0: Refactoring does not improve software Completeness.
H1: Refactoring improves the software Completeness.

For validating all the hypothesis of this paper the relation between the values of software metrics and Refactoring methods is given below in Table 1. Where '↓' shows decrease in the value of metric,'↑' means increase in the value of the metric and '-' shows no change in the value of metric.

Table 2. Relation between Refactoring methods and software quality metrics.

Refactoring Method	**WMC**	**Vg**	**LOC**	**NOM**	**CBO**	**LCOM**	**DIT**	**MPC**	**CCavg**	**AHF**	**AIF**	**CF**	**MHF**	**MIF**	**RFC**
Extract Delegate	↑	↓	↓	↑	↑	↓	↓	↑	↑	↓	↓	↓	↑	↓	↑
Encapsulate Field	↑	↑	↑	↑	↑	↑	↑	↑	↓	↑	↓	–	↓	↑	↑
Inheritance To Delegation	↑	↓	↑	↑	↑	↓	↓	↓	↑	↑	↓	↑	↑	↓	↑
Extract Interface	↓	↓	↑	↑	↑	↑	↑	↑	↑	–	↑	↓	↓	↓	↑
Extract Method	↑	↓	↑	↑	↑	↑	↑	–	↓	–	–	–	↑	↓	↑
Push Method Down	↓	↓	↑	↑	↓	↑	↑	↑	↓	↑	↓	↓	↑	↓	↑
Move Method	↑	↓	↑	–	↑	↑	↑	↑	↑	–	↑	–	↑	–	↑
Extract Parameter	↑	↓	↑	↑	↑	–	↑	–	↓	–	–	–	↑	↓	↑
Safe Delete	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↑	↓	↓	↓
Inline	↑	↑	↑	↑	↑	↑	–	↑	↑	↑	↓	↑	↓	↓	↓
Static	–	↑	↑	↑	–	↑	↑	–	↑	–	↓	–	–	↑	↓
Wrap Method	↓	↓	↑	↑	↓	↓	↓	↓	↓	↑	–	↓	↑	↓	↓

Return Value															
Replace Constructor with factory method	↑	↓	↑	↑	↓	↑	↑	–	↓	–	↓	↓	↑	↑	↓
Replace Constructor with Builder	↓	↓	↑	↑	↓	↓	↓	↓	↓	↑	↓	↓	↑	↓	↓

After Analyzing Table 2, it is concluded that the following methods give desirable result for every metrics [20] and hence improves the quality attribute of software:

1. Wrap Method Return Value

2. Static method

As indicated in hypothesis, we are attempting to find out the refactoring method which improves a particular category of software metrics. The metrics are divided according to the type of impact they make on the software. Table 3 summarizes the relation between metrics and their categories.

Table 3. Relation between metrics and their Categories.

Category	**Attributes**	**Method**	**Coupling /Cohesion**	**Inheritance**
MOOD[27]	AHF, AIF	MHF, MIF ,PF		MIF, AIF
C & K[17]	LCOM	LCOM,WMC, RFC	CBO	DIT
Li and Henry[19]		MPC ,NOM	MPC	

The various refactoring methods show random effect on the metric values. So, we classify the refactoring methods according to the desirable effects they make on the categories of Table 3: attributes, methods, coupling/cohesion and inheritance based metrics. From Table 2 and Table 3 the analysis result is shown in Table 4.

Table 4. Desirable refactoring for the particular category of metrics.

Category	**Refactoring Method**
Attributes	Inheritance to delegation, Wrap return value method and Constructor to Builder
Methods	Wrap Method Return Value
Coupling/Cohesion	Safe Delete ,Replace constructor with builder method , Replace constructor with factory method and Wrap Method Return Value
Inheritance	Extract Delegate, Inline, Safe Delete and Inheritance To Delegation

We used the previously published research work to make the correlation between the software metrics and external quality attributes. We used work of Dandashi [9] to assess the adaptability, maintainability, understandability and reusability quality attributes. The following table summarizes the relationship between software metrics and external quality attributes which can be helpful to find out the direct effect of refactoring on the external software quality attributes.

In this relationship (+) shows positive correlation, the attributes improve as the metric value increases, (-) shows negative correlation, the attributes degrade as the metric value decreases and (0) shows neutral effect.

Table 5. Relation between metrics and external quality attributes.

External Quality	**DIT**	**CBO**	**RFC**	**WMC**	**NOM**	**LOC**	**LCOM**
Adaptability[9,6]	-	-	-	+	0	+	0
Maintainability[22,16,9,25,6]	-	-	-	+	-	+	-
Understandability[9,16,6]	-	-	-	+	0	+	-
Reusability[22,9,16,6]	+	-	-	+	0	+	-
Testability[21,22,6]	-	-	-	-	-	-	-
Complexity[21]	+	+			+	+	+
Fault Proneness[22,24]	+	+	+	+		+	+
Stability[23]	-	-	-	-			-
Completeness[9]	-	-	-	+	-	+	0

To validate the hypothesis, we took the relation between Table 2 and Table 5 and come to the following conclusion:

Table 6. Particular refactoring method for certain quality attributes.

Refactoring Method	**Quality Attribute**
Wrap Return value	Testability
Safe Delete	Adaptability, Understandability, Less fault proneness and Stability
Replace Constructor with Builder method	Stability

1. “Wrap Return value” refactoring improves testability of the program.
2. “Safe Delete” makes program more adaptable, understandable, less fault proneness and stable.
3. “Replace Constructor with Builder method” makes program more stable.

From Table 2 and Table 5, we found that for other quality attributes inconsistent results are coming where some metrics values are needed to be ignored to improve the quality to certain limit.

1. “Wrap return method” makes the program less fault proneness if increased LOC effect is ignored.
2. “Wrap Return Method” makes system more adaptable when WMC is ignored.

Summing up the analysis part, we concluded that from Table 6 there are few refactoring methods which improve certain quality attributes and hence Hypothesis 1, Hypothesis 3, Hypothesis 5, Hypothesis 7 and Hypothesis 8 are rejected.

From the analysis part of Table 2 “wrap return method value” refactoring changes most of the metric values to desirable state and hence to certain limit improves every quality attribute. Therefore the Hypothesis 2, Hypothesis 4, Hypothesis 6 and Hypothesis 9 are rejected.

6. THREATS TO VALIDITY

There are some limitations to extend the result to general case. There are possible numbers of threats to validity as the few selective classes are taken from the project. The results may vary when implemented on the whole system and when the scenario is changed. We have applied the refactoring on class level not on the system level.

Another possible threat is the correlation between the internal metrics and the external software quality attributes; we have not put validation from our side and directly took the result of previous research.

7. CONCLUSION

Refactoring methods are applied to improve the software quality attribute but the effect of refactoring on particular quality attribute is still ambiguous. In this paper, we applied fourteen refactoring methods and noticed that they effect randomly on different software quality attributes. We classified the refactoring methods which improve a set of metrics which belongs to the attribute, method, coupling, cohesion and inheritance category of software. We focused on different external quality attributes, which are Reusability, Complexity, Maintainability, Testability, Adaptability, Understandability, Fault Proneness, Stability and Completeness and found the effect of refactoring methods on them. By looking at the results, we found that there are few refactoring methods which particularly improve a certain quality attributes of software, which can help the developer to choose them. Our work concludes that refactoring improves the quality of software but developers need to look for the particular refactoring method for desirable quality attribute.

Future research can also test and verify the result on bigger projects and can come up with general relation between refactoring and quality attributes.

ACKNOWLEDGMENT

I would like to acknowledge the support and guidance of my teachers for the excellent guidance throughout the entire work. I am thankful to university resource center for providing the resources for the work.

I would like to thank my parents and friends for all their support during my studies.

REFERENCES

[1] Cinnéide, Mel Ó., Dermot Boyle, and Iman Hemati Moghadam, (2011), "Automated refactoring for testability" , In Software Testing Verification and Validation Workshops (ICSTW), IEEE Fourth International Conference, pp. 437-443, IEEE.

[2] Francisco Zigmund Sokal, Mauricio Finavaro Aniche and Marco Aurelio Gerosa, (2013), “Does The Act Of Refactoring Really Make Code Simpler?, A Preliminary Study”.

[3] Elish, Karim O., and Mohammad Alshayeb. (2009), "Investigating the Effect of Refactoring on Software Testing Effort" In Software Engineering Conference, APSEC'09, Asia-Pacific, pp. 29-34, IEEE.

[4] Bruntink, Magiel, and Arie van Deursen, (2006), "An empirical study into class testability", Journal of systems and software 79, no. 9, pp. 1219-1232.

[5] Kataoka, Y., Imai, T., Andou, H. and Fukaya, T., (2002), "A quantitative evaluation of maintainability enhancement by refactoring", Software Maintenance, Proceedings International Conference, pp.576-585.

[6] Mohammad Alshayeb, (2009), “Empirical Investigation Of Refactoring Effect On Software Quality”, Volume 51, Issue 9, Pages 1319-1326, Elsevier.

[7] M.Fowler, K. Beck, J. Brant, W.Opdyke and D. Roberts, (1999), “Refactoring: Improving the Design of Existing Code”, Addison Wesley.

[8] W.C Wake, (2003), “Refactoring Workbook”, Addison Wesley.

[9] Dandashi Fatma, (2002) "A method for assessing the reusability of object-oriented code using a validated set of automated measurements", In Proceedings of the 2002 ACM symposium on applied computing, pp. 997-1003, ACM.

[10] www.jhotdraw.org.

[11] www.sourceforge.net.

[12] www.jetbrains.com

[13] Opdyke, William F., (1990) "Refactoring: An aid in designing application frameworks and evolving object-oriented systems", In Proc. 1990 Symposium on Object-Oriented Programming Emphasizing Practical Applications (SOOPPA).

[14] IEEE, (1991), Std. 610.12 – IEEE Standard Glossary of Software Engineering Terminology, The Institute of Electrical and Electronics Engineers.

[15] ISO/IEC, (1991), 9126 Standard, Information Technology – Software Product Evaluation – Quality Characteristics and Guidelines for their Use, Switzerland, International Organization for Standardization.

[16] Kayarvizhy, N. and Kanmani, S., (2011) "Analysis of quality of object oriented systems using object oriented metrics," Electronics Computer Technology (ICECT), 3rd International Conference on, vol.5, no., pp.203-206.

[17] Chidamber, S.R and Kemerer, C.F., (1994) “A metrics suite for object oriented design," Software Engineering, IEEE Transaction, vol.20, no.6, pp.476-493.

[18] Www. Refactoring.com.

[19] Li, W and Henry, S., (1993) "Maintenance metrics for the object oriented paradigm," Software Metrics Symposium, Proceedings, First International, pp.52-60.

[20] Daniel Rodriguez and Rachel Harrison, (2001),“An Overview of Object-Oriented Design Metrics”.

[21] Khalid, Sadaf, Saima Zehra and Fahim Arif, (2010) "Analysis of object oriented complexity and testability using object oriented design metrics", In Proceedings of the 2010 National Software Engineering Conference, ACM.

[22] Srivastava, Sandeep, and Ram Kumar, (2013) "Indirect method to measure software quality using CK-OO suite." In Intelligent Systems and Signal Processing (ISSP), 2013 International Conference on, pp. 47-51, IEEE.

[23] Elish, Mahmoud O. and David Rine, (2003) "Investigation of metrics for object-oriented design logical stability", In Software Maintenance and Reengineering Proceedings, Seventh European Conference on, pp. 193-200, IEEE.

[24] Basili, Victor R., Lionel C. Briand and Walcélio L. Melo, (1996) "A validation of object-oriented design metrics as quality indicators", Software Engineering, IEEE Transactions on 22, no. 10, pp. 751-761.

[25] Jehad Al Dallal, (2013) "Object-oriented class maintainability prediction using internal quality attributes", Information and Software Technology 55, no. 11.

[26] Subramanian, Nary, and Lawrence Chung, (2001) "Metrics for software adaptability", Proc. Software Quality Management (SQM 2001).

[27] Abreu, Fernando B, (1995) "The MOOD Metrics Set," Proc. ECOOP'95 Workshop on Metrics.

[28] Stroggylos, Konstantinos, and Diomidis Spinellis., (2007) "Refactoring--Does It Improve Software Quality?” proceedings of the 5th International Workshop on Software Quality, IEEE Computer Society.

[29] Vasudeva Shrivastava, S.,and V. Shrivastava. (2008) "Impact of metrics based refactoring on the software quality: a case study". TENCON 2008 IEEE Region 10 Conference, IEEE.

[30] Sharma, Tushar., (2012), "Quantifying Quality of Software Design to Measure the Impact of Refactoring”. Computer Software and Applications Conference Workshops, IEEE 36th Annual.

MANAGING UNCERTAINTY OF TIME IN AGILE ENVIRONMENT

Rashmi Popli[1] and Priyanka Malhotra[2] and Naresh Chauhan[3]

[1]Assistant Professor,Department of Computer Engineering,
YMCAUST, Faridabad
rashmimukhija@gmail.com
[2]M.Tech,Scholar,Department of Computer Engineering,
YMCAUST, Faridabad
jayant.malhotra1@gmail.com
[3]Professor, Department of Computer Engineering, YMCAUST, Faridabad
nareshchauhan19@gmail.com

ABSTRACT

Agile software development represents a major departure from traditional methods of software engineering. It had huge impact on how software is developed worldwide. Agile software development solutions are targeted at enhancing work at project level. But it may encounter some uncertainties in its working. One of the key measures of the resilience of a project is its ability to reach completion, on time and on budget, regardless of the turbulent and uncertain environment it may operate within. Uncertainty of time is the problem which can lead to other uncertainties too. In uncertainty of time the main issue is that the how much delay will be caused by the uncertain environment and if the project manager comes to know about this delay before, then he can ask for that extra time from customer. So this paper tries to know about that extra time and calculate it.

KEYWORDS

Slack, Optimistic time, Pessimistic time, Probability of Delay

1. INTRODUCTION

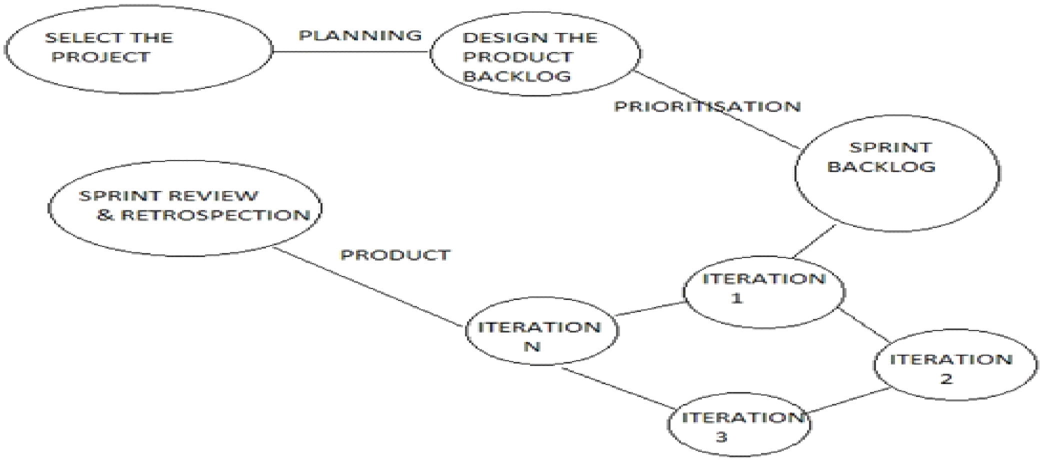
Agile software development methodologies become increasingly popular as the word spreads about the benefits they provide under certain project conditions. A key characteristic of any agile approach is its explicit focus on time estimation and business value for the clients. The goal of time estimation is typically to develop potentially shippable product. The accurate estimations of time is critical for both developer and customer. Ignorance of estimation methods may cause serious effects like exceeding the budget, poor quality and not right product. The key factor which is causing the problem in estimation is time uncertainty, so there is a need of some mechanism for minimizing uncertainty of time.

In Section II the life cycle of agile is described. Section III describes what uncertainty is. In Section IV related work in this field is discussed, section V proposes scenario calculation of slack time and uncertainty in time in agile. In Section VI shows evaluation and results of proposed algorithm, section VII concludes the paper.

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 47–56, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4506

2. AGILE LIFE CYCLE

Meaning of Agile is “moving quickly”. When applied to software development, it means that delivering the software that meets the customer requirements in shortest possible time. The success and failure of a software project is determined by accurate estimation. This is the process to calculate the time that will be taken to finish the project, cost of the project and the effort required to complete the project.



Estimation is very important task as improper estimation may lead to failure of the software project. It may also increase the budget of the customer and sometimes the nature of the project is also affected. The estimation in the Agile environment is a difficult task due to the changing requirements. The figure 1 shows the Agile software development life cycle. Agile software lifecycle is an iterative process where software is ready at each iteration but can always be improved in next iteration. Or in agile terms a part of the project is ready at each iteration and that part itself can be improved at each iteration or can be free from bugs at each iteration.

3. UNCERTAINTY

Meaning of Agile is “moving. In releasing a particular plan or **user story**, it is needed to fix a set of release dates and then determine how much functionality can be achieved by those dates. Also this can be done by deciding the functionality first and then deriving the release date. In either case the functionality value is accessed against the cost and time to develop the system, In previous used methods cost and time both get neglected in assessing the functionality which lead to uncertainty in both time and cost. Also the size of the user story is not certain. These all factors leads to poor estimation in agile project and hence time uncertainty. In this paper there is an attempt to find solution to problem of these uncertainties and calculating the percentage of uncertainty.

4. RELATED WORK

McDaid, D. Greer, F. Keenan, P.Prior, P. Taylor, G. Coleman[8] proposed a set of key practices which includes the practice, termed “Slack”, of only signing up to for what the team is confident of achieving. Within this approach it is always possible to add more stories, time permitting, thus delivering more than was actually promised. This practice acknowledges that there is a significant amount of uncertainty in the estimated time to complete releases.

Rashmi Popli, Anita and Dr. Naresh Chauhan[5] proposed a common life cycle approach that is applicable for different kinds of teams. This approach describes a mapping function for mapping of traditional methods to agile method.

Siobhan Keaveney and Kieran Conboy[1], gave the study of the applicability of current estimation techniques to more agile development approaches by focusing on four case studies of agile method use across different organizations. The study revealed that estimation inaccuracy was a less frequent occurrence for these companies. The main estimation techniques used were expert knowledge and analogy to past projects also the Component of the process; fixed price budgets can prove beneficial for both developers and customers, and experience and past project data should be documented and used to aid the estimation of subsequent projects.

S.Bhalerao and Maya Ingle[4] presented the study of both traditional and agile estimation methods with equivalence of terms and differences. This study investigated some vital factors affecting the estimation of an agile project with scaling factor of low, medium and high. Also, an algorithm Constructive Agile Estimation Algorithm (*CAEA*) is proposed for incorporating vital factors.

Daniel D. Galorath[3] proposed a 10-step estimation process that begins by addressing the need for project metrics and the fundamental software estimation concepts. It shows how to build a viable project estimate, which includes the work involved in the actual generation of an estimate, including sizing the software, generating the actual software project estimate, and performing risk/uncertainty analysis.

5. PROPOSED WORK

Client gives customer the requirements in form of user stories, and a backlog is created with those requirements. The time required for completion of project depends upon size of the user story. It needs to be certain about time taken by the project to be completed. The uncertainties in the timing is a big problem so project takes some marginal time called as slack that will compensate for extra time taken in the project work other then the optimal development time. However there is no formula for calculating this Slack. The proposed algorithm and formula suggests that how to calculate the slack time. For calculations, the time taken as hours rather than days because that will lead to actual result.

Effective time per day for Sprint Related work = Average work day time - Time allocated for other activities.

For example Average work day time = 10 hours
Time Allocated for other activities = 5 hours
Emails and Phone: 1 Hrs
Lunch: 1 Hrs
Meetings: 2 Hrs
Bug fixes: 1 hrs
Available time for Sprint Related work = 5 hours

5.1 Proposed Diagram

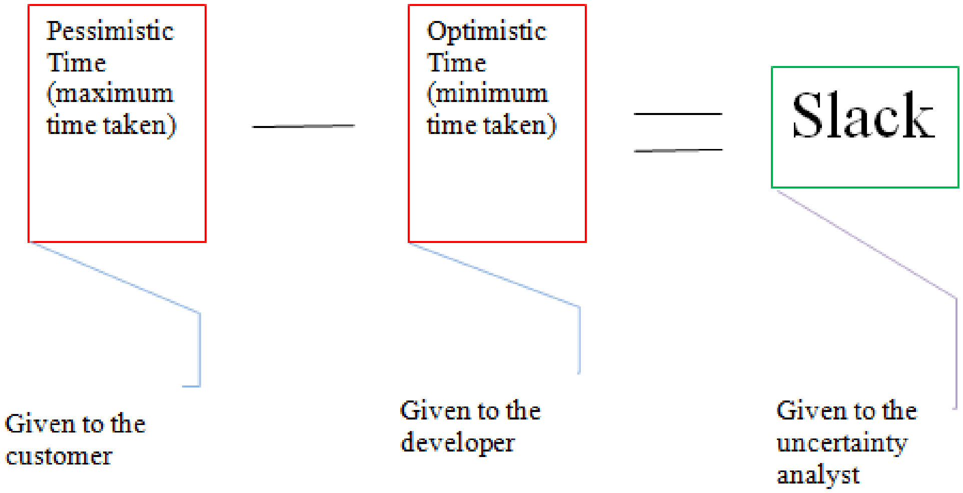

Pessimistic Time (maximum time taken)
Optimistic Time (minimum time taken)
Slack
Given to the customer
Given to the developer
Given to the uncertainty analyst

Figure 2: Showing what actually happening in this scenario of calculating slack

In the project there is a team of about 7 persons, one is project manager, four developers and two testers who are doing pair programming. For four developers total development time will be the ideal available time for development related work * 4.

The optimistic value of time period for one iteration is given to the developers so that they have to complete task in that period, and the pessimistic value is given to the customer or client. The calculation for pessimistic timing will be done by using probability of delay.

5.2 Proposed Formulae

The difference between these two timings will be helpful for exactly calculating the slack time.

1. ***Duration for developer*** *= =* $\sum_{i=1}^{n} Ti$ *[optimistic time] / work per time*
2. ***Duration for customer*** *= =* $\sum_{i=1}^{n} Ti$ *[pessimistic time] / work per time*
3. ***pessimistic time for each task*** *=(percent probability of delay *optimistic time)/100+optimistic time*
4. ***Slack time for a task*** *= duration for developer - duration for customer*
5. ***Time Uncertainty****= (slack/duration for developer)*100*
6. ***Total slack time****= (Total pessimistic time-Total optimistic time)/(Total working hours per day*Number of developers)*

5.3 Proposed Algorithm

The proposed algorithm explains the various steps involved in removing the uncertainty in time in agile environment.

1. Identify the tasks of each user story based on the requirements, suppose for each user story there are n number of tasks so for each user story the tasks are t_1, t_2, t_3.......t_n.
2. Identify the optimistic time for each task and probability of delay for each task. Metric for Optimistic time value is hours and probability of delay will be in percentage.

3. Calculate the time (pes) for each user story by using the formula $\sum_{i=1}^{n} T_{\mathrm{I}}$ [pessimistic time] where "i" denotes the tasks and pessimistic time for each task =(percent probability of delay *optimistic time)/100+optimistic time
4. Compute the overall Slack time for all the user stories using the formula Slack time = duration for developer - duration for customer
   Duration for developer= $\sum_{i=1}^{n} Ti$[optimistic time] / Effective working hrs
   Duration for customer = $\sum_{i=1}^{n} Ti$[pessimistic time] / Effective working hrs
5. Calculate the uncertainty percentage which is
   = (slack time/duration for developer)*100

5.4 Proposed Activity Diagram

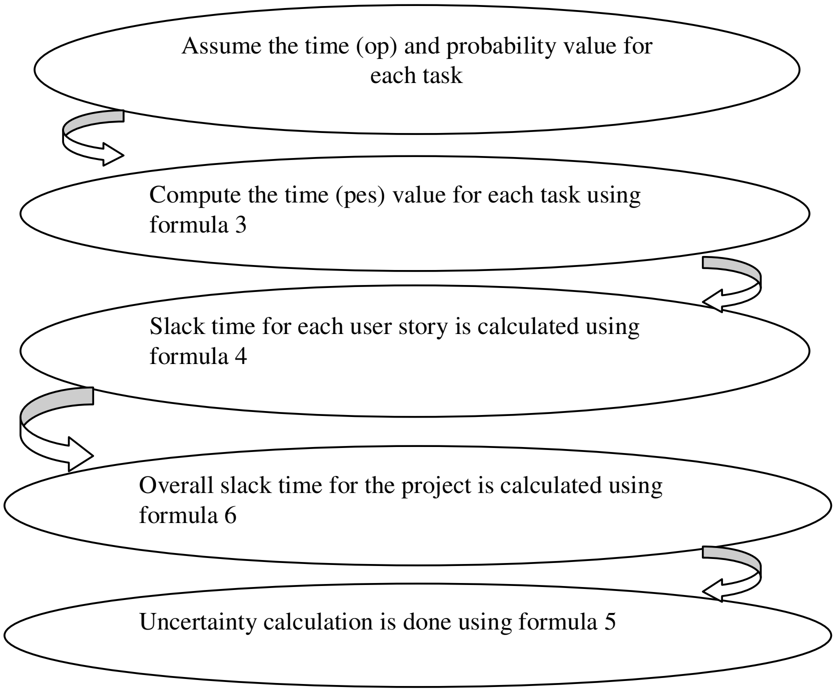

Assume the time (op) and probability value for each task
Compute the time (pes) value for each task using formula 3
Slack time for each user story is calculated using formula 4
Overall slack time for the project is calculated using formula 6
Uncertainty calculation is done using formula 5

Figure 3: Steps involved in calculation of slack and uncertainty

6. EVALUATION AND RESULTS

In this section the feasibility of our algorithm is shown by calculating the estimated values of pessimistic time using probability of delay, value of slack and percentage of uncertainty for a project. We had considered the number of effective working hours as 5 per day .In this section the feasibility of our algorithm is shown by a case study in which the values are being calculated. We have considered the user stories of project which is **Letters of Credits**; the client is HP-client. A graph can be drawn taking the optimistic time and pessimistic time values against each other, the bar graph and line graph both show the difference between both the values.

Table 1: **Showing tasks of the project under taken and related values of optimistic time and probability of delay**

N O.	DEVELOPMENT TASKS	OPTIMISTIC TIME	PROBABILITY OF DELAY
1	FSD REVIEW EXPORT OPENING	20	30
2	FSD REVIEW EXPORT REVIEW	10	20
3	FSD REVIEW OPENING CREATION	26	20
4	FSD IMPORT REVIEW CREATION	42	10
5	FSD EXPORT REVIEW CREATION	33	10
6	PH-2 REQUIREMENT STUDY	20	30
7	FSD EXPORT REVIEW CREATION	30	10
8	FSD IMPORT OPENING SIGN-OFF	45	20
9	FSD IMPORT REVIEW SIGN-OFF	50	10
10	FSD EXPORT OPENING SIGN-OFF	54	30
11	FSD EXPORT REVIEW SIGN-OFF	32	10
12	DEVELOPMENT REVIEW 1	34	20
13	DEVELOPMENT REVIEW 2	65	10
14	BACKUP ARCHIVE	30	20
15	PROJECT MONITORING	40	30
16	CONFIGURATION	23	20
17	UNIT TESTING	25	10

18	INTEGRATION TESTING	35	10
19	SYSTEM TESTING	25	20
20	TRAINING	20	20
21	PH1 UAT	5	30
22	PH1 UAT SIGN OFF	15	20
23	PH2 UAT SIGN OFF	20	10
24	PH2 DEVELOPMENT REVIEW1	20	20
25	PH2 DEVELOPMENT REVIEW 2	19	20
26	PH2 UAT	7	10

Table 2: showing user stories along with there corresponding associated tasks

NO.	USER STORY	ASSOCIATED TASKS
1	SRS	1-5
2	SRS REVIEW	6
3	DOCUMENTATION	6,7
4	FSD REVIEW IMPORT OPENING	8-10
5	NON-FUNCTIONAL DATA COLLECTION	20-23
6	PRE ENGAGEMENT SUPPORT	13-16
7	REWORK CODING	11,12
8	TESTING	17-19
9	CODE REVIEW	24,25
10	GO LIVE SUPPORT	21,26

Table 3: user stories with their optimistic time and calculated pessimistic time

USER STORY NO.	**TOTAL OPTIMISTIC TIME**	**TOTAL PESSIMISTIC TIME**
1	131	152
2	20	26
3	50	59
4	149	179.2
5	60	70.5
6	158	187.1
7	66	76
8	85	96
9	39	46.8
10	12	14.5

6.1 Numerical analysis

Total slack time= (Total pessimistic time-Total optimistic time)/Total working hours per day*Number of developers
Here total working hours per day are=5 hours
Number of developers=4
Total Slack time = (907.1-770)/20=6.85
Now the Percent uncertainty in the project= (Slack time/Pessimistic time)*100= (6.85/45.35)*100=15.10%

6.2 Graphs

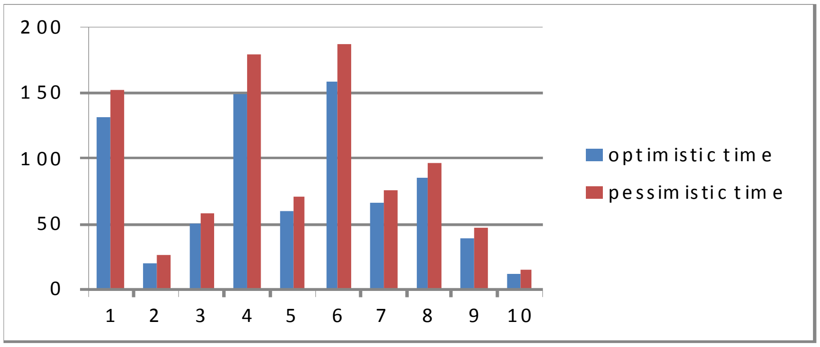


Figure 4: Bar graph representation of optimistic and pessimistic timing values

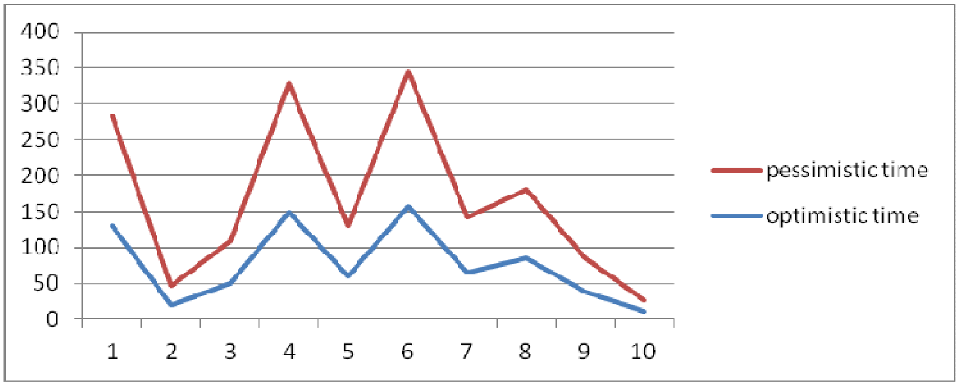


Figure 5: Difference between optimistic time and pessimistic time shown by lines

7. CONCLUSION

This uncertainty percentage tells us that by this percentage there are the chances of the project to get faulty. If this percentage value is much higher, then this may lead the developers think about removing the uncertainties first and then start the project. The slack value helps in improving the project as we can remove some causes which are leading to the delay. Also the value of slack is important because now the project manager is sure about the completion time of the project and the relations with the customer can be improved as customer is well satisfied with the project completion at the given dead line. The approach used in the paper is very easy to understand and do not need any hard and fast calculations. After having exact values it would be easier for the manger to read out the reports because the work is shifted from the theoretical portion to the exact numerical data. Hence this paper is an approach to solve the problem of uncertainty of time.

REFERENCES

[1] Siobhan Keaveney and Kieran Conboy , "Cost Estimation and agile development projects" Product-Focused Software Process Improvement, 435-440

[2] Du, G., McElroy, J., &Ruhe, G. (2006). Ad hoc versus systematic planning of software releases–a three-staged experiment. Product-Focused Software Process Improvement, 435-440.

[3] Daniel D. Galorath, "The 10 Step Software Estimation Process For Successful Software Planning, Measurement and Control"galorth incorporated 2006.

[4] S. Bhalerao and Maya Ingle, "Incorporating Vital Factors In Agile Estimation Through Alogorithmic Method" International Journal of Computer Science and Applications, Ó2009 Technomathematics Research Foundation ,Vol. 6, No. 1, pp. 85 – 97

[5] Rashmi Popli, Anita and Naresh Chauhan. "mapping of traditional software development methods to agile methodology"

[6] Logue, K., &McDaid, K. (2008). Agile Release Planning: Dealing with Uncertainty in Development Time and Business Value. Engineering of Computer Based Systems, 2008. ECBS 2008. 15th Annual IEEE International Conference and Workshop on the (pp. 437-442). IEEE.

[7] McDaid, K., Greer, D., Keenan, F., Prior, P., Taylor, P., & Coleman, G. (2006). Managing Uncertainty in Agile Release Planning.Proc. 18th Int. Conference on Software Engineering and Knowledge Engineering (SEKE'06) (pp. 138-143).

[8] Logue, K., &McDaid, K. (2008). Agile Release Planning: Dealing with Uncertainty in Development Time and Business Value. Engineering of Computer Based Systems, 2008. ECBS 2008. 15th Annual IEEE International Conference and Workshop on the (pp. 437-442). IEEE.

[9] RashmiPopli, NareshChauhan," Research Challenges of Agile Estimation" Journal of Intelligent Computing and Applications" July- Dec 2012.

[10] RashmiPopli, NareshChauhan," "Scrum- An Agile Framework", International Journal of Information Technology and Knowledge Management (IJITKM) ISSN: 0973-4414", Vol-IV, Number-I, 20 Aug 2010.

AUTHORS

Rashmi Popli is pursuing her Ph.D in Computer Engineering from YMCA University of Science & Technology, M.Tech(CE) from M.D University in year 2008,,B.Tech(IT) from M.D University in the year 2004.She has 9 years of experience in teaching. Presently she is working as an Assistant Professor in department of Computer Engineering in YMCA University of Science &Technology, Faridabad, Haryana, India. Her research areas include Software Engineering, Software Testing and Software Quality.

Priyanka Malhotra is a research scholar pursuing her M.tech in computer engineering (Compter Networks) from YMCA University of Science & Technology, Faridabad, Haryana, India. Completed B.tech (Cse) from Kurukshetra Unviersity, Kurukshetra

Dr. Naresh Chauhan received his Ph.D in Computer Engineering in 2008 from M.D University, M.Tech(IT) form GGSIT,Delhi in year 2004,B.Tech(CE) from NIT Kurukshetra in 1992.He has 22 years of experience in teaching as well as in industries like Bharat Electronics and Motorola India Pvt. Ltd. Presently he is working as a Chairman and Professor in the department of Computer Engineering ,YMCA University of Science and Technology, Faridabad, Haryana, India.. His research areas include Internet Technologies, Software Engineering, Software Testing and Real Time Systems.

.

.

RECONSTRUCTION OF TOLLAN-XICOCOTITLAN CITY BY AUGMENTED REALITY (EXTENDED)

M. en C. Martha Rosa Cordero López, M. en C. Marco Antonio Dorantes González.

Escuela Superior de Cómputo,
I.P.N, México D.F.
Tel. 57-29-6000 ext. 52001 y 52021.
mcorderol@ipn.mx
mdorantesg@ipn.mx

ABSTRACT

Work In Terminal presents the analysis, design, implementation and results of Reconstruction Xicocotitlan Tollan-through augmented reality (Extended), which will release information about the Toltec capital supplemented by presenting an overview of the main premises of the Xicocotitlan Tollan city supported dimensional models based on the augmented reality technique showing the user a virtual representation of buildings in Tollan phase.

KEYWORDS

Databases, Visual Programming, Augmented Reality, Virtual Reconstruction, Archaeological Site.

1. INTRODUCTION

The Archeological Zone of Tula, is the most important of tolteca culture. It's conformed by a set of buildings with a religious symbolism, for example the Central Altar, the Coatepantli (wall of Snakes), Burnt Palace, ball games and the Tzompantli. The National Institute of Anthropology and History (INAH) opened in Tula a museum about Tolteca Culture.

Thanks to science and technology have made great discoveries and changes in society over time [1]. Nowadays, it is possible to combine virtual and real objects within the same environment, to create supplemented views from somewhere that people are viewing [2]. This process is called Augmented Reality (AR) [3] 4].

The project applies AR together with archaeological knowledge of the Tollan-Xicocotitlan city, in Tula, Hidalgo. In order to obtain a system that models projecting three-dimensional (3D) showing the architecture of the buildings constructed there and complemented with written information about each campus.

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 57–67, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4507

This paper describes how to reconstruct a building using three-dimensional models design based on AR. We validate our approach using the buildings of the Xicocotitlan city of Tollan.

Indeed, the reconstruction allows to display any building that is in ruins, presenting it in three-dimensional model of the structure information. Besides the system provides support in order to have better idea of the constructed buildings in the past. The system can be applied in various places, with desired display information from the Toltec culture, a museum, exhibition or educational institutions where they are taught subjects related to the teaching of the Hispanic cultures.

AR supports markers located on a fixed surface, such as the ruins of a temple, a pyramid or a display in a museum. Such markers are detected by the input devices that should be placed in a specific position for the brand to be recognized and to be viewed on virtual model for the whole environment.

The viewer appreciates a virtual city by means of the system which builds boom in architectural, or reconstruction of events occurred in the past.

As a result, AS does not absolve the user from the reality, all experiences become more interesting for visitors, who are immerse in a particular event occurred in the past.

2. PRINCIPLE

Augmented reality

AR adds virtuality to real parts, staying in the world where they below, and enhancing them with other elements, without disconnecting altogether without leave to travel from other virtual environments. Moreover isolates virtual reality world in which we live, i.e., the individual is disconnected from the real environment and go onto another world.

AR environment adds more information to the real one observed by users.

AR and virtual reality are related each other. Firstly, we want to clarify some concepts that distinguish both (see Figure 1). Virtual Reality (VR) is defined as "a computer-generated environment, interactive, three-dimensional in which the person is immersed" [8]. While AR provides efficient location to interact with the space. VR provides experiences where space and time can be completely controlled, allowing users to interact simultaneously on multiple types of spaces (RA&VR). At the same time, the environments can be beneficial for a large number of applications, like architecture, chemistry, marketing.

Virtual Reality. Augmented Reality.

Fig. 1 Virtual vs Augmented Reality

Operation of reality increased

Three basic key elements of AR are:

- Display (output),
- Location of virtual objects in the real world (registration),
- Methods interaction (input).

Multimedia information plays a principal role of character, handled through photos, videos, extra-sounds, and with the three dimensions models, to present virtually acclimate.

The main point in the development of an AR application is a motion tracking system. RA technique relies on "Bookmarks" or an array of markers within the field of view of the cameras, such that a computer system has a benchmark on which superimpose images.

These markers are predefined by the system and the pictogram can be unique for each image to be superimposed or simple shapes, such as picture frames, or textures within the field of view.

A computer system can be more intelligent, able to recognize simple shapes, such as the floor, objects like chair, table, simple geometric shapes, to name a cell phone on the table that can be used with a brand or even with the human body that can be used with the same purpose. The following figure shows an example of the marks described in the previous paragraph.

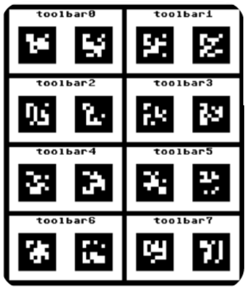


Fig. 2 Joint monitoring card use in typical RA

Artoolkit

ARToolKit is a set of libraries for C / C + +, that are useful for building AR applications. It includes a number of computer vision techniques for video capture and pattern searching for capturing images.

Users believe that only in the real world it is possible to perform transformation on objects. But, we want to show that it is possible to perform this kind of transformation on virtual objects. Users are able to see this transformation via the camera or by capturing them, taking into account position, size, orientation, and lighting, as these objects would be perceived by the user in the real world, if they were actually there. This is possible thanks to the libraries of ARToolKit.

A square-shaped templates is used, which are composed of a black square with a white square four times smaller at its center, and a simple picture inside the white square (see Fig. 2). The application, using the features and functionality provided by ARToolKit, is able to spot one of these templates in the video images captured.

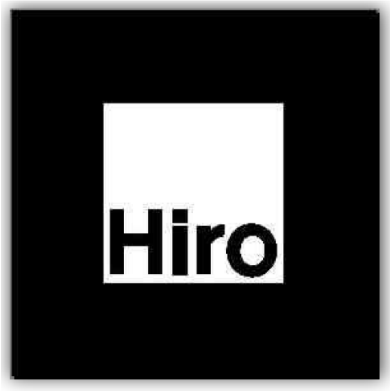


Fig. 2 RA mark detected by ARToolKit

Once a template is detected within an image, studying the orientation, position, and size of the template, the application is able to calculate the relative position and orientation of the camera, and relative to the template. Using this information, you can draw the corresponding object on the captured image by means of the ARToolKit external libraries (e.g., GLUT and OpenGL). In this way, the object appears on the template, in the position, orientation, and size corresponding to the view taking by the camera (see Fig. 3). Due to the number of possibilities are big, the application take a decision to select one, taking into account the information of other various operations.

Operation of an application artoolkit

The basic operations of ARToolKit application are as follows:

- Firstly, a frame captures real world through a camera.

- The image is modified taking into account a certain threshold value. Thus, the pixels whose intensity exceeds the threshold are converted into white pixels. The remainder is transformed into black pixels.
- They seek and find all black frames as the existing brands in the image.
- Compare the inside of the frame with the markings of the stored information.
- If the shape of the brand and the brand analyzed stored matches, using the size and orientation information of the mark stored for comparison with the brand that has been detected in order to calculate the position and orientation of the camera relative to the mark, and stored in an array.
- The matrix establishes the position and orientation of the virtual camera (processing chamber view), equivalent to a transformation of the coordinates of the object to draw.
- Having put the virtual camera in the same position and orientation as the real camera, the virtual object is drawn on the brand, and renders the resulting image is displayed, containing the image of the real world and the virtual object superimposed, aligned on mark
- It performs the same process with the following frames.

Nyartoolkit

ARToolKit, NyARToolkit provide a trail marker based AR. However, the software has been optimized for easy portability among different programming languages. In order to develop an application running AR on different platforms and operating systems, NyARToolkit libraries are the best option.

NyARToolkit include some key features, like:

- Bookmarks AR based tracking.
- Support for desktop and mobile platforms.
- Scoreboard optimized and enhanced survey.

Blender

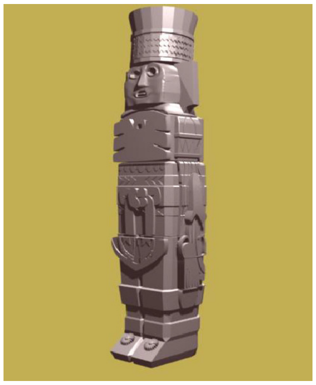
Blender is a tool for creating mainly modeling animation and creation of three-dimensional graphics. Some features are:

- It is a cross-platform tool, is free software and complies with the functionality provided similar commercial tools.
- Along with the animation tools including inverse kinematics, armature or grid deformations, loading and particle vertices static and dynamic.
- Features interactive games such as collision detection, dynamics and logic recreations

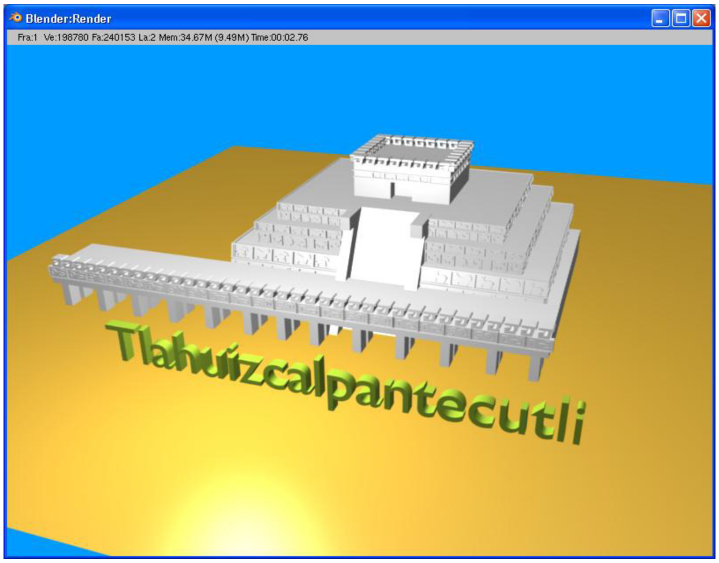
3. Experimental Results

Modeling with Blender

We have made some 3D models of the city of Tollan. The following images have developed in Blender:

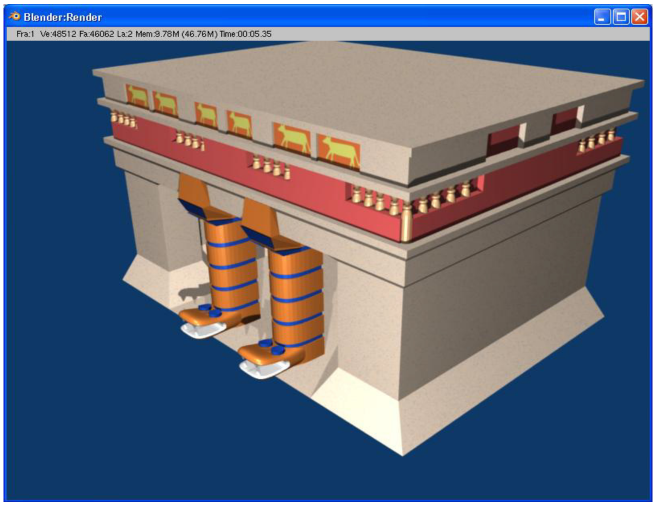


3D Model of “Atlante de Tula”



```
Fra:1  Ve:198780 Fa:240153 La:2 Mem:34.67M (9.49M) Time:00:02.76
```

Prueba de la Pirámide B



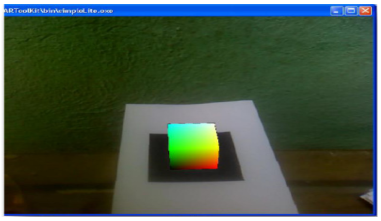
```
Fra:1  Ve:48512 Fa:46062 La:2 Mem:9.78M (46.76M) Time:00:05.35
```

Tlahuizcalpantecuhtli Temple.

Tests with ARToolKit

Various tests were performed to understand the operation of ARToolKit, like markers included within the environment of the working tool.

We used a VRML file to check the brand recognition. ARToolKit is responsible for recognizing the associated brand and rendering a three-dimensional model. The result is as follows.

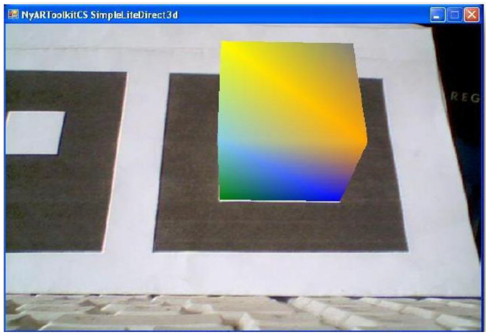


Test art whit ARToolKit

Note the superposition of a three dimensional object (a cube) on the mark before the House RA.

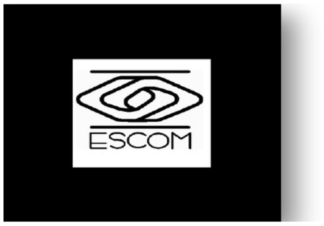
Tests NyARToolkit

Test was conducted in NyARToolkit development environment C #. where it was possible to load a three-dimensional model on screen. The result is shown in the following diagram.

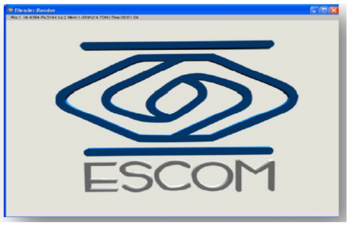


Test whit ARToolKit

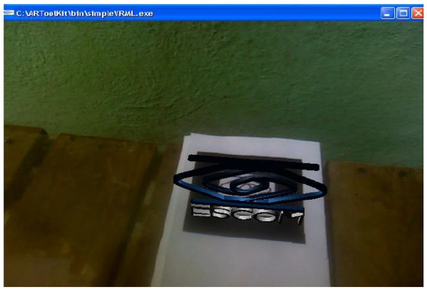
After understanding the operation of the libraries were established own brands of RA and generated three-dimensional models that would be superimposed on these markers. The results are shown below:



Own brand of RA.



Three Dimensional model



Final Result of RA using own brand and model.

Development

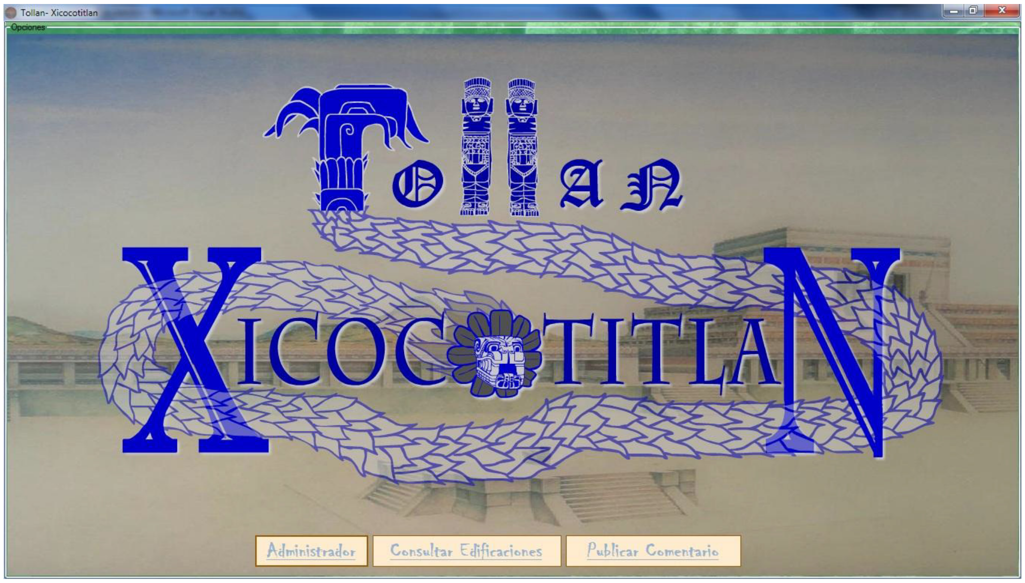
The system is divided into two main modules User and Manager.

The user module is in charge for presenting 3D models of each building, in this module, users can visualize a pyramid in 3D and can also comment on the experience that let them use this type of system.

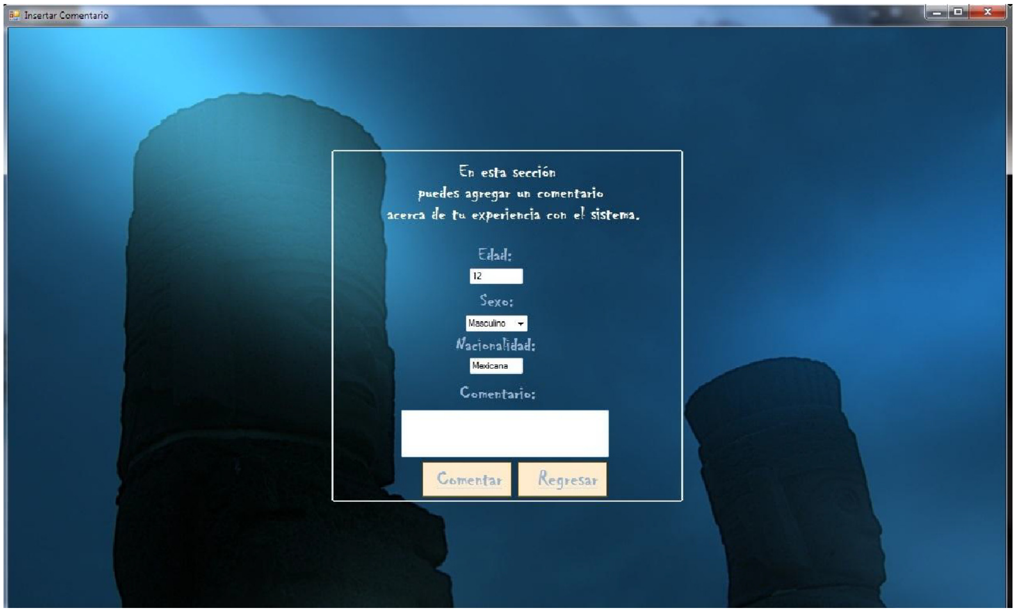
The Administrator module allows administrators to upload new handling system, just as you can modify the information associated with each building, this section administrators perform the query of comments made by users of the system.

Here are some screens that make up the system and a brief description as presented.

Main menu



Add comment


Insertar Comentario

En esta sección puedes agregar un comentario acerca de tu experiencia con el sistema.

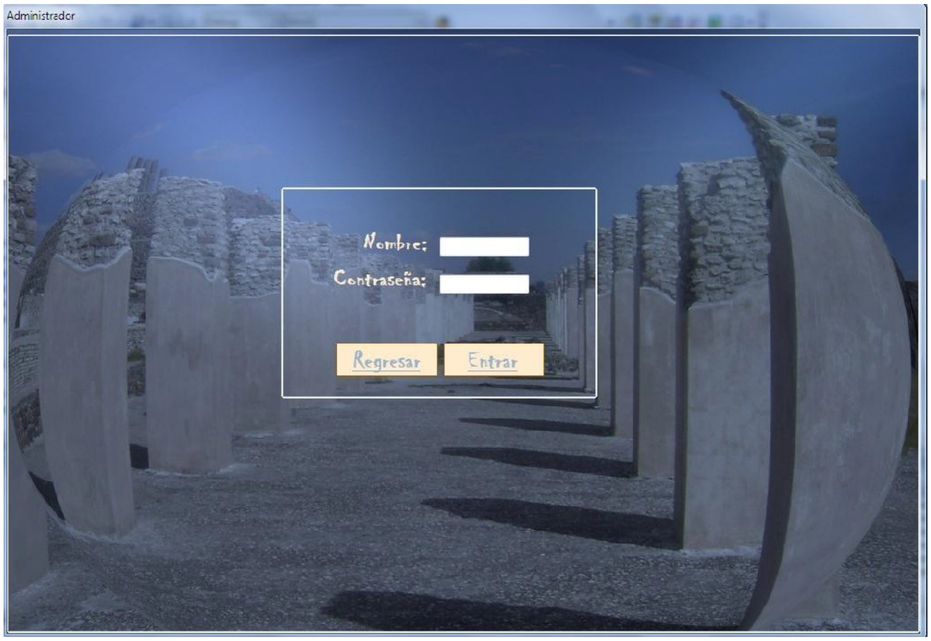
Edad: 12

Sexo: Masculino

Nacionalidad: Mexicana

Comentario:

Manage Menu


Administrador

Nombre:

Contraseña:

4. CONCLUSIONS

Our approach fulfills with the aim of presenting three-dimensional models of the major archaeological sites of the city of Tollan Xicocotitlan. Augmented Reality technology has been used to present a model to show the marks of RA defined for the system and having the display city in its architectural boom, achieving user interactivity, in a nice and easy way to manipulate objects.

By means of our approach, it is possible to travel through archeology museums, exhibitions or in the same archaeological site as presented to the general public or as ancient cultures and civilizations had been developed. Old civilization can be shown its culture.

To validate our application, we choose Xicocotitlan Tollan, that was one of the most important cities in the history of Mexico and served as the basis for the development of other cultures, as the Mayan culture.

AR places virtual objects in a real environment, allowing users to get a view of what is supplemented watching and with the possibility to transform these virtual objects, such as observing the virtual object from different perspectives views.

The aim of augmented reality is to set virtual objects of the real world, complementing what the user is watching and he can manipulate the virtual objects. In this case, Augmented Reality presents an interactive way to know the architecture of the Archeological Site Tollan, making a friendly system for the user to enrich the knowledge about this Culture.

ACKNOWLEDGEMENTS

The authors thank the school of Computing National Polytechnic Institute (ESCOM-IPN, México), for the economical support and the facilities provided for the development of this research work.

REFERENCES

[1] RODRIGUEZ, Yanet; SOLER, Dulce. La ciencia y la tecnología en América Latina: su impacto en el desarrollo de la medicina natural y tradicional. En línea. Centro Nacional de Sanidad Agropecuaria (CENSA); Grupo de Desarrollo y Biotecnología Industrial.

[2] GG, Zaira. Entre lo tangible e intangible, RA (Realidad Aumentada) + Realidad + Entorno. Un acercamiento a una Realidad Aumentada. [Diciembre 2009].

[3] KATO, H., BILLINGHURST, M. "Marker tracking and hmd calibration for a video-based augmented reality conferencing system.", In Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality (IWAR 99), [Octubre 1999].

[4] MONSALVE Manuela; CASTILLON, Adriana; CUARTAS, José. Exploración Teórica de la realidad aumentada para determinar su incidencia en el diseño visual. [Julio 2007].

[5] KENDALL, Kenneth E. KENDALL, Julie E. Análisis y diseño de sistemas. 3° ed. México, Prentice Hall, 1997.

[6] SCHACH, Stephen R.; Ingeniería de software clásica y orientada a objetos, 6ª ed. Mëxico 2006.

[7] PAREDES G, Blanca. Tula, Hidalgo Zona Arqueológica. [Centro INAH Hidalgo; Departamento de Difusión].

[8] SILVA Eliud. Encuesta a públicos de museos 2008-2009. [Sistema de Información Cultural (sic) Coordinación Nacional de Desarrollo Institucional].

[9] PRESSMAN, Roger S. Ingeniería de Software: Un enfoque Práctico. Cuarta Edición. s.l. : McGraw-Hill.

AUTHORS

M. Sc. Marco Antonio Dorantes González. Was born at Córdoba, Veracruz on 28 June, 1968. He had done his graduation in Electronics from ITO, Veracruz, México in 1990. After that he had completed his M. Sc. Degree in Computing in CINVESTAV in 1996 and M. Sc. of computing technologies in CIDETEC-IPN in 2008, research professor of ESCOM (IPN). He has been research Professor since 1996. He is interested in: Mobile Computing, Software Engineering, Data Bases. He has directed more than 70 engineering degree theses. Technical reviewer of interested areas books of publishers (McGraw Gill, Thompson, Pearson Education), He has participated in several research projects and has held some administrative positions in the IPN, also has experience in the industrial sector in the area of instrumentation and electronics; has done graduate studies in some fields, he has participated in several television programs and publications in scientific journals.

M. Sc. Martha Rosa Cordero Lopez. Was born at México D.F on 25 March, 1972. He had done his graduation in degree informatics from ITO, Veracruz, México in 1994. After that he had complete his Master Science Degree in Computing in CINVESTAV (IPN) in 1996, Master of computing technologies in CIDETEC-IPN in 2008, research professor of ESCOM (IPN) since 1995, her areas of interest are: Software engineering, Mobile Computing, Data Bases, affective computing, she has been director of in more than 70 theses to date, technical reviewer of interested areas books of publishes (McGraw Gill, Thompson, Pearson Education, among others). He has participated in various research projects and has held various administrative positions in the IPN also has experience in the private sector in the area of systems development; has done graduate studies in some areas, has been assistant mMcanager of technology intelligence unit in the technological development of the IPN, has participated in some television programs and publications in scientific journals.

INTENTIONAL BLANK

OFFICIAL VOTING SYSTEM FOR ELECTRONIC VOTING: E-VOTE

Marco Antonio Dorantes González,
Martha Rosa Cordero López, Jorge Benjamín Silva González

Escuela Superior de Cómputo I.P.N México D.F.
Tel. 57-29-6000 ext. 52000 y 52021.

mdorantesg@ipn.mx, mcorderol@ipn.mx, jorge.ben.silva@gmail.com

ABSTRACT

This paper describes the Official voting system by electronic ballot: E-Vote, which aims to streamline primary electoral processes performed in the country, beginning with the District Federal benefits and improvements. The principal benefices are economic and ecological time, taking into account process security features and the integrity of the captured votes. This system represents an alternative to the currently devices and systems implemented in countries like Venezuela, Brazil and the United States, as well formalized as a prototype able to compete with others developed by the Institute Federal Electoral District (IEDF).

KEYWORDS

Biometrics, Privacy, Fingerprint, Security, Electronic voting, Voting, Vote.

1. INTRODUCTION

The use of computerized systems in electoral processes is not new. Although certain actions are still made by hand, others have sophisticated technology. For example, aggregation of results is typically done electronically, although remaining paper backing can be checked with the provided data.

Thus, the electronic voting studies normally do not cover the phases and the computing process. But the introduction of electronics in the electoral process kernel, is the moment at which citizen people emit their vote. Currently, this is done by introducing a paper sheet vote into an urn. It can be possible that such operation can be computerized. Precisely, our approach adopts narrowly this kind of electronic voting and analyzes various forms to perform it.

While a controlled environment, as current boxes, we can not exclude the possibility of immediate coercion, voting from home or from the workplace leaves the door open to possible extortion.

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 69–80, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4508

Electronic voting present many advantages compared to current processes vote. Are ecological, they make faster and more agile counts and ratings long are cheaper.

Despite all the benefits, many experts believe that the main vulnerability of electronic voting is the integrity of the vote, that is, the voter is satisfied that Your vote will be counted as the did. Having taken into account this problem, have sought various solutions to this, ranging from the total suspension of use of electronic voting to implementation and testing of better security systems.

That's why we propose, through a study of the problem, an accurate and economically viable solution. The proposed system aims to meet the security needs and counting of votes from a number of electronic modules. These modules will be presented below.

PRINCIPLE

Our methodology uses the spiral method. Basically, it consists in repetitive spiral series of cycles starting from the center (see Fig. 1). Usually, it is interpreted as within each cycle of the spiral method follows a waterfall, but it is not like this.

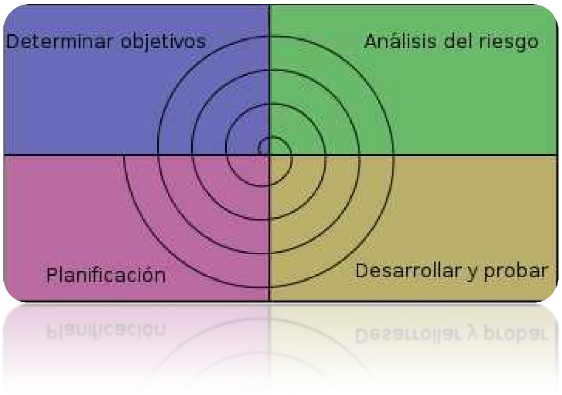


Fig. 1 The Spiral Methodology

The spiral evolutionary method combines the iterative nature of MCP model with the controlled and systematic aspects of the waterfall model, adding the risk management.

We designed our system with three layered architecture: Presentation, Business, and Data.

1) Presentation Layer

The presentation layer serves as the interface among users with the system. The layer processes carried out bio-data capture, deployment, and user data, as well as configuration ballots and the summary of the electoral exercise activities.

2) Business layer

The business layer takes the collected Data by means of the presentation layer, performing operations related to the voting exercise. This layer authenticates the processes by taking the voter registration and the vote counting. This is where the interfaces are contained in the management of database users and voting and voting and candidates.

3) Data Layer

This layer are contained in the database voters and users, as well as the candidates' database and votes. The layer is accessible only through the functions and processes established in the Business layer.

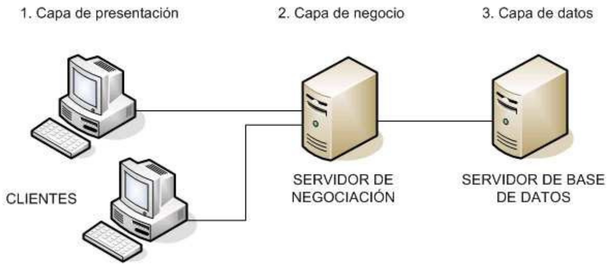


Fig. 2 Three layers architecture

Likewise, the Official Voting System for E-Vote electronic ballot box, our system uses specific modules to the following functions: recognition, authentication, digital signature, encryption, and decryption.

1. Identification RFID Module

The radio frequency identification system Frequency (RFID) stores and retrieves data using devices like remote labels, cards, transponders, or by RFID tags. The fundamental purpose of RFID tag is an object's identity (a unique serial number) using radio waves.

Fig. 3 Reader, cards and tags RFID

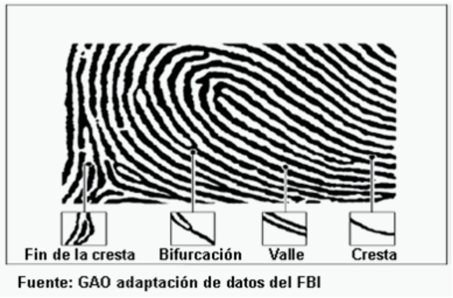
We used RXTX Java library to implement the RFID module. It serves as the communication interface between the serial and parallel ports with our development toolkit in Java or JDK.

Currently there is no way to access the serial or parallel ports with the standard Java API. This includes all versions up to 1.6 of the JDK. The communication of Java API provides the necessary support for the communication with the Serial and parallel port. Currently, CXR is the most complete implementation of this API.

2. Fingerprint Authentication digital Module

We think that human has ID cards integrated, easily accessible and virtually with unique design : the fingerprints.

Fingerprints allow to grab things more easily, because they have tiny "ridges and valleys" of skin. These “valleys and ridges” are very useful until nowadays. They are produced from the combination of genetic and environmental factors, like the fetus position at a particular moment, the exact composition and density of surrounding amniotic fluid.



Fuente: GAO adaptación de datos del FBI

Fig. 4: Features of the fingerprints

Fig. 5: Pattern of fingerprint

A fingerprint reader function performs two tasks:

1) To get a picture of the fingerprint.

2) To compare the pattern of “ridges and valleys” with image patterns stored in the traces DB.

The reading or the scanning capacitance are the two main methods for obtaining fingerprint images.

The module fingerprint recognition implemented in our system has been developed using the U.are.U 4000 model.

The Digital Person Sensor Company produces. A scanning device, offering an application programming interface or API that allows to integrate the following functions: the fingerprint reader, the fingerprint Registration, the fingerprint Verification, and the fingerprint Baja.

3. Data Security Module

Our system, **E-Vote** has a unique module for ensuring that certain information, such as database or public keys are known only to the charge of the polls.

Thanks to the cryptographic algorithm called RSA, is possible to generate two keys (public and private) and to encrypt/decipher these information. RSA uses the prime factorization and the arithmetic functions.

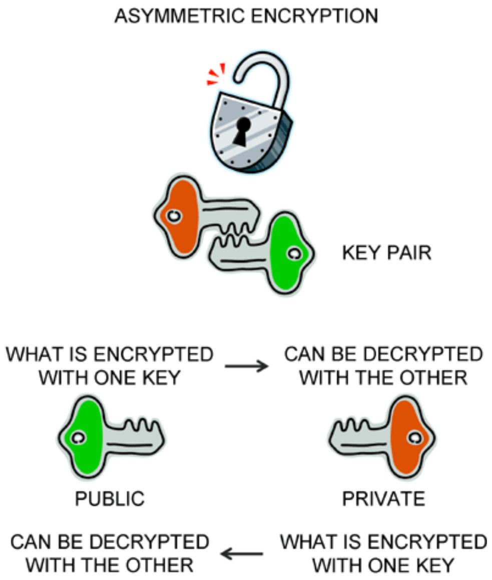

ASYMMETRIC ENCRYPTION
KEY PAIR
WHAT IS ENCRYPTED WITH ONE KEY
CAN BE DECRYPTED WITH THE OTHER
PUBLIC
PRIVATE
CAN BE DECRYPTED WITH THE OTHER
WHAT IS ENCRYPTED WITH ONE KEY

FIG. 5 Asymmetric encryption RSA

The the most safest and efficient cryptographic algorithm is RSA created by Rivest, Shamir, and Adleman . However, recently the RSA algorithm has suffered different attacks, because not only can be broken when using keys of 1024 bytes, although this problem can be easily solved just by extending the key size to 2048 bytes.

The digital signature is a mathematical scheme that perimenopausal verify the integrity and authenticity of a message. Thus, we can identify whether or not our the key database has not undergone any change over his transfer. Yielding a digital signature is a mathematical residue which is compared to the original that the representative can confirm if the message is corrupt.

The digital signature module can be made to different files and obtain a residue which we verified whether or not there are drawbacks.

Electronic voting and Operation Scheme: E-Vote

Official Voting System for E-Vote electronic ballot includes two operation schemes: -the overall system, including the involvement of the central shrine system and the electronic voting; -the operation of the scheme as such electronic ballot. The latter is located within the former.

1) Operation E-Vote System

The voting system is composed of four phases: History, Home, and End Exercise.

At the stage of **history** voters? will go to the central shrine system to be discharged by an authorized officer.

In this phase, the voters, together with personal data, provide the fingerprint. The official in turn, gives high associating RFID card with biometric voter registration, for electronic voting later use. In the initial phase, which is based on a streamlined electoral process, candidates are set to choose, and the criteria by which biometric references to candidates will be split to stay in each of the deployed electronic voting machines. These references divided fragments of the database, packed, encrypted and signed electronically to be stored on USB storage devices that can have its own security system fingerprint, to add additional insurance to the operation of the data transported.

In the exercise phase, once the polls and storage devices have been transferred to the place of voting, the officer assigned to the operation of the urn will identify it with your card, fingerprint and password. Only in this way will be able to set the time of voting, attempts to identify voters and begin and end the exercise.

To set the total time and start voting the same, voters will go to cast their vote by the scheme transaction narrate later.

At the end of time, the votes will be packed, encrypted and digitally signed to return to the central shrine described by the media before. The results of the choice of the particular electronic ballot box displayed on the screen.

In the final stages USB storage devices with the votes of the polls deployed will be checked, decrypted and imported by the central shrine in the database, which will host the final count and the issuance of the results.

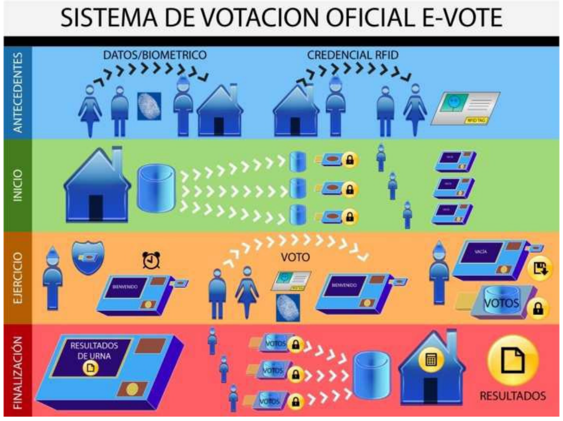


Fig 6 Operation E-Vote System

2) Operation of the electronic ballot box E-Vote

The operation of the electronic ballot box has action in the **Exercise** phase of the system, and comprises three phases: authentication, the election and confirmation of the vote, which will be described below.

In the **authentication** phase currently voter presents the identification card with RFID chip. Then, a view confirmed this card existence within the chunk of data belonging to the specific urn. For complete identification the fingerprints are submitted.

In the **choice** phase, an electronic template candidates is selected for the E**xercise** will be displayed on screen, the voter may well choose the desired or cast a blank vote.

Once the choice is performed, in the **confirmation** phase, the voter has the opportunity to correct your choice, when you are sure, the urn will tell you the number of ballot box and the exact time of your choice. Through this information, voter can ensure the vote at the end of the year without being publicly linked with it. Thus, fulfilling the electoral exercise as confidentiality.

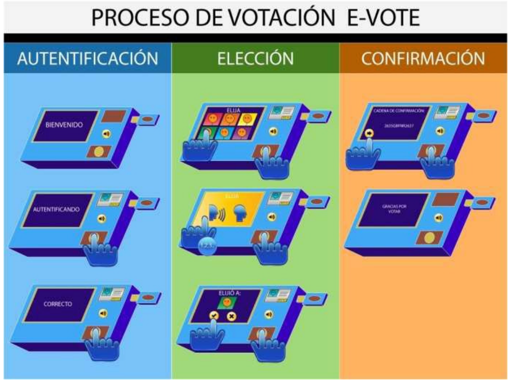


Fig 7 Operation of the electronic ballot box E-Vote

3. Experimental Results

According to the extension of this project as opposed to the time set for execution, prototypes were developed central shrine and electronic ballot on two laptop computers, one with a touch screen, which is a housing manufactured allow a modular host computer and reading devices and RFID fingerprint.

Tests were performed with data fragmentation modules described above, in which successful results obtained in the 90% of cases to compress, encrypt and sign the content, as well as a 95% success to verify electronic signatures, decrypt and sign the votes generated by the electronic ballot box.

Were tested for reading RFID tags in electronic voting, the maximum reading distance of 10cm was with direct line of sight and interference 5cm with housing, which was more than enough for identification purposes.

As for fingerprint reading we test registration and authentication, check that the device performed successful readings in the 98% of cases regardless of the lighting conditions. Regarding the identification all successful tests established with FAR 2.0% error.

As for the average voting time per person, the amount was accounted for 3 minutes, so as an exercise of the federal elections of 2006, having four polls for the 130, 488 boxes, and the total time of the vote would decrease by 40% from 10 hrs to 6 hrs. Looking at the whole electoral roll, which usually is not presented in its entirety.

From the economical point of view, each urn costs $10,000.00 M.N using 4 polls for the box, for 10 years of lifetime. By contrast, for electronic voting we consider investments of $521,952,000 M.N for electronic voting machines, $50,604,380 for the credentialing of the whole electoral roll. In addition, 7.49% and 0.72% of the average annual budget allocated $6966.44 IFE MDP.

Some results are

- The process of authentication, in this image shows the message of "welcome".
- The voter inserts the RFID in this urn.
- The voter inserts the fingerprint in this urn.

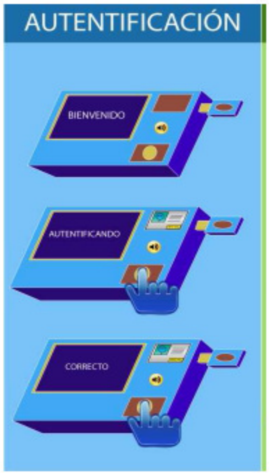
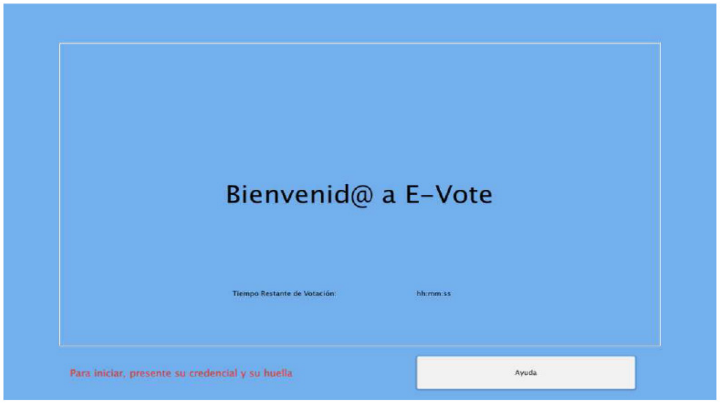


Fig 8 authentication

The following image shows the general project:



Bienvenid@ a E–Vote

Tiempo Restante de Votación: hh:mm:ss

Para iniciar, presente su credencial y su huella

Fig. 9 Welcome a "E-Vote"

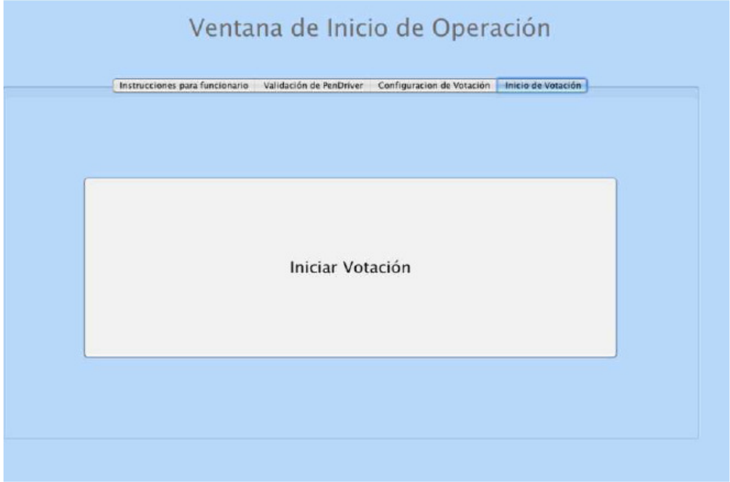
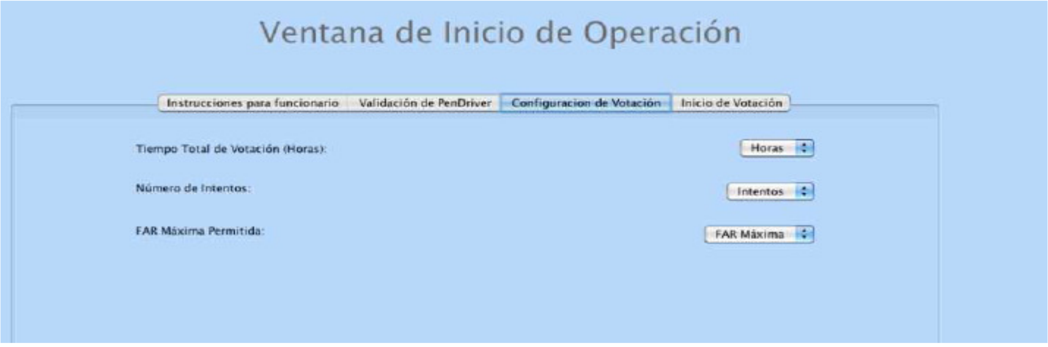

Ventana de Inicio de Operación
Instrucciones para funcionario
Validación de PenDriver
Configuracion de Votación
Inicio de Votación
Iniciar Votación

Fig. 10 Begin vote


Ventana de Inicio de Operación
Instrucciones para funcionario
Validación de PenDriver
Configuracion de Votación
Inicio de Votación

Tiempo Total de Votación (Horas): Horas

Número de Intentos: Intentos

FAR Máxima Permitida: FAR Máxima

Fig. 11 Window of begin of operation

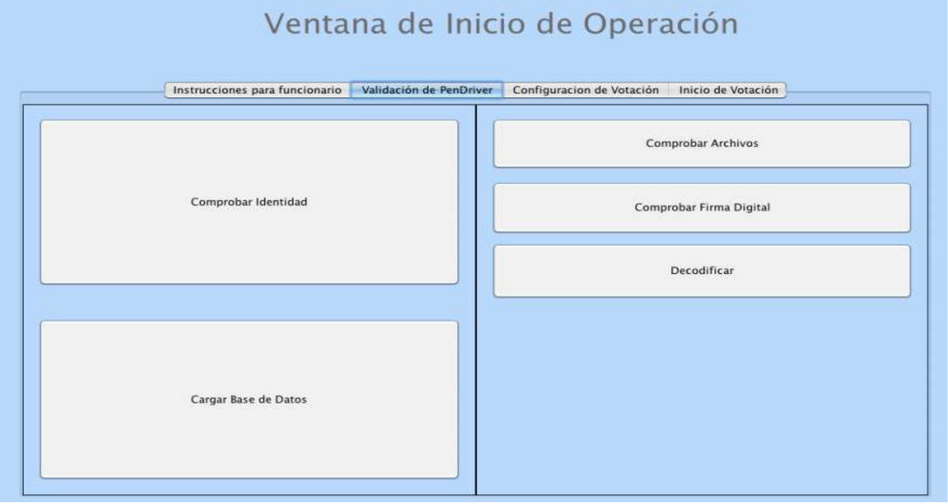


Fig. 12 Window of begin of operation

4. CONCLUSIONS

Electronic voting systems are gaining acceptance in the world. Our E-Vote system has been validated for the Mexican Electoral Institute, in the Coahuila state and the Jalisco state. They have declared to be satisfied with the results and they will plan to use the E-Vote system for the upcoming local elections. As a future research work, we will redesign the engineering process, including the identification and authentication mechanisms for RFID. The fingerprint will be complemented with better electronic devices.

During this study, we observed that these technologies are viable, affordable, and secure. This approach preserves the right of choice and national sovereignty for any country. Besides, this kind of systems, significantly reduce the alarming environmental impact, like that represented by the more than 60 tons of printed paper sheet vote as well as urn votes generated each elections without any dedicated computer system.

With regard to the restructuring of the voting process , though perhaps being the most ambitious of our project aspect , I have to say that is even more efficient than the current , carries in itself the same difficulty for implementation, however, sometimes a complete change for improvement is necessary.

In developing the system , we note that , despite the mistrust arising electronic voting procedures and counting, capture system itself may be more accurate , fast and economical long-term paper procedures . In addressing the risks posed in the analysis , we note that most of the causes of failure, as in the ballot paper , is represented by malice and desire premeditated to boycott the elections , a factor that is beyond the scope of any computer system in catastrophic events such as theft or destruction of equipment . There are also related failures inexperience or lack of user training , which in the case of our system are provided with the materials necessary training for operators and voting , as well as a simple and user-friendly interface.

ACKNOWLEDGE

The authors thank the School of Computing, National Polytechnic Institute (ESCOM-IPN, Mexico) for the economical support and the facilities provided for the development of this research work.

REFERENCES

[1] Kendall, Kenneth E.; Kendall, Julie E. Análisis y diseño de sistemas. 3° ed. México, Prentice Hall, 1997.

[2] Calculan 27.6 mdp operación de urnas electrónicas [online].
Available: http://www.elimparcial.com/EdicionEnLinea/Notas/Noticias/07102010/472309.aspx .

[3] El IFE registra cobertura del 99.57% de votantes en México [online].
Available:http://www.elsiglodetorreon.com.mx/noticia/427523.el-ife-registra-cobertura-del-99-57x-de-votan.html.

[4] El universal IFE estudia instalación de urnas electrónicas [online].
Available: http://www.eluniversal.com.mx/notas/660728.html.

[5] Hackers brasileños no pudieron vulnerar urnas electrónicas [Online].
Available:http://www.fayerwayer.com/2009/11/hackersbrasilenos-no-pudieron-vulnerar-urnas-electronicas/.

[6] El voto electrónico [online]. Available: http://www.larepublica.pe/archive/all/larepublica/20101017/6/node/295583/todos/15.

[7] Número de habitantes [Online]. Available: http://cuentame.inegi.gob.mx/monografias/informacion/df/poblacion/default.aspx?tema=me&e=09

[8] Ahorro millonario en comicios con la urna electrónica [Online]. Available: http://eleconomista.com.mx/notasimpreso/politica/2009/10/04/ahorro-millonariocomicios-urna-electronica-tellez.

[9] The History of Voting Machines By Mary Bellis [Online]. Available: http://inventors.about.com/library/weekly/aa111300b.htm

[10] Remote Voting Technology, Chris Backert e-Government Consulting

[11] U.S. Election Assistance Commission: 2005Voluntary Voting System Guidelines. Manual deprocedimientos para el sistema de votaciónvoluntaria de 2005.

[12] U.S. Federal Election Commission: DirectRecording Electronic. http://www.fec.gov/pages/dre.htm

[13] Instituto Tecnológico de Informática: Líneas I+D+I:Biometría [Online]. Available: http://www.t2app.com/index.php?derecha=ayuda/controlbiometrico.htm

[14] SAB - Sociedad Avanzada de Biometría. Available: http://www.sabiometria.net/

[15] Aplicación con Biometría Vascular. Available: http://www.kimaldi.com/aplicaciones/control_de_acceso/control_de_presencia_y_acceso_mediante_biometria_vascular_a_entornos_ofimaticos_y_de_pc_de_alta_seguridad

[16] Investigación y desarrollo de lectores biométricos [Online]. Available: http://www.by.com.es/es/lectores-deproximidad.html

[17] M1 – Biometrics About This Committee INCITS/M1, Biometrics Technical Committee [Online]. Available: http://standards.incits.org/a/public/group/m1

[18] The BioAPI Consortium [Online]. Available: http://www.bioapi.org/

[19] Identificación biométrica por huellas digitales [online]. Available: http://www.inegi.gob.mx/inegi/contenidos/espanol/ciberhabitat/hospital/huellas/textos/identificacion.htm

[20] Cómo Funcionan los Lectores de Huella [online]. Available: Digitalhttp://www.tecmex.com.mx/promos/bit/bit0903-bio.htm

[21] Roger Smith: RFID: A Brief Technology Analysis, CTO Network Library, 2005.RFID Journal [Online]. Available: http://www.rfid.org/

[22] Tipos de tags RFID [Online]. Available: http://www.idautomatica.com/informaciontecnica/tipos-de-tags-rfid.php

[23] Tipos de tags o etiquetas RFID [Online]. Available: http://www.rfid-a.com/index.php/2008/05/06/tiposde-tags-o-etiquetas-rfid/

[24] Lemmons, Phil; Robertson, Barbara (October 1983)."Product Review: The HP 150".

[25] Investigaciones del Laboratorio de Investigación Eléctrica de Mitsubishi (MERL) en interacciones con pantallas táctiles.[online]. Available: http://diamondspace.merl.com/

[26] Criptografia simetrica y asimetrica Dr. José de Jesús Ángel Ángel [Online]. Available: http://www.virusprot.com/Art1.html

[27] Seguridad en JAVA, Sergio Talens-Oliag, Instituto Tecnológico de Informática (ITI) [Online]. Available: http://www.uv.es/sto/cursos/seguridad.java/html/sjava-13.html

[28] Funciones Hash Criptografia [Online]- Available: http://www.monografias.com/trabajos76/funcioneshash-criptografia/funciones-hash-criptografia2.shtml

[29] Curso Seguridad de Redes y Sistemas. Autor: Msc.Walter Baluja García. CUJAE.“MD5 by Professor Ronald L. Rivest of MIT” [Online]. Available: http://userpages.umbc.edu/~mabzug1/cs/md5/md5.html

[30] The Legion of the Bouncy Castle [Online]. Available: http://www.bouncycastle.org/

[31] http://www.truecrypt.org/docs/

[32] How to use Model-View-Controller (MVC), Steve Burbeck, Ph.D. Roger Pressman, “Ingeniería de Software” Ed.8”

AUTHORS

M. Sc. Marco Antonio Dorantes González. Was born at Córdoba, Veracruz on 28 June, 1968. He had done his graduation in Electronics from ITO, Veracruz, México in 1990. After that he had completed his M. Sc. Degree in Computing in CINVESTAV in 1996 and M. Sc. of computing technologies in CIDETEC-IPN in 2008, research professor of ESCOM (IPN). He has been research Professor since 1996. He is interested in: Mobile Computing, Software Engineering, Data Bases. He has directed more than 70 engineering degree theses. Technical reviewer of interested areas books of publishers (McGraw Gill, Thompson, Pearson Education), He has participated in several research projects and has held some administrative positions in the IPN, also has experience in the industrial sector in the area of instrumentation and electronics; has done graduate studies in some fields, he has participated in several television programs and publications in scientific journals.

M. Sc. Martha Rosa Cordero Lopez. Was born at México D.F on 25 March, 1972. He had done his graduation in degree informatics from ITO, Veracruz, México in 1994. After that he had complete his Master Science Degree in Computing in CINVESTAV (IPN) in 1996, Master of computing technologies in CIDETEC-IPN in 2008, research professor of ESCOM (IPN) since 1995, her areas of interest are: Software engineering, Mobile Computing, Data Bases, affective computing, she has been director of in more than 70 theses to date, technical reviewer of interested areas books of publishes (McGraw Gill, Thompson, Pearson Education, among others). He has participated in various research projects and has held various administrative positions in the IPN also has experience in the private sector in the area of systems development; has done graduate studies in some areas, has been assistant mMcanager of technology intelligence unit in the technological development of the IPN, has participated in some television programs and publications in scientific journals.

Engineer Jorge Benjamín Silva González: Computing Systems Engineer from ESCOM (IPN) México D.F.

REORGANIZATION OF LINKS TO IMPROVE USER NAVIGATION

Deepshree A. Vadeyar[1] and Yogish H.K[2]

[1,2]Department of Computer Science and Engineering, EWIT Bangalore
deepshri.b@gmail.com
yogishhk@gmail.com

ABSTRACT

Website can be easily design but to efficient user navigation is not a easy task since user behavior is keep changing and developer view is quite different from what user wants, so to improve navigation one way is reorganization of website structure. For reorganization here proposed strategy is farthest first traversal clustering algorithm perform clustering on two numeric parameters and for finding frequent traversal path of user Apriori algorithm is used. Our aim is to perform reorganization with fewer changes in website structure.

KEYWORDS

Farthest-first, web mining, website design, web logs.

1. INTRODUCTION

WWW is large source of information and lakhs of user uses internet as search engine and website, website is used for providing information and also it is big source of commercialization but even user not get proper information over website and seeking for page what user wants, its happening due to reason that user view to use website is different than developer and as user has different characteristics than other user [1].As user have different requirement so how to naviage use effectively one of the way is by changing link structure[2] ,by changing structure we can say that we are performing web transformation .Web transformation is not done manually but ats automatic process by providing some learning method to websites, we used un-supervised learning method Clustering and intelligent method Association to find frequent links by traversing path of user.

As we performing data mining techniques over website we can define it as web mining which categorized as web usage mining in this technique weblogs are collected according to user behavior, web content mining is about content on web pages and it is also called as Text mining and web structure mining about changing link structure. As we are performing link mining but it depends on the web usage data and we also performed some of pre-process task.

Our work is to perform clustering for this we chooses two parameter Average duration of user we can say time for which user on website and number of clicks on URL of web pages, we perform clustering for the parameter more than some defined threshold ,as more click on URL we can get

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 81–87, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4509

cluster with high frequency of page access, due to clustering large data space say 6000 rows divide is some clusters and each cluster may conation 100,1000 rows according to similarity measures, as we used here farthest first traversal clustering algorithm our clustering performed on Euclid and Manhattan similarity measures and objects are assign at smallest distance of objects from centroid and max distance between centroid actually centrod is chosen as randomly bust with maximum Euclid distance between two centroid and distance between centroid to object is minimum distance.

To find which link structure to be change we will user identification and we also collect traversing path of user for which after performing association by using Apriori algorithm, we will get frequent links, if frequent links are available in cluster means user are on page for more time and we have pages with high frequency, we will check it for categorization and with out-degree threshold and if both condition satisfies link will be reorganized.

Categorization is nothing but storing link structure as binary 0 or 1 as absence and presence of structure respectively so we can check whether link already present, we will use this as matrix so out-degree can be calculates easily.

As we used Apriori algorithm with threshold min-support, max-support, min-confidence and delta later we will get best item sets say x-item set ‘L(X)’,and best rule found which can be used to find best link for which reorganization can be performed in later section we will see how the item sets and formed and that search on cluster to perform link mining., at last web pages should be reorganized in such way that it should full fill the user needs.

In later section we shown how links will be structured and we also performed comparison of other clustering algorithm. We found that farthest first is fastest in building model. We can also conclude that when links which to be reorganized available in clusters and frequent item set as data is reduced before clustering by data reduction and outlier mining then time to reorganized also reduced. For website navigation as our first step is on weblogs so before actual link mining we need to perform some pre-processing and for this we considered some threshold values like session-threshold ‘α’ and click-threshold ‘β’ here α defines time spent more than threshold considered in clustering and β define click on links more than threshold consider for clustering. Following are steps for pre-processing :

1) Data cleaning: Removal of session value and clicks less than define α and β.
2) User Identification: Finding user with unique ip even user on same network so if we get frequent user we can provide priority to user .
3) Session Identification: Method of finding average time of user spent on website.
4) Path Completion: Finding missing links but which are important and frequently request.
5) Formatting: Converting logs into data which can be mined. It includes removal of binary value for clustering and removal of numeric parameter for frequent mining.

This preprocessing task actually removes irrelevant links before start of miming so time of building model and reorganizing of links will be reduced.

2. RELATED WORK

In this section we are providing information about previous work done for web site intelligence in terms of web usage mining and web transformation, our main focus on link mining for which we first need input parameters as web logs and data mining technique over web logs called as web usage mining. Introduction of web usage mining by Cooley in [3]explains about usage mining techniques, for working with usage mining author proposed several mining techniques, their main focus is on some data preparation and pre-processing techniques for web logs.

Perkowitz in [4] focus on problem of page indexing and for this proposed technique is clustering like novel and conceptual clustering algorithm(COBWEB).In this paper clustering of objects are achieved by means of some common content shared between objects, here author also proposed quality measures like user looking for page and efforts by user to get desired page later they used evaluation by measures like how much website improved and how many users are benefitted.

As our main goal is structure mining and many work done on structure mining as Joy Shalom in [5] consider website structure as graph and website reorganization can be achieved by means of server logs ,proxy server and cookies, other main criteria suggested by author is browsing efficiency that is ratio of shortest path from index page to desired page to cost of operation. Here main aim is to guide user for accessing website effectively.

Lin in [6] proposed here a model for reducing load on searching desired page for user. Their approach is based on session and pages which occurred together like if ‘A’ access than ‘B’ access than sequence from A to B is important here.

Fu in [7]proposed reorganization of website based on used access pattern and here to achieve change in structure strategies used are pre-processing and classification. Here in pre-processing step information about website and web server logs is maintained. Second step include page classification based on content of website here we can say content mining is involved on parameters like file type, total links present ,session present and user time on website.

Gupta in[8] proposed reorganization by classification techniques based on type of file extension, number of links page, ratio of session on last page to the total session on web site and average time for which user on websites or user is login.

Min chen in [9] proposed model for reorganizing website structure for reorganizing website structure with minimum change ,here to know how much changes required out-degree threshold is used, their model also consist of path threshold,their model work on traversal path required to reach target page.

Yitong in [10] proposed k-means clustering algorithm using cosine similarities for link analysis and objects are cluster according to common links, words or phrases shared between documents. They used here concept of co-citation and coupling with concept of Hubs and Authority.

Amar Singh In [11] proposed clustering approach over links for improving searching of links in search engines and to accomplish this they used k-means and page-rank algorithm. Later they also shown results for weighted page rank algorithm using k-means.

Mobasher In[12][13][14] proposed clustering methods based on usage mining. In [12] author first find frequent item set and then performed clustering on user profiles. In[13]author reverse process first cluster and then finf frequent item set on usage data so imp data not lost from analysis. In [14] author proposed clustering algorithm on parameter like user transaction and page views.

3. OUR MODEL

We consider websites as graph and each page as node and redirecting URL between pages as edges figure-1 shows our website with 15 pages and many links, links as edges represent as 1 or 0 say we have source node is i and node on which out-link connect to j node with link X_{ij} so this can be represent as equation 1.

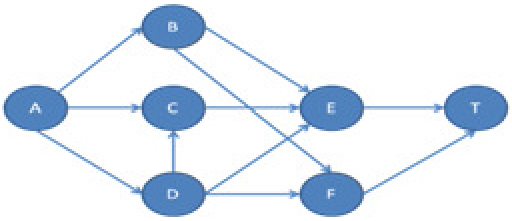


Figure-1: Website structure

$$X_{ij} = 1 \quad \text{Or } X_{ij} = 0 \quad \text{--------------------------------------(1)}$$

In equation 1 represents that there is links from page i to page j and zero represents there is no links between pages ,it means user can traverse to connected links but if links is not there and page is most visited or user spent more time tan link can be created by reorganizing its website structure.

Here we will form cluster for which we have object are URL on which user is active and URL which satisfies threshold criteria for cluster of links between node i and j for which cluster can be represent as K_{ij} ,to perform clustering we will use the similarity measures by farthest distance and we use distance measures is Euclid , here we have two parameter session 'S' and Clicks 'C' on which cluster is performed for most far value from mean as shown in equation 2.

$$d((i,j) = \sqrt{(S_i - S_j)^2 + (C_i - C_j)^2} \quad \text{--------------------------------------}(2)$$

We also considered the user behaviour and time for which user spent more time on pages we consider that page as target page and we saves user navigation path. Say 'I' is item set contain 'n' item set like $I = I_{1,} \ I_2 ,---- \ I_n$ and each item shows navigation path of user which can be $I_1 = (A,B,E,K), I_2 = (A,C,J,K), I_n = (A,B,E,A,J,K)$ here we can show that 'K' is more frequent and user traversing back from 'E' to 'A' that is start page, as on user navigated path we perform Apriori algorithm, if we have user with say 'I' item set with 'N' rows after association we get frequent URL of item sets which is less than original set sat L(k)

After association we need to check the links which are available in association A_{ij} as well as in clusters C_{ij} so we can say there is mapping between Links 'i and j' in cluster and in Item-set that can be represent as equation 3. $A_{ij} -> \quad C_{ij}$.---(3)

If links are matches between frequent item set and the clusters than we considered that links as links to be reorganized. To achieve minimum changes in website structure we considered an out-degree threshold which defines how many links can be change that can be determine by categorization .For example out degree threshold is four so only four out- links are allowed on page if already four links present we can't make out-links from that page.

3.1 Farthest First Algorithm

Farthest first algorithm proposed by Hochbaum and Shmoys 1985 has same procedure as k-means, this also chooses centroids and assign the objects in cluster but with max distance and initial seeds are value which is at largest distance to the mean of values, here cluster assignment is different, at initial cluster we get link with high Session Count, like at cluster-0 more than in cluster-1, and so on.

Farthest first algorithm need less adjustments and basic for this explained in [15].

Working as described here, it also defines initial seeds and then on basis of 'k' number of cluster which we need to know prior. In farthest first it takes point P_i then chooses next point P_1 which is at maximum distance. p_i is centroid and $p_1, p_2,........p_n$ are points or objects of dataset belongs to cluster from equation 4.

$$\min\{\max dist(p_i , p_1), \max dist(p_i , p_2)......\} \;\text{---}\; (4)$$

Farthest first actually solves problem of k-centre and it is very efficient for large set of data.In farthest first algorithm we are not finding mean for calculating centroid ,it takes centrod arbitrary and distance of one centroid from other is maximum figure-2 shows cluster assignment using farthest –first. When we performed outlier detection for our dataset we get which objects is outlier.

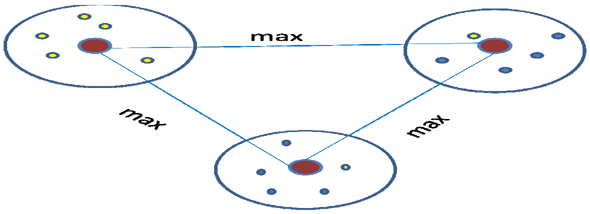


Figure-2 Object assignment in cluster

4. RESULT

We collected data set from Depaul university on which first we compare clustering algorithm using tool weka, fig-3 shows comparison between all algorithms and we found that farthest –first is fastest algorithm is fastest among others for building model. After clustering we performed association on transactional database, we get frequent item sets, here we performed association with delta=0.05 so time of finding frequent item-set is reduced since by using delta minimum support decrease by 0.05 in each iteration.

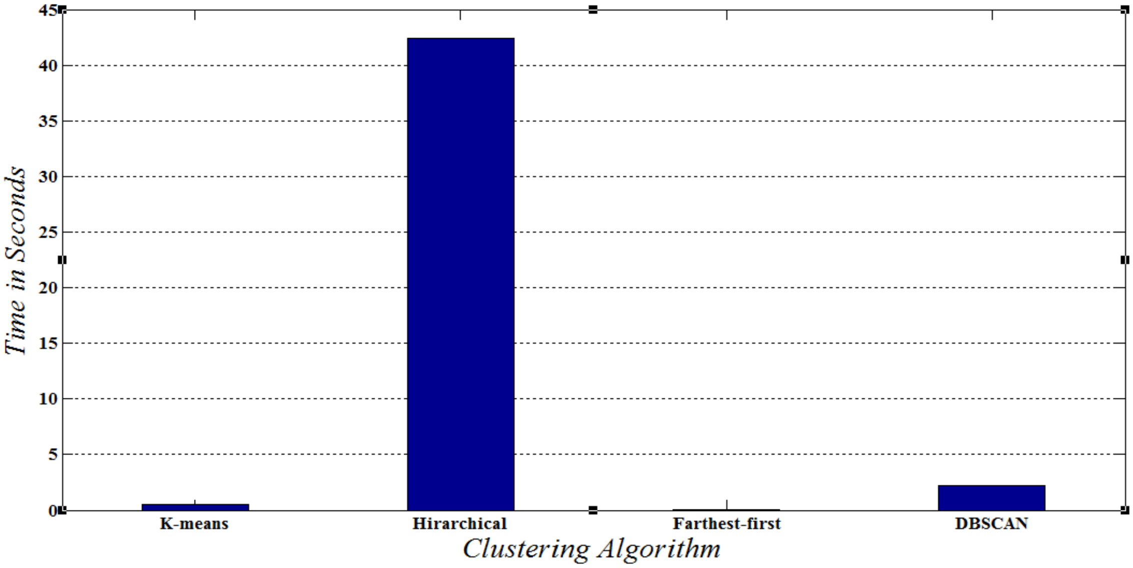


Figure-3: Comparison of clustering algorithm

Benefit of using cluster is search time for finding links to be reorganized is less than search time from transactional data base and we also get best links since links at initial clusters are links of largest duration of user on page.

Outlier analysis performed in weka tool by using filter interquartile ranges, here using farthest first outliers are more specific since this procedure not take high session_count or clicks as outliers but for k-means it takes ,since our main gal is reorganization on high count and clicks so we can say we get better outliers detection with farthest –first algorithm.

For reorganizing website structure we use real dataset on website like songs.pk , parameters session and clicks used for which links are clustered. After getting frequent item set and cluster data we found that time of reorganization using farthest first is reduced shown since time required to execute algorithm is 0.2 sec and for k-means it is 0.42 seconds, if we got assignment of links in cluster similar for farthest first and k-means even then total time of reorganization is reduced since farthest first execution time is less then k-means. Theoretical complexity for farthest first traversal algorithm is O (nk), where n is number of objects in the dataset and k is number of desired clusters. Similarly for k-means complexity is O (nkt) where n and k same as farthest first and‘t’ is number of iteration.

5. CONCLUSION

As we simulated and from result we got farthest first algorithm is fastest even than k-means and it also solves k-centre problem other aspect is better result for outlier detection we used, so it is more advantageous, hence we adopt that and when we search for links to be reorganized, we used set of traversing path for which we get matching links in cluster for which links are reorganized.

Second usefulness of farthest first is we get links to be reorganized most near to the max parameters since it works on largest distance so we get best links. Future work remain is to find frequent item sets for particular user which we get by identification due to which mostly visiting user of website will be benefited.

At last we can conclude as proposed algorithm gives some similar object assignment as k-means but in less time. It solves k-centre problem. If we perform evaluation then also proposed algorithm provides more correct instances for clustering.

This paper can be extended with outlier mining since proposed algorithm is faster than k-means but k-means is not robust for outliers so this future work will leads for good clustering algorithm selection.

ACKNOWLEDGEMENT

We would like to express our gratitude to all those who gave their valuable times to guide and solving doubts. We also want to thanks Computer science and Engineering department of East West Institute of Technology, Bangalore for giving me permission to continue my research on this topic.

REFERENCES

[1] T. Nakayama, H. Kato, and Y. Yamane, "Discovering the Gap between Web Site Designers' Expectations and Users' Behavior,"Computer Networks, vol. 33, pp. 811-822, 2000.

[2] J. Lazar, User-Centered Web Development. Jones and Bartlett Publishers, 2001.

[3] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava" Data Preparation for Mining World Wide Web Browsing Patterns".

[4] Mike Perkowitz and Oren Etzioni "Towards adaptive Web sites: Conceptual framework and case study" Artificial Intelligence 118 (2000) 245–275, 1999.

[5] Joy Shalom Sona, Asha Ambhaikar "Reconciling the Website Structure to Improve the Web Navigation Efficiency" June 2012.

[6] C.C. Lin, "Optimal Web Site Reorganization Considering Information Overload and Search Depth," European J. Operational Research.

[7] Y. Fu, M.Y. Shih, M. Creado, and C. Ju, "Reorganizing Web Sites Based on User Access Patterns," Intelligent Systems in Accounting, Finance and Management.

[8] R. Gupta, A. Bagchi, and S. Sarkar, "Improving Linkage of Web Pages," INFORMS J. Computing, vol. 19, no. 1, pp. 127-136, 2007.

[9] Min Chen and young U. Ryu "Facilitating Effective User Navigation through Website Structure Improvement" IEEE KDD vol no. 25. 2013.

[10] Yitong wang and Masaru Kitsuregawa "Link Based Clustering of Web Search Results" Institute of Industrial Science, The University of Tokyo.

[11] Amar Singh,navjot Kaur,"To Improve the Convergence Rate of K-Means Clustering Over K-Means with Weighted Page Rank Algorithm"IJARCSE August 2013.

[12] Bamshad Mobasher,Robert Cooley, Jaideep Srivastava "Creating Adaptive Web Sites Through Usage-Based Clustering of URLs"

[13] Bamshad Mobasher,Robert Cooley, Jaideep Srivastava"Automatic Personalization Based on Web Usage Mining"

[14] Bamshad Mobasher, Honghua Dai, Tao Luo, Miki Nakagawa" Discovery and Evaluation of Aggregate Usage Profiles Web Personalization.

[15] Zengyou He "Farthest-Point Heuristic based Initialization Methods for K-Modes Clustering.

INTENTIONAL BLANK

ANALYSIS OF INDIAN WEATHER DATA SETS USING DATA MINING TECHNIQUES

T V Rajini kanth[1], V V SSS Balaram[2] and N.Rajasekhar[3]

[1]Professor, CSE, SNIST, Hyderabad
rajinitv@gmail.com
[2]Professor & HOD, IT, SNIST, Hyderabad
vbalaram@sreenidhi.edu.in
[3]Assistant Professor, VNRVJIET,Hyderabad
n rajasekhar_n@vnrvjiet.in

ABSTRACT

India has a typical weather conditions consisting of various seasons and geographical conditions.Country has extreme high temperatures at rajasthan desert, cold climate at Himalayas and heavy rainfall at chirapunji. These extreme variations in temperatures make us to feel difficult in inferring / predictions of weather effectively. It requires higher scientific techniques / methods like machine learning algorithms applications for effective study and predictions of weather conditions. In this paper, we applied K-means cluster algorithm for grouping similar data sets together and also applied J48 classification technique along with linear regression analysis.

KEYWORDS

Geographical conditions, Temperatures, weather, clustering, classification

1. INTRODUCTION

Farming is the background of the economy; every person requires food for their survival. The farmers must be helped, so that they will come to know which crop to grow under various circumstances. Farming not only depends on manpower but also on various aspects like water, type of soil, fertilizers used, climate etc. Our intention through this project is to guide the farmers in choosing a crop[1,2,3,4] for cultivation that has the most productive yield thereby being beneficial to them.

In this project, an attempt has been made to review the research studies on application of data mining techniques in the field of agriculture [1, 2, and 3]. This project being a research oriented one; we have analyzed data of various regions, read several papers for reference and implemented suitable data mining techniques to achieve our goal of predicting the weather. Most of the databases contain information that is accumulated for years. These databases can become valuable information for analysts who use the data to perform various operations on data. Analysis was done on the weather data sets using machine learning algorithms [4, 5, 6].

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 89–94, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4510

It is important to remember that none of predictive techniques gives 100% accurate results. The main aim of data mining is giving help in decision making, but the final decision is always after you. A BI application gives you an interpretation of data, but it is important to remember that all results you will obtain are an aid in decision making, and the final decision is always after you. And that there is no technology that is able to give 100% accurate results.

2. LITERATURE SURVEY

Data mining, a branch of computer science, is the process of extracting patterns from large data sets by combining methods from statistics and artificial intelligence with database management. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery.

The related terms data dredging, data fishing and data snooping refer to the use of data mining techniques to sample portions of the larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These techniques can, however, be used in the creation of new hypotheses to test against the larger data populations.

If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data preprocessing includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set.

2.1.K-means Algorithm:

- K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships.
- The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

2.2. Working of k means algorithm

1. Place K points into the space represented by the objects that are being clustered.
2. These points represent initial group centroids. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

2.3. Decision tree:

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are classification trees or regression trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making. J48 are the improved versions of C4.5 algorithms or can be called as optimized implementation of the C4.5. The output of J48 is the Decision tree. A Decision tree is similar to the tree structure having root node, intermediate nodes and leaf node. Each node in the tree consist a decision and that decision leads to our result. Decision tree divide the input space of a data set into mutually exclusive areas, each area having a label, a value or an action to describe its data points. Splitting criterion is used to calculate which attribute is the best to split that portion tree of the training data that reaches a particular node.

3. PROPOSED APPROACH

In this we apply the data mining technique Kmeans cluster algorithm on the data set which was modified in to suitable format from the raw format after preprocessing stage. After that J48 algorithm was applied on to it. Over that Regression techniques were applied.

4. IMPLEMENTATION OF PROPOSED APPROACH

The data sets with min temperature was clustered and kept in a table 3.1 for further analysis. From this table one can conclude that there are 5 clusters namely cluster0, cluster 1, cluster2, cluster3 and cluster4.

Cluster0: The annual Min. temperature went up to 19.5^{0}C. There is temperature variation across seasons i.e. it is low during winter (14^{0}C) and slowly raised to summer season (23.4^{0}C)and again fallen down in rainy season(16.5^{0}C).

Same Phenomena has appeared in all the remaining clusters. There are low temperature values in Annual, Jan-Feb, Mar-May, Jun-Sep and Oct-Dec duration in cluster2 and high in cluster4. The minimum temperature is raising year by year but slight downfall in the duration 1960 – 1975 but again rose after that duration. That means warming of earth is taking place year by year due to many factors.

The data sets with max temperature was clustered and kept in a table 3.2 for further analysis. From this table one can conclude that there are 5 clusters namely cluster0, cluster 1, cluster2, cluster3 and cluster4.

Cluster0: The annual Max. temperature went up to 30^{0}c. There is a temperature variation across seasons i.e. low during winter (25^{0}c) and raised to peak during Mar-May season (32^{0}c), downfall starts from Jun-Sep season (31^{0}c) and further downfall starts in rainy season (28^{0}c). The mean of max. temperature was raised from 1900 year to 2012. Same is the case happened across the seasons Jan-Feb, Mar-May, Jun-Sep, Oct-Dec and also along annual. There are low temperature values in Annual, Jan-Feb, Mar-May, Jun-Sep and Oct-Dec duration in cluster4 i.e. in the year 1905 and high in cluster0. The maximum temperature is increasing year by year and there is no downfall except in 1920 -25 years during Jun-Sep. That means warming of earth is taking place year by year due to many factors indicated by Annual- seasonal Max. Temperature data. The data sets with mean temperature was clustered and kept in a table 3.3 for further analysis. From this table one can conclude that there are 5 clusters namely cluster0, cluster 1, cluster2, cluster3 and cluster4.

Cluster0: The annual mean temperature went up to 24^0c. There is a temperature variation across seasons i.e. it is it is low during winter (19^0C) and slowly raised to summer season (27^0C)and again fallen down in rainy season(21.4^0C).

Same Phenomena has appeared in all the remaining clusters. There are low temperature values in Annual, Jan-Feb, Mar-May, Jun-Sep and Oct-Dec duration in cluster1 and high in cluster 4.

5. Results And Analysis

The mean temperature is raising year by year but slight downfall in the duration 1955 – 1965 but again rose after that duration. That means warming of earth is taking place year by year due to many factors. J48 algorithm was applied on that data set and constructed a decision tree which is shown in Fig. 3.2. The graph represented below by Fig.3.1 was plotted with years along x-axis and minimum temperature along y-axis.

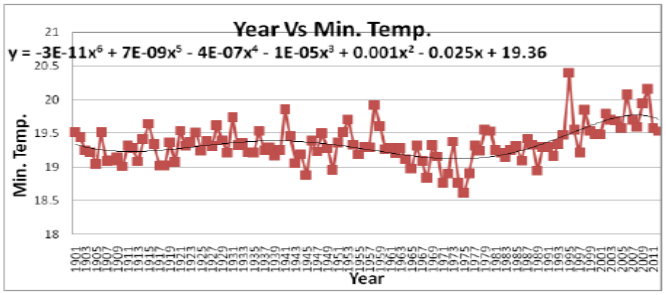


Fig3.1: Annual and seasonal minimum temperature for the years 1900-2012

Annual and seasonal minimum (night) temperatures is averaged over the country as a whole for the period 1901- 2012. It is based on the surface air temperature (i.e. 1.2 m above sea level) data from more than 350 stations spread over the country. In this in year 1995 it is showing 20.3 as highest min. temp and in 1975 lowest min. temp is 18.61. The regression trend line was drawn with equation is a polynomial equation.

$$\mathbf{y = -3E\text{-}11x^6 + 7E\text{-}09x^5 - 4E\text{-}07x^4 - 1E\text{-}05x^3 + 0.001x^2 - 0.025x + 19.36}$$

We can predict the value of y based on required x value.

The mean temperature data set was classified under the classifier function called linear regression and got the Linear Regression Model equation a

ANNUAL = -0.0002 * YEAR + 0.1732 * JAN-FEB + 0.2519 * MAR-MAY + 0.3064 * JUN-SEP + 0.2733 * OCT-DEC + 0.4846

By using this equation we can able to predict the Annual mean temperature based on year and seasonal temperature values. Only based on year also we can predict the Annual Mean temperature.

ANNUAL = 0.0069 * YEAR +10.7018

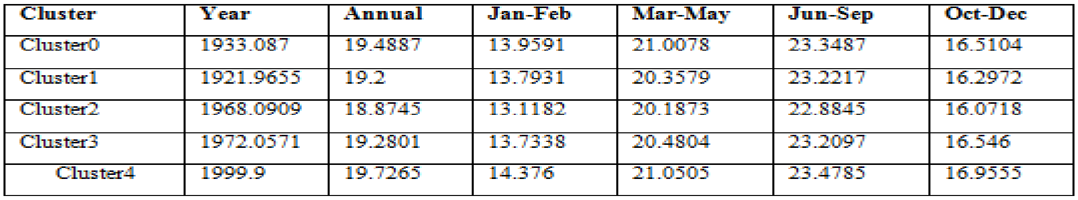
Only based on year also we can predict the Annual Min temperature.

ANNUAL = 0.0025 * YEAR + 14.3979

Only based on year also we can predict the Annual Max temperature.

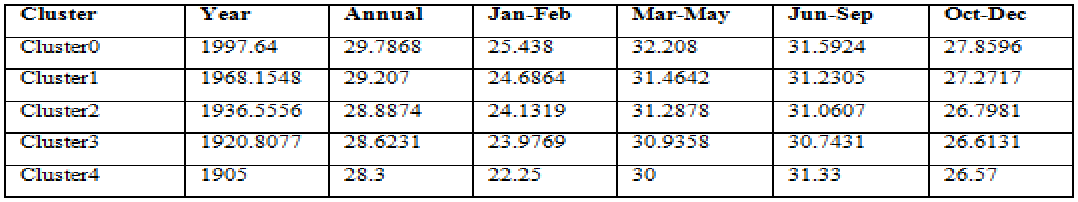
ANNUAL = 0.0116 * YEAR +6.394

6. FIGURES AND TABLES



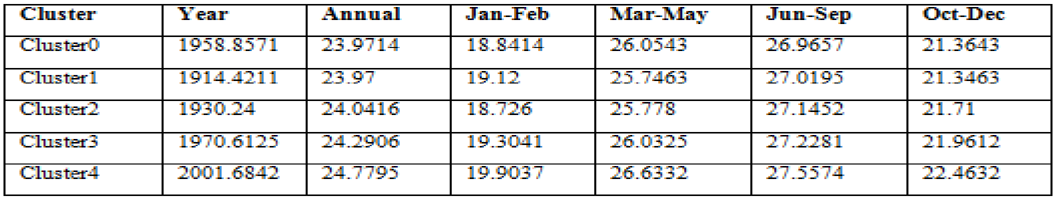
Cluster	Year	Annual	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec
Cluster0	1933.087	19.4887	13.9591	21.0078	23.3487	16.5104
Cluster1	1921.9655	19.2	13.7931	20.3579	23.2217	16.2972
Cluster2	1968.0909	18.8745	13.1182	20.1873	22.8845	16.0718
Cluster3	1972.0571	19.2801	13.7338	20.4804	23.2097	16.546
Cluster4	1999.9	19.7265	14.376	21.0505	23.4785	16.9555

Table 3.1: Annual- Seasonal Min temperatures



Cluster	Year	Annual	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec
Cluster0	1997.64	29.7868	25.438	32.208	31.5924	27.8596
Cluster1	1968.1548	29.207	24.6864	31.4642	31.2305	27.2717
Cluster2	1936.5556	28.8874	24.1319	31.2878	31.0607	26.7981
Cluster3	1920.8077	28.6231	23.9769	30.9358	30.7431	26.6131
Cluster4	1905	28.3	22.25	30	31.33	26.57

Table 3.2: Annual- Seasonal Max temperatures



Cluster	Year	Annual	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec
Cluster0	1958.8571	23.9714	18.8414	26.0543	26.9657	21.3643
Cluster1	1914.4211	23.97	19.12	25.7463	27.0195	21.3463
Cluster2	1930.24	24.0416	18.726	25.778	27.1452	21.71
Cluster3	1970.6125	24.2906	19.3041	26.0325	27.2281	21.9612
Cluster4	2001.6842	24.7795	19.9037	26.6332	27.5574	22.4632

Table 3.3: Annual –Seasonal Mean temperatures

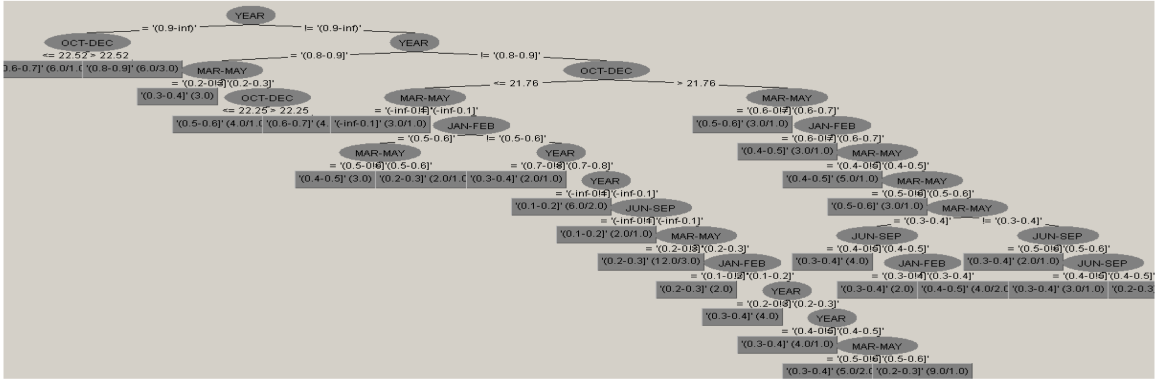


Fig.3.2: J48 tree diagram Mean Temperature

7. CONCLUSION

It is found that over 112 years of temperature data that temperature is increasing gradually i.e. there is an indication of global warming taking place. Temperature in terms of min or max or mean irrespective of it is increasing gradually and is found through k-means cluster analysis. The predictions can be done using the linear regression line equations that are found in an effective manner. The future scope of this is it can be extended to any huge data sets with various attributes /parameters for effective analysis and accurate prediction.

REFERENCES

[1] Ananthoju Vijay Kumar, T. V. Rajini Kanth, Estimation of the Influence of Fertilizer Nutrients Consumption on the Wheat Crop yield in India- a Data mining Approach, 30 Dec 2013, Volume 3, Issue 2, Pg.No: 316-320, ISSN: 2249-8958 (Online).

[2] Ananthoju Vijay Kumar, T. V. Rajini Kanth, A Data Mining Approach for the Estimation of Climate Change on the Jowar Crop Yield in India, 25Dec2013,Volume 2 Issue 2, Pg.No:16-20, ISSN: 2319-6378 (Online).

[3] A. Vijay Kumar, Dr. T. V. Rajini Kanth "Estimation of the Influential Factors of rice yield in India" 2nd International Conference on Advanced Computing methodologies ICACM-2013, 02-03 Aug 2013, Elsevier Publications, Pg. No: 459-465, ISBN No:978-93-35107-14-95.

[4] J Rajanikanth, Dr. T.V. Rajinikanth, T V K P Prasad, B Radha Krishna, "Analysis on Spatial Data Clustering Methods - A Case Study" Pg.no:51-54, IJCST Vol. 3, ISSue4, OCT- DeC2012, ISSN: 0976-8491 (Online) | ISSN: 2229-4333 (Print).

[5] Tarun Rao , N Rajasekhar, Dr T V Rajinikanth, "An efficient approach for Weather forecasting using Support Vector Machines", 2012 International Conference on Intelligent Network and Computing (ICINC 2012), IPCSIT Vol. 47 (2012) © (2012) IACSIT Press Singapore. DOI 10.7763/IPCSIT. 2012. V 47. 39. Pg. No: 208 – 212

[5] Dr.T.V.Rajini Kanth, K Anuradha, P.Premchand, I.V. Murali Krishna, "Weather Data Analysis Of Rajasthan State Using Data Mining Techniques", Journal of Advanced Computing Vol3, Issue2, Pg: 82-86, April 2011, ISSN: 0975-7686.

[6] Dabberdt, W., Weather for Outdoorsmen: A complete guide to understanding and predicting weather in mountains and valleys, on the water, and in the woods. Scribner, New York, 1981.

[7] Prema K.V., "A Multi Layer Neural Network Classifier", Journal of Computer Society of India, Volume 35, Issue no: 1, Jan- Mar 2005.

[8] Philip D. Wasserman, Neural Computing Theory and Practice, Van nostrand reinhold, New York.

[9] Badhiye S. S., Wakode B. V., Chatur P. N. "Analysis of Temperature and Humidity Data for Future value prediction", IJCSIT Vol. 3 (1), 2012, 3012 – 3014

[10] Sarah N. Kohail, Alaa M. El-Halees, "Implementation of Data Mining Techniques for Meteorological Data Analysis", IJICT Journal Volume 1 No. 3, July 2011

[11] S. Kotsiantis and et. al., "Using Data Mining Techniques for Estimating Minimum, Maximum and Average Daily Temperature Values", World Academy of Science, Engineering and Technology 2007 pp. 450-454

[12] Thair Nu Phyu, "Survey of classification techniques in Data Mining", IMECS 2009 Volume 1 Hong Kong pp. 1-5

[13] G. D'souza, E.C. Barrett, C.H. Power (1990): ―Satellite rainfall estimation techniques using visible and infrared imagery", Remote Sensing Reviews, 4:2, 379-414

[14] J. K. Mishra, O. P. Sharma, Cloud top temperature based precipitation intensity estimation using INSAT-1D data, International Journal of Remote Sensing 2001, 22:6, 969-985

[15] Tao Chen, Milcio Talagi, ―Rainfall prediction of geostationary meteorological satellite images using artificial neural network", International Geoscience and Remote Sensing Symposium 1993

[16] E. C. Barrett, M. J. Beaumont, ―Satellite rainfall monitoring: An overview", International Journal of Remote Sensing Reviews, 1994 11:1-4, 23-48

[17] Indian Meteorological Department, http://www.imd.gov.in

[18] MacKay, David (2003). "Chapter 20. An Example Inference Task: Clustering". Information Theory, Inference and Learning Algorithms. Cambridge University Press. pp. 284–292. ISBN 0-521-64298-1. MR 2012999.

ENERGY EFFICIENT NEIGHBOUR SELECTION FOR FLAT WIRELESS SENSOR NETWORKS

Saraswati Mishra[1] and Prabhjot Kaur[2]

Department of Electrical, Electronics and Communication Engineering, ITM University, Gurgaon, India
[1]saraswati12ecp020@itmindia.edu
[2]prabhjotkaur@itmindia.edu

ABSTRACT

In this paper we have analyzed energy efficient neighbour selection algorithms for routing in wireless sensor networks. Since energy saving or consumption is an important aspect of wireless sensor networks, its precise usage is highly desirable both for the faithful performance of network and to increase the network life time. For this work, we have considered a flat network topology where every node has the same responsibility and capability. We have compared two energy efficient algorithms and analyzed their performances with increase in number of nodes, time rounds and node failures.

KEYWORDS

Flat Topology, Negotiation Based Routing, Routing Protocol, Wireless Sensor Networks

1. INTRODUCTION

Wireless sensor networks consist of number of small nodes deployed in an area under supervision. Each node has limited storage, computational and sensing capability and limited energy resource, as nodes are battery operated. Since energy is the main concern in wireless sensor networks (WSN) to maximize the performance and to increase the lifetime of a network, various approaches are implemented to reduce energy consumption in a network. Most of the energy is consumed during idle period and during transmission of data from one node to another i.e. routing. An efficient media access control (MAC) and routing protocol should be designed to save energy. While MAC protocol targets at reduction of energy in scanning and accessing the channel, routing protocol helps to reduce the energy requirement for end-to-end transmission.

In WSN, there are number of routing protocols that have been proposed for different network criteria. Based on the network topology WSN protocols have been categorized as – flat network protocols and hierarchical protocols as shown in Figure 1. Protocols that fall under hierarchical class select one head amongst all and form a hierarchy [1]. This hierarchy may be a cluster, a chain or a grid. Cluster head or a leader collects the information from all the other nodes in its region and sends it to the sink node or gateway node. Some examples are low energy adaptive clustering hierarchy (LEACH), power efficient gathering in sensor information systems (PEGASIS), virtual grid architecture (VGA), etc. The work presented in this paper considers flat

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 95–102, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4511

protocols and thus we are not including descriptions of hierarchical protocols and will only be focusing on flat protocol strategies in the rest of this paper. In a flat network, every node is treated equally in terms of responsibility and capability. There is no master and no slave. Flat network protocols are further classified into – quality of service (QoS) based protocols, data centric protocols and location based protocols. Some examples of this type of protocols are sensor protocol for information via negotiation (SPIN), directed diffusion (DD), gradient based routing (GBR) and geographic and energy aware routing (GEAR) [2].

Most of the protocols mentioned above implement energy saving as an important feature and accordingly, before delivering the data packet to the next hop neighbour, the source or intermediate node checks residual energy or consumed energy to decide the neighbour for the data forwarding. Protocols working on this approach are known as energy centric or energy aware protocols. In data centric protocols, sink sends queries and waits for data. Attribute-based naming is necessary to specify the property of data that data can be requested through queries. QoS based routing is different from address based routing mechanism used in data centric protocols. It selects the path based on some previous knowledge of resource availability and maximum tolerable delay as well as QoS requirement of a network. Also for optimum routing it adaptively allocates the available resources to maintain QoS. Location based protocols are used for routing queries towards targeted region of sensor network. Location information of next hop neighbour should be known to each sensor node. This information is used to calculate the distance between two nodes so that energy consumption can be determined [7]. Out of these techniques, we have considered data centric approach for the analysis and analyzed two different strategies to reduce energy consumption during routing. The detail description of these techniques is given in the next section.

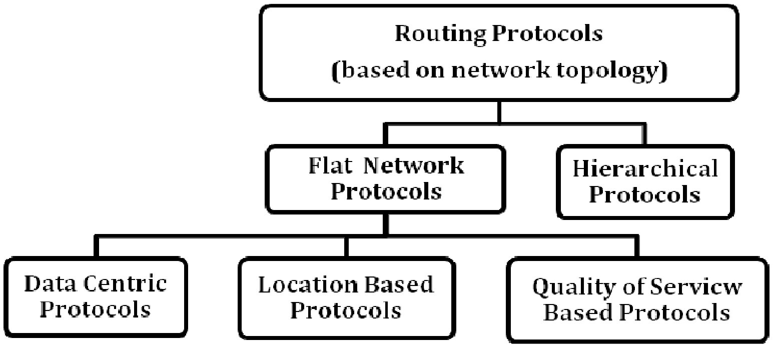


Figure 1. Classification of routing protocols based on network topology

The rest of this paper is organized as follows. Section 2 briefs the routing algorithms considered for analysis. Section 3 presents different approaches that have been implemented on the chosen routing protocol to attain energy efficiency. The simulation setup and results are discussed in section 4 and the paper concludes with section 5.

2. DATA CENTRIC PROTOCOLS

As we have discussed, the data centric protocols work in a flat networks. The working principle of these routing protocols is based on query (or a request) [4]. Query may be generated either by a sink node or source node.

In first case, sink broadcasts query to get specific type of data, any node having that specific data replies back. In second case, source sends the signal to specify that it is having some specific data, interested node can receive that by sending request. As we can observe that the routing is taking place via negotiation, it is important here to mention that the negotiation based protocols are the special class of data centric protocols. Negotiation based protocols may be of two stages – query and data or it may be of three stages – metadata, query and data. Metadata is a packet that contains information regarding the data of the node. Format of metadata may vary with the variation in application. Traditional flat protocols like flooding and gossiping have various drawbacks and limitations like implosion, data redundancy and resource blindness [3]. These can be overcome by use of data centric protocols. According to the stages best example of three stages negotiation based protocol is SPIN and example of two stage protocol is DD. Their brief description is given below.

2.1. Sensor Protocol for Information via Negotiation

SPIN is a data dissemination protocol that disseminates its information to all the nodes in its vicinity. This protocol works in three stages. First the node having data sends the advertising message (ADV) to the single hop neighbour. ADV acts as a metadata here. The interesting neighbour replies with request message REQ to indicate that it needs the data and finally the data is sent to requesting node. SPIN is classified in different classes like point-to-point (SPIN-PP), broadcast (SPIN-BC) reliable (SPIN-RL) and energy centric (SPIN-EC) and are used depending upon the application [4].

2.2. Directed Diffusion

The DD is again a flat network protocol that works on a principle of flooding. Here for a need of specific data sink node floods the interest signal in the network through the neighbours. After receiving a request every node maintains an interest cache. This is maintained till the gradient is not formed. The gradient is a reply link through which a request was received. Gradient contains all the information about the path i.e. data rate, duration etc. among all the paths formed from sink to source the best path is selected through the reinforcement process that means data is sent through selected shortest path and hence prevents further flooding [4].

From above discussed protocols we have analyzed three stage negotiation based protocol with the addition of subroutine that makes it energy efficient.

3. ENERGY EFFICIENT APPROACH

There are various approaches to minimize the energy consumed by the routing protocol. To make flat routing protocols more energy efficient - 1) we can select the neighbour which is closest to base station so that the number of hops to perform routing is minimum, this will save energy 2) another approach is to select the neighbour or next hop having maximum residual energy among all and 3) select the path towards the destination or sink that consumes minimum energy. Among all the approaches mentioned above we have applied second and third approach for the analysis. Detailed specification is given below.

3.1. Selection of Neighbour having Highest Energy

In this selection approach, when source want to transmit data to the destination which is multiple hop away from the source then the source checks the energy level of all the neighbours and selects the one having highest energy among all. Similarly, all intermediate nodes find out the neighbour with highest energy and deliver the data to that node and finally the data packet reaches to the destination. In the rest of this paper, we have referred this technique as highest energy (HE).

3.2. Selection of Path that Consumes Minimum Energy

In this process the source node or sending node at first, estimates the total energy that will be consumed by all possible paths formed in multipath communication scenario and then selects the best path toward base station which will consume minimum energy amongst all paths during transmission of data packets. This technique is referred as minimum energy consumption route (MECRT).

4. SIMULATION SET UP AND ANALYSIS

A flat sensor network was created and above said routing protocols were compared using SENSORIA – a Graphical User Interface (GUI) based simulator [6]. The details of simulation setup are given below in table1.

Routing protocol selected is a negotiation based protocol that works on three stages as we have discussed earlier (replica of SPIN). Simulation takes place till there is a path (nodes are alive) between source and destination to forward packets, otherwise it gets terminated. The nodes are randomly deployed and are dynamic in nature.

Table 1. Simulation parameters and values

Parameters	**Values**
Number of nodes	50
Energy / node	0.5 J (homogenous)
Simulation area	50m X 50m
Transmission range of each node	15 m
Sensing range (each node)	8m
Location of base station	25m X 150 m
Data packet	2000 bits
Control packet	248 bits
Data transmission speed	100 bits/sec.
Bandwidth	5000 bits/sec.

4.1. Analysis and comparison of routing protocols for their energy consumption

First, we have studied the impact of energy awareness on routing protocol. Selected protocol has been made energy efficient by the application of minimum energy consuming path selection algorithm, MECRT. The comparison of this protocol with its counterpart i.e. the technique that does not account energy consumption of the path during the transmission is done. The result shows that the energy saving capability of a network is more in which MECRT is implemented, as shown in figure 2. Life time of a network is also increased by the application of MECRT as compared to the life of a network that works without MECRT routing protocol. We can also see from the graph that energy consumption is less in MECRT hence it works for comparatively long time rounds.

Further, we have compared the performance of HE and MECRT that makes routing more energy efficient. Highest Energy neighbour selection and MECRT are applied on three stage negotiation based protocol like SPIN, where communication takes place through the exchange of metadata, query (or request) and then data, as we discussed earlier. These energy aware techniques used in our scenario helped in increasing life time and its performance.

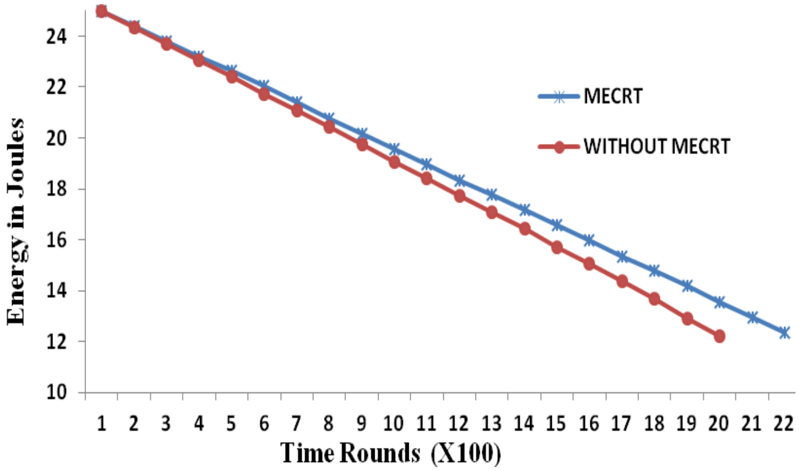


Figure 2. Comparison of algorithms (to select the route among all the available routes) with and without energy consideration

Both the protocols were applied on the same network with same parameters as described above and their performances were recorded. The performance analysis represents that MECRT gives better results than HE in terms of reduction of energy consumed by the network during routing as shown in Figure 3. Moreover when network route was discovered using HE algorithm, its life time was shortened as compared to the network that has MECRT as energy saving mechanism. Nodes death rate frequently increased in case of HE after certain time round compared to the MECRT algorithm which is shown in Figure 4. Both plots represent the energy decay and node failure with respect to the time round.

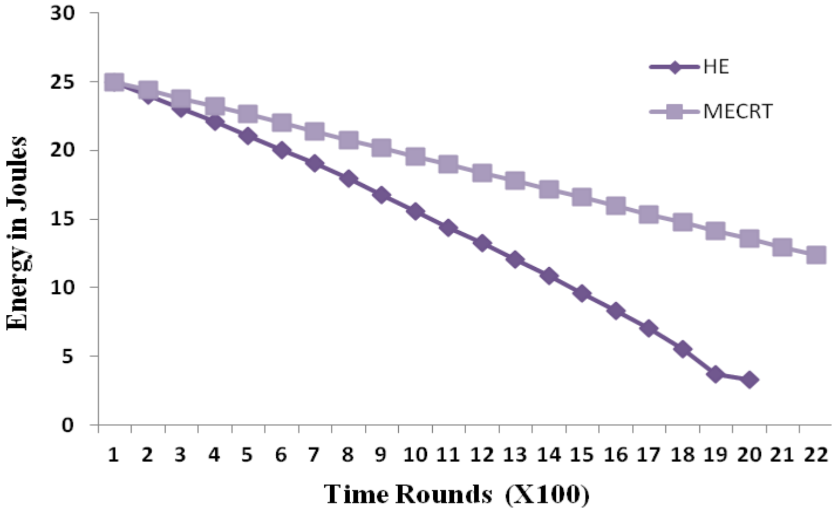


Figure 3. Comparison of algorithms HE and MECRT in terms of energy degradation of a network

In small network size, i.e. with few numbers of nodes, the difference was not significant. As flat routing protocols are mainly designed for small and medium size networks, we have simulated their performance on networks having nodes ranging from 10 to 200.

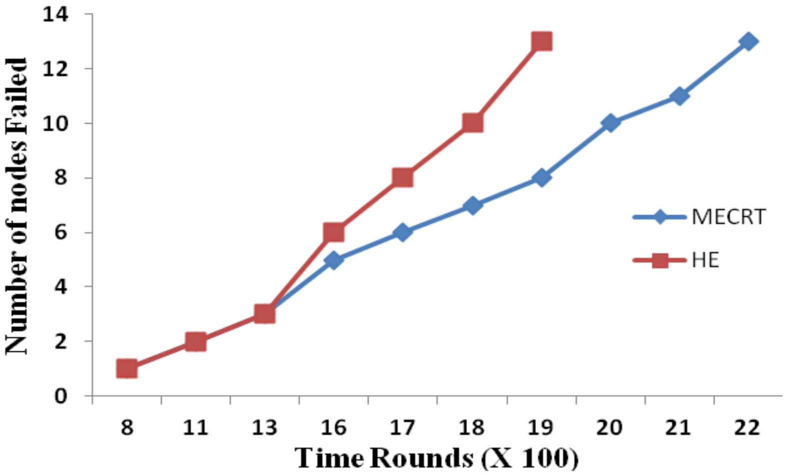


Figure 4. Comparison of algorithms HE and MECRT in terms of number of nodes failure

As the number of nodes increased to 25 to 150 we can easily identify the difference and can conclude that MECRT gives better results than HE for the static network topology in terms of energy consumption as well as network life time. The comparison with increase in number of nodes in a scenario is shown in figure 5.

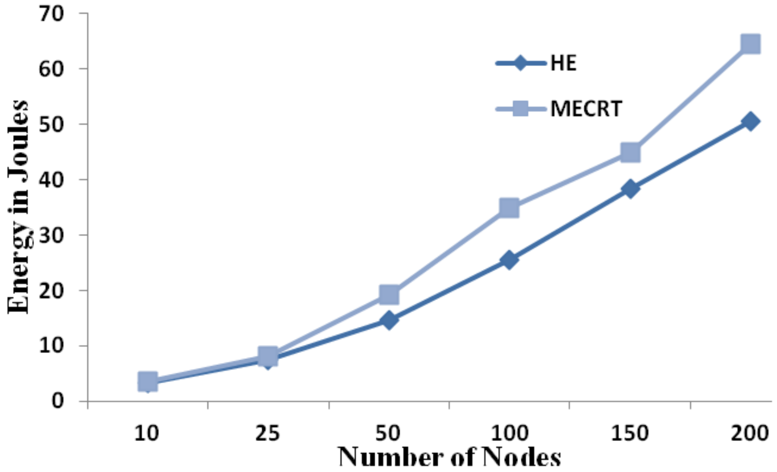


Figure 5. Comparison of algorithms HE and MECRT when number of nodes are increased

5. CONCLUSIONS

Through this paper, it is clear that energy efficient routing protocols helps to save energy in wireless sensor network and should be used in scenarios where energy consumption of sensors is a constraint. We have compared HE and MECRT in a flat network topology, to reduce energy consumption during routing. In a network that contains 10 or 20 nodes, any approach either HE or MECRT will give almost similar result. Hence any of the algorithms can be implemented in a routing protocol. Through the experimental analysis we can conclude that MECRT is better for medium to large network size, where node selects a path that consumes minimum energy among all available paths for data forwarding as compared to the HE algorithm where node delivers the data to the neighbouring node having highest energy. However, both these techniques do not guarantee the shortest route selection or fast routing mechanism. These protocols only deal with less expenditure of energy during routing. Hence more effective routing algorithm can be designed in future that will tend to select shortest path while assuring least energy consumption.

ACKNOWLEDGEMENTS

We would like to thank G. Al.-Mashaqbeh for providing permission and guidance to install and understand SENSORIA simulator.

REFERENCES

[1] K. Akkaya and M. Younis,(2005) “A Survey on Routing Protocols for Wireless Sensor Networks,” *Elsevier Ad Hoc Network Journal*, Vol. 3, No. 3, pp. 325-349.

[2] I.F. Akyildiz, M.C. Vuran, O. Akan and W. Su, (2004) “Wireless Sensor Networks: A Survey Revisited,” *Elsevier Journal of Computer Networks,* Vol. 45, No. 3.

[3] Jamal N. Al-Karaki and Ahmed E. Kamal,(2004) “Routing Techniques in Wireless Sensor Networks: A Survey,*” IEEE Wireless Communications*, Vol.11, No.6, pp:6-28.

[4] Sohrabi K., Gao J., and Ilawadhi V.,(2000) “Protocols for Self-organization of a Wireless Sensor Network,” *IEEE Personal Communications*, Vol. 7, No. 5, pp. 16-27.

[5] J. N. Al-Karaki and G. Al-Mashaqbeh, (2006) “Energy Centric Routing in Wireless Sensor Networks,” *Proceedings of IEEE International symposium on comuputer and Communications*. pp. 948-954.

[6] J. N. Al-Karaki and G. Al-Mashaqbeh, (2007) "SENSORIA: A New Simulation Platform for Wireless Sensor Networks*," Internatiuonal conferednce on Sensor Technologies and Applications, SensorComm*. pp. 424 – 429.

[7] Sohrabi K, Daniel Minoli and Taieb Znati, (2007) *Wireless Sensor Networks: Technology,Protocols and Applications*., Wiley and sons Publication.

AUTHORS

Saraswati Mishra is pursuing post graduation in Electronics and Communications from Institute of Technology and Management University, Gurgaon, India. She did her graduation in Electronics and Software Technology from LAD College, Nagpur University, Nagpur, Maharashtra, India. She is currently working on Cognitive radio technology.

Prabhjot Kaur did her Ph.D. on MAC models for Dynamic Spectrum Access in Cognitive Radio Networks from National Institute of Technology, Jalandhar, and her Masters of Engineering from Punjab University, Chandigarh India. She is currently working as Associate Professor and Deputy Dean (RDIL) with ITM University, Gurgaon, India. Her research interests include dynamic spectrum allocation, Ad-hoc Networks, Green Networks, MIMO, software defined radios and Cognitive Radios. She has completed a research project funded under Research Grant from AICTE, Govt. of India under research promotion scheme in April 2010 and an international travel grant for attending IEEE conference by Department of Science and Technology, Govt. of India. She received the `Best Emerging Researcher Award` of the year 2012 at ITM University, India and Best Paper recognition at an International conference, Malaysia. She is member, IEEE and life member of IETE and ISTE societies

DESING ON WIRELESS INTELLIGENT SENEOR NETWORK ON CLOUD COMPUTING SYSTEM FOR SMART HOME

Tsung-Han Tsai, Chih-Chi Huang, and Chih-Hao Chang

Department of Electrical Engineering, National Central University, Taiwan

ABSTRACT

Sensors on (or attached to) mobile phones can enable attractive sensing applications in different domains such as environmental monitoring, social networking, healthcare, etc. In this paper we propose a cloud computing system dedicated on smart home applications. We design the proposed wireless vision sensor network (WVSN) with its algorithm and hardware implementation. In WVSN, The partial-vision camera strategy is applied to allocate the computation task between the sensor node and the central server. Then we propose a high performance segmentation algorithm. Meanwhile, an efficient binary data compression method is proposed to cope with the result on labeling information. The proposed algorithm can provide high precision rate for the smart home applications such as the gesture recognition and humanoid tracking. To realize the physical system, we implement it on the embedded platform and the central server with their transmission work.

1. INTRODUCTION

Mobile devices (smart phones, tablets, laptops, embedded boards, robots) can serve as terminals for cloud computing services over intelligent network. Mobile cloud has emerged as a new cloud computing platform that ‘puts cloud into a pocket’. Most of current devices are equipped with a rich set of embedded sensors such as camera, GPS, WiFi/3G/4G radios, accelerometer, digital compass, gyroscope, microphone and so on. These sensors can enable attractive sensing applications in various domains such as environmental monitoring, social networking, healthcare, transportation, safety, etc.

Smart home is a trend on modern human life. It is dominated by different kinds of sensors over the wireless network. The smart appliance is controlled by the central processor to achieve electrical automation features such as: lights, TV, air conditioning, security systems, etc. With a success on smart home, many functions can be explored, especially at home care, convenient and energy-saving environmental protection respectively. The concept of smart home with the versatile human-style applications is shown in Fig. 1.

To realize a smart home as a cloud computing system with the relative applications, sensor network technique is deployed as a service platform. Most deployed wireless sensor networks measure scalar physical phenomena such as temperature, pressure, humidity, or location of

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 103–111, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4512

objects. More recently, intelligent monitoring system began to have a new breakthrough in image recognition. The original image by the person viewing the monitor mode is gradually developed into automatically monitored by a computer. Then the system can directly give a warning to person. Some applications are mostly eager to this solution, including: vehicle license plate recognition, falls among the elderly, face department identification, fire identification, etc. With this technology, many semantic applications on a smart home are realized.

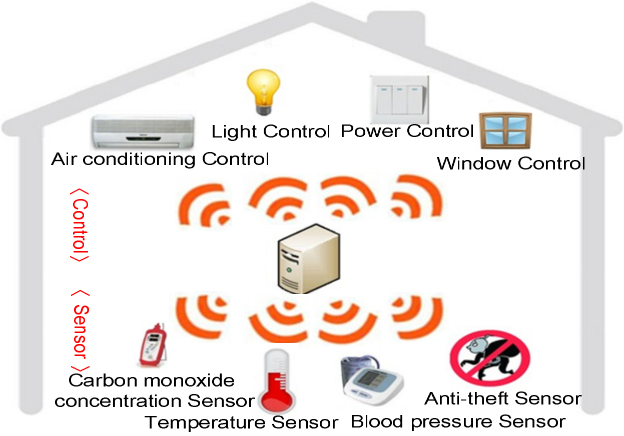


Fig. 1. Schemes of smart home.

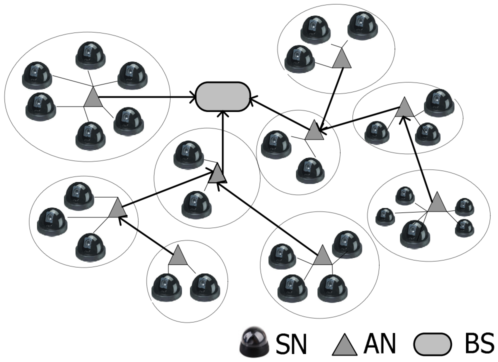


Fig. 2. Topology of WVSN.

In this paper, we design a wireless intelligent sensor network on cloud computing system for smart home applications. First, the Wireless Vision Sensor Network (WVSN) platform is investigated and the design issue for smart home is concerned. Here we propose a strategy which is more dedicated to smart home scenario. Second, we propose an object segmentation algorithm with a high performance and low complexity considerations. Third, to cope with the limited bandwidth on transmission, the foreground data is still needed to compress into a smaller amount of data. This paper is organized as follows. Section II briefly reviews the relevant literature in the aspect of smart home and the design challenges. Section III explores the whole algorithm in detail. Section IV presents the experimental results. Finally, Section V draws a conclusion.

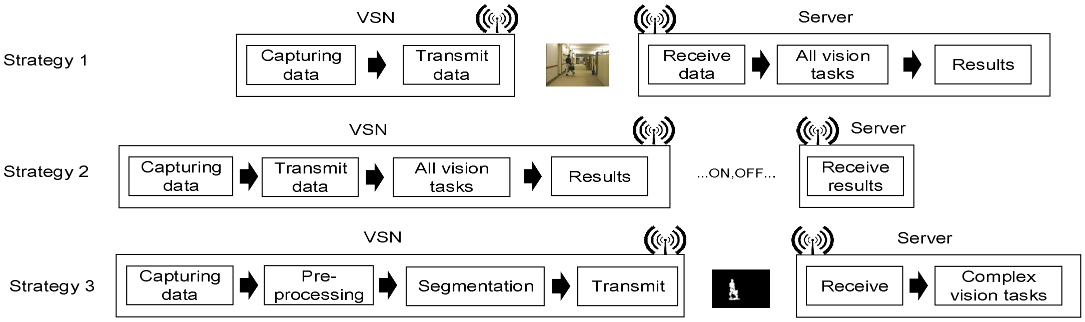


Fig. 3. Three strategies on WVSN architecture

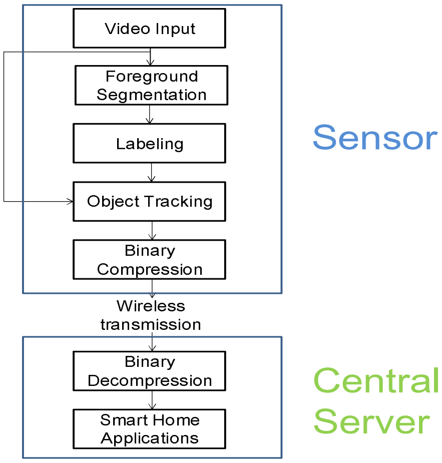


Fig. 4. Flowchart of the whole system on smart home.

2. Background on Related Work

This paper targets on the smart home scene captured by the wireless video sensor nodes. To accomplish it, the concept on WVSN is first constructed. Then the related techniques on vision processing with their challenges are discussed later.

A. WVSN

Video data on wireless sensor network is accomplished as Wireless Vision Sensor Network [1]. It is constructed based on the sensor network, as shown in Fig. 2. In WVSN, each sensor node is tasked to capture video data and is capable of performing specific content analysis tasks to extract information from the video. The captured video and the extracted information are delivered to an aggregation node (AN). The role of the AN is to process the collected data and deliver important information to the base station (BS) [1]. WVSN has been envisioned for a wide range of important applications, including security monitoring, environment tracking, and the assisted living [2]. A design with distributed interactive video arrays for situation awareness of traffic and incident monitor was presented in [3]. A multicamera tracking adjacent cameras was proposed in [4].

Currently a well-established VSN, Zigbee, is widely applied. Zigbee had flexible network structure; it could support the topology of star shape, tree shape and mesh shape [5]. Mesh structure maintains ad-hot feature on VSN however it is not easily realized on real environment. Star and binary tree structure contains less variation but the robust is controllable.

B. Design Challenges

WVSN will be enabled by the convergence of communication and computation with signal processing and several branches of control theory and embedded computing. To realize the WVSN, it often exposes to a number of challenges. Generally, researchers employ three strategies for WVSN implementation [4]-[7], as shown in Fig. 3.

In the first strategy, no local processing is performed on the VSN and raw data is directly transmitted to the server for vision processing. However, this strategy consumes large transmission time because the large amount of data is communicated [8]. Moreover, in a house environment, privacy should be the most concerned issue. The video data is transmitted on air and thus this could not be accepted in many personal and private environments.

In the second strategy, all vision tasks are performed on the VSN and only the final features are transmitted to the server for analysis [4]. The advantage is that no visible data is transmitted on air and consequently the privacy is fully reserved. This strategy forces the VSN to consume large processing effort on the currently available software platforms and has high design complexity on the hardware platforms.

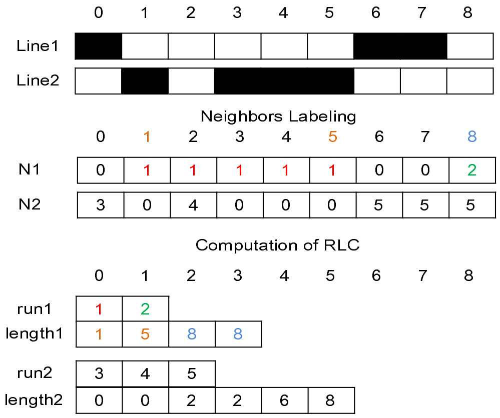
Referring to the third strategy, it moves the complex tasks i.e., labeling, feature extraction and classification, to a server. This strategy will reduce both the processing consumption and the design complexity because the complex tasks are moved to a generalized platform on the server.

3. PROPOSED ALGORITHM

In this paper, we adopt the third strategy to realize the WVSN on a smart home. We propose a high performance object segmentation algorithm dedicated to this scenario. As shown in Fig. 4, the proposed system can be divided into foreground segmentation, object labeling, object tracking, binary data compression, and the smart home applications. Details are shown as follows.

A. Foreground Segmentation

The foreground detection is meant to separate the interest objects in video. It is always the most crucial and important task in an object segmentation flow. The first step is the background modeling where we consider the rate control issue. The second step is the background subtraction and the third step is the area filter. Background subtraction method is a comprehensive concept and successfully utilized on background modeling and moving-object detection. We modified the algorithm on [9] to perform the background modeling with a more rapid convergence rate. The proposed algorithm utilizes Gaussian mixture models (GMM) and adaptive mixture learning. By this technology, each pixel is framed as a mixture of Gaussians. Then the on-line approximation is used to improve it. Here we evaluate these Gaussian distributions to judge the most likely result.



	0	1	2	3	4	5	6	7	8
Line1	■	□	□	□	□	□	■	■	□
Line2	□	■	□	■	■	■	□	□	□

Neighbors Labeling

	0	1	2	3	4	5	6	7	8
N1	0	1	1	1	1	1	0	0	2
N2	3	0	4	0	0	0	5	5	5

Computation of RLC

	0	1	2	3	4	5	6	7	8
run1	1	2							
length1	1	5	8	8					
run2	3	4	5						
length2	0	0	2	2	6	8			

Fig. 5. Neighbors labeling and RLC.

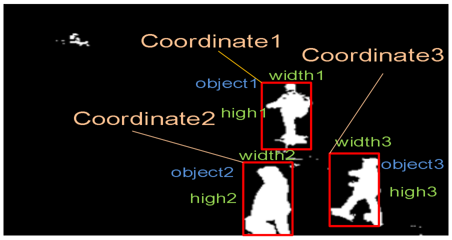
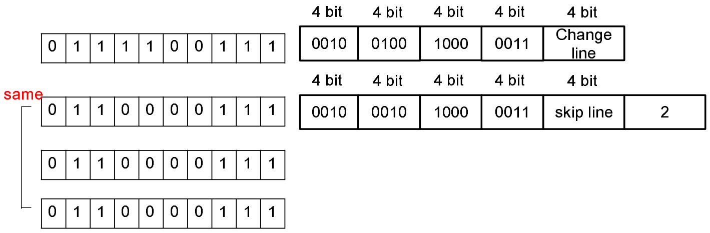


Fig. 6. Object information.



0	1	1	1	1	0	0	1	1	1
0	1	1	0	0	0	0	1	1	1
0	1	1	0	0	0	0	1	1	1
0	1	1	0	0	0	0	1	1	1

4 bit	4 bit	4 bit	4 bit	4 bit	
0010	0100	1000	0011	Change line	
4 bit	4 bit	4 bit	4 bit	4 bit	
0010	0010	1000	0011	skip line	2

Fig. 7. RLC with skip-line algorithm.

B. Object Labeling

After we have the binary foreground mask, we have to label the connect pixel. We need the object labeling algorithm with fast and low complexity issue. We use the approach where it is a combination of run-length encoding (RLC) and object labeling of algorithms. Fig. 5 shows the neighbors labeling and calculation on RLC. Our approach will label line by line. Only the foreground data, representing as white, is needed with labeling. As shown in Fig. 5, in the first line N1, the data on position 1-5 are connected together. They will be given as the same number, where 1 (marked as red) is the run of position 1-5, and 2 (marked as green) is run of position 8. The second line N2 is performed with the same procedure. Then we derive a run-length representation for N1 (run1, length1) and N2 (run2, length2) respectively.

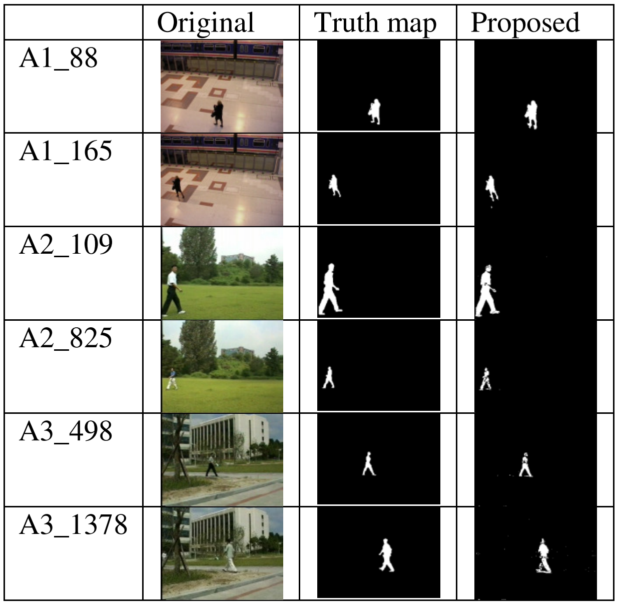


Fig. 8. Result of foreground segmentation

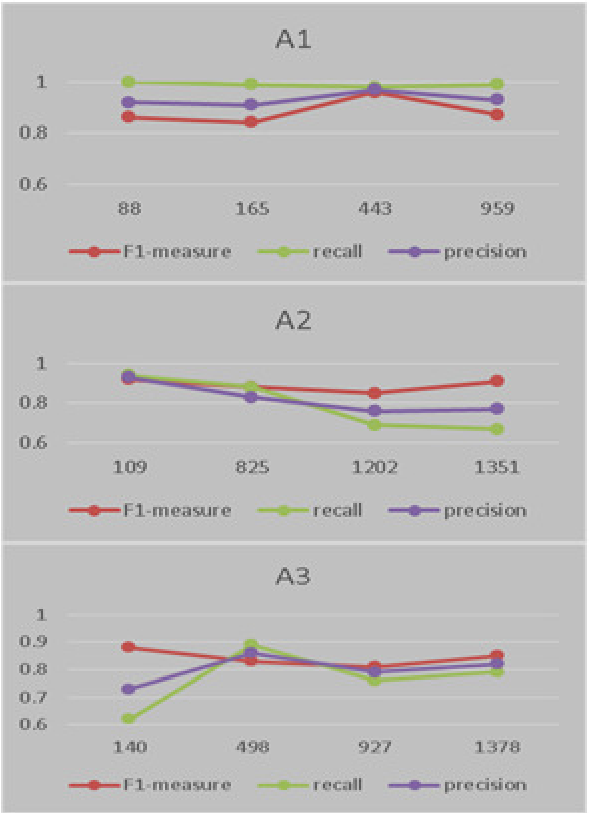


Fig. 9. The evaluation of the foreground segmentation.

C. Object Tracking

We use the color feature to establish a color distribution model. Then we use a Kernel-based algorithm. If the model already exists, we can just track it. If this model does not exist, we create a new model. When an object moves, we can directly use the object labeling information to find the object. When the object is stationary, we can use the mean-shift algorithm to find the candidate model. Kernel-based tracking must meet a number of conditions. Meanwhile, the object cannot move faster than a certain speed and the kernel of the object cannot change too much.

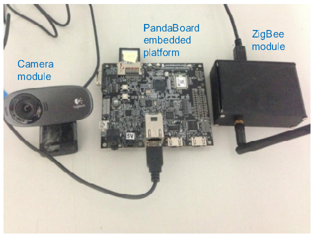


Fig. 10. View of the embedded system

D. Binary Data Compression

Now we have successfully tracked all objects, and each object is given with a label and recorded with the coordinates. This information facilitates our binary data compression. Taking into account the issue of privacy and data size, we will send the foreground object mask and size also the coordinate information, as shown in Fig. 6. Since we have tracked each object separately, we can use this information to calculate the size of each object.

The main compression technique is run-length coding with an additional concept on skip-line method. As a foreground image, two concessive rows usually have high degree of similarity. When many lines of data are the same, if we use this part of a symbol to represent. As shown Fig. 7, since the data in the first and second line is different, the data on line 2 is kept completely. From the second line to the forth line, all the data are identical. In that case we just add a skip-line symbol. The value of this symbol is 2 representing that there are two more identical rows existed.

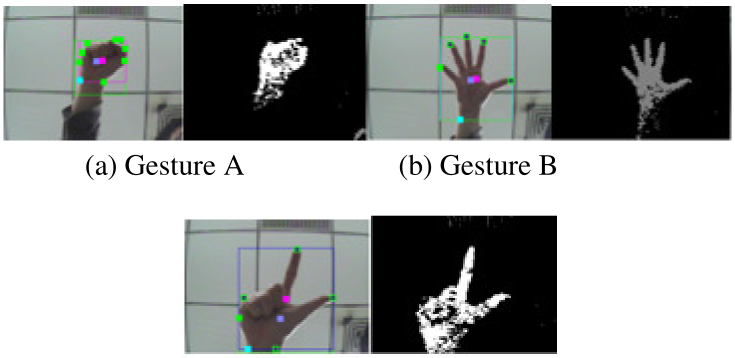
4. Experimental Results

In order to verify the results of our algorithm, we will simulate the individual algorithm with the experimental result. Some evaluation and comparison are provided.

We dedicate many experiments on foreground segmentation. Obviously foreground segmentation will directly affect the accuracy of performance. If the some unnecessary portions are included, the system will waste time on scanning and judgment. Also if the cut is less the truth case, it will lead to the failure or error on judgment. We analyze the results and draw their own truth map. Fig. 8 shows the truth map and the prospect the proposed algorithm cut out. By this subjective view, the similarity between them is very high. To measure the accuracy, the recall rate, precision and F1-measure are the most comprehensive assessments. In our three test sequences, F1-measure can reach above 0.8. The evaluation is shown in Fig. 9.

Referring to the compression rate analysis, because our binary data compression method is a lossless technique, no quality loss is existed between the source and decoded result. The compression ratio of our method varies from 20 to 500 in some frame. In average 50 times of compression ratio is achieved.

The coordinator is as our central server. All end devices are the sensors which can only return data to a central server.



(a) Gesture A (b) Gesture B

(b) Gesture C

Fig. 11. Hand gesture recognition result.

Then the server can determine which end device is now transmitting the data. In the VSN, a camera is plug at an embedded system board. Here we use the PandaBoard embedded platforms with camera and Zigbee modules. Fig. 10 shows the view of the embedded system. This board is based on ARM Cortex-A9 dual-core processor and has 1GB memory for usage. We use USB cameras and USB Zigbee modules. Several gestures are identified by the server. Fig. 11 shows the different gesture recognition results.

5. CONCLUSION

In this paper, we realize a smart home based on the implementation on WVSN as a cloud computing system. Due to different design philosophies on WVSN, this paper applies a low complexity approach which concerns the characteristics on the VSN and the server. For the algorithm level, a complete system with efficient vision-based tasks is provided. Versatile test sequences are simulated to prove the feasibility of the propose algorithm. Furthermore, a physical system of WVSN is constructed and some intelligent applications are realized.

REFERENCES

[1] F. Akyildiz, T. Melodia, and K. R. Chowdhury, “A survey on wireless multimedia sensor networks,” Elsevier Comput. Netw., vol. 51, pp. 921–60, Mar. 2007.

[2] M. Valera and S. Velastin, “Intelligent distributed surveillance systems: A review,” IEEE Proc. Vis. Image, Signal Process., vol. 152, no. 2, pp. 192–204, Apr. 2005.

[3] M. M. Trivedi, T. L.Gandhi, and K. S. Huang, “Distributed interactive video arrays for event capture and enhanced situational awareness,” IEEE Intell. Syst., vol. 20, no. 5, pp. 58–66, Oct. 2005.

[4] M. Quaritsch, M. Kreuzthaler, B. Rinner, H. Bischof, and B. Strobl, “Autonomous multicamera tracking on embedded smart cameras,” in EURASIP J. Embed. Syst., 2007.

[5] F. Akyildiz, T. Melodia, and K. R. Chowdhury, “A survey on wireless multimedia sensor networks,” Elsevier Comput. Netw., vol. 51, pp. 921–60, Mar. 2007.

[6] Imran, M.; Ahmad, N.; Khursheed, K.; Waheed, M.A.; Lawal, N.; O'Nils, M., "Implementation of Wireless Vision Sensor Node With a Lightweight Bi-Level Video Coding," Emerging and Selected Topics in Circuits and Systems, IEEE Journal on , vol.3, no.2, pp.198,209, June 2013.

[7] M. Imran, K. Khursheed, M. O’Nils, and N. Lawal, “Exploration of target architecture for a wireless camera based sensor node,” in IEEE Norchip Conf., Nov. 2010, pp. 1–4.

[8] S. Soro and W. Heinzelman, “A survey of visual sensor networks,” Adv. Multimedia, pp. 1–22, May 2009.

[9] D.-S. Lee, "Effective Gaussian Mixture Learning for Video Background Subtraction," IEEE Transactions on Pattern Analysis and Maching Intelligence, vol. 27, no. 5, MAY 2005.

INTENTIONAL BLANK

A CROSS-LAYER DELAY-AWARE MULTIPATH ROUTING ALGORITHM FOR MOBILE ADHOC NETWORKS

Mahadev A. Gawas[1], Lucy J.Gudino[2], K.R. Anupama[3], Joseph Rodrigues[4]

[1,2] Department of Computer Science BITS PILANI K.K. Birla Goa campus.
[1]mahadev@goa.bits-pilani.ac.in
[2]lucy@goa.bits-pilani.ac.in
[3]Department of EEE/EI BITS PILANI K.K. Birla Goa campus.
[3]anupkr@goa.bits-pilani.ac.in
[4]Department of Electronics ATEC Verna goa.
Joseph1_x_r@rediffmail.com

ABSTRACT

Mobile Ad Hoc Networks (MANETS) require reliable routing and Quality of Service(QoS) mechanism to support diverse applications with varying and stringent requirements. Routing protocols such as AODV, AOMDV, DSR and OLSR use minimum hop count as the metric for path selection, hence are not suitable for delay sensitive real time applications. To support such applications delay constrained routing protocols are employed. These Protocols makes path selection based on the delay over the discovered links during routing discovery and routing table calculations. We propose a variation of a node-disjoint Multipath QoS Routing protocol called Cross Layer Delay aware Node Disjoint Multipath AODV (CLDM-AODV) based on delay constraint. It employs cross-layer communications between MAC and routing layers to achieve link and channel-awareness. It regularly updates the path status in terms of lowest delay incurred at each intermediate node. Performance of the proposed protocol is compared with single path AODV and NDMR protocols. Proposed CLDM-AODV is superior in terms of better packet delivery and reduced overhead between intermediate nodes.

KEYWORDS

AODV, Cross Layer, MANET, MAC, NS2, QoS.

1. INTRODUCTION

MANETs are self-organizing, rapidly deployable wireless network that require no fixed infrastructure. It is composed of wireless mobile nodes that can be deployed anywhere, and can dynamically establish communications using limited network management. Real time applications have been most popular among the applications run by ad hoc networks. It strictly adheres to the QoS requirements such as overall throughput, end-to-end delay and power level. Traditionally multihop wireless network protocol design is largely based on a layered approach. Here each layer in the protocol stack is designed and operated independently with interfaces between layers that are rather static. This paradigm has greatly simplified network design and led to the robust scalable protocols on the internet. However, the rigidity of this paradigm results in poor performance for multihop wireless networks in general, especially when the application has high bandwidth requirements and/or stringent delay constraints [1]-[4].

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 113–123, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4513

1.1 RELATED WORK

To meet these QoS requirements, recent study on multihop networks has demonstrated that cross-layer design which can significantly improve the system performance [5]-[6]. To guarantee QoS in MANETs for delay sensitive applications two factors are considered. Firstly, route selection criterion must be QoS-aware i.e., it must consider the link quality before using the link to transmit. Secondly, the instantaneous response to the dynamics of MANET topology changes must be considered so that the route changes are seamless to the end user over the life time of a session. Generally, a QoS model defines the methodology and architecture by which certain types of services can be provided in the network. Protocols such as routing, resource reservation signaling and MAC must cooperate to achieve the goals set by the QoS model. QoS routing is one of the most essential parts of the QoS architecture [7]–[9]. Multipath approach has many advantages such as load balancing, QoS assurance and fault tolerance [10]- [12]. Several multipath routing protocols have been proposed so far in the literature. One of the earliest multipath routing protocols is Ad hoc On demand Multipath Distance Vector (AOMDV) [13]. AOMDV is a variant of Ad Hoc On Demand Distance Vector (AODV) [14] which establishes loop-free and link-disjoint paths based on the minimum hop count. QoS AODV (QS-AODV) in [15] extended the basic AODV routing protocol to provide QoS support in MANETs. It uses hop count as criterion for choosing the route with an assumption that NODE_TRAVERSAL_TIME (NTT) is constant. Stephane Lohier et al.[16] proposed reactive QoS routing protocol that also deals with delay and bandwidth requirements. In his proposal, QoS routes are traced by node to node and NTT is an estimate of the average one-hop traversal time, which includes queue, transmission, propagation, and other delays.

Cross-layered multipath AODV (CM-AODV)[17]**,** selects multiple routes on demand, based on the signal-to-interference plus noise ratio (SINR) measured at the physical layer**.** Load Balancing AODV (LBAODV)[18] is a new multipath routing protocol that uses all discovered path simultaneously for transmitting data. By using this approach data packets are balanced over discovered paths and energy consumption is distributed across many nodes throughout the network.

Xuefei Li et al. [19] proposed Node-Disjoint Multipath Routing protocol (NDMR) by modifying and extending AODV to enable the path accumulation feature of DSR in route request packets. Multiple paths between source and destination nodes are discovered with low broadcast redundancy and minimal routing latency. A delay aware protocol proposed in Boshoff et al. [20], uses end-to-end delay, instead of hop count, as metric for route selection. Upon route failure, the route table which contains multiple paths, along with the end-to-end delay is first searched for an alternative route to the destination before a new route discovery process is initiated. Even though it reduces both routing overhead and end-to-end packet delay, the route delay information might not always be upto date. Perumal Sambasivam et al. [21] modified the AODV protocol's route discovery mechanism by incorporating multiple node-disjoint paths for a particular source node along with mobility prediction**.**

Thus, it is found that most approaches to multipath routing protocols consider the end-to-end delay. They do not emphasize on considering the processing delay incurred at each node which may indicate the congestion or link quality along the path which is node disjoint. They also do not have a mechanism to handle expiry of stale cached routes in the route table before making their selection. Hence we propose a new algorithm CLDM-AODV with cross-layer communications between MAC and Routing layers to achieve link and channel-awareness. In section II, we describe the proposed algorithm. We present simulation results in section III followed by conclusion.

2. PROPOSED CLDM-AODV ROUTING ALGORITHM

The proposed algorithm considers only node disjoint routes which satisfy the end-to-end delay specified in the route request. For calculating end-to-end delay, the algorithm estimates inter-node packet processing delay at each node. Source node makes a selection of primary path out of available multiple QoS enable paths. The proposed algorithm includes calculation of inter-node packet processing delay at each mobile node, initiation of route discovery and route reply processes.

2.1. END-TO-END DELAY

In general, total latency or delay experienced by a packet to traverse the network from source to destination may include routing delay, propagation delay and processing or node delay. Routing delay is the time required to find the path from source to destination. Propagation delay is related to propagating bits through wireless media. Processing delay involves the protocol processing time at node x for link between node x and node y. The end-to-end delay of a path is the sum of all the above delays incurred at each link along the path [17]. For MANETs, propagation delays are negligibly small and almost same for each hop along the path. The major factors involved in computation of processing delay are the queuing delay and delay incurred at the MAC layer processing.

In the proposed method, we have named processing delay as Packet Processing Delay (PPD) which includes queuing delay and delay incurred at the MAC contention. IEEE 802.11 MAC with the distributed coordination function (DCF) is used as MAC protocol and the access method is Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) with acknowledgments. To transmit packets, nodes make use of request-to-send (RTS), clear-to-send (CTS), data and acknowledgement (ACK) packets. The amount of time between the receipt of one packet and the transmission of the next is called a short inter frame space (SIFS). Average queuing Delay at the node i is $\overline{D}_i$ is given by equation [22],

$$\overline{D_i} = \alpha\overline{D_{j-1}} + (1-\alpha)\overline{D_j} \qquad (1)$$

where,

$$\alpha = \frac{(queue_{size} - queue_{length})}{queue_{size}} \qquad (2)$$

$queue_{size}$ is the current size of the queue at node *i*, $queue_{length}$ is the length of the queue at node *i* and j is the current period.
The channel occupation due to MAC contention is given by,

$$T_{mac} = T_{RTS} + T_{CTS} + 3 * T_{SIFS} + T_{acc} \qquad (3)$$

T_{RTS} and T_{CTS} are the time periods on RTS and CTS respectively and T_{SIFS} is the SIFS period. T_{acc} is the time for channel contention. The Packet Processing Delay (PPD) is given by:

$$PPD = \overline{D}_i + T_{mac} \qquad (4)$$

2.2. ROUTE DISCOVERY

Generally in reactive protocols[1], when a source node ‘S’ has to communicate with destination node ‘D’, it initiates path discovery by broadcasting a route request packet RREQ to its neighbours. The <source-address, broadcast-id> pair is used to identify the RREQ uniquely. In the proposed system, during initial route discovery phase, more than one node disjoint path between the source and destination is determined and optimal path which satisfies QoS delay requirement is chosen for the data transmission. When this primary path breaks due to nodes mobility or path fails to satisfy QoS requirement, then one of the alternate path is chosen as the next primary path and data transmission can continue without initiating another route discovery thus reducing the overhead of additional route discovery. In the proposed algorithm, the RREQ packet is modified to contain the address of the source through which it is forwarded. The packet header contains additional field for PPD and Thresh_Delay. PPD is initialized to zero and subsequently updated at each intermediate node as per Eq.(4). Thresh_Delay is set to the maximum allowable time delay for any path from source to destination. Since RREQ is flooded network-wide, a node may receive multiple copies of the same RREQ. After receiving the first RREQ, an intermediate node can receive and collect subsequent RREQ copies for the predetermined time duration, RREQ_WAIT_TIME, which is assumed as 20ms. The intermediate node also maintains RREQcounter to limit the number of RREQ that it can receive. In our proposed system, we initialize RREQcounter to three which is as shown in Figure 1. On receiving up to three RREQs, the route with minimum PPD selected which ensures the path with highest quality. Before forwarding the RREQ, intermediate node computes its PPD and compares it with Thresh_Delay**.** If the difference between the Thresh_Delay and current value of its PPD is zero or negative, it drops the RREQ packet avoiding unnecessary flooding into the network. If it satisfies, node broadcasts the packet by updating Thresh_Delay value less by currently computed PPD value of the node. Since every intermediate node forwards only one RREQ towards the destination, each RREQ arriving at the destination has traveled along a unique path from source to destination. Figure 2 shows an example of the delay based route discovery.

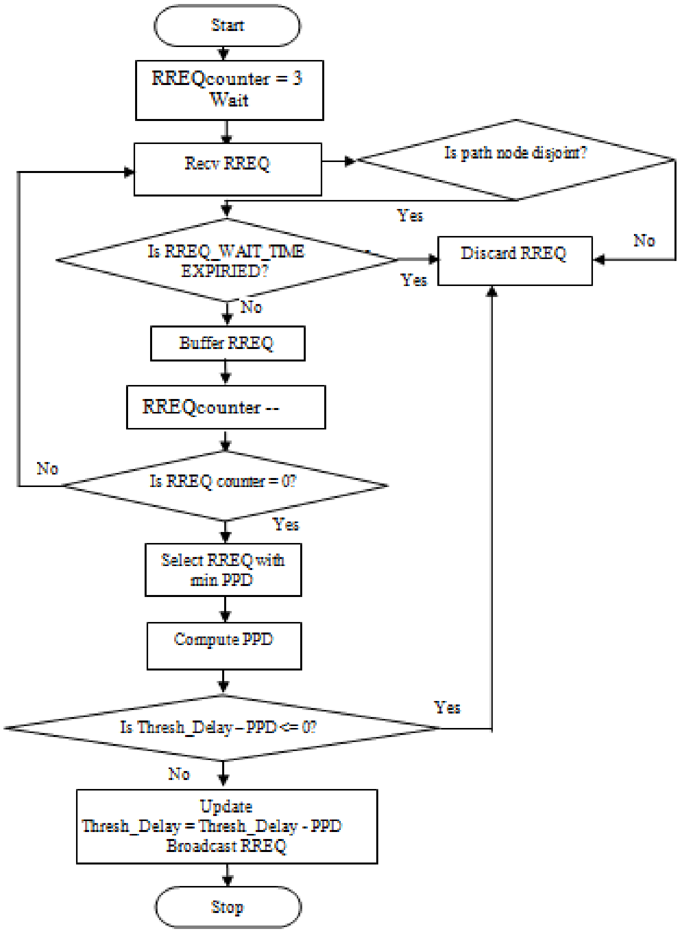


Fig. 1: RREQ Flowchart of proposed CLDM-AODV

Source node S initiates route request by updating the Thresh_Delay in RREQ Packet to acceptable delay say 100ms and PPD to zero. On receiving RREQ, node 1, 2 and 3 computes their PPD and updates Thresh_Delay in respective RREQ packet. Node 4 receives three RREQs, from node 1, node 2 and node 3 respectively. PPD values of these RREQs are compared and minimum PPD path from node 2 is chosen. Node 4 broadcast the RREQ, since it's computed PPD value satisfies the QoS constraint i.e. the difference between Thresh_Delay and PPD of node 4 is greater than zero. On the other hand, at node 5, RREQ packet gets dropped as difference between Thresh_Delay and PPD at node 5 do not satisfy the QoS criteria. Destination node D receives two RREQs from node 6 and node 4 respectively. D buffers both the paths for the route reply.

2.3. ROUTE REPLY

In proposed CLDM-AODV destination node D can collect up to RREQcounter times RREQ packets within time duration RREQ_WAIT_TIME, which is assumed to 20 ms. Node D generates a route reply RREP packets in response to every RREQ copy that arrives from the source S via loop-free and node disjoint paths to the destination. RREP packet is an extension of AODV RREP packet with additional field Max_PPD, which will hold the maximum packet processing time at intermediate nodes along the reverse path. Before destination node forwards the RREP, it computes the PPD and updates it in the Max_PPD field as shown in Figure 3. On reaching the next node, the intermediate node computes its PPD and compares it with the value in the Max_PPD field of RREP packet if current PPD computed is more than value in the Max_PPD. On receiving the RREP from all the disjoint routes, the source selects the primary route with minimum Max_PPD value. This signifies that the packet travelled through the less congested network, and possibility of packet incurring extra delay or getting dropped on the path is very low. Figure 4 shows an example of node disjoint route reply procedure. Destination node D calculates its PPD which is 25 ms and initializes Max_PPD with that PPD. Node D then sends RREP packets to all QoS qualified RREQ routes. Intermediate nodes 4 and 6, on receiving the RREP compute their own PPD i.e. 45 ms and 15 ms respectively. This value is compared with Max_PPD field of RREP packet. If PPD value is less or equal to Max_PPD, it ignores else it replaces the Max_PPD value in RREP packet. Node 6 does not modify Max_PPD as its computed PPD value is less than Max_PPD whereas node 4 replaces Max_PPD with 45 ms as its computed PPD value is greater than Max_PPD. Source node S on receiving the multiple RREP, it buffers them in the route table. Source S chooses the path with minimum value of Max_PPD as primary path i.e. path which source receives from node 1 as its Max_PPD value is 25 ms. If source does

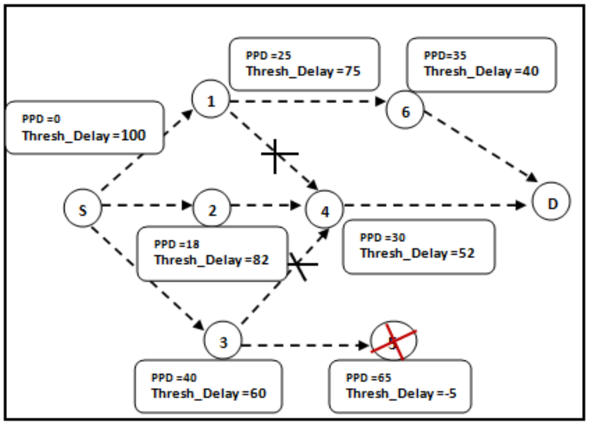


Fig. 2: Route Discovery of proposed CLDM-AODV

not receive RREP in RREP WAIT_TIME from destination, then it restart route discovery with new session Id.

2.4. ROUTE MAINTENANCE

Route maintenance is very essential as there are high chances of route failure and QoS constraint violation due to mobility. Route failure due to link breakage is handled by the method using periodic *Hello* packets [15]. Any node which detects either a QoS violation or a link failure, informs the source by sending a route error packet (RERR). If a source node itself moves, restart the route discovery procedure to find a new route to the destination. If a node along the route moves so that it is no longer reachable, its upstream neighbor sends a link failure notification message to each of its active upstream neighbors through RERR until reaches the source node. QoS violation due to end-to-end delay constraint is detected by the intermediate nodes by computing one way delay experienced by the data packets from the sender's timestamp on the received data packets. During data transmission, source node appends the Thresh_Delay information to the data packets. Intermediate nodes on receiving the data packets, finds the difference between current time and time stamp of data packet. If value is less than Thresh_Delay, it generates the RERR packet to the source, or else forwards the packet to the next hop in the route table.

In our proposed CLDM-AODV, we introduce a method to validate other alternate node disjoint paths already discovered. At regular interval of time, Life Line Packets (LLP) is forwarded

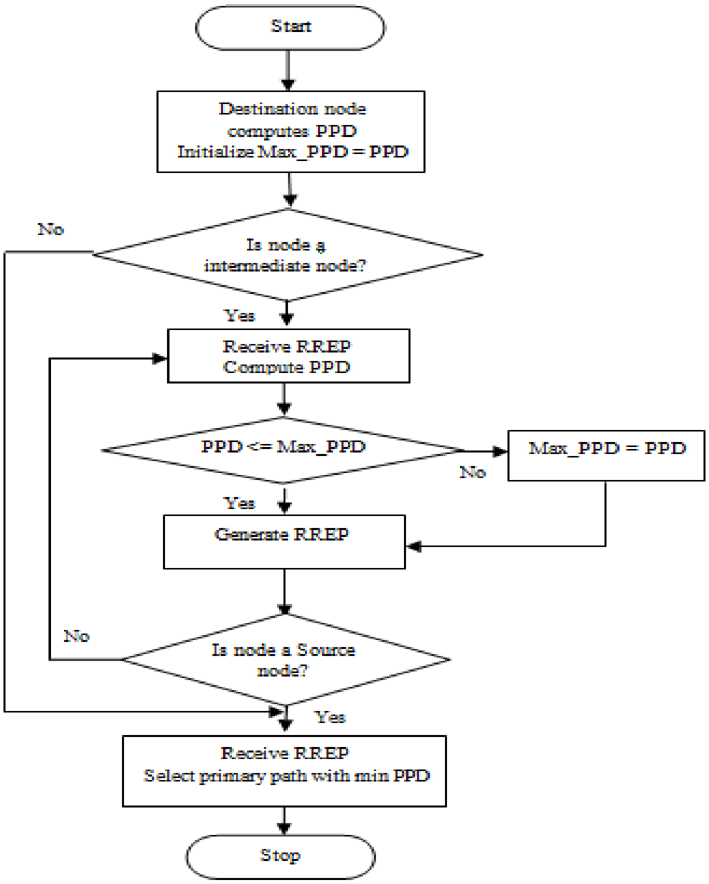

Start
Destination node computes PPD Initialize Max_PPD = PPD
Is node a intermediate node?
No
Yes
Receive RREP Compute PPD
PPD <= Max_PPD
No
Max_PPD = PPD
Yes
Generate RREP
Is node a Source node?
No
Yes
Receive RREP Select primary path with min PPD
Stop

Fig. 3: RREP Flowchart of proposed CLDM-AODV

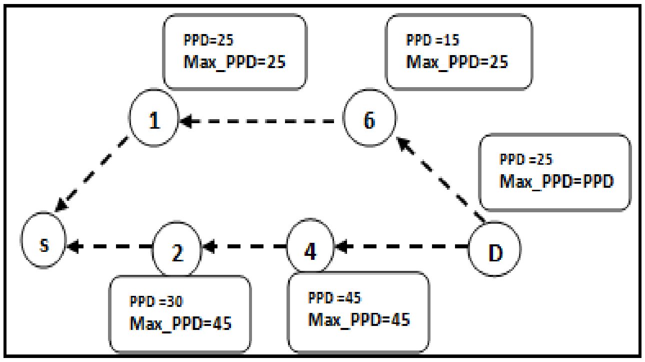


Fig. 4: Route Reply of proposed CLDM-AODV

through alternate paths which contain Thresh_Delay. Intermediate nodes on receiving LLP, verifies the eligibility of packet forwarding by computing difference between current time and time stamp of data packet. If it is less than Thresh_Delay, it generates the RERR packet to the source indicating that the path is no longer QoS compliance link and corresponding path entry is deleted from route table. Destination node replies to these LLP by the same procedure as followed during RREP packets. On receiving the fresh route quality, source updates the primary path with highest quality.

3. Simulation Experiments

3.1 Simulation Environment

The performance of the proposed CLDM-AODV protocol is evaluated and compared with AODV and NDMR. Simulations are conducted on the Network Simulator (ns-2) with network comprising of 50 wireless ad hoc nodes moving over an area of 1500m x 300m for 900s of simulated time. Physical layer is a bi-directional link and channel transmission rate is 2Mbps. At MAC layer, the DCF of IEEE 802.11 standard for wireless LANs is assumed. RTS and CTS packets are exchanged before the transmission of data packets. The channel propagation model we used two-ray ground reflection model. Constant Bit Rate (CBR) traffic is used. A 512-byte data packet with 2 packets/second sending rate is assumed for all the experiments. Inter packet time is assumed to be 35 ms. Radio transmission range of each node is set to 250m. The initial placement of nodes is random and random waypoint mobility model [24] is used to simulate node movements. Simulation is run for seed value of 1 to 9.
The simulation parameters are shown in table 1.

Table 1

Parameters	Value
NS version	Ns –allinone-2.35
Number of nodes	50
Simulation Time	900 sec
Radio transmission range	250m
Traffic	CBR(Constant Bit Rate)
CBR Packet size	512 bytes
Simulation Area size	1500m * 300 m
Node Speed	4m/s to 20 m/s
Mobility model	Random WayPoint mobility

We compare the performance of AODV, NDMR and CLDM-AODV using the following three metrics:

1. *Control Overhead* is the ratio of the number of protocol control packets transmitted to the number of data packets received.
2. *Packet Delivery Fractions (PDF)* is the ratio of the data packets delivered to the destination to those generated by the CBR sources.
3. *Average end-to-end delay* is an average end to delay of all successfully transmitted data packets from source to destination.

3.2 Simulation Results

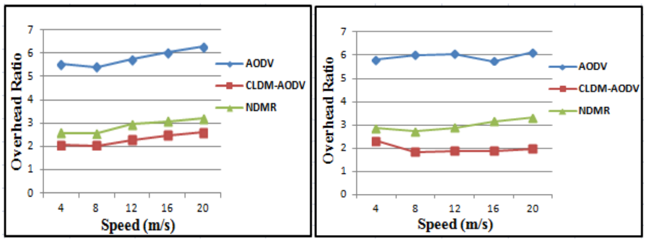
Example 1*:* In this example, we analyze the effect of speed on control overhead, PDF and average end to end delay for different number of source nodes in the network. In the simulation we assume the number of sources to be 30 and 35 and mobility of nodes is 4 meters/sec to 20 meters/sec.

Figure 5(a)-(b) shows the plot of control overhead vs. speed. It is evident from the result that CLDM_AODV has minimum control overhead compared to AODV and NDMR. In Figure 6, average control overhead ratio for sources 30, and 35 is plotted. It is easily inferred that CLDM_AODV has smaller overhead than AODV and NDMR in harsh operation environments. This improvement is mainly because multiple QOS compliance routes are discovered in single route discovery phase, which significantly reduces frequent route discovery on route failure.

Figure 7(a)-(b) shows the plot for End-to End delay vs. speed. It can be seen from the plot corresponding to AODV that there is an increase in delay which is due to high mobility of nodes which in turn results in increased probability of link failure that causes an increase in the number of routing rediscovery processes. This makes data packets to wait for more time in its queue until a new routing path is found. Average end-to-end delay in NDMR does not show much variation over varying speed and shows better results compared to AODV.

In Figure 8, average End-to End delay vs. speed for sources 30, and 35 is plotted. In proposed CLDM-AODV protocol, delay curve remains consistently low compared to AODV and NDMR even though extra waiting time, RREQ_WAIT_TIME, is added in route discovery process. Addition of RREQ_WAIT_TIME has little effect on the overall performance since CLDM-AODV has multiple alternate node disjoint paths satisfying the delay constraint, leads to less route discoveries. Also source regularly uses the primary path with optimal quality.

A packet delivery ratio for AODV, NDMR and CLDM_AODV is as shown in Figure 9(a)-(b). In Figure 10, average Packet delivery ratio vs. speed for sources 30, and 35 is plotted. Since CLDM_AODV attempts to use optimal QoS enabled node disjoint path among available multiple alternate paths for data delivery, the protocol is able to deliver more packets to the destination compared to AODV and NDMR.



(a) 30 source nodes (b) 35 source nodes

Fig. 5(a)-(b): control packet overhead vs. speed (m/s).

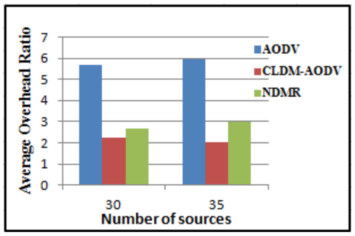
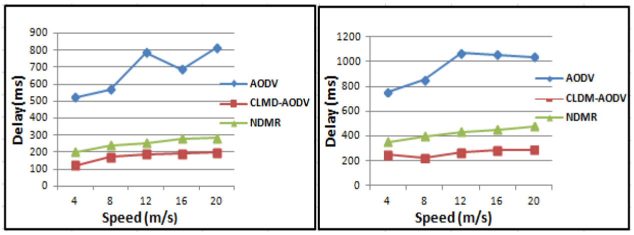


Fig. 6: Average Control overheads for varying number of sources for speed 4m/s to 20 m/s



(a) 30 source nodes (b) 35 source nodes

Fig. 7(a)-(b): End-to-End delay (ms) vs. speed (m/s).

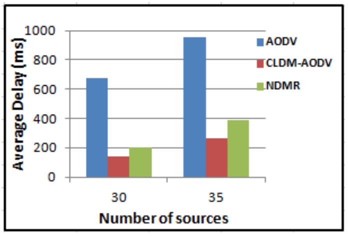
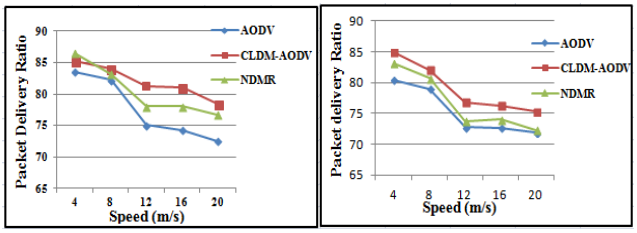


Fig. 8: Average end-to-end for varying number of sources from speed 4m/s to 20 m/s



(a) 30 source nodes (b) 35 source nodes

Fig. 9(a)-(b): Packet delivery ratio vs. speed (m/s).

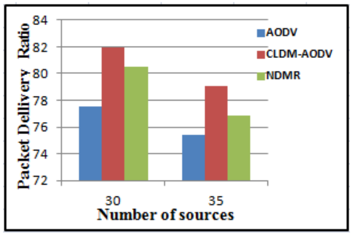


Fig. 10 Average Packet delivery ratio for varying number of sources from speed 4m/s to 20 m/s

AODV simply drops data packets when routes are disconnected, as it has to resort to a new discovery when the only path fails. Proposed CLDM_AODV algorithm performs better as data packets travel through less congested and delay compliance.

4. CONCLUSION

A new algorithm CLDM-AODV suitable for delay sensitive application is presented. Proposed CLDM-AODV algorithm with multipath capability effectively deals with high mobility traffic route failures in MANET. Proposed algorithm ensures that the multiple paths are loop-free and is node disjoint. Comparative study of CLDM-AODV, classical AODV and NDMR is performed using ns-2 simulations under varying mobility and traffic scenarios. The results indicate that CLDM-AODV has lower average end-to-end delay even by including extra fields to RREQ and RREP packets to provide QOS support. The routing overhead is low compared to its counter parts as route discovery process is minimized by providing QOS compliance alternate routes. The added advantage of the proposed algorithm is, it periodically checks the paths obtained during route discovery process and uses optimal link for data communication.

REFERENCES

[1] R. Jurdak, Wireless Ad Hoc and Sensor Networks 9th ed. Springer Series, United States of America.2006.

[2] C.K.Toh, Ad Hoc Mobile Wireless Networks: Protocols and Systems. 2nd ed. Printice Hall, China, 2001.

[3] C. E. Perkins, E. M. Belding-Royer, “Quality of Service for Ad hoc On- Demand Distance Vector Routing, Internet Draft,” October, 2003.

[4] X. Li and L Cuthbert, “Multipath QoS routing of supporting Diffserv in Mobile Ad hoc Networks,” Proceedings of SNPD/SAWN'05, 2005.

[5] Z. Ye, S. V. Krishnamurthy and S. K. Tripathi, “Framework for Reliable Routing in Mobile Ad Hoc Networks,” IEEE INFOCOM 2003.

[6] P. Sambasivam, A. Murthy, and E. M. Belding-Royer, “Dynamically Adaptive Multipath Routing based on AODV,” proc of the 3rd Annual Mediterranean Ad Hoc Networking Workshop (MedHocNet), Bodrum, Turkey, June 2004.

[7] P. Macharla, R. Kumar, A. Kumar Sarje, “A QoS routing protocol for delay-sensitive applications in mobile ad hoc networks,” COMSWARE. 2008, pp. 720-727.

[8] S.Chakrabarti, and A. Mishra, “QoS issues in ad-hoc wireless networks,” IEEE Communication Magazine, Feb 2001,Vol. 39, No. 2, pp.142 148.

[9] M.K. Gulati and , K. Kumar, “A review of QoS routing protocols in MANETs”, 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Jan. 4 - 6, 2013.

[10] C. Chen, W. Wu Z. Li, “Multipath Routing Modeling in Ad Hoc Networks,” Proc. of IEEE ICC, May 2005, pp.2974-2978.

[11] P. Wannawilai, C. Sathitwiriyawong, "AOMDV with Sufficient Bandwidth Aware," Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on Computer and Information Technology, June 2010, pp.305-312.

[12] C. Ahn, S. Chung, T. Kim, and S. Kang, “A node-disjoint multipath routing protocol based on aodv in mobile ad hoc networks,” In Information Technology: New Generations (ITNG), 2010 Seventh International Conference, April 2010, pp. 828 –833.

[13] M.K.Marina , and S.R.Das, “On-Demand multipath distance vector routing in ad hoc networks,” Proceedings of the 9th IEEE International Conference on Network Protocols (ICNP), 2001.

[14] C. E. Perkins, E. M. Royer, and S. R. Das, Ad hoc on-demand distance vector routing, Internet Draft, 2002.

[15] C.E. Perkins, and E.M. Belding-Royer, “Quality of Service for Ad Hoc on Demand Distance Vector Routing,” draft-perkins-manet-aodvqos-02.txt, Mobile Ad Hoc Networking Working Group Internet Draft, 14 October 2003.

[16] S. Lohier, S. Senouci, Y. M. Ghamri Doudane and G. Pujolle, “A reactive QoS Routing Protocol for Ad Hoc Networks,” European Symposium on Ambient Intelligence (EUSAI'2003), Eindhoven, Netherlands, November 2003.

[17] J. Park, S. Moh†, and I. Chung, “A Multipath AODV Routing Protocol in Mobile Ad Hoc Networks with SINR-Based Route Selection,” ISWCS '08. Wireless Communication Systems, October 2008, pp. 682–686.

[18] E. Mehdi, E.MohammadReza, D.Amir, Z.Mehdi, and Y. Nasser, "Load Balancing and Route Stability in Mobile Ad Hoc Networks base on AODV Protocol," Proceeding of International Conference on Electronic Devices, Systems and Applications(ICEDSA2010), 11-14 April 2010, pp. 258 – 263.

[19] X. Li and L. Cuthbert, "Multipath QoS Routing of supporting DiffServ in Mobile Ad hoc Networks," SNPD-SAWN '05 Proceedings of the Sixth International Conference on Software Engineering, IEEE Computer Society Washington, DC, USA 2005, pp. 308-313.

[20] Boshoff, J.N., Helberg, A.S.J. (2008), " Improving QoS for Real-time Multimedia Traffic in Ad-hoc Networks with Delay Aware Multi-path Routing ," IEEE, Wireless Telecommunication Symposium, WTS 2008, pp. 1-8.

[21] P. Sambasivam, A. Murthy, E. M. Belding-Royer, "Dynamically Adaptive Multipath Routing based on AODV," Med-Hoc- Net, (2004), pp. 16-28.

[22] M. Obaidat, "A Novel Multipath Routing Protocol for Manets," Wireless Communications, Networking and Mobile Computing (WiCOM), 2011, pp.1-6.

AUTHORS

[1]Mahadev Anant Gawas
Email: mahadev@goa.bits-pilani.ac.in
Phone: 0832 – 2580330
Department of CS & IS,
BITS Pilani, K. K. Birla, Goa Campus,
NH 17-B, Bypass Road, Zuarinagar,
PIN - 403 726, Goa, India.

Research domain:

- Wireless Networking
- Mobile Ad-Hoc Networks.
- Cross Layer

INTENTIONAL BLANK

HYBRID MAC PROTOCOL FOR WIRELESS SENSOR NETWORKS USED IN TIME CRITICAL APPLICATIONS

Pandeeswaran Chelliah[1], Pappa Natarajan[2] , and Jayesh Sundar Gopinath[3]

[1]Research Scholar, Dept of Instrumentation Engineering, MIT Campus,
Anna University, Chennai- 44
cpandees@gmail.com
[2]Professor, Dept of Instrumentation Engineering,
MIT Campus, Anna University, Chennai- 44
npappa@rediffmail.com
[3]PG Scholar, Dept. of Electronics and Instrumentation,
St.Joseph's College Of Engineering
gjayeshsundar91@gmail.com

ABSTRACT

In this paper a H-MAC protocol (Hybrid Medium Access Control protocol) has been proposed, which is an energy efficient and low latency MAC protocol which uses node ID method to assign priority for certain wireless sensor nodes that are assumed to be present in critical loops for an industrial process control domain. H-MAC overcomes some of the limitations in the existing approaches. In the case of industrial automation scenario, certain sensor loops are found to be time critical, where data's have to be transferred without any further delay, as failure in immediate transmission leads to catastrophic results for humans as well as machinery in industrial domain. The proposed H-MAC protocol is simulated in NS2 environment, from the result it is observed that the proposed protocol provides better performance compared to the conventional MAC protocols mentioned in the recent literature for the conceded problem.

A MAC protocol which provides both energy saving mechanism and that can handle emergency situation is the most desired for any industry. In any industry time and mission critical scenarios requires strict timeliness and reliability along with the energy efficiency. However there are dynamic and harsh environmental conditions for which the MAC protocol must survive and do transmission accordingly. The dynamic changes in topology must also be adapted so that the nodes are in constant link to the destination. Most of the existing MAC protocols have been identified as they face a number of limitations for industrial application domain.

KEYWORDS

MAC protocol, Industrial wireless sensor networks (IWSN), Time critical applications and Energy efficiency.

1. INTRODUCTION

Research of the past years has led to numerous novel development and approaches for wireless sensor networks. Energy efficiency is a critical issue for sensor networks, where nodes work with resource constraint battery power. Recent advancement in wireless communication and device technology has enabled the development of low cost sensor networks composed of tiny sensors.

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 125–135, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4514

The sensor nodes are typically capable of sensing, processing, and networking. Since a sensor node is a small, lightweight, un-tethered, battery-powered device, its energy is limited [1], [2], [3]. As a result, energy efficiency is a critical issue for sensor networks. Many researchers have focused on the development of power saving schemes for wireless sensor networks [10], [11], [12], [13]. These schemes include power saving hardware and topology design, power-efficient MAC (medium access control) layer protocol/ network layer routing protocol. Even though the research field of wireless sensor networks and in particular the MAC protocols is relatively new, there exist numerous MAC protocols proposed in the recent literature, designed specifically for wireless sensor networks.

Nowadays WSN have been extended to support many application domains such as military target tracking, industrial automation or patient monitoring. The aforementioned conventional MAC protocol is no longer adequate for these application domains. For example sensor observing pressure in pipes must deliver messages to an actuator connected to a valve in a timely and reliable fashion. Another example in power plant boiler process control, apart from other parameters, pressure and level must be controlled in timely and reliable fashion. Hence, to support time critical and mission critical applications a necessary first step is to find a MAC protocol that is capable of supporting performance bounds on data transport delay and reliability.

A large number of MAC protocols for wireless sensor networks have been proposed in literatures [10], [11], [12], [13]. It is believed that the data transport delay and reliability are two important objectives most relevant in the context of mission critical and time critical applications, while energy efficiency could be addressed additionally if required.

Communication in wireless sensor networks can be divided into several layers like other communication infrastructure. The MAC layer, which is primarily responsible for providing accessibility to the channel for communication. It is described by a MAC protocol, which tries to ensure that no two nodes interfere with each other during communication using a proper coordination mechanism. In general, the main design goal of typical MAC protocols is to provide high throughput, minimized latency, fairness, and quality of service. In addition, the MAC protocol for wireless sensor network needs to consider energy efficiency because of the limited energy of constituent sensor nodes. The primary design issue for the MAC protocol of wireless sensor network is thus, how to support the basic functions of MAC protocol while minimizing energy consumption of the sensor nodes to maximize the lifespan of the network.

In Bluetooth or 802.11, Energy conservation is not a primary objective, because mostly nodes are charged every day or mains powered. The commercial standards like IEEE 802.11 define a power management scheme for ad hoc networks, wherein the nodes remain in idle listening state to conserve the energy in low traffic condition. It was shown that a significant amount of energy can be wasted even in the idle listening mode.

Hence, IEEE 802.11 is not suitable for sensor networks. S-MAC is a MAC protocol designed specifically for wireless sensor networks. It forces the sensor nodes to operate with low duty cycle and take periodic sleep instead of idle listening. The sensor nodes also sleep during overhearing period to save the energy [2], [3]. Although S-MAC can save more energy than IEEE 802.11 protocol, it cannot efficiently adapt to the network traffic condition since it uses a fixed duty cycle for all the sensor nodes. A duty cycle tuned for high traffic loads results in a waste of energy when the traffic is low, while tuning for low traffic loads results in low throughput under high traffic loads. The Timeout-MAC protocol (T-MAC) improves the S-MAC protocol by employing the approach of adaptive duty cycle. If there is no activity in the vicinity of a node for a while, it sleeps. Such an adaptation frees the application from the burden of selecting an appropriate duty cycle. T-MAC displays the same performance as S-MAC under constant traffic loads, but saves more energy under variable traffic [4].

In this paper a H-MAC protocol proposed, which determines end to end delay and adaptively determines the transmission schedule according to the buffer condition and the context of the packets. The existing approaches designed for wireless sensor network improve energy efficiency by controlling the duty cycle. The proposed protocol reduces energy consumption by letting each node stay in the sleep mode if the number of packets in the buffer is smaller than the threshold, while the threshold value is decided according to the distance of the node to the sink node. The variable threshold for each switch node may cause increased latency, Since certain sensor loops in industries are found to be of critical loops data's have to be transferred without any further delay, thus the contention access period for these prioritized nodes can be decided if the sensed information is of time critical, and transferring the data immediately to the sink.

The rest of the paper is presented as follows. Section 2 reviews the related work. Section 3 presents the proposed approach that use H-MAC protocol. Section 4 reveals about results acquired and discussion. Finally section 5 concludes the paper and outlines the future research direction.

2. RELATED WORK

Due to the energy constrained environment, the MAC protocol for sensor networks has to take energy efficiency as one of its primary concerns. The existing wireless MAC protocols such as Bluetooth and 802.11 MAC protocols [6] cannot be directly applied to the sensor networks since none of them take energy conservation as the primary design objective.

There have been several MAC protocols specially designed for sensor networks. S-MAC is a MAC protocol with periodic listen/sleep scheduling based on local synchronization. In the S-MAC protocol, the listen and sleep period are set to be a fixed length. During the listen period, SYNC and RTS/CTS control packets are transmitted based on the CSMA/CA mechanism for the purpose of synchronization and announcement of the succeeding data packet transmission. Any two nodes exchanging the RTS (Request-to-Send)/CTS (Clear-to-Send) packet in the listen period stays in the wake state and start data transmission during the sleep period of other nodes. All other nodes can enter the sleep mode to conserve the energy. Generally, periodic listen/sleep has the trade-offs between energy saving and latency. To improve the performance, S-MAC uses an adaptive listening scheme in which the node receiving NAV information remains awake and tries to communicate in the sleep mode without waiting for the next listen/sleep cycle. In order to decrease the latency of S-MAC, DSMAC [5] supports multiple duty cycles automatically adjusted according to the energy consumption level and delay. T-MAC improves the energy efficiency of S-MAC by using a very short listening window at the beginning of each active period. The length of active time is adaptive, and the timer is defined by equation (1)[4].

$$T_{out} = C + R + T \qquad (1)$$

C is the contention interval, R is the length of RTS packet, and T is a very short time interval between RTS and CTS that is identical to SIFS (Short Inter Frame Space) in 802.11 MAC. If no data is transmitted during Timeout, the active nodes enter the sleep mode for saving the energy until the beginning of the next listening period. If no activity occurs in that period, the node returns to the sleep mode by adapting the duty cycle. T-MAC thus saves the energy at a cost of reduced throughput and increased latency. With the same workloads, T-MAC and S-MAC perform equally, while T-MAC suffers from the same complexity and scaling problem of S-MAC. Shortening the active window in T -MAC reduces the ability to snoop on the surrounding traffic and adapt to the changing network condition.

2.1. Unique Requirements of Industrial WSN

At the Media Access Control (MAC) layer Energy efficiency becomes primary concern in designing MAC protocol to maximize network lifetime. Deterministic MAC layer design required to achieve low latency and reliable delivery of messages to the destination.ISA SP 100 Working group classified industrial process control in to six different classes based on latency

Class 0: Emergency action (in terms of micro seconds)
Class 1: Closed-loop regulatory control (in terms of milliseconds)
Class 2: Closed-loop supervisory control (in terms of seconds)
Class 3: Open-loop control (in terms of minutes)
Class 4: Monitoring with short-term operational consequences (in terms of hours)
Class 5: Monitoring without immediate operational consequences. (in terms of days)

3. THE PRIORITIZING APPROACH

In order to satisfy the unique requirements of sensor and control devices a suitable MAC protocol must be devised. In the proposed prioritizing approach, it is planned to give priority only to sensor nodes which are in the critical loop during contention access period of super frame structure, more importantly collision avoided by giving channel access to nodes in critical loop. Reduced collisions and transmissions in turn will consequently reduce power consumption. Worst case delay for the urgent packets is the one cycle time. The contexts collected by the sensors are diverse because of the inherent characteristics of wireless sensor network. Also, the types and importance of the contexts are all different. There might be some context data requiring urgent transmission, while the urgency varies according to the location of the sensors. On the other hand, there are some sensors located in such places requiring little monitoring. We use the new approach deciding the operation of the nodes based on the contexts.

H-MAC protocol changes the state of a node according to the quantity of accumulated data in the buffer and the importance of the context before the contention period begins.

3.1. Proposed Approach

The basic mechanism involved in this Hybrid MAC PROTOCOL involves a series of steps after deploying the nodes and they are as follows,

Step 1- The protocol identifies the near one hop neighbors by broadcasting the ping message from its location. The ping message informs the network nodes begin by first identifying the one hop neighbors from its location. This is feasible by broadcasting ping message once in a while. The ping message exchanges neighbor's location to each other and also to inform the sleep and wake cycles. This is done every 30 seconds in this simulation [4]. And the ping memory is then exchanged between nodes, thus enabling the nodes to have two hop neighbors list.

Step 2- Initially CSMA mechanism is followed as the traffic will be low and thus data transfer occurs by sensing the carrier. However the traffic is never the same and as the traffic increases beyond the threshold the CSMA mechanism is dropped.

Step 3- The node that is present in a high priority loop must be given first priority, thus using the node ID the current transmission is dropped and changes over to TDMA thereby giving the first slot to the node that is in the high priority loop.

Step 4- If two nodes that are present in the same high priority region, then slots are assigned to the nodes one after the other. And after the transmission the network changes back to CSMA.

Step 5-The fast transmission using the node id is done only for critical loops, this node id information is passed along the header packet before the data packets being sent.

However, this approach does not considers the importance of important nodes getting access of channel in case of emergency or time critical situation, because in many industries most nodes are powered by wire and thus energy efficiency is not the only constraint. A time critical information have to be transferred quicker than any other data, thus the best approach is by getting a hold of the channel, that is the contention access period of the node should be flexible in case of time critical information. This feature is deployed to certain nodes only, i.e. nodes that are present in critical loops in and around the industry. Thus the data is transferred to the sink quicker as the channel is being taken over for quicker transfer.

3.2. A Decision of Threshold Value

In the H-MAC protocol the packets are transmitted only when the height of the buffer exceeds the specified threshold value. This can assure the energy efficiency. Each node switches to the sleep mode if the number of packets in the buffer is smaller than the specified threshold value or an RTS packet has not been received when the timer is on. To reduce unnecessary idle listening time after data transmission, the proposed H-MAC lets the node switch to the sleep mode and conserve the energy when there is no data to receive. This will substantially reduce the energy consumption since a node has more chances of sleep. However, if all nodes have the same threshold value, the effectiveness will be valid in only the single-hop network. For example, assume that a node is waiting for transmission until its buffer exceeds the specified threshold value. As soon as the buffer reaches the threshold value, the node transmits the packets to the next node. However, the buffer of the next node will be full as soon as it receives them since they have the same buffer threshold.

The figure shown in the figure 2 explains an industrial scenario that is filled with both high priority and low or normal priority loops. The network consists of several end devices and routers that transfer the data from one loop to the other towards the sink. There is a single central co coordinator node that governs the network protocols and proper time slotting for all the nodes. The nodes in high priority loops are the ones that are given the top priority in transferring the data; the node ID as mentioned in the algorithm enables this feature. The router holds the routing table in order to establish route on demand.

Therefore, the protocol decides the threshold value of a node according to the hop-count from the sink node to maximize the energy efficiency and also to apply to the multi-hop networks. Fig. 2 shows that each node has a different threshold value according to the hop-count from the sink in the proposed H-MAC protocol. If node-D transmits a packet to node-

C after its buffer exceeds the threshold value, node-C is able to get a chance of sleep since its threshold value is bigger than the preceding node, node-D. As a result, each node can improve the energy efficiency. We use two parameters, α and λ, to decide the threshold value, where α is the parameter reflecting the hop count while λ is the degree of the change of threshold value. When α is bigger than or equal to the difference between the hop-count of the source packet(Ntotal) and the hop count to the sink(Nown), the threshold value, Qthresh, is determined by Equation-(2) below the figure 2. Both these equations are used in the network based upon the nodes position.

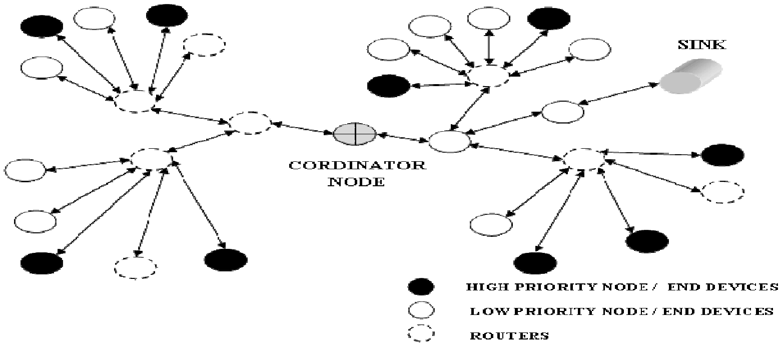

SINK
CORDINATOR NODE
HIGH PRIORITY NODE / END DEVICES
LOW PRIORITY NODE / END DEVICES
ROUTERS

Fig 1. Network Scenarios in Industries.

$$Q_{\text{threshold}} = [(\lambda\text{-}\alpha)10] \qquad \alpha < \lambda \tag{1}$$

$$Q_{\text{threshold}} = [(\alpha)10] \qquad \alpha = \lambda \tag{2}$$

The existing approach uses greedy based algorithm for routing, which has a major setback of not proceeding in the most optimal way, whereas here we use an optimal link state routing which improves the network performance than the existing approach.

The first equation is considered for nodes that are except for the last node in the network towards the sink. And the second equation is used for the node that was left in the first. Thus we save energy by the varying buffer level.

In the above two equations, α is the current node in the gird, while λ is the total number of nodes in the sensor network. If there are 10 nodes then λ is 10, then the 10th node or first node from sink then it will have 100 has its threshold level. If current node α is 8 or the third node from sink, then $Q_{threshold}$ for λ of 10 will be 80.

This feature can be easily visualized for grid topology, however in the case of random since any nodes can be in the path of network the data's are being transferred via the shortest possible route to the destination. In the case of an emergency data the priority is authorized to that particular nodes information. The following section explains the result of the proposed hybrid MAC protocol.

4. RESULTS AND DISCUSSIONS

The simulation is performed using network simulator 2 for two different scenarios, they are grid and random topology. The grid topology is that the nodes are connected together in a matrix form and thus named as grid topology. The random topology has no predefined order or manner for the nodes, and thus they are deployed in random locations. The proposed protocol is compared with the some of the existing protocols such as S MAC T MAC. The Figures explains energy consumed by nodes linearly increases for all the protocols with respect to number of nodes. The grid topology has a fixed distance between the nodes and the network size increases as the number of nodes increases. Gradually the number of nodes increased from 16 to 100. For any IWSN the two factors that play the key roles are reliability and timeliness, to satisfy this real time link state routing protocol is used in this approach which ensures reliability and timeliness.

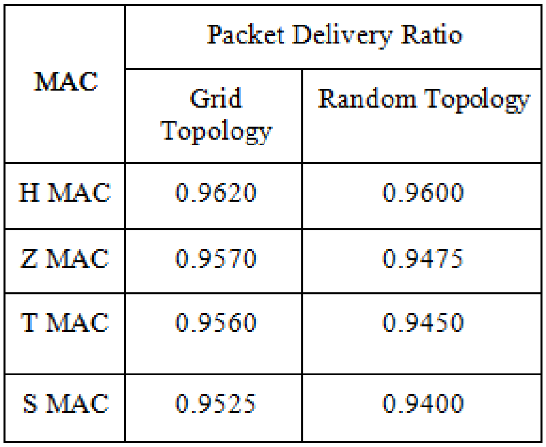
H-MAC consumes significantly lesser energy; this is done by reducing and setting threshold value for deciding the transmission. Lower power consumption is also achieved avoiding CAP collision by giving the bandwidth access to the nodes on high priority nodes. Fig 3(b)illustrate the control packet overhead, specifically the number of RTS, CTS and ACK packets used by H-MAC with that of SMAC and TMAC. It is very clear from the graphs that the total number of packet overhead is significantly reduced. It is also noted that transmission of packets are also reduced by transmitting only when the specified threshold value is exceeded, thus conserving energy in every possible way. This mechanism greatly reduces the control packet overhead.

The latency in the case of H MAC for normal transmission is found to be the most delayed transmission when compared to that if the existing, however this is done to conserve energy as much as possible. On the other hand there are time critical data's and for important packets that are marked as H-MAC, the latency is much lower than SMAC and TMAC protocols, which is another critical requirement of IWSN. The latency of H-MAC for important packets is close to that of MAC with no sleep, since the data packets of important context are allowed to be immediately transmitted to the next node.

4.1. PERFORMANCE ANALYSIS

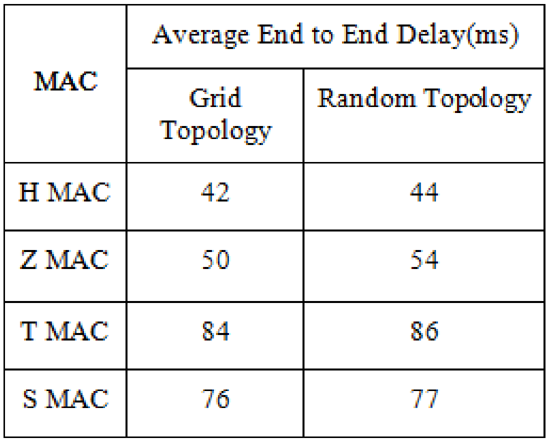
The performance analysis is evaluated using the graph obtained from the simulation. The values are the ones that are present at the end of simulation. The simulation lasts for 120 ms and the below values are adapted from the respective graphs.

TABLE 1: PACKET DELIVERY RATIO

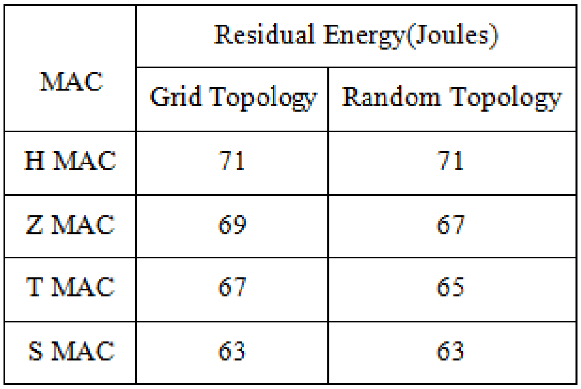


MAC	Packet Delivery Ratio	
	Grid Topology	Random Topology
H MAC	0.9620	0.9600
Z MAC	0.9570	0.9475
T MAC	0.9560	0.9450
S MAC	0.9525	0.9400

TABLE 2: AVERAGE END TO END DELAY

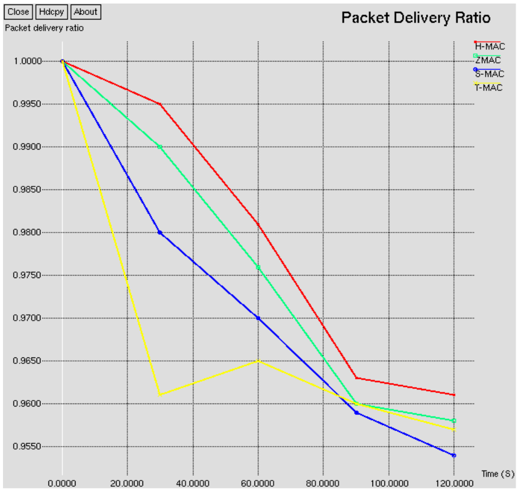


MAC	Average End to End Delay(ms)	
	Grid Topology	Random Topology
H MAC	42	44
Z MAC	50	54
T MAC	84	86
S MAC	76	77



MAC	Residual Energy(Joules)	
	Grid Topology	Random Topology
H MAC	71	71
Z MAC	69	67
T MAC	67	65
S MAC	63	63

TABLE 3: RESIDUAL ENERGY



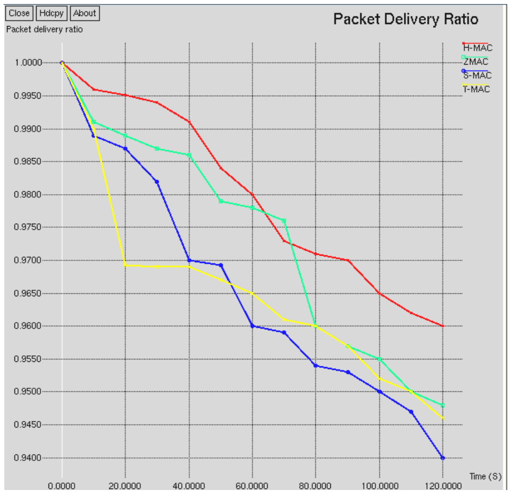
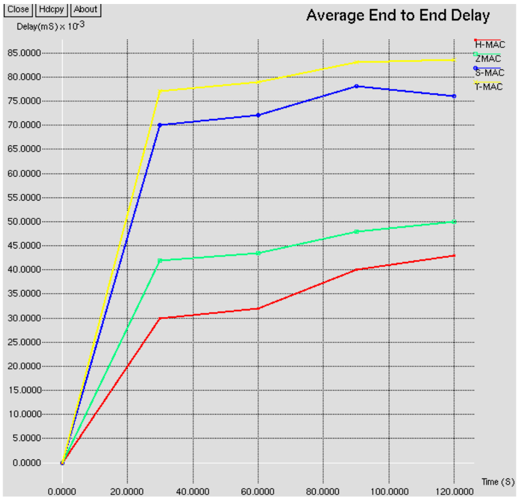


Fig 2. Packet Delivery Ratio (grid)

Fig 3. Packet Delivery Ratio (random)

Packet Delivery Ratio-The packet delivery ratio in general is the ratio between "the total numbers of packets sent from the source to that of the total number of packets received in the destination". This ratio indicates the successful ability of the protocol developed in any scenario as the ultimate aim is to transfer data.

From the graph analysis it can be clearly understood that the H MAC are being capable of separating the high priority information and transfer the data to the destination. The comparison clearly indicates that the proposed H MAC on comparison with the existing protocol is much better and has a marginal difference to that of the existing protocol as seen in the graph. The simulation shows different colour for different protocols, the simulation time is for 120 seconds with 100 nodes at any instant.



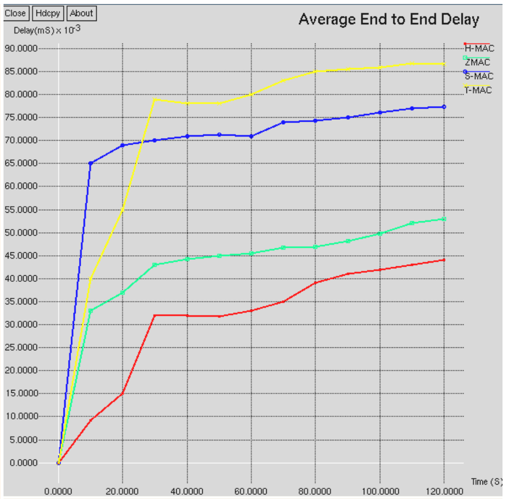
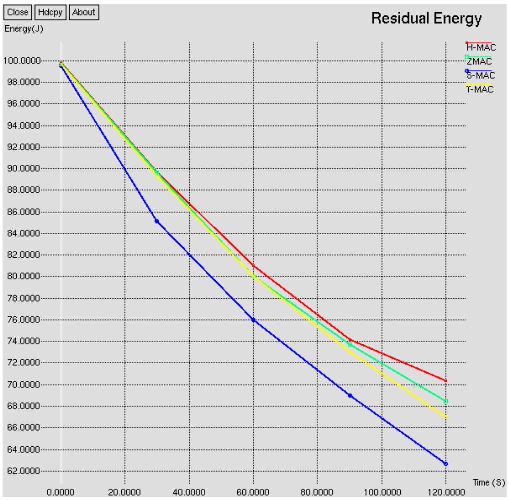


Fig 4. Average End to End delay (grid)

Fig 5. Average End to End delay (random)

Average End to End Delay- The delay that occurs as the data travels through the network from the source to the destination is the end to end delay. It is clearly observed that the H MAC has the least end to end delay of ~ 43(approx.). The major advantage in the hybrid MAC protocol is that information here is immediately transferred however in the existing protocols had to wait and transfer data only in the next wake cycle or only by informing the other nodes in prior. In the case of H MAC the delay is greatly reduced with the aid of the buffer memory and prioritizing nodes which enables quick and safe delivery of information.



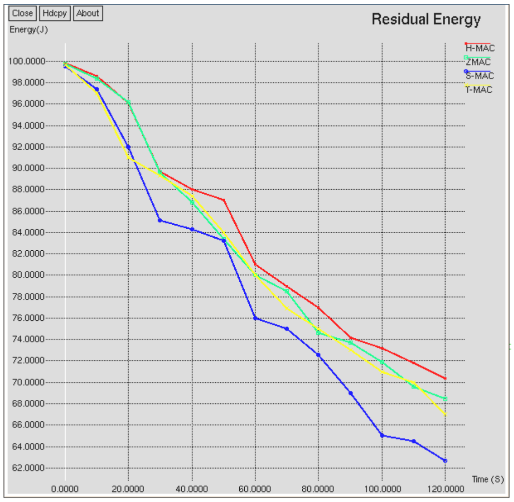


Fig 6. Residual Energy (grid) Fig 7. Residual Energy (random)

Residual energy- The energy left in any node at any instant is said to be the residual energy. The initial energy of each node is 100 joules. The transmission, route discovery, priority checking, and reception are some of the major reasons for energy loss in a network, it is thus found that this protocol is efficient as it consumes lesser energy.

From the graph it is clearly observed that the S MAC with a 50 % duty cycle consumes the most energy while H MAC protocol consumes the least amount with the aid of the changing buffer level in each node, which reduces the energy consumption.

5. Conclusion And Future Work

The Carrier sense medium access provides the network to transfer data in low traffic scenario to conserve energy; however the time division medium access have the ability to transfer in high traffic scenario there by transferring data's as efficient as possible.

The proposed protocol with the above features together enables it to outperform the existing protocol. The reason why H MAC outperforms when compared to that of the Z MAC is that its feature of changing buffer (memory) level. The reliability of the path may be improved by link state routing and end to end delay may be further reduced by using cross layer approach, thereby saving more energy.

However the protocol has to be advanced by reducing the time consumption in changing over from TDMA to CSMA. In order to further reduce the energy consumption variable memory level has to be set to change from that instant, based upon the distance from that node to the sink, which will enable the nodes anywhere in the network to save energy.

REFERENCES

[1] Akyildiz, W. Su, Y. Sankarasubramaniam and E. Cayirci, A survey on sensor networks. IEEE Communications Magazine, 40(8), pages 102–114, 2002.

[2] G W. Ye, J. Heidemann, and D. Estrin, An energy-efficient MAC protocol for wireless sensor networks. In IEEE INFOCOM, volume 3, pages 1567-1576, June 2002.

[3] W. Ye, J. Heidemann, and D. Estrin, "Medium Access Control With Coordinated Adaptive Sleeping for Wireless Sensor Networks". IEEE/ACM Transactions on Networking, Volume: 12, Issue: 3, Pages: 493 - 506, June 2004.

[4] T. V. Dam and K. Langendoen, "An Adaptive Energy-Efficient MAC Protocol for Wireless Sensor Networks". SenSys'03, Los Angeles, Pages 171 - 180. Nov 2003.

[5] P. Lin, C. Qiao, and X. Wang, "Medium Access Control With A Dynamic Duty Cycle For Sensor Network". IEEE WCNC'04, March 2004.

[6] LAN MAN Standards Committee of the IEEE Computer Society, "Wireless LAN medium access control (MAC) and physical layer (PHY) specification". IEEE, New York, NY, USA, IEEE Std 802.11-1997 edition, 1997.

[7] T. Zheng, S. Radhakrishnan, and V. Sarangan, "PMAC: An adaptive energy-efficient MAC protocol for wireless sensor networks," in Proc.IPDPS, 2005.

[8] V. Shnayder, M. Hempstead, B. Chen, G. W. Allen, and M.Welsh, "Simulating the power consumption of large-scale sensor network applications". In SenSys '04 pages 188–200, New York, NY, USA, 2004. ACM Press.

[9] S. Narayanaswamy, V. Kawadia, R. S. Sreenivas, and P. R. Kumar, Power Control in Ad-Hoc Networks: "Theory, Architecture, Algorithm and Implementation of the COMPOW protocol". In European Wireless 2002, February 2002.

[10] G. Simon, P. Volgyesi and A. Ledeczi, "Simulation-based optimization of communication protocols for large-scale wireless sensor networks". IEEE Aerospace, March 2003.

[11] K.T. Kim, H.S. Kim, and H.Y. Youn, "Optimized Clustering for Maximal Lifetime of Wireless Sensor Networks", EUC 2006, LNCS 4097, pp. 465 – 474.

[12] K.T. Kim and H.Y. Youn, "PEACH: Proxy-Enable Adaptive Clustering Hierarchy for Wireless Sensor network", Proceeding of the 2005.

[13] Pei Huang, Li Xiao, Soroor Soltani,Matt W. Mutka, and Ning Xi."The Evolution of MAC Protocols in Wireless Sensor Networks: A Survey" IEEE Communications Surveys & Tutorials, 2012.

[14] Abdelmalik Bachir, Mischa Dohler, Thomas Watteyne, K. Leung "MAC Essentials for Wireless Sensor Networks", IEEE Communications Surveys & Tutorials, Vol. 12, No. 2, Second Quarter 2010

[15] Vehbi C. Gungor, P. Hancke,"Industrial Wireless Sensor Networks: Challenges,Design Principles, and Technical Approaches" , IEEE Transactions on Industrial Electronics, Vol. 56, No. 10, October 2009.

[16] Demirkol and C. Ersoy, "Energy and delay optimized contention for wireless sensor networks",Computer Networks: The International Journal of Computer and Telecommunications Networking,, Vol. 53,Issue 12, pp. 2106-2119, August 2009.

[17] Khaldoun Al Agha, Marc-Henry Bertin, Tuan Dang,"Which Wireless Technology for Industrial Wireless Sensor networks? The Development of OCARI Technology" IEEE Transactions on Industrial Electronics, Vol. 56, No. 10, October 2009.

[18] Rhee, I., Warrier, A., Aia, M., and Min, J "ZMAC: A hybrid MAC for wireless sensor networks". Proc. Of the 3rd ACM Conference on Embedded Networked Sensor Systems (2005)

AUTHORS

PANDEESWARAN CHELLIAH

Graduated from Madurai Kamaraj University with B.E (Instrumentation and Control). M.Tech (Applied Electronics) from Dr.M.G.R University, Chennai. He has worked as faculty in Electronics and Instrumentation Engg in Jaya Engg College, Chennai for more than 8 years. He is with St.Joseph's College of Engineering, Chennai, Tamil Nadu, India for more than 7 years. He is a part time research scholar in MIT Anna University Chennai. His field of Interests are Microprocessor and Microcontrollers, Embedded systems, Wireless Sensor Networks and Industrial Automation.

PAPA NATARAJAN

Graduated from Annamalai University with B.E (Electronics and Instrumentation). M.Tech (Digital Electronics) from Cochin University. She has worked as faculty in Instrumentation Engg in Annamalai University for 5 years. She is with Anna University for more than 10 years. She has carried out Ph.D in "Nonlinear control of Heat Exchanger". Her field of interests are process control, Industrial Automation and VLSI Design.

JAYESH SUNDAR GOPINATH

Graduated from Thangavelu Engineering College(affiliated to Anna University), Chennai, with B.E in Electronics and communication and currently pursuing final year in M.E Control and Instrumentation at St.Joseph's College Of Engineering(affiliated to Anna University), Chennai, Tamil Nadu, India. He has been a part of the research team and presented papers in the field of Industrial wireless sensor network. His field of interest are industrial wireless sensor networks and industrial instrumentation.

INTENTIONAL BLANK

COMPARATIVE ANALYSIS OF FILTERS AND WAVELET BASED THRESHOLDING METHODS FOR IMAGE DENOISING

Anutam[1] and Rajni[2]

[1]Research Scholar SBSSTC, Ferozepur, Punjab
anutam.bansal@gmail.com
[2]Associate Professor SBSSTC, Ferozepur, Punjab
rajni_c123@yahoo.co.in

ABSTRACT

Image Denoising is an important part of diverse image processing and computer vision problems. The important property of a good image denoising model is that it should completely remove noise as far as possible as well as preserve edges. One of the most powerful and perspective approaches in this area is image denoising using discrete wavelet transform (DWT). In this paper comparative analysis of filters and various wavelet based methods has been carried out. The simulation results show that wavelet based Bayes shrinkage method outperforms other methods in terms of peak signal to noise ratio (PSNR) and mean square error(MSE) and also the comparison of various wavelet families have been discussed in this paper.

KEYWORDS

Denoising, Filters, Wavelet Transform, Wavelet Thresholding

1. INTRODUCTION

Applications of digital world such as Digital cameras, Magnetic Resonance Imaging (MRI), Satellite Television and Geographical Information System (GIS) have increased the use of digital images. Generally, data sets collected by image sensors are contaminated by noise. Imperfect instruments, problems with data acquisition process, and interfering natural phenomena can all corrupt the data of interest. Transmission errors and compression can also introduce noise [1]. Various types of noise present in image are Gaussian noise, Salt & Pepper noise and Speckle noise. Image denoising techniques are used to prevent these types of noises while retaining as much as possible the important signal features [2]. Spatial filters like mean and median filter are used to remove the noise from image. But the disadvantage of spatial filters is that these filters not only smooth the data to reduce noise but also blur edges in image. Therefore, Wavelet Transform is used to preserve the edges of image [3]. It is a powerful tool of signal or image processing for its multiresolution possibilities. Wavelet Transform is good at energy compaction in which small coefficients are more likely due to noise and large coefficients are due to important signal feature. These small coefficients can be thresholded without affecting the significant features of the image.

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 137–148, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4515

This paper is organized as follows: Section 2 presents Filtering techniques. Section 3 discusses about Wavelet based denoising techniques and various thresholding methods. Finally, simulated results and conclusion are presented in Section 4 and 5 respectively.

2. Filtering Techniques

The filters that are used for removing noise are Mean filter and Median filter.

2.1. Mean Filter

This filter gives smoothness to an image by reducing the intensity variations between the adjacent pixels [4]. Mean filter is also known as averaging filter. This filter works by applying mask over each pixel in the signal and a single pixel is formed by component of each pixel which comes under the mask. Therefore, this filter is known as average filter. The main disadvantage of Mean filter is that it cannot preserve edges [5].

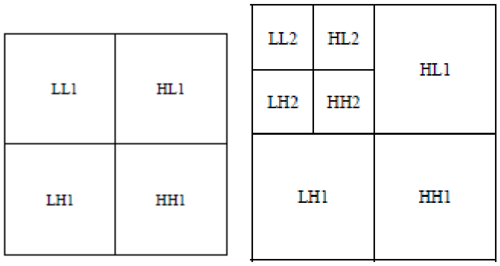
2.2. Median Filter

Median filter is a type of non linear filter. Median filtering is done by, firstly finding the median value across the window, and then replacing that entry in the window with the pixel's median value [6]. For an odd number of entries, the median is simple to define as it is just the middle value after all the entries are made in window. But, there is more than one possible median for an even number of entries. It is a robust filter. Median filters are normally used as smoothers for image processing as well as in signal processing and time series processing [5].

3. Wavelet Transform

In Discrete Wavelet Transform (DWT) , signal energy is concentrated in a small number of coefficients .Hence, wavelet domain is preferred. DWT of noisy image consist of small number of coefficients having high SNR and large number of coefficients having low SNR. Using inverse DWT, image is reconstructed after removing the coefficients with low SNR [3]. Time and frequency localization is simultaneously provided by Wavelet transform. In addition, Wavelet methods are capable to characterize such signals more efficiently than either the original domain or transforms such as the Fourier transform [7].

The DWT is identical to a hierarchical sub band system where the sub bands are logarithmically spaced in frequency and represent octave-band decomposition. When DWT is applied to noisy image, it is divided into four sub bands as shown in Figure 1(a).These sub bands are formed by separable applications of horizontal and vertical filters. Finest scale coefficients are represented as sub bands LH1, HL1 and HH1 i.e. detail images while coarse level coefficients are represented as LL1 i.e. approximation image [8] [3]. The LL1 sub band is further decomposed and critically sampled to obtain the next coarse level of wavelet coefficients as shown in Fig. 1(b).



(a) One- Level (b) Two- Level

Figure1. Image Decomposition by using DWT

LL1 is called the approximation sub band as it provides the most like original picture. It comes from low pass filtering in both directions. The other bands are called detail sub bands. The filters L and H as shown in Fig.2 are one dimensional low pass filter (LPF) and high pass filter (HPF) for image decomposition. HL1 is called the horizontal fluctuation as it comes from low pass filtering in vertical direction and high pass filtering in horizontal direction. LH1 is called vertical fluctuation as it comes from high pass filtering in vertical direction and low pass filtering in horizontal direction. HH1 is called diagonal fluctuation as it comes from high pass filtering in both the directions. LL1 is decomposed into 4 sub bands LL2, LH2, HL2 and HH2. The process is carried until some final scale is reached. After L decompositions a total of D (L) = 3 *L +1 sub bands are obtained .The decomposed image can be reconstructed using are construction filter as shown in Figure 3. Here, the filters L and H represent low pass and high pass reconstruction filters respectively.

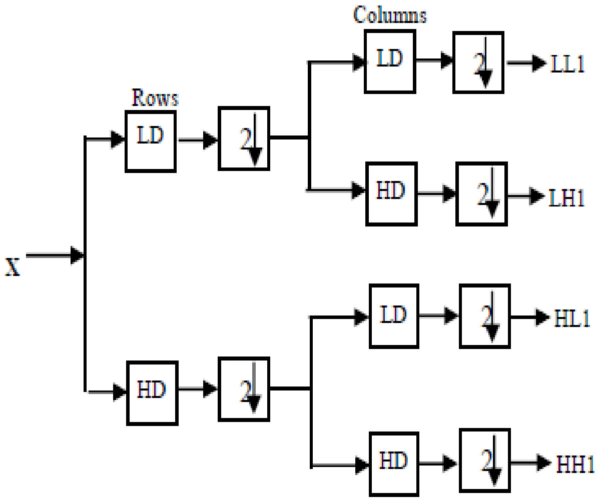


Figure2. Wavelet Filter bank for one-level Image Decomposition

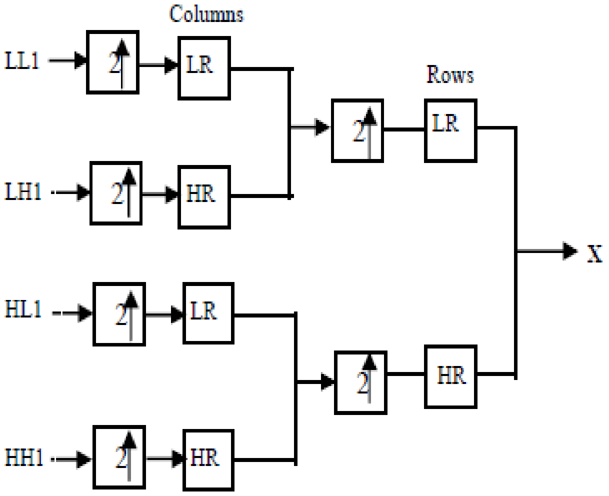


Figure3. Wavelet Filter bank for one-level Image Reconstruction

3.1 Wavelet Based Thresholding

Wavelet thresholding is a signal estimation technique that exploits the capabilities of Wavelet transform for signal denoising. It removes noise by killing coefficients that are irrelevant relative to some threshold [8] .Several studies are there on thresholding the Wavelet coefficients. The process, commonly called Wavelet Shrinkage, consists of following main stages:

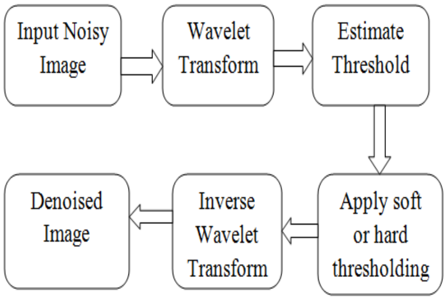


Figure 4. Block diagram of Image denoising using Wavelet Transform

- Read the noisy image as input
- Perform DWT of noisy image and obtain Wavelet coefficients
- Estimate noise variance from noisy image
- Calculate threshold value using various threshold selection rules or shrinkage rules
- Apply soft or hard thresholding function to noisy coefficients
- Perform the inverse DWT to reconstruct the denoised image.

3.1.1 Thresholding Method

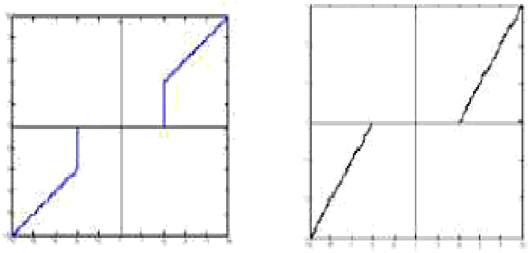
Hard and soft thresholding is one of the thresholding techniques which are used for purpose of image denoising. Keep and kill rule which is not only instinctively appealing but also introduces artifacts in the recovered images is the basis of hard thresholding [9] whereas shrink and kill rule which shrinks the coefficients above the threshold in absolute value is the basis of soft thresholding [10]. As soft thresholding gives more visually pleasant image and reduces the abrupt sharp changes that occurs in hard thresholding, therefore soft thresholding is preferred over hard thresholding [11] [12].

The **Hard Thresholding** operator [13] is defined as,

$$D\ (U, \lambda) = U \text{ for all } |U| > \lambda \qquad = 0 \text{ otherwise} \tag{1}$$

The **Soft Thresholding** operation the other hand is defined as ,

$$D\ (U, \lambda) = \operatorname{sgn}(U) * \max(0, |U| - \lambda) \tag{2}$$



(a) Hard Thresholding (b) Soft Thresholding [14]

Figure 5. Thresholding Methods

3.1.2 Threshold Selection Rules

In image denoising applications, PSNR needs to be maximized , hence optimal value should be selected [8]. Finding an optimal value for thresholding is not an easy task. If we select a smaller threshold then it will pass all the noisy coefficients and hence resultant images may still be noisy but larger threshold makes more number of coefficients to zero, which provides smoothness in image and image processing may cause blur and artifacts, and hence the resultant images may lose some signal values [15].

3.1.2.1 Universal Threshold

$$T = \sigma\sqrt{2logM} \tag{3}$$

where σ^2 being the noise variance and M is the number of pixels [16] .It is optimal threshold in asymptotic sense and minimizes the cost function of difference between the function. It is assumed that if number of samples is large, then the universal threshold may give better estimate for soft threshold [17].

3.1.2.2 Visu Shrink

Visu Shrink was introduced by Donoho [18]. It follows hard threshold rule. The drawback of this shrinkage is that neither speckle noise can be removed nor MSE can be minimized .It can only deal with additive noise [19]. Threshold T can be calculated using the formulae [20],

$$T_v = \hat{\sigma}\sqrt{2\log N} \tag{4}$$

$$\hat{\sigma}^2 = \left[\frac{median(|X_{ij}|)}{0.675}\right]^2, X_{ij} \in HH1 \tag{5}$$

Where σ is calculated as mean of absolute difference (MAD) which is a robust estimator and N represents the size of original image.

3.1.2.3 Bayes Shrink

The Bayes Shrink method has been attracting attention recently as an algorithm for setting different thresholds for every sub band. Here subbands refer to frequency bands that are different from each other in level and direction [21]. Bayes Shrink uses soft thresholding. The purpose of this method is to estimate a threshold value that minimizes the Bayesian risk assuming Generalized Gaussian Distribution (GGD) prior [12]. Bayes threshold is defined as [22],

$$t_B = \sigma^2 / \sigma_s \tag{6}$$

Where σ^2 is the noise variance and σ_s is signal variance without noise.

From the definition of additive noise we have,

$$\mathrm{w}(\mathrm{x}, \mathrm{y}) = \mathrm{s}(\mathrm{x}, \mathrm{y}) + \mathrm{n}(\mathrm{x}, \mathrm{y}) \tag{7}$$

Since the noise and the signal are independent of each other, it can be stated that ,

$$\sigma_w{}^2 = \sigma_s{}^2 + \sigma^2 \tag{8}$$

$\sigma_w{}^2$ can be computed as shown below:

$$\sigma_w{}^2 = \frac{1}{n^2} \sum_{x,y=1}^{n} w^2(\mathrm{x}, \mathrm{y}) \tag{9}$$

The variance of the signal, $\sigma_s{}^2$ is computed as

$$\sigma_s = \sqrt{\max(\sigma_w{}^2 - \sigma^2, 0)} \tag{10}$$

4. Simulation Results

Simulated results have been carried on Cameraman image by adding two types of noise such as Gaussian noise and Speckle noise. The level of noise variance has also been varied after selecting the type of noise. Denoising is done using two filters Mean filter and Median filter and three Wavelet based methods i.e. Universal threshold, Visu shrink and Bayes shrink. Results are shown through comparison among them. Comparison is being made on basis of some evaluated parameters. The parameters are Peak Signal to noise Ratio (PSNR) and Mean Square Error (MSE).

$$\mathrm{PSNR} = 10 \log_{10}\left(\frac{255^2}{MSE}\right) \mathrm{db} \tag{11}$$

$$\mathrm{MSE} = \frac{1}{MN} \sum_{i=1}^{M} (\mathrm{x}, \mathrm{y}) \sum_{j=1}^{N} (\mathrm{X}(\mathrm{i}, \mathrm{j}) - P(i, j))^2 \tag{12}$$

Where, M-Width of Image, N-Height of Image
P- Noisy Image , X-Original Image

Table 1 and Table 2 show the comparison of PSNR and MSE for cameraman image at various noisevariancies. Figure6 and Figure 7 shows that bayes shrinkage has better PSNR and low MSE than filtering methods and other wavelet based thresholding techniques.

Table1. Comparison of PSNR for Cameraman image corrupted with Gaussian and Speckle noise at different Noise variances using db1 (Daubechies Wavelet)

PSNR (Peak Signal To Noise Ratio)						
Noise	Noise Variance	Mean Filter	Median Filter	Universal Threshold	Visu Shrink	Bayes Shrink
Gaussian Noise	0.001	24.0598	25.4934	27.2016	28.2978	33.7031
	0.002	23.2251	24.3480	25.1748	26.1439	29.9001
	0.003	22.5261	23.4147	24.0062	24.8430	27.7650
	0.004	21.9796	22.6049	23.1590	23.8149	26.0865
	0.005	21.4536	22.0205	22.5099	23.0527	25.1235
	0.01	19.5569	19.7703	20.3580	20.5660	22.0446
Speckle Noise	0.001	24.8274	26.6157	28.4073	32.6526	44.0220
	0.002	24.5114	26.1260	26.8834	30.4768	40.0535
	0.003	24.2207	25.6708	25.9557	29.3585	38.3935
	0.004	23.9316	25.2771	25.3274	28.1881	35.6827
	0.005	23.7015	24.8599	24.8691	27.5283	34.3460
	0.01	22.6357	23.4053	23.3231	25.1853	30.9207

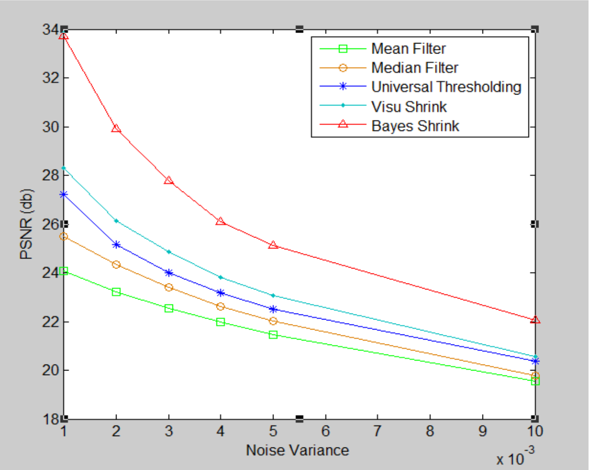


Figure6. Comparison of PSNR for cameraman image (corrupted with Gaussian noise) at different noise variance

Table2. Comparison of MSE for Cameraman image corrupted with Gaussian and Speckle noise at different Noise variances using db1

MSE (Mean Square Error)						
Noise	**Noise Variance**	**Mean Filter**	**Median Filter**	**Universal Threshold**	**Visu Shrink**	**Bayes Shrink**
Gaussian Noise	0.001	255.3265	183.5446	123.8560	96.2288	27.7188
	0.002	309.4321	238.9368	197.5136	158.0136	66.5377
	0.003	363.4693	296.2178	258.5006	213.1975	108.7875
	0.004	412.2133	356.9362	314.1828	270.1428	160.1160
	0.005	465.2894	408.3482	364.8271	321.9641	199.8629
	0.01	720.1005	685.5656	598.8007	570.7912	406.0842
Speckle Noise	0.001	213.9645	141.7451	93.8319	35.3036	2.5756
	0.002	230.1138	158.6638	133.2721	58.2642	6.4229
	0.003	246.0413	176.1971	165.0083	75.3748	9.4130
	0.004	262.9796	192.9158	190.6971	98.6903	17.5716
	0.005	277.2851	212.3693	211.9193	114.8823	23.9047
	0.01	354.4109	296.8613	302.5347	197.0393	52.6035

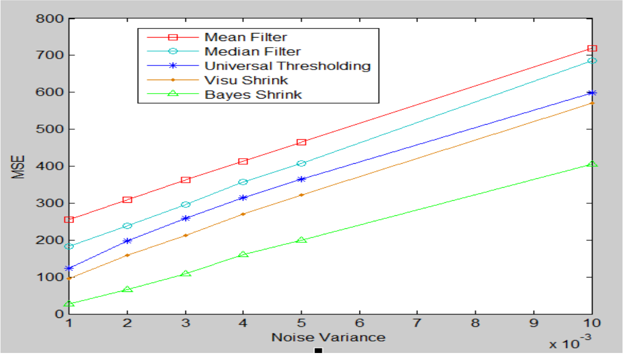


Figure 7. Comparison of MSE for cameraman image (corrupted with Gaussian noise) at different noise variances

The cameraman image is corrupted by gaussian noise of variance 0.01 and results obtained using filters and wavelets have been shown in Figure 8.

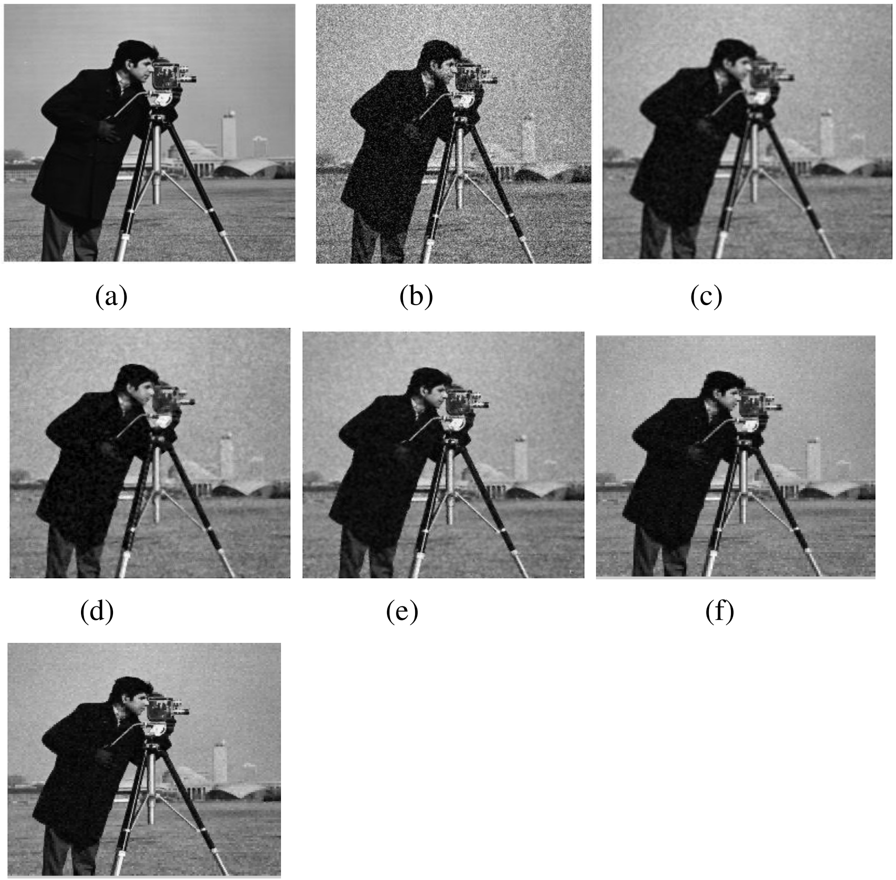


Figure 8. Denoising of cameraman image corrupted by Gaussian noise of variance 0.01 (a) Original image (b) Noisy image (c) Mean Filter (d) Median Filter (e) Universal Thresholding (f) Visu Shrink (g) Bayes shrink

A Comparative study of various wavelet families viz. Daubechies, Symlet, Coiflet, Biorthogonal and Reverse Biorthogonal using the Matlab Wavelet Tool box function *wfilters* is done and results have been tabulated in Table 3. Almost all the wavelet families perform in a much similar fashion.

Table3. Comparison of MSE and PSNR for Cameraman image (with Gaussian noise of variance 0.001) using various Wavelet families namely Daubechies, Symlet, Coiflet, Biorthogonal and Reverse Biorthogonal.

Wavelet Families		MSE			PSNR		
		Universal Threshold	Visu Shrink	Bayes Shrink	Universal Threshold	Visu Shrink	Bayes Shrink
DAUBECHIES	db2	118.9888	92.7006	27.8870	27.3757	28.4600	33.6768
	db5	116.0008	91.0493	29.1175	27.4862	28.5380	33.4893
	db7	114.5742	93.8306	32.3802	27.5399	28.4074	33.0280
	db9	117.1231	96.3611	33.6797	27.4444	28.2918	32.8571
	db10	117.7054	97.1057	33.8515	27.4228	28.2584	32.8350
Symlets	sym2	118.9952	93.2712	30.7511	27.3755	28.4333	33.2522
	sym4	114.9689	91.2290	29.3524	27.5250	28.5295	33.4544
	sym6	113.4957	92.9196	30.9472	27.5810	28.4497	33.2246
	sym7	112.3352	89.5128	29.1537	27.6256	28.6120	33.4839
	sym8	111.7177	90.6427	30.6893	27.6496	28.5575	33.2609
Coiflet	coif1	119.0472	93.1594	27.9323	27.3736	28.4385	33.6697
	coif2	113.9656	89.6841	29.1131	27.5631	28.6036	33.4899
	coif3	112.4675	92.3045	29.8983	27.6205	28.4786	33.3743
	coif4	112.3909	91.2025	31.0492	27.6235	28.5307	33.2103
	coif5	112.2086	90.1873	30.9109	27.6305	28.5794	33.2297
Biorthogonal	bior1.3	124.8644	99.1098	28.1472	27.1664	28.1696	33.6365
	bior2.2	125.0148	79.3262	22.2066	27.1612	29.1366	34.6660
	bior3.1	145.9058	85.0012	28.1984	26.4901	28.8366	33.6286
	bior4.4	114.5491	88.4300	29.0607	27.5409	28.6648	33.4977
	bior6.8	114.2567	88.5645	29.8665	27.5520	28.6582	33.3790
REV ERS	rbio1.5	117.1884	98.8170	35.9098	27.4420	28.1825	32.5787

	rbio2.4	106.7042	109.6627	47.6843	27.8490	27.7302	31.3470
	rbio3.3	104.6786	155.5353	75.9330	27.9322	26.2125	29.3265
	rbio5.5	119.0634	82.0170	22.4013	27.3730	28.9918	34.6281
	rbio6.8	111.1183	94.8413	31.7120	27.6729	28.3608	33.1186

5. CONCLUSION

In this paper, an analysis of denoising techniques like filters and wavelet methods has been carried out. Filtering is done by Mean and Median Filter. And three different wavelet thresholding techniques have been discussed i.e. Universal Thresholding, Bayes Shrink and Visu Shrink. From the simulation results, it is evident that Bayes shrinkage method has high PSNR at different noise variance and low MSE. This concludes that this method performs better in removing Gaussian noise and Speckle noise than filters and other wavelet methods.

REFERENCES

[1] Rajni, Anutam, "Image Denoising Techniques –An Overview," International Journal of Computer Applications (0975-8887), Vol. 86, No.16, January 2014.

[2] Akhilesh Bijalwan, Aditya Goyal and Nidhi Sethi, "Wavelet Transform Based Image Denoise Using Threshold Approaches," International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249-8958, Vol.1, Issue 5, June 2012.

[3] S.Arivazhagan, S.Deivalakshmi, K.Kannan, "Performance Analysis of Image Denoising System for different levels of Wavelet decomposition," International Journal of Imaging Science and Engineering (IJISE), Vol.1, No.3, July 2007.

[4] Jappreet Kaur, Manpreet Kaur, Poonamdeep Kaur, Manpreet Kaur, "Comparative Analysis of Image Denoising Techniques," International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Vol. 2, Issue 6, June 2012.

[5] Pawan Patidar, Manoj Gupta,Sumit Srivastava, Ashok Kumar Nagawat, "Image De-noising by Various Filters for Different Noise," International Journal of Computer Applications, Vol.9, No.4, November 2010.

[6] Govindaraj.V, Sengottaiyan.G , "Survey of Image Denoising using Different Filters," International Journal of Science, Engineering and Technology Research (IJSETR) ,Vol.2, Issue 2, February 2013.

[7] Idan Ram, Michael Elad, "Generalized Tree-Based Wavelet Transform," IEEE Transactions On Signal Processing, Vol. 59, No. 9, September 2011.

[8] Rakesh Kumar and B.S.Saini,"Improved Image Denoising Techniques Using Neighbouring Wavelet Coefficients of Optimal Wavelet with Adaptive Thresholding," International Journal of Computer Theory and Engineering, Vol.4, No.3, June 2012.

[9] Sethunadh R and Tessamma Thomas, "Spatially Adaptive image denoising using Undecimated Directionlet Transform," International Journal of Computer Applications, Vol.84, No. 11,December 2013

[10] S.Kother Mohideen Dr. S. Arumuga Perumal, Dr. M.Mohamed Sathik , " Image De-noising using Discrete Wavelet transform," IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.1, January 2008.

[11] Savita Gupta, R.C. Chauhan and Lakhwinder Kaur, "Image denoising using Wavelet Thresholding," ICVGIP 2002, Proceedings of the Third Indian Conference on Computer Vision, Graphics Image Processing, Ahmedabad, India, 2002

[12] S.Grace Chang, Bin Yu, Martin Vetterli , "Adaptive Wavelet Thresholding for image denoising and compression," IEEE Transaction On Image Processing, Vol.9, No.9, September 2000

[13] Nilanjan Dey, Pradipti Nandi, Nilanjana Barman, Debolina Das, Subhabrata Chakraborty ,“ A Comparative Study between Moravec and Harris Corner Detection of Noisy Images Using Adaptive Wavelet Thresholding Technique,” International Journal of Engineering Research and Applications (IJERA), ISSN: 2248-9622 , Vol. 2, Issue 1, Jan-Feb 2012.

[14] Tajinder Singh, Rajeev Bedi, “A Non - Linear Approach For Image De-Noising Using Different Wavelet Thresholding,” International Journal of Advanced Engineering Research and Studies, ISSN-2249-8974,Vol.1,Issue3,April-June,2012

[15] Abdolhossein Fathi and Ahmad Reza Naghsh-Nilchi, “Efficient Image Denoising Method Based on a New Adaptive Wavelet Packet Thresholding Function,” IEEE Transaction On Image Processing, Vol. 21, No. 9, September 2012

[16] Virendra Kumar, Dr. Ajay Kumar, “Simulative Analysis of Image denoising using Wavelet ThresholdingTechnique,” International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), Vol.2 , No.5, May 2013

[17] Mark J.T. Smith and Steven L. Eddins, “Analysis/SynthesisTechniques for subband image coding," IEEE Trans. Acoustic Speech and Signal Processing, Vol.38, No.8, Aug 1990

[18] D.L. Donoho and I.M. Johnstone, “Denoising by soft thresholding,” IEEE Trans. on Information Theory, Vo.41, 1995

[19] Raghuveer M. Rao, A.S. Bopardikar Wavelet Transforms: Introduction to Theory and Application published by Addison-Wesley, 2001

[20] S.Sutha, E. Jebamalar Leavline, D. ASR Antony Gnana Singh, “ A Comprehensive Study on Wavelet based Shrinkage Methods for Denoising Natural Images,” WSEAS Transactions on Signal Processing, Vol. 9, Issue 4, October 2013

[21] E.Jebamalar Leavline, S.Sutha, D.Asir Antony Gnana Singh, “Wavelet Domain Shrinkage Methods for Noise Removal in Images: A Compendium,” International Journal of Computer Applications,Vol.33, No.10, November 2011

[22] G.Y. Chen, T.D. Bui, A. Krzyak, “Image denoising using neighbouring Wavelet coefficients,” Acoustics Speech and Signal processing, IEEE International Conference, Vol.2, May 2004

AUTHORS

Anutam

She is currently pursuing M.Tech from SBS State TechnicalCampus, Ferozepur, India. She has completed B.Tech from PTUJalandhar in 2012. Her areas of interest includes Wireless Communication and Image Processing.

Mrs. Rajni

She is currently Associate Professor at SBS StateTechnical Campus Ferozepur, India. She has completed her M.E. from NITTTR, Chandigarh, India and B.Tech from NIT,KurukshetraIndia. She has fourteen years of academic experience.She has authored a number of research papers in International journals,National and International conferences. Her areas of interes include Wireless communication and Antenna design.

EFFECT OF MOBILITY MODELS ON THE PERFORMANCE OF MULTIPATH ROUTING PROTOCOL IN MANET

Indrani Das[1] , D.K Lobiyal[2] and C.P Katti[3]

[1,2,3]School of Computer and Systems Sciences
Jawaharlal Nehru University, New Delhi, India.
[1]indranidas2000@gmail.com, [2]lobiyal@gmail.com,[3]cpkatti@yahoo.com

ABSRACT

In this paper, we have analyzed the performance of multipath routing protocol with various mobility models for Mobile Ad Hoc Networks. The basic purpose of any multipath routing protocol is to overcome various problems occurs while data delivery through a single path routing protocol. For high acceptability of routing protocol, analysis of routing protocol in ad hoc network only with random way point mobility model is not sufficient. Here, we have considered Random waypoint, Random Direction and Probabilistic Random Walk mobility Model for proper analysis of AOMDV routing protocol. Results obtained show that with increasing node density, packet delivery ratio increases but with increasing node mobility Packet delivery ratio decreases.

KEYWORDS

AOMDV, Multipath Routing, Ad hoc Network, Packet delivery ratio, Mobility models.

1. INTRODUCTION

A Mobile Ad-Hoc Network (MANET) is a network where more than two autonomous mobile hosts (mobile devices i.e. mobile phone, laptop, iPod, PDAs etc) can communicate without any mean of infrastructure i.e. on the fly. When source (*S*) node want to send some data toward the destination (*D*), if they are in the same transmission range can directly communicate with each other otherwise intermediate nodes help to relay data from source to destination. In MANETs individual node can leave and join the network on its own, therefore the physical structure of the network frequently changes dynamically. Battery power of mobile device is also important aspect, because depletion of battery power may affect the lifetime of a node . Node movements differ for mobile nodes are different, the topology also depend on the speed and direction of nodes. Due to dynamic topology of the network routing in MANET is a challenging issue. Single path routing is not always sufficient to disseminate data to the destination. Therefore; multipath routing comes into existence to overcome the problem of single path routing.

In this paper we have considered various mobility models for proper and in depth analysis of AOMDV protocol. In literature we have discussed various works related to AOMDV protocol and brief about various multipath routing protocols. Most of the work carried out based on random waypoint mobility model. So we tried to analyze AOMDV protocol with various network parameters and mobility models.

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 149–155, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4516

The rest of the paper is organized as follows. In section II we have discussed various works related to multipath routing. In section III, various mobility models and AOMDV routing protocol briefly discussed. Results analysis and simulation work is presented in Section IV and finally, we have concluded the paper in Section V.

2. RELATED WORKS

Multipath routing overcomes various problems occurs while data delivered through a single path. The multipath routing protocols are broadly classified based on on-demand, table driven, and hybrid. The following multipath routing protocols are used in MANETs. In [1] authors have compared the performance of AOMDV and OLSR routing protocol with Levy-Walk and Gauss-Markov Mobility Model. For the analysis they have considered varying mobility speed and the traffic load in the network. Their results show that AOMDV protocol achieved higher packet delivery ratio and throughput compared to OLSR. Further, OLSR has less delay and routing overhead at varying node density. In [2] authors only compared AOMDV and AODV routing protocol with random way point mobility model. Different traffic source like TCP and CBR is considered. The result shows that with increasing traffic both routing protocols performance degraded. In M-DSR (Multipath Dynamic Source Routing) [5, 21] is an on demand routing protocol based on DSR [12] is a multipath extension of DSR. In SMR (Split Multipath Routing) [5, 15] is an on demand routing protocol and extension of well- known DSR protocol. The main aim of this protocol is to split the traffic into multiple paths so that bandwidth utilization goes in an efficient manner. In GMR (Graph based Multipath Routing) [5, 9] protocol based on DSR, a destination node compute disjoint path in the network using network topology graph. In MP-DSR [5, 13, 16] is based on DSR; it is design to improve QoS support with respect to end-to-end delay. In [10,19] authors have proposed an on-demand multipath routing protocol AODV-BR. But to establish multipath it does not spend extra control message. This protocol utilizes mesh structure to provide multiple alternate paths. In [8] authors have considered node-disjoint and link-disjoint multi-path routing protocol for their analysis. The various mobility model considered are Random Waypoint, Random Direction, Gauss-Markov, City Section and Manhattan mobility models. Through the thorough analysis they have shown that in Gauss markov mobility model multipath formation is less but path stability is high. The random direction model form larger number of multipath. In [14] authors have considered AODV and AOMDV protocol for their performance analysis with random waypoint model. The result shows that AOMDV has more routing overhead and average end to end delay compared to AODV. But AOMDV perform better in term of packets drops and packet delivery. In [17] various energy models with Random Waypoint Mobility Model-Steady State mobility model is used to analyze the energy overhead AOMDV, TORA and OLSR routing protocols. Results show that TORA protocol has highest energy overhead in all the energy models.

3. DESCRIPTION OF ROUTING PROTOCOL AND MOBILITY MODELS

In this section we have discussed brief about AOMDV routing protocol and various mobility models considered for simulation work.

3.1 Ad Hoc On Demand Multipath Distance Vector (AOMDV)

Ad Hoc On Demand Multipath Distance Vector (AOMDV) [3, 5, 6, 11] protocol is a multipath variation of AODV protocol. The main objective is to achieve efficient fault tolerance i.e. quickly recovery from route failure. The protocol computes multiple link disjoint loop free paths per route discovery. If one path fails the protocol choose alternate route from other available paths. The route discovery process is initiated only when to a particular destination fails. When a source needs a route to destination will floods the RREQ for the destination and at the intermediate nodes all duplicate RREQ are examined and each RREQ packet define an alternate route. However, only link disjoint routes are selected (node disjoint routes are also link disjoint). The destination node replies only k copies of out of many link disjoint path, i.e. RREQ packets arrive through unique neighbors, apart from the first hop are replied. Further, to avoid loop 'advertised hop count' is used in the routing table of node .The protocol only accepts alternate route with hop count less than the advertised hop count. A node can receive a routing update via a RREQ or RREP packet either forming or updating a forward or reverse path .Such routing updates received via RREQ and RREP as routing advertisement.

3.2 Mobility Models

Mobility pattern of node plays a vital role in evaluation of any routing protocol in MANET. We have considered various categories mobility models for acceptability of routing protocol. The following mobility model we have considered in simulation work.

3.2.1 Random Waypoint Model

The Random Waypoint (RWP) mobility model [4,7] is the only model which is used in maximum cases for evaluation of MANET routing protocols. In this model nodes movement depends on mobility speed, and pause time. Nodes are moving in a plane and choose a new destination according to their speed. Pause time indicate that a node to wait in a position before moved to new position.

3.2.2 Probabilistic Random Walk Model

In this model [4,7] nodes next position is determined by set of probabilities. A node can be move forward, backward or remain in x and y direction depends on the probability defined in probability matrix. There are three state of node is defined by 0 (current position), 1 (previous position) and 2 (next position). Where, in the matrix P (a,b) means the probability that an node will move from state a to state b.

3.2.3 Random Direction Model

The random direction model [4,7] is the further modification of Random waypoint mobility model. This model overcome the density wave problem occur in random waypoint model, where clustering of nodes occur in a particular area of simulation. In Random Waypoint model this density occurs in the center of the simulation area. Here, nodes are move upto the boundary of the simulation area before moving to a new location with new speed and direction. When nodes are reached to the boundary of simulation area, before changing to new position it pauses there for sometimes. The random direction it chooses from 0 to 180 degrees. The same process is continued till the simulation time.

4. Simulation Setup And Result Analysis

For the simulation works we have used Bonn-Motion mobility generator [18] to generate the mobility of nodes based of various mobility models. The most popular network simulator NS-2.34 [20] is used to simulation work. Finally, Matlab [22] is used to compute the results. In table-1 and table-2 shows different simulation parameters and their values respectively. We have computed packet delivery ratio as a parameter to analyze the performance of AOMDV protocol.

Table 1. Simulation Parameters

Parameter	Specifications
MAC Protocol	IEEE 802.11 DCF
Routing Protocol	AOMDV
Radio Propagation Model	Two-ray ground reflection model
Channel type	Wireless channel
Antenna model	Omni-directional
Mobility Models	Random waypoint, Random Direction, Probabilistic Random Walk

Table 2. Values of Simulation Parameters

	Values
Simulation Time	1000s
Simulation Area (X *Y)	1000 m x1000 m
Transmission Range	250 m
Bandwidth	2 Mbps
No. of Nodes	10,20,30,40,50,100
Node speed	10,20,30 m/s

Fig.1 shows the packet delivery ratio at node mobility 10 m/s in various mobility models. In Probabilistic Randomwalk model AOMDV gives better packet delivery ratio with increasing node density. In Random direction model AOMDV protocol perform better at node density 70 onwards. Except Probabilistic Random walk model in rest of the model PDR value decreases in high node density. The highest PDR value achieved 77.8.

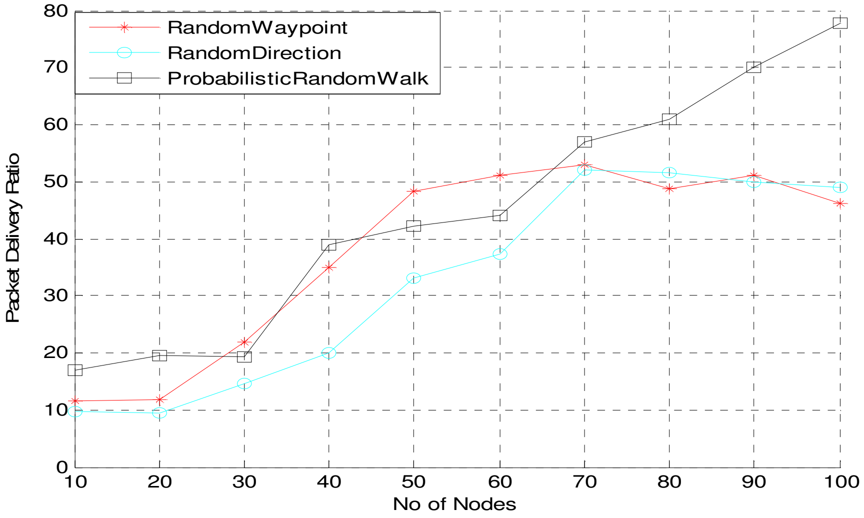


Figure 1. Packet delivery ratio with node speed at 10m/s.

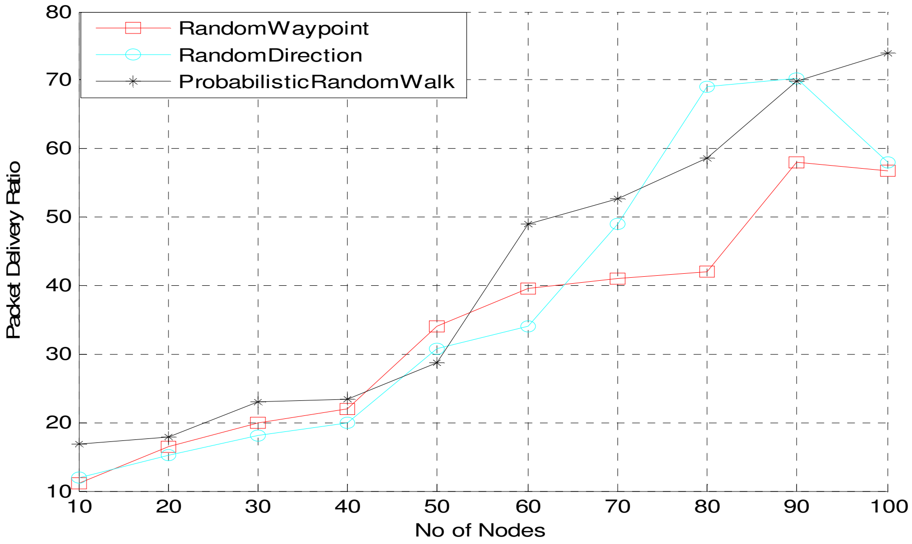


Figure 2. Packet delivery ratio with node speed at 20m/s.

Fig.2 shows the packet delivery ratio at node mobility 20 m/s in various mobility models. In this scenario up to node density 50 protocol perform quite same, but there is slight improvement is noticed in all the models till node 90. After node density 90 only in probabilistic random model protocol perform better.

Fig.3 shows the packet delivery ratio at node mobility 30 m/s in various mobility models. In Probabilistic Randomwalk model AOMDV gives better packet delivery ratio after node density 80. The protocol perform better in Randomway point model as compare to others till node density 40, but after that slight decrease in noticed in PDR values till node density 60 in Randomway point model. In node density 40 to 70 the protocol performs better with random direction model.

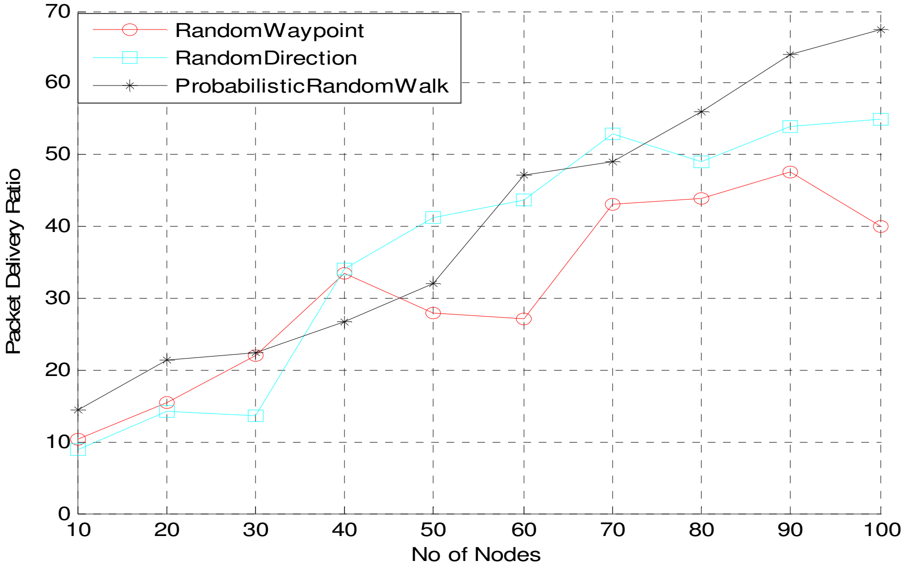


Figure 3. Packet delivery ratio with node speed at 30m/s.

The results show that with high node mobility in all models PDR value decreases. When node density is 100, PDR value decreases almost 13% is noticed in probabilistic Random walk and Randomway point model. But in random direction model increase of 12% in PDR value noticed.

5. CONCLUSIONS

We have evaluated the performance of multipath routing protocol with different mobility models. We have generated various node movements with varying node speed and number node based on mobility models. For analysis the performance of the protocol packet delivery ratio is computed. It is evident from the results that AOMDV protocol perform better in term of PDR in Probabilistic Randomwalk model in low node mobility, and for higher node mobility except random direction model in other models PDR decreases. In future, this multipath protocol can be investigated with various other network topologies.

REFERENCES

[1] Gowrishankar. S, et al. (2010) “Analysis of AOMDV and OLSR Routing Protocols under Levy-Walk Mobility Model and Gauss-Markov Mobility Model for Ad Hoc Networks”, (IJCSE) In-ternational Journal on Computer Science and Engineering, Vol. 02, No. 04, 2010, pp. 979-986.

[2] Vivek B. Kute et al., (2013) “Analysis of Quality of Service for the AOMDV Routing Protocol”, ETASR - Engineering, Technology & Applied Science Research Vol. 3,No. 1, pp.359-362.

[3] Jiazi Yi , AsmaaAdnane, Sylvain David, and Benoît Parrein, (2011) “Multipath optimized link state routing for mobile ad hoc networks”, Ad Hoc Networks, Vol. 9, No.1, pp. 28-47 .

[4] Radhika Ranjan Roy, (2011) Handbook of Mobile Ad Hoc Networks for Mobility Models, First Edition, Springer, New York Dordrecht Heidelberg London, ISBN 978-1-4419-6048-1 e-ISBN 978-1-4419-6050-4.

[5] Tsai, J., & Moors, T., (2006) “A review of multipath routing protocols: from wireless ad hoc to mesh networks”, In Proceedings of ACoRN early career researcher workshop on wireless multi-hop networking, Sydney.

[6] M. K. Marina and S. R. Das, (2006) “Ad-hoc on-demand multi-path distance vector routing”, Wireless Communication Mobile Computing, Vol. 6, No. 7, pp. 969–988.

[7] Camp, Tracy et al., (2002) “A Survey of Mobility Models for Ad Hoc Network Research”, wire-less communications & mobile computing (WCMC): special issue on mobile ad hoc networking: research, trends and applications, Vol.2, No.5, pp. 483-502.

[8] Nicholas cooper et al., (2010) “Impact of Mobility models on multipath routing in mobile Ad hoc Networks”, International Journal Of Computer Networks & Communications (IJCNC), Vol. 2, No.1, pp.185-194.

[9] Gunyoung Koh, Duyoung Oh and Heekyoung Woo, (2003) “A graph-based approach to com-pute multiple paths in mobile ad hoc networks”, Lecture Notes in Computer Science Vol. 2713, Springer, pp. 3201–3205.

[10] M.T.Toussaint, (2003) “Multipath Routing in Mobile Ad Hoc Networks”, TU-Delft/TNO Traineeship Report.

[11] S. Das, C. Perkins and E. Royer, “Ad Hoc On Demand Distance Vector (AODV) Routing”, IETF RFC3561, July 2003.

[12] D. Johnson, (2003) “The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR)”, IETF Internet Draft, draft-ietf-manet-dsr-09.txt.

[13] E. Esmaeili, P. Akhlaghi, M. Dehghan, M.Fathi,(2006) “A New Multi-Path Routing Algorithm with Local Recovery Capability in Mobile Ad hoc Networks”, In the Proceeding of 5th Interna-tional Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP 2006), Patras, Greece, pp. 106-110.

[14] R.Balakrishna et al., (2010) “Performance issues on AODV and AOMDV for MANETS”, International Journal of Computer Science and Information Technologies, Vol. 1, Issue.2, pp. 38-43.

[15] S. J. Lee and M. Gerla, (2001) “Split Multipath Routing with Maximally Disjoint Paths in Ad Hoc Networks”, In Proceedings of the IEEE ICC, pp. 3201-3205.

[16] R. Leung, J. Liu, E. Poon, Ah-Lot. Chan, B. Li, (2001) “MP-DSR: A QoS-Aware Multi-Path Dynamic Source Routing Protocol for Wireless Ad-Hoc Networks”, In Proc. of 26th Annual IEEE Conference on Local Computer Networks (LCN), pp. 132-141.

[17] Gowrishankar.S et al., (2010) “Simulation Based Overhead Analysis of AOMDV, TORA and OLSR in MANET Using Various Energy Models”, Proceedings of the World Congress on Engi-neering and Computer Science,San Francisco, USA, Vol.1.

[18] Bonn Motion, http://net.cs.uni-bonn.de/wg/cs/applications/bonnmotion/

[19] Sung-Ju Lee and Mario Gerla, (2000) “AODV-BR: Backup Routing in Ad hoc Networks”, IEEE Conference on Wireless Communications and Networking Conference (WCNC- 2000),Vol.3, PP. 1311-1316 .

[20] The Network Simulator. http://www.isi.edu/nsnam/ns/.

[21] A. Nasipuri and S. R. Das, (1999)“On-demand multipath routing for mobile ad hoc networks”, In the Proceedings of Eight International Conference on Computer Communications and Net-works, Boston, MA, .

[22] The Math Works: http://www.mathworks.com

AUTHORS

Indrani Das did her B. E. and M.Tech in Computer Science. She is working as Assistant Professor in Computer Science depart-ment in Assam University (A Central University), Assam, India. Presently, she is perusing her Ph.D from School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, In-dia. Her current research interest includes Mobile Ad hoc Net-works and Vehicular Ad hoc Networks.

Daya K. Lobiyal Received his Bachelor of Technology in Com-puter Science from Lucknow University, India, in 1988 and his Master of Technology and Ph.D both in Computer Science from Jawaharlal Nehru University, New Delhi, India, 1991 and 1996, respectively. Presently, he is an Associate Professor in the School of Computer and Systems Sciences, Jawaharlal Nehru University, India. His areas of research interest are Mobile Ad hoc Networks, Vehicular Ad Hoc Networks, Wireless Sensor Network and Vid-eo on Demand.

C. P. Katti is a professor of computer science at Jawaharlal Nehru University. He received his Ph.D from IIT Delhi. He has published over 30 papers in journals of international repute. His area of research includes parallel computing, ad hoc networks and numerical analysis.

INTENTIONAL BLANK

UNIFORMLY SPACED PLANAR ANTENNA ARRAY OPTIMIZATION USING CUCKOO SEARCH ALGORITHM

A.Sai Charan[1], N.K.Manasa[2], Prof. N.V.S.N Sarma[3]

[1,2,3]Department of Electronics and Communication Engineering,
National Institute of Technology, Warangal
[1]charanadd@gmail.com
[2]manasank1992@gmail.com
[3]sarma@nitw.ac.in

ABSTRACT

In this modern era a great deal of metamorphism is observed around us which eventuate due to some minute modifications and innovations in the area of Science and Technology. This paper deals with the application of a meta heuristic optimization algorithm namely the Cuckoo Search Algorithm in the design of an optimized planar antenna array which ensures high gain, directivity, suppression of side lobes, increased efficiency and improves other antenna parameters as well[1], [2] and [3].

KEYWORDS

Meta-Heuristic, Side Lobe Suppression, Gain, Directivity, Side Lobe Level (SLL).

1. INTRODUCTION

Antenna optimization techniques have made a breakthrough in the Communication domain. They have contributed vividly to modern wireless communications in the form of smart antennas which are antenna arrays that adjust their own beam pattern to accentuate signals of interest and concurrently reducing the radio frequency interference. In the field of antennas, Cuckoo Search Algorithm (CSA) was first applied for side lobe suppression in linear antenna array by distance modulation.

Large arrays are complex to build, have increased fabrication and set up cost and are heavier at the same time. Therefore reducing antenna element weight from the array is desirable without degrading the performance of the array. But here we are not reducing the mass of the antenna array elements, only the weight of the antenna elements(current)are adjusted in order to achieve minimum side lobe level.

We opt a technique based on density tapering to lower side- lobes in the array by monotonically decreasing the magnitude of weights away from the centre of the array.

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 157–167, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4517

2. REVIEW OF VARIOUS TECHNIQUES

Owing to high adaptability and ability to optimize multi-dimensional problems, several evolutionary algorithms have been proposed such as Particle Swarm Optimization (PSO), Invasive Weed Optimization (IWO), Genetic Algorithm (GA),etc. These algorithms are associated with some drawbacks which make them unreliable. The PSO could not work out the problem of scattering and optimization, the IWO require the genes of minimum one parent species to be forwarded to next generation and the GA has a poor fitness function which generates bad chromosome blocks in spite of the fact that only good chromosome blocks cross over. Also no assurance is given whether the GA will find a global optimum solution [4].

This paper has explored a choice of antenna array synthesis, the (CSA) [5], to overcome the above mentioned problems and to yield promising results.

3. PLANAR ARRAYS

Planar array is a two dimensional configuration of elements arranged to lie in a plane. The planar array may be thought of as an array of linear arrays. The elements are arranged in a matrix form having a phase shifter. The planar arrangement of all antenna elements forms the complete phased array antenna. There are wide spread applications of planar antenna arrays which involve the suppression of side lobes. The signals radiated by individual antennas determine the effective radiation pattern of the array. They are used to point a fixed radiation pattern or to scan a region rapidly in the azimuthal plane. Several methods have been developed for the design of planar antenna array but all those methods pertain to other nature inspired optimization algorithms.

Planar antenna array optimization has been implemented earlier using Fuzzy GA [6]. Direction angle (reference angle) is considered with the plane of planar antenna array. This paper deals with the design of a planar antenna array by using CSA.

4. CUCKOO SEARCH ALGORITHM (CSA)

CSA is one of the modern nature inspired meta-heuristic algorithms. The Greek terms “meta” and “heuristic” refer to “change” and “discovery oriented by trial and error” respectively. Various techniques are used to minimise the constraints associated with the problem in order to obtain a global optimum solution.

Cuckoos are attractive birds. The attractiveness is owing to the beautiful sounds produced by them and also due to their reproduction approach which proves to be combative in nature. These birds are referred to as brood parasites as they lay their eggs in communal nests. They remove the eggs in the host bird nest in order to increase the hatching probability of their own eggs.

There are three types of brood parasites - the intraspecific brood parasite, cooperation breed and nest take over type. The host bird involves in direct combat with the encroaching cuckoo bird. If the host bird discovers the presence of an alien egg, it either throws away the egg or deserts the nest. Some birds are so specialized that they have the characteristic of mimicking the colour and the pattern of the egg which reduces the chances of the egg being left out thereby increasing their productivity [7].

The timely sense of egg laying of cuckoo is quite interesting. Parasitic cuckoo birds are in search of host bird nests which have just laid their own eggs. In general the cuckoo birds lay their eggs earlier than the host bird's eggs in order to create space for their own eggs and also to ensure that a large part of the host bird feed is received by their chicks.

5. PRINCIPLE BEHIND CUCKOO SEARCH ALGORITHM

Each cuckoo bird lays a single egg at a time which is discarded into a randomly chosen nest. The optimum nest with great quality eggs is carried over to next generations. The number of host nests is static and a host can find an alien egg with a probability (Pa) [0, 1], whose presence leads to either throwing away of the egg or abandoning the nest by the host bird [8].

One has to note that each egg in a nest represents a solution and a cuckoo egg represents a new solution where the objective is to replace the weaker fitness solution by a new solution.

The flowchart for CSA is as shown which involves the following steps:

Step (1) - Introduce a random population of n host nests, X_i.

Step (2) - Obtain a cuckoo randomly by Levy flight behaviour, i.

Step (3) - Calculate its fitness function, F_i.

Step (4) - Select a nest randomly among the host nests say j and calculate its fitness, F_j.

Step (5) - If $F_i < F_j$, $_{then}$ replace j by new solution else let j be the solution.

Step (6) - Leave a fraction of Pa of the worst nest by building new ones at new locations using Levy flights.

Step (7) - Keep the current optimum nest, Go to Step (2) if T (Current Iteration) < MI (Maximum Iteration).

Step (8) - Find the optimum solution.

Important Stages involved in CSA are:

i) Initialization: Introduce a random population of n host nest (X_i = 1, 2, 3...n).

ii) Levy Flight Behaviour: Obtain a cuckoo by Levy flight behaviour equation which is defined as follows:

$$X_i(t+1) = X_i(t) + \alpha \square \text{ Levy } (\lambda), \alpha > 0 \quad (1)$$

$$\text{Levy } (\lambda) = t^{(-\lambda)}, 1 < \lambda < 3 \quad (2)$$

iii) Fitness Calculation: Calculate the fitness using the fit- ness function in order to obtain an optimum solution. Select a random nest, let us say j. Then the fitness of the cuckoo egg (new solution) is compared with the fitness of the host eggs (solutions) present in the nest. If the value of the fitness function of the cuckoo egg is less than or equal to the fitness function value of the randomly chosen nest then the randomly chosen nest (j) is replaced by the new solution.

$$\text{Fitness Function} = \text{Current Best Solution} - \text{Previous Best Solution} \qquad (3)$$

Since the Fitness function = Current best solution - Previous best solution, the value of the fitness function approaching the value zero means that the deviation between solutions decreases due to increase in the number of iterations.

The conclusion is that if the cuckoo egg is similar to a normal egg it is hard for the host bird to differentiate between the eggs. The fitness is difference in solutions [10] and the new solution is replaced by the randomly chosen nest. Otherwise when the fitness of the cuckoo egg is greater than the randomly chosen nest, the host bird recognizes the alien egg, as a result of which it may throw the egg or forsake the nest.

The various stages involved in the working of this algorithm are explained in the flow chart.

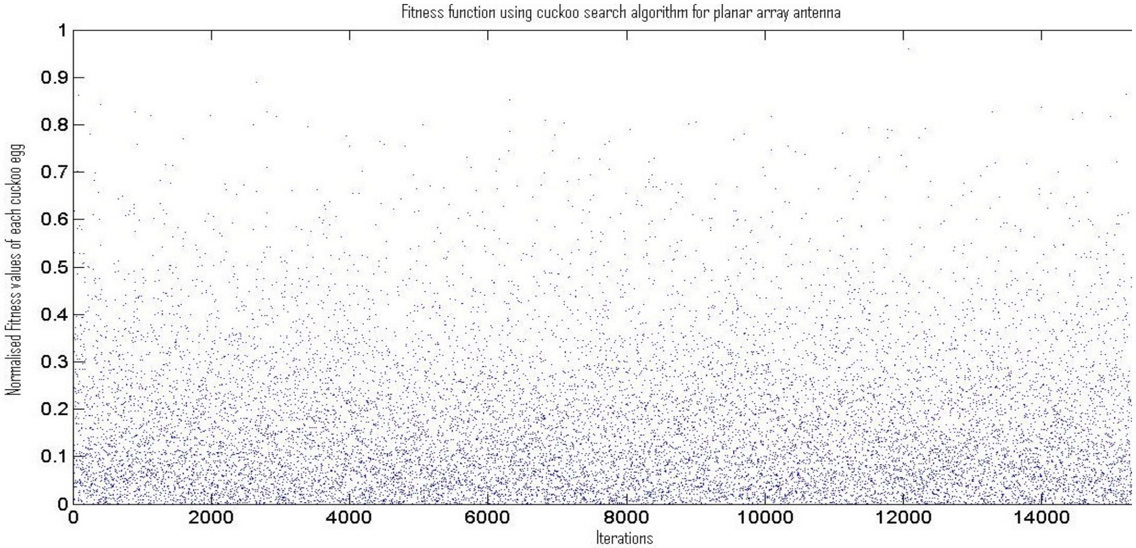
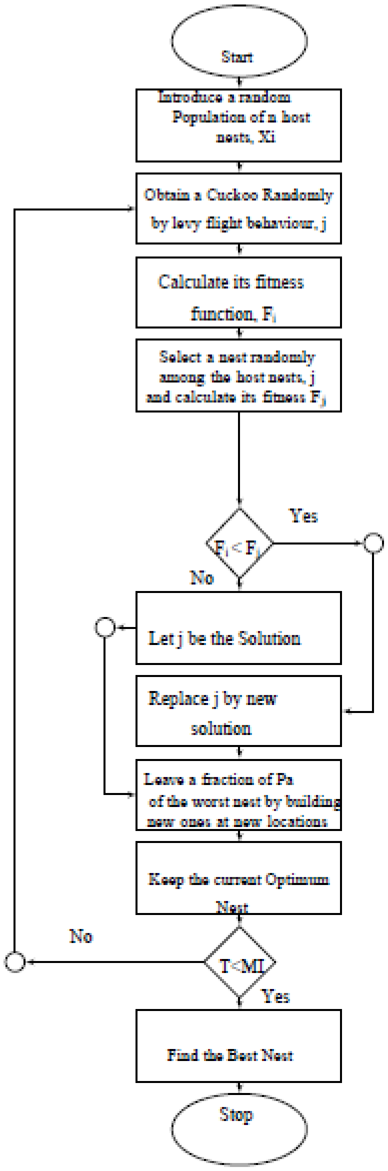


Figure 1. Fitness function values for planar antenna array of 18x18 elements

From the fitness function graph it can be observed that as the number of iterations increases, the value of the fitness function graph approaches to zero.


Start
Introduce a random Population of n host nests, Xi
Obtain a Cuckoo Randomly by levy flight behaviour, j
Calculate its fitness function, F_i
Select a nest randomly among the host nests, j and calculate its fitness F_j
Yes
$F_i < F_j$
No
Let j be the Solution
Replace j by new solution
Leave a fraction of Pa of the worst nest by building new ones at new locations
Keep the current Optimum Nest
No
T<MI
Yes
Find the Best Nest
Stop

iv) Termination: In the current iteration the solution is compared and the best solution is only passed further which is done by the fitness function. If the number of iterations is less than the maximum then it keeps the best nest.

After the execution of the initialisation process, the levy flight and the fitness calculation processes, all cuckoo birds are prepared for their next actions. The CSA will terminate after maximum iterations [MI], have been reached.

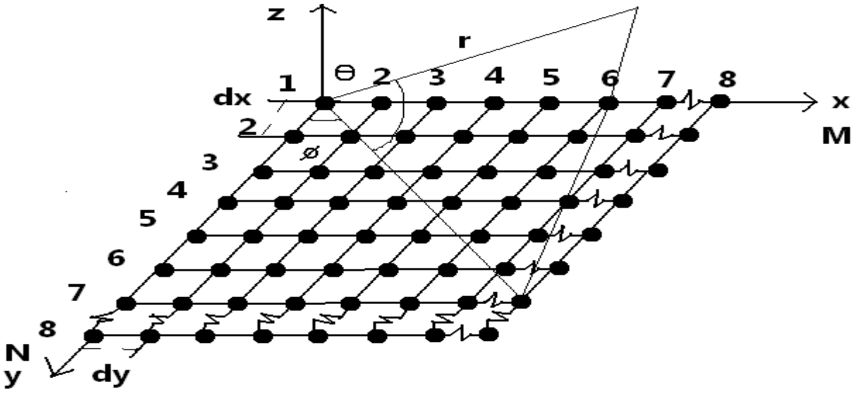


Figure 2. Planar Antenna Array Set–Up

6. Technical Details

6.1. Synthesis of Planar Antenna Array

Consider a planar antenna array which consists of M-by-N rectangular antennas which are spaced equally [10]. They have been arranged in a regular rectangular array in the x-y plane. The inter-element spatial arrangement is

$$d = dx = dy = \lambda/2 = R_0$$

Where λ is the wavelength

The outputs are summed up in order to provide a distinct output.

$$F_s(\theta,\varphi) = \frac{f(\theta,\varphi)}{F_{s\,msx}} \sum_{m=1}^{M} I_m \, e^{j((m-1)k_x dx)+\psi_m)} \sum_{n=1}^{N} I_n e^{j\left((n-1)k_y dy\right)+\psi_n)}$$

Where $k_x = \frac{2\pi}{\lambda}\sin\theta\cos\varphi$, $k_y = \frac{2\pi}{\lambda}\sin\theta\sin\varphi$

6.2 Number of Cuckoo Birds

This parameter decides number of Cuckoo birds being initialized in the field space.

6.3. Step Size

In case of CSA, step size refers to the distance covered by a cuckoo bird for a fixed number of iterations. It is preferred to have an intermediate step size in order to obtain an effective solution. If the step size is too large or too small it leads to deviation from the required optimum solutions [7].

7. FIGURES AND TABLES

Table 1: SLL values for various sizes of Planar Antenna Array

M	N	P_a	SLL (in dB)	Main Lobe Range	Φ(in degrees)	Figure No.
11	11	.25	-28.8	[79.8, 100.2]	90	4
13	13	.25	-26.5	[81.5, 100.5]	0	6
15	15	.25	-31.7	[82.2, 97.8]	90	7
16	16	.25	-29.2	[83, 97]	0	8
18	18	.25	-29.8	[83.7, 96.3]	0	9
20	20	.25	-32.3	[84.2, 95.8]	90	10

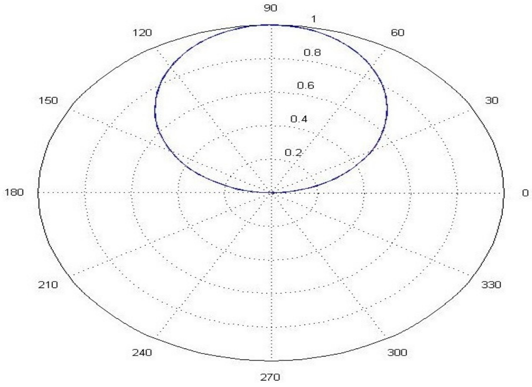


Figure 3. Polar pattern of a single antenna array element

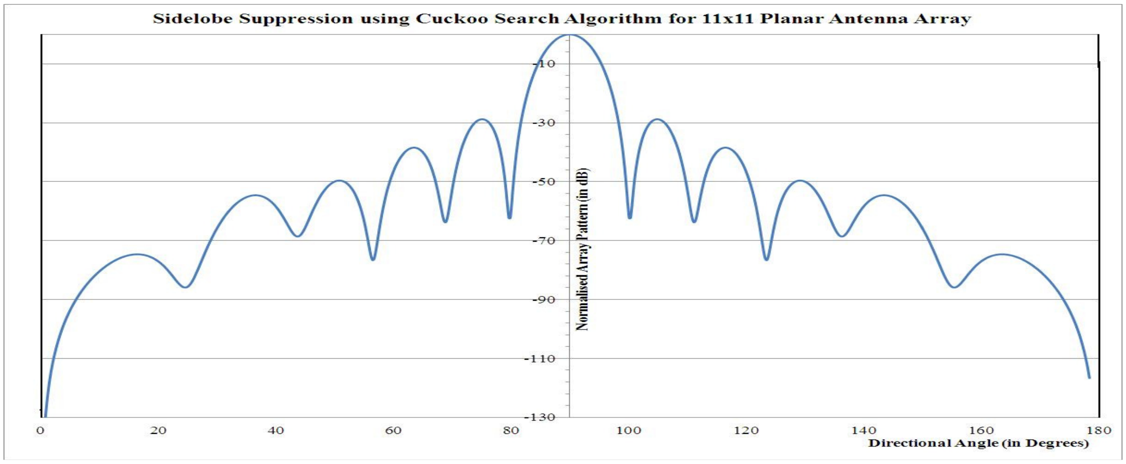


Figure 4. Radiation Pattern for a planar antenna array of 121 elements and φ= 0 degrees

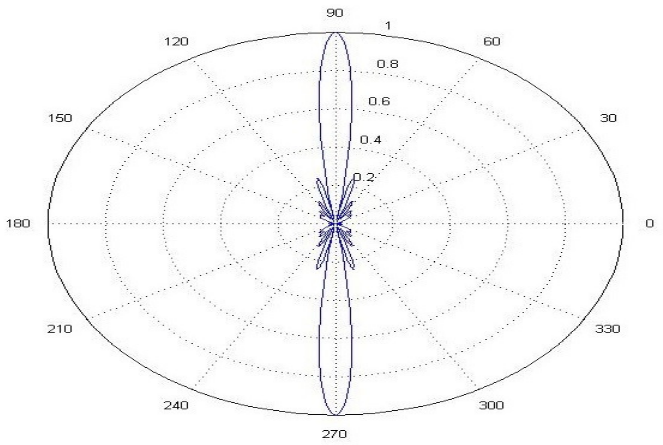


Figure 5. Polar pattern for the radiation of an 11X11 planar antenna array

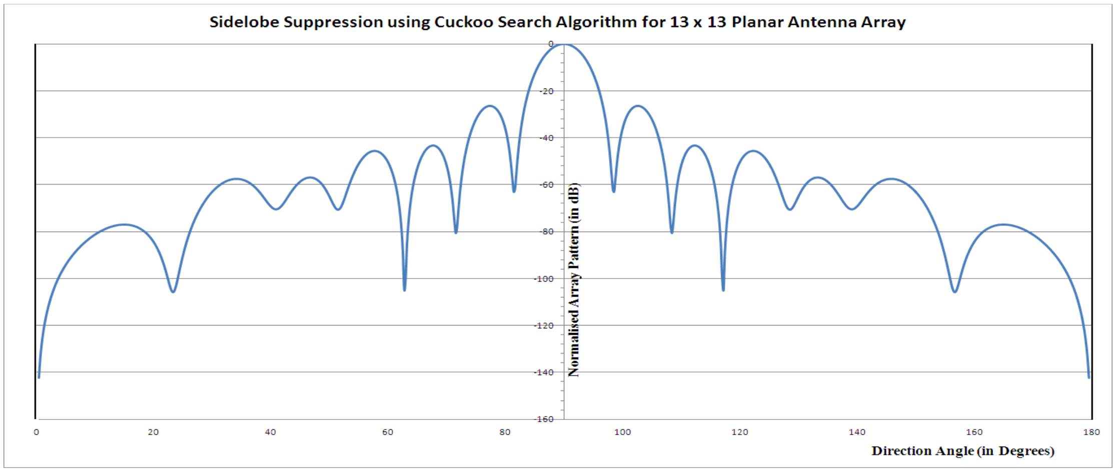


Figure 6. Radiation Pattern for a planar antenna array of 169 elements and φ= 0 degrees

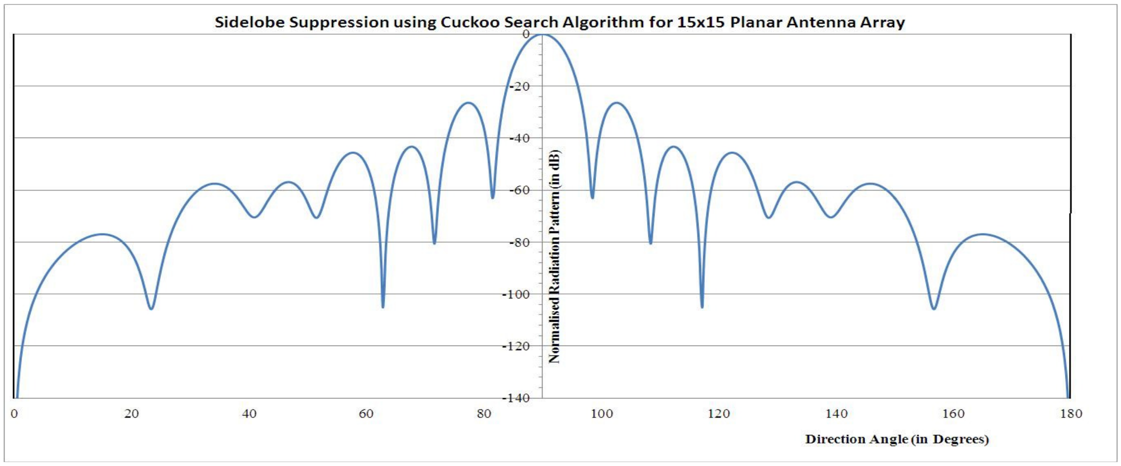


Figure 7. Radiation Pattern for a planar antenna array of 225 elements and φ=90 degrees

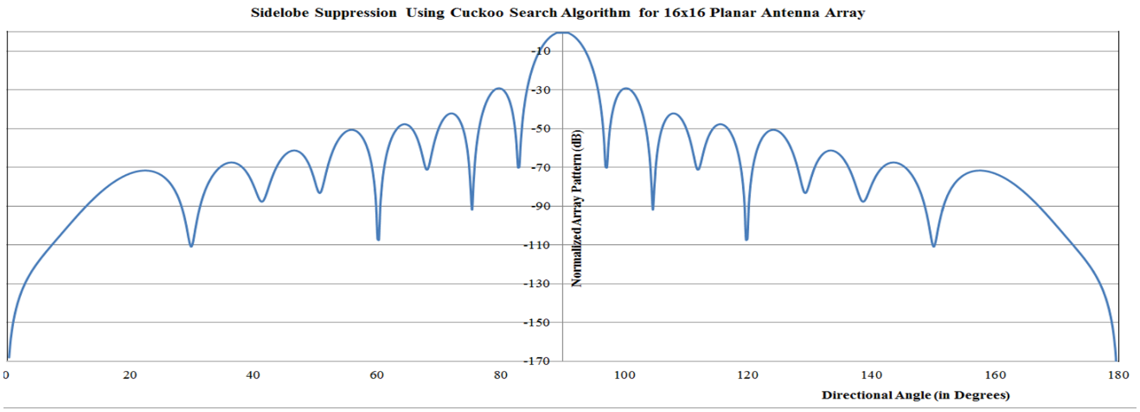


Figure 8. Radiation Pattern for a planar antenna array of 256 elements and φ= 0 degrees

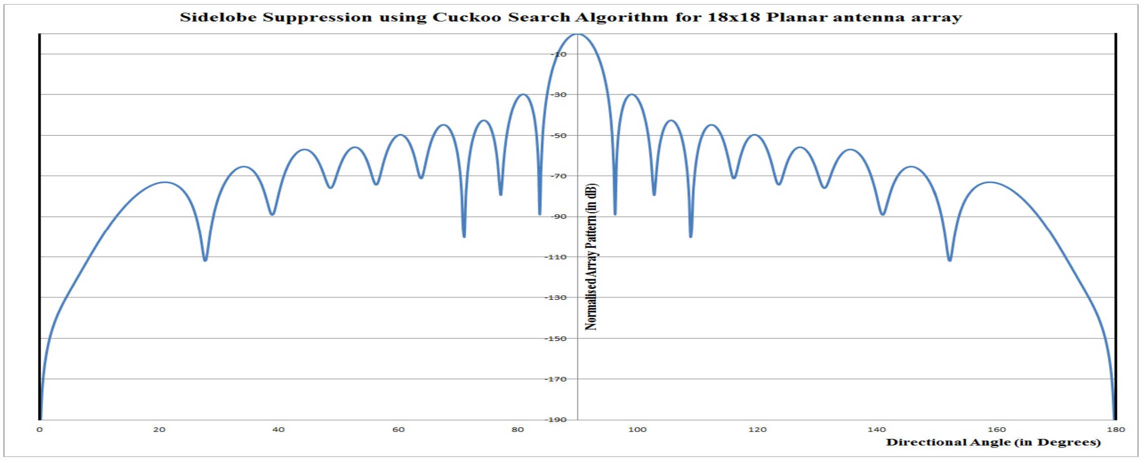


Figure 9. Radiation Pattern for a planar antenna array of 324 elements and φ= 0 degrees

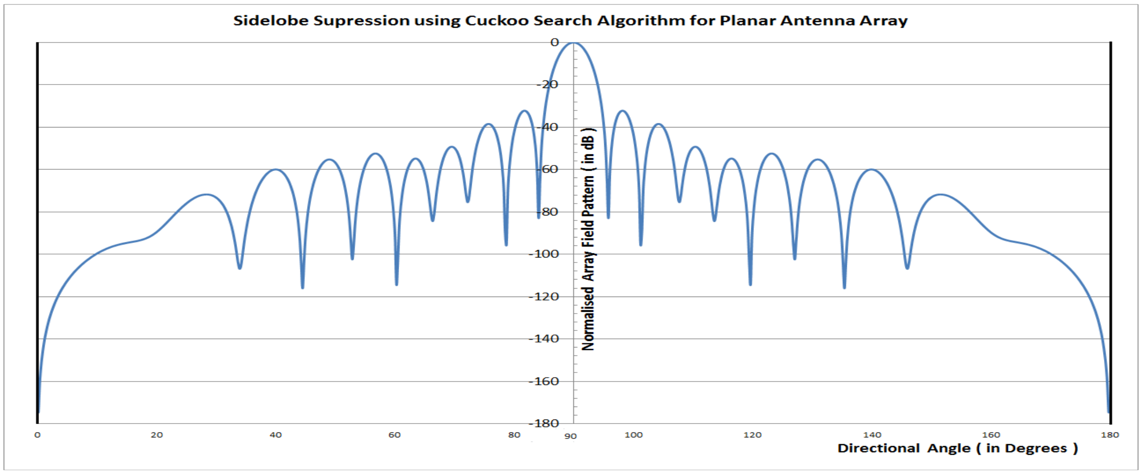


Figure 10. Radiation Pattern for a planar antenna array of 400 elements and φ = 90 degrees

8. OBSERVATIONS

When φ = 0 degrees, Maximum iterations = 500, a narrow beam is obtained as the best optimum solution for a large value of the number of iterations (optimum value) [11]. When maximum iteration is 150, main lobe appears to be spread over a wide range of direction angle (θ). For increase in the number of antenna elements in a planar antenna array a narrow beam is achieved correspondingly. The same field pattern is obtained for φ=π/2 degrees, maximum iterations of 500.

The directivity of an isotropic antenna is unity as power is radiated equally in all directions [Fig 3]. In case of other sources such as omnidirectional antennas, sectoral antennas, directivity is greater than unity. Directivity can be considered as the figure of merit of directionality as it is an indication of the directional properties of the antenna with respect to an isotropic source. This shows that for any alignment of planar antenna array the same field pattern will be obtained which promotes beam steering in RADAR applications.

9. CONCLUSIONS

CSA is very easily applicable among all the nature inspired meta-heuristic algorithms since it provides the optimum solution. The implementation of CSA led to a tremendous increase in directivity [Fig 4] which promotes long distance communication. Gain of Antenna Array is also increased by using CSA.

ACKNOWLEDGEMENTS

We are very thankful to Prof. N.V.S.N. Sarma for guiding us throughout this project.

REFERENCES

[1] Ehsan Valian, Shahram Mohanna, Saeed Tavakoli, Improved Cuckoo Search Algorithm for Feed Forward Neural Network Training, International Journal for Artificial Intelligence and Applications (IJAIA) ,Vol.2, No.3, Pp. 36-43, July 2011.

[2] Pallavi Joshi, Nitin Jain, Optimization of Linear Antenna Array Using Genetic Algorithm for Reduction in Side Lobe Level and to Improve Directivity, International Journal of Latest Trends in Engineering and Technology(IJLTET), Vol.2, Issue 3, Pp. 185-191, May 2013.

[3] Khairul Najmy ABDUL RANI, Mohd.Fareq ABD MALEK, Neoh SIEW-CHIN, Nature Inspired Cuckoo Search Algorithm for Side Lobe Suppression in a Symmetric Linear Antenna Array, RADIO ENGINEER-ING, Vol.21, No.3, Pp. 865-974, September 2012.

[4] Ch.Ramesh, P.Mallikarjuna Rao, Antenna Array Synthesis for Suppressed Side Lobe Level Using Evolutionary Algorithms, International Journal of Engineering and Innovative Technology(IJEIT), Volume 2, Issue 3, Pp. 235-239, September 2012.

[5] Ehsan Valian, Shahram Mohanna, Saeed Tavakoli, Improved Cuckoo Search Algorithm for Global Optimization, International Journal of Communications and Information Technology, IJCIT-2011-Vol.1-No.1, Pp. 31-44, Dec 2011.

[6] Boufeldja Kadri, Miloud Boussahla, Fethi Tarik Bendimerad, Phase-Only Planar Antenna Array Synthesis with Fuzzy Genetic Algorithms, IJCSI, International Journal of Computer Science Issues, Vol.7, Issue 1, No.2, Pp. 72-77, January 2010.

[7] Xin-She-Yang, Nature Inspired Meta-heuristic Algorithm, 2nd Edition, Luniver Press, 2010.

[8] Monica Sood, Gurline Kaur, Speaker Recognition Based on Cuckoo Search Algorithm, International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-2, Issue-5, Pp. 311-313, and April 2013.

[9] Moe Moe Zaw, Ei Ei Mon, Web Document Clustering Using Cuckoo Search Clustering Algorithm based on Levy Flight, International Journal of Innovation and Applied Studies, ISSN 2028-9324, Vol.4, No.1, Pp182-188, Sep 2013.

[10] Constantine A.Balanis, Antenna Theory Analysis and Design, 2nd Edition, Pp. 310.

[11] Robert S. Elliott, Antenna Theory and Design, Revised Edition, IEEE press.

[12] Randy I. Haupt, Wiley, Antenna Array - A Computational Approach, IEEE press.

INTENTIONAL BLANK

EFFICIENT ASIC ARCHITECTURE OF RSA CRYPTOSYSTEM

Varun Nehru[1] and H.S. Jattana[2]

VLSI Design Division, Semi-Conductor Laboratory,
Dept. of Space, S.A.S. Nagar.
[1]nehruvarun@gmail.com, [2]hsj@scl.gov.in

ABSTRACT

This paper presents a unified architecture design of the RSA cryptosystem i.e. RSA crypto-accelerator along with key-pair generation. A structural design methodology for the same is proposed and implemented. The purpose is to design a complete cryptosystem efficiently with reduced hardware redundancy. Individual modular architectures of RSA, Miller-Rabin Test and Extended Binary GCD algorithm are presented and then they are integrated. Standard algorithm for RSA has been used. The RSA datapath has further been transformed into DPA resistant design. The simulation and implementation results using 180nm technology are shown and prove the validity of the architecture.

KEYWORDS

RSA, cryptosystem, crypto-accelerator, public key, private key, Extended Binary GCD, Stein, Miller-Rabin, modular inverse, DPA resistance

1. INTRODUCTION

The RSA algorithm [1] is a public key algorithm and is extensively in security and authentication applications. Being computationally intensive, use of separate crypto-accelerator hardware to accelerate the computations is common. The communication between the main processor (32-64 bit) and the RSA crypto-accelerator (1024-2048 bit) requires a protocol for data exchange and a FIFO register bank can implemented for the same. This paper describes an architecture design for the RSA cryptosystem useful for both the Encryption/Decryption and for the Key-Pair Generation which may be required due to security. The number to be tested as prime is fed as input to the system and the random numbers for Miller-Rabin test are generated using Pseudo-Random Number Generator (PRNG).

The paper is organized as follows: Section 2 introduces the basics of RSA algorithm. Section 3 describes fundamental algorithms, with modular architecture around which the top level system was developed. Section 4, discusses top-level implementation. Section 5 briefs about power analysis attacks. In Section 6, implementation results have been shown. In Section 7, conclusion is drawn.

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 169–183, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4518

2. Basics of RSA

RSA involves the use of a public key-pair {e, n} and a private key {d, n} for encryption and decryption respectively. Messages encrypted with the public key can only be decrypted using the private key. For digital signatures private key is used. The proof of the algorithm can be found in [1]. The steps for Key Generation and Encryption/Decryption are reproduced below:

2.1. Key-Pair Generation

1. Choose primes, p and q.
2. Compute modulus $n = p*q$. Its length is called the key length.
3. Compute Euler's totient function, $\varphi(n) = (p - 1)(q - 1)$.
4. Choose a public key, e, such that $1 < e < \varphi(n)$ and $\gcd(e, \varphi(n)) = 1$.
5. Determine d as $d-1 \equiv e \pmod{\varphi(n)}$.

2.2. Encryption and Decryption

Cipher text(C) is obtained as a number theory equivalent to the public key (e) exponentiation of message (M) in modulus n

$$C = M^e \bmod \{n\}.$$

Similarly, message can be recovered from cipher text by using private key exponent (d) via computing

$$M = C^d \bmod \{n\}.$$

3. Modular Design Architectures

This section describes the architectures developed for various modules used in the design of RSA cryptosystem.

3.1. Modular Multiplication

The binary interleaving multiplication and reduction algorithm is the simplest algorithm used to implement the modular multiplication [2]. The algorithm can be obtained from the expansion,

```
$P = 2 ( . . . 2 ( 2 ( 0 + A*B_k ) + A*B_{k-1} ) + . . . ) + A*B_1$, as :
Input: A, B
$R \leftarrow 0$
for {i = 0 to k-1} {
        $R \leftarrow 2R + A*B_{k-1-i}$
        $R' \leftarrow R-n$
        if {[R'] >= 0} {$R \leftarrow R'$}
                $R' \leftarrow R-n$
        If {[R']  >= 0} {$R \leftarrow R'$} }.
```

The hardware implementation of the datapath core is shown as in the Fig. 1. Signed subtractors have been used. The word-length of the subtractors and adders used is one and two bits more respectively.

3.2. Modular Exponentiation

The binary method for computing M^e (mod n) has been implemented using Left-to-Right (LR) algorithm. [2]

```
Input: M; e; n
if {$e_{h-1} = 1$} {$C \leftarrow M$} else {$C \leftarrow 1$}
for {$i = h-2$ to 0} {
    $C \leftarrow C*C$ (mod n)
    if {$e_i = 1$} {$C \leftarrow C*M$ (mod n)} }
```

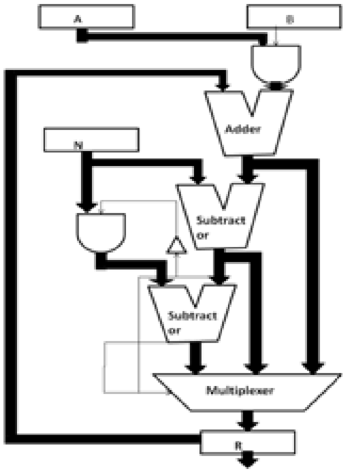


Figure 1. Architecture of RSA Datapath

The above algorithm is specific to the design of control unit for the RSA module. For the purpose of hardware optimization, it has been assumed that the MSB of exponent bit-word is always 1 i.e. the exponent always starts with the MSB.

The datapath core of RSA, as depicted in Fig. 1, is combined with some additional digital design blocks for complete RSA module. The state diagram for the same is given in Fig. 2. The states s0, s1, s2 are used for initialization and directing the primary input into the registers.

The states s4, s5 perform the binary multiplication; s5a checks the LSB of the exponent bit and if the LSB is HIGH it directs controller to another binary multiplication with changed inputs. The second binary multiplication is performed in state s9. If the LSB was LOW, the controller loops back to state s3. The state machine essentially performs binary modular multiplication. When the signal for completion of exponentiation is received, the state s11 is jumped to.

3.3. Miller-Rabin Primality test

Miller-Rabin Primality test is the most widely used primality testing algorithm [3][4]. The design for Miller-Rabin algorithm, shown in Fig. 3, is built around the RSA module described above with some additional control signals. The same RSA module has been used for exponentiation and squaring purposes.

This test provides advantages over other primality tests given by the Fermat and Euler [5]. The algorithm is reproduced below from [4][5] in an implementation friendly, Register Transfer Language (RTL), format.

```
Input: K, N
Output: P_Cb
For {i = 0 to K-1} {
        $D \leftarrow N-1$
        $S \leftarrow 0$
        While {$[D_0] = 0$} {
```

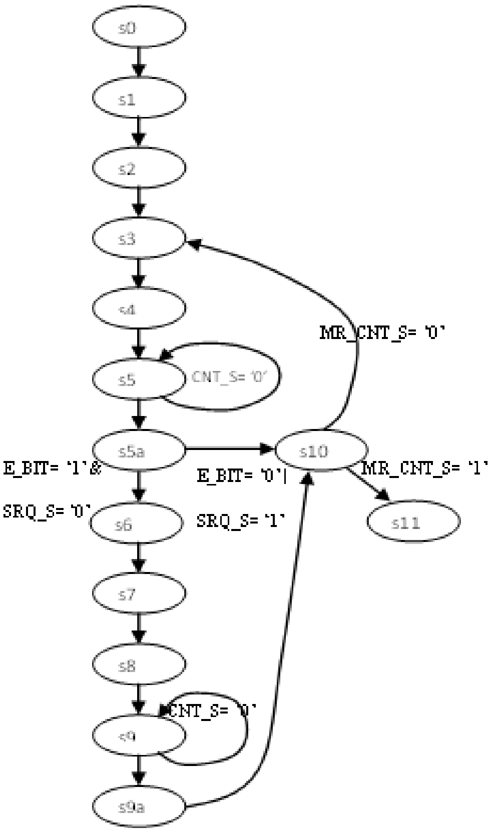


Figure 2. State Diagram for Modular Exponentiation

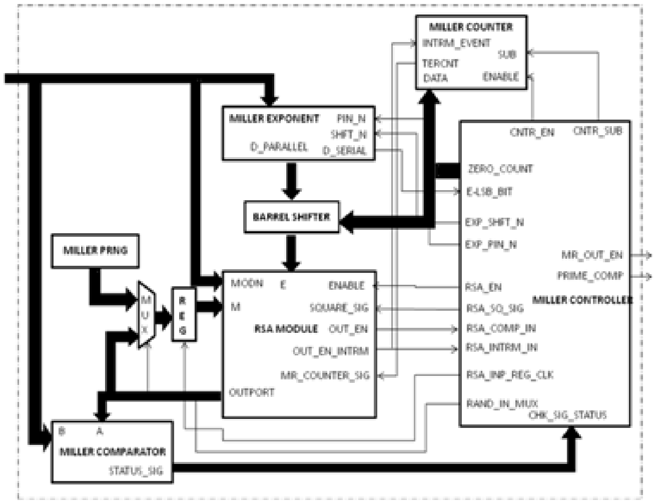


Figure 3. Architecture for Miller-Rabin Test Algorithm

```
D ← shr (D, 1)
S ← S + 1 }
        A ← RB (Random Base) {  RB ϵ [2, N-2]}
        $X ← A^D$ mod (N)
        if {[X] = 1 || [X] = N-1} {continue}
        for {r = 1 to S - 1} {
                $X ← X^2$ mod (N)
                if {[X] = 1} {P_Cb ← 0}
                if {[X] = N-1} {continue} }
        P_Cb ← 0 }
P_Cb ← 1
```

K is selected as per target accuracy and is sufficed at 7 for 512 bit primes and at 4 for 1024 bit primes [6].

The Miller exponent block, which is a modification over PI-P/SO shift register is used to calculate the ‘S' and 'D' values in the algorithm. The Miller controller detects the zeros in the exponent using shifting. A PRNG has been used to feed the random seed value to the RSA module for random base number. The counter counts a RSA intermediate event as clock. Miller controller serves as the master control unit of the system. The signal from the Miller controller further controls the events/states controlled by a separate RSA module controller which acts as a slave control unit.

The state diagram for Miller-Rabin primality test is given in Fig. 4. States s0, s1, s2 are used for initialization purposes. State s0 enables the exponent register to take input exponent, N, which is the number to be tested for primality. State s1 and s2 are used to count the number of trailing zeros in the exponent. It is to be ascertained that the exponent bit-string must begin with the MSB.

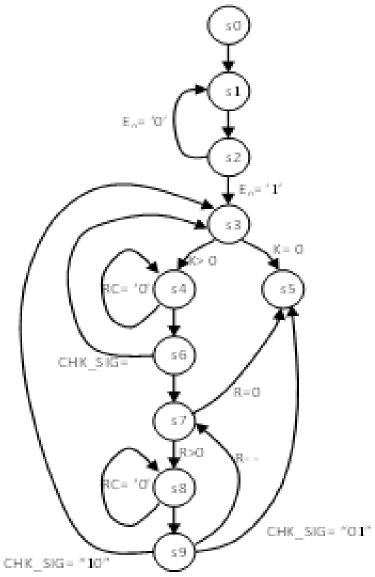


Figure 4. State diagram for Miller-Rabin Primality test

After all the trailing zeros have been counted, state s3 takes a random number from instantiated PRNG and while the number of iterations, K, for which the Miller-Rabin test is to be run is not equal to zero, it calls the state s4, which performs exponentiation.

When the exponentiation is complete state s6 checks the status in the miller comparator. If the status signal from miller comparator is "10" or "01", the controller goes back to state s3. Status "10" denotes that the result from the exponentiation is equal to N-1 and status "01" denotes the result to be unity.

For other status signals, the state s6 jumps to s7 which send a square signal to RSA module and performs the squaring operation in state s8. State s9 again checks the status and jumps of the consequent state.

3.4. Extended Binary GCD Algorithm

The binary GCD algorithm, also known as Stein's algorithm, computes the GCD of non-negative numbers using shifts, subtraction and comparisons rather than division used in Extended Euclidean algorithm. The binary GCD algorithm given in [7] can be implemented as shown in Fig. 5. The extended version of the same algorithm for calculating modular inverse has been presented below, for implementation, in RTL as

```
Inputs: A, B
Outputs: GCD, INV_OUT
Initialize: U ← 1; V ← 0; S ← 0; T ← 1; P ← A;
Q ← B
While {[B] ~= 0} {
        If {[B] = [A]} {
                        GCD ← shl (A,[R])
                INV_OUT ← S }
        Else if {[B] < [A]} {
                A ↔ B
                U ↔ S
                V ↔ T }
        Else if {[$A_0$] = 0 & [$B_0$] = 0} {
                A ← shr (A, 1)
                A ← shr (B, 1)
                R ← R + 1 }
        Else if {[$A_0$] = 0 & [$B_0$] = 1} {
                A ← shr (A, 1)
                If {[$U_0$] = 0 & [$V_0$] = 0} {
                        U ← shr (U, 1)
                        V ← shr (V, 1) }
                Else {
                        U ← shr (U + Q)
                        V ← shr (V – P) } }
         Else if {[$A_0$] = 1 & [$B_0$] = 0} {
                B ← shr (B, 1)
                If {[$S_0$] = 0 & [$T_0$] = 0} {
                        S ← shr (S, 1)
                        T ← shr (T, 1) }
                Else {
                        S ← shr (S + Q)
                        T ← shr (T – P) } } }
GCD ← shl (A, [R])
INV_OUT ← S
```

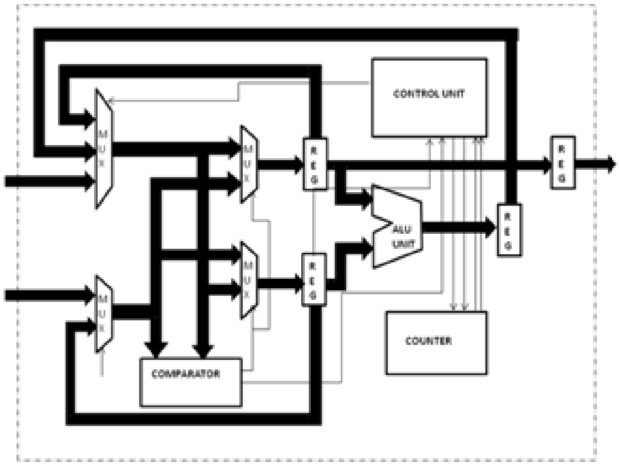


Figure 5. Architecture for BCD Algorithm

The above extended algorithm can be implemented by augmenting the architecture given in Fig. 5 with addition of few multiplexers, registers, subtraction units and control signals, as in Fig. 6.

The state diagram for Extended Binary Greatest Common Divisor (EBGCD) is given in Fig. 7. State s0 is the initialization state in which the inputs A & B are read in the various registers. In

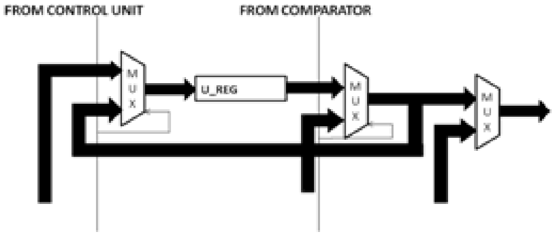


Figure 6. Additional structures required for Extended Binary GCD algorithm

state s1, the values and LSBs of both the inputs are compared. When LSBs of both A and B are LOW, the state s1 jumps to s3. The registers of both the inputs are right shifted and a counter is incremented.

When LSB of only either of the input is LOW, the state s4 or s5 are traversed to. The states s4, s4a, s4b, s4c and s5, s5a, s5b, s5c are used to perform the required computations. The states s6 through s6d operate when LSBs of both the inputs are HIGH. When both the inputs are equal, the state s1 jumps to s2 or s2b depending on whether the count for bit-shifts is zero or not. The state s2a and s2 are used to left-shift the output required number of times.

When value of B is less than A, the signal from the comparator to various MUXs goes HIGH and the interchange between various register is performed within that clock cycle.

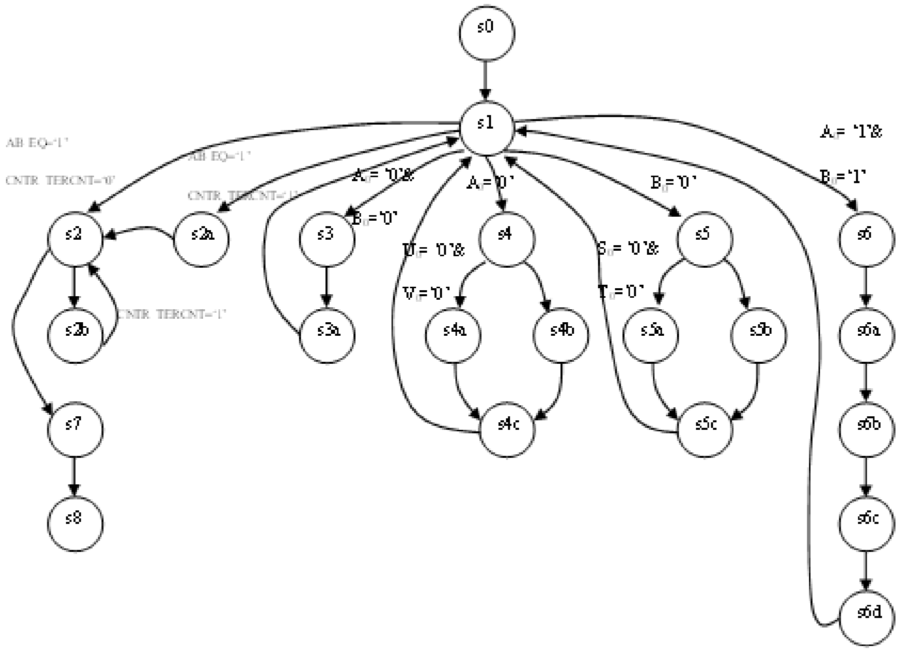


Figure 7. State diagram of Extended Binary GCD Algorithm

The Fig. 8 gives the complete architecture of the Extended Binary GCD algorithm. The signals from the comparator and EBGCD controller are used to control the data flow inside the register loops.

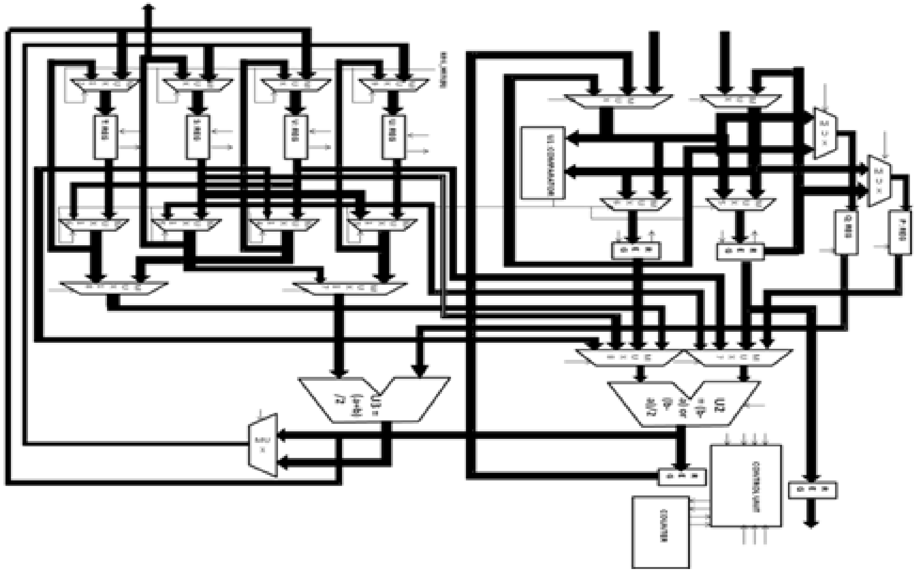


Figure 8. Detailed Architecture of Extended Binary GCD Algorithm

4. TOP-LEVEL DESIGN

After the individual design is completed for various modules, these are integrated in top-level design of RSA cryptosystem.

The cryptosystem can be run in either of the two modes:
(i) RSA encryption/ decryption (RSA mode) and,
(ii) Key-Pair Generation (GKP mode).

The design of the complete cryptosystem as implemented is shown in Fig. 9. The modes are controlled by GKP_RSAb control input. The system has an EXPONENT_BIT_CNTR counter which counts the intermediate RSA event and sends the signal for RSA completion. The input to the counter is number of bits of exponent bit-word that are to be used for exponentiation. The number for primality test may be supplied from memory or True-RNG as input.

During RSA computation, the controller after enabling the RSA module and directing the input MUXs to feed from Primary inputs waits for a signal from RSA module for completion. A signal from the exponent bit counter is sent to RSA module to indicate last bit the exponentiation.

During generation, the top system controller runs the Miller-Rabin controller twice to obtain two primes. In case the test fails and the random number is composite, the system keeps on taking the random numbers as input till both the prime numbers are determined. The product of primes and their Euler totient function are computed in two cycles using single combinational multiplier. The values computed are fed in to the EBGCD module the output of which is compared to the unity. If the output is not unity, another random number is taken as input. If the result is unity, the random number taken as input serves as the public key and the modular

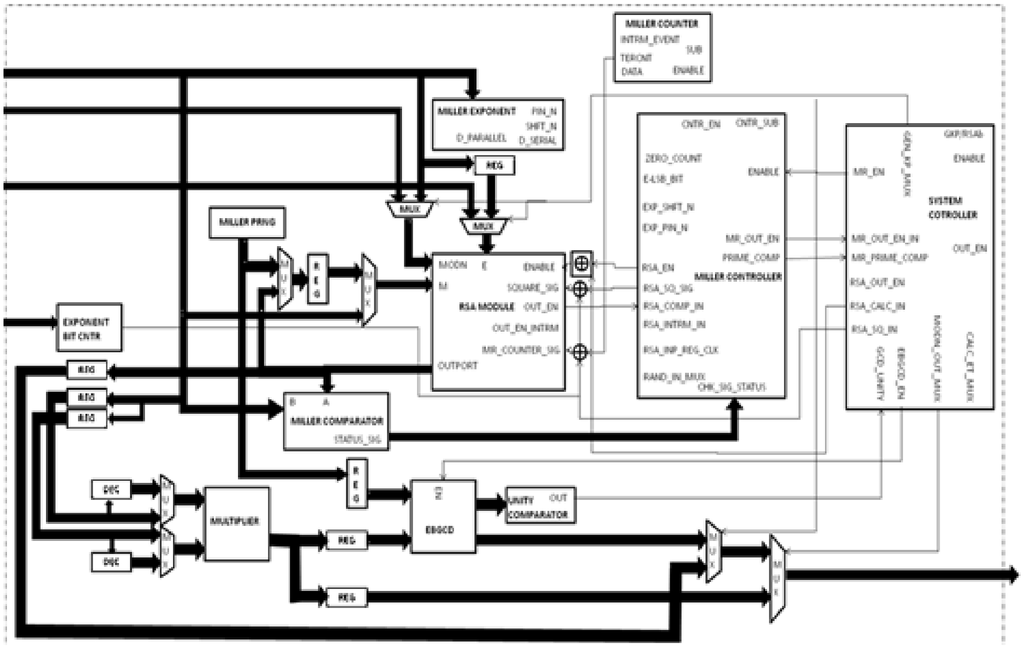


Figure 9. Top-level Architecture of RSA Cryptosystem

inverse output from the EBGCD module serves as the private key with modulus being the product of the primes.

The Miller PRNG has been used to generate a public key exponent; however, desired key may be provided externally with use of an additional multiplexer. The unity comparator block is implemented by a using a series of the OR gates.

5. Power Analysis Resistance

Power analysis attacks exploit the fact that the instantaneous power consumption of a cryptographic device depends on the data it processes and the operations it performs.

Simple power analysis (SPA) involves directly interpreting power consumption measurements collected during cryptographic operation. Differential power analysis (DPA) attacks, which require large number of power traces for analysis, are used due to the fact that these do not require detailed knowledge about the attacked device.

In CMOS technology, it is a fact that transitions are affiliated and determined by statistics of gate inputs and previous outputs, to the differing way energy is consumed between a 0→VDD and VDD→0 transitions.

To counter DPA, the device needs to be built in such a way that every operation requires approximately the same amount of energy, or it can be built in such a way that the power consumption is more or less random. To the effect of first technique a custom EDA flow was developed for transforming the synthesized design into a design compliant to Differential Power Balancing DPA resistant technique called Delay Insensitive Minterm Synthesis-3 (DIMS-3) [8]. Fig. 10 shows the typical transformation methodology used for improving the DPA resistance of the RSA datapath.

6. IMPLEMENTATION

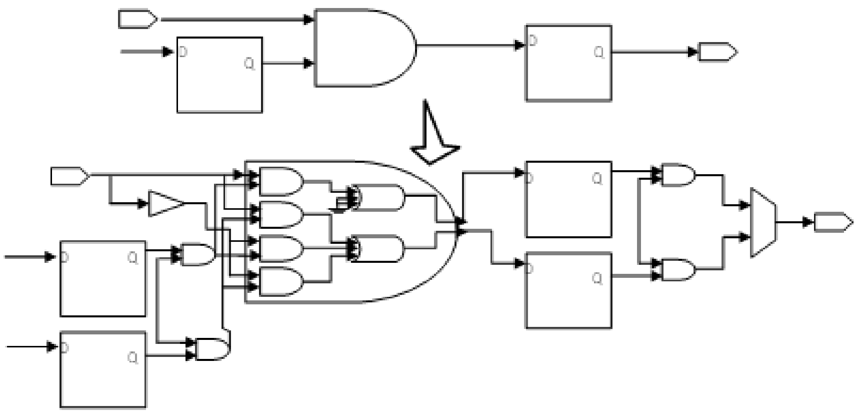


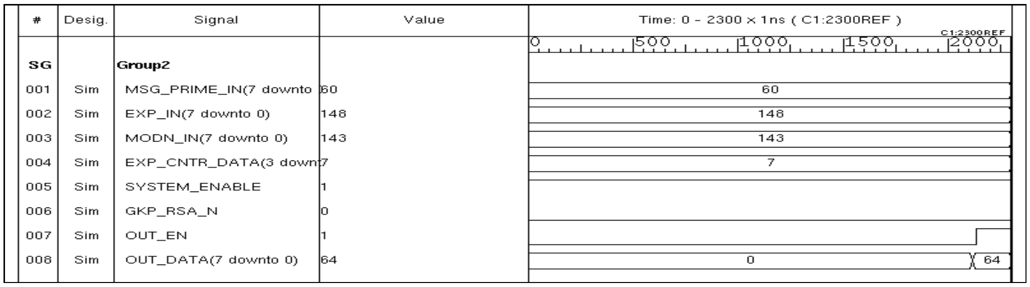
Figure 10. Delay Insensitive Minterm Synthesis-3 compliant transformation

This work describes the architecture of RSA cryptosystem built with the individual modules in the beginning to the top-level system in the end. The code of the described architecture was written in VHDL. The code for 8-bit system was synthesized and simulated using Tower 180nm digital library in Synopsys tools.

6.1. Simulation Results

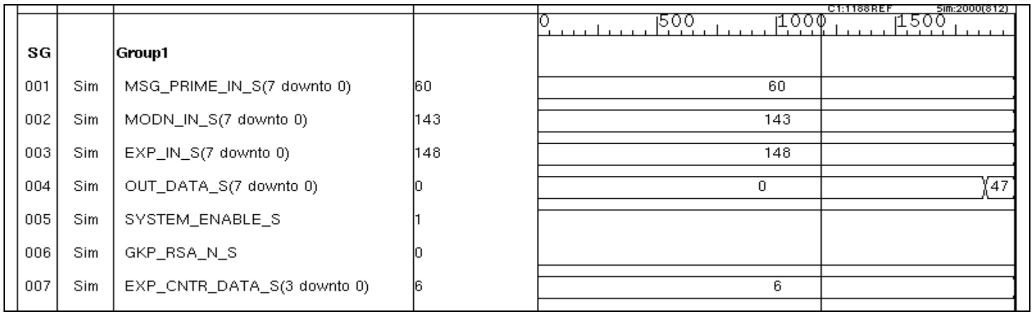
Fig. 11 and Fig. 12 show the simulation result of the above said architecture for RSA encryption/decryption. Though both of figures use the same input bit-strings, their EXP_CNTR_DATA_S input to EXPONENT_BIT_CNTR is different. Thus, in Fig. 11, effective exponent is 74("1001010") and in Fig. 12 effective exponent is 37("100101").
Fig. 13 shows the output sequencing of private key and modulus, when the system is used for key pair generation with primes 11 and 13.

Fig. 14 and Fig. 15 show the power signatures for a computation of Differential Power Balancing DIMS-3 compliant RSA datapath transformed using custom EDA flow at positive and negative clock edges respectively.



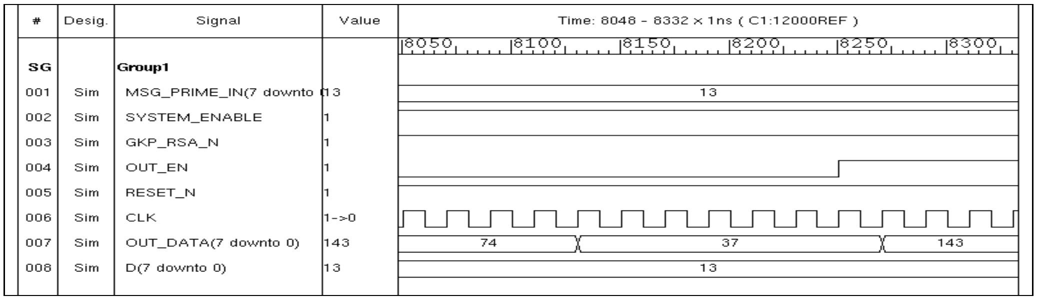
#	Desig.	Signal	Value
SG		Group2	
001	Sim	MSG_PRIME_IN(7 downto	60
002	Sim	EXP_IN(7 downto 0)	148
003	Sim	MODN_IN(7 downto 0)	143
004	Sim	EXP_CNTR_DATA(3 down	7
005	Sim	SYSTEM_ENABLE	1
006	Sim	GKP_RSA_N	0
007	Sim	OUT_EN	1
008	Sim	OUT_DATA(7 downto 0)	64

Figure 11. Simulation of RSA Cryptosystem for RSA Encryption/Decryption with Exponent bits count = 7



SG		Group1	
001	Sim	MSG_PRIME_IN_S(7 downto 0)	60
002	Sim	MODN_IN_S(7 downto 0)	143
003	Sim	EXP_IN_S(7 downto 0)	148
004	Sim	OUT_DATA_S(7 downto 0)	0
005	Sim	SYSTEM_ENABLE_S	1
006	Sim	GKP_RSA_N_S	0
007	Sim	EXP_CNTR_DATA_S(3 downto 0)	6

Figure 12. Simulation of RSA Cryptosystem for RSA Encryption/Decryption with Exponent bits count = 6



#	Desig.	Signal	Value
SG		Group1	
001	Sim	MSG_PRIME_IN(7 downto 0	13
002	Sim	SYSTEM_ENABLE	1
003	Sim	GKP_RSA_N	1
004	Sim	OUT_EN	1
005	Sim	RESET_N	1
006	Sim	CLK	1->0
007	Sim	OUT_DATA(7 downto 0)	143
008	Sim	D(7 downto 0)	13

Figure 13. Output sequence of private key and modulus during Key-Pair generation

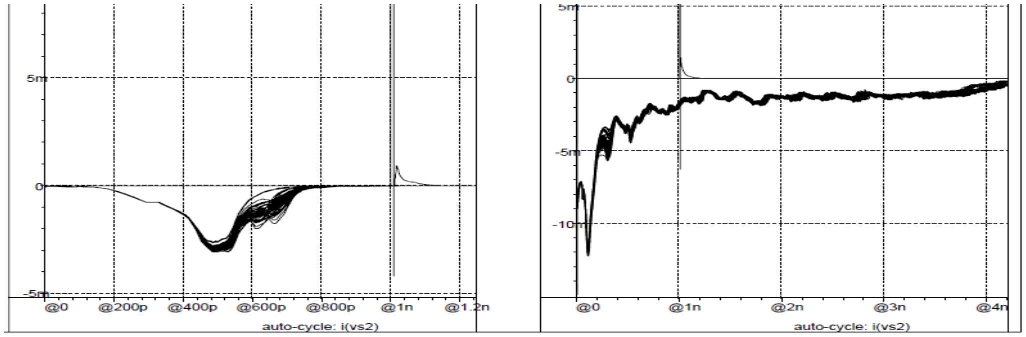


Figure 14. Power signature comparison between pre-transformed (left) and post-transformed (right) RSA datapath for various input

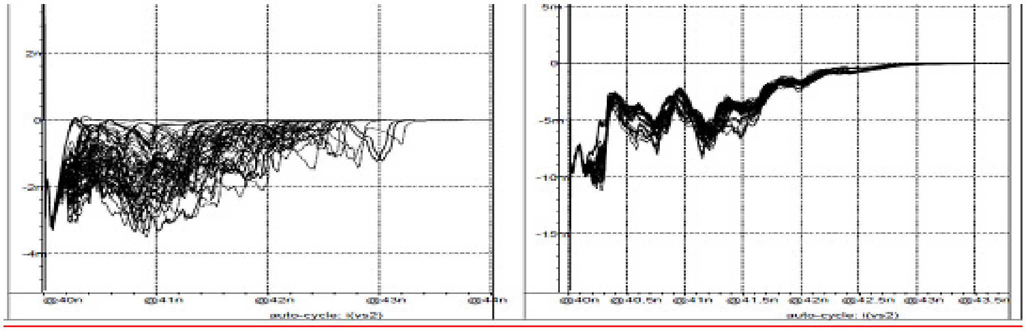


Figure 15. Power signature comparison between pre-transformed (left) and post-transformed (right) RSA datapath for various input

6.2. Implementation Results

Table I, II & III present the implementation results of synthesis of the RSA cryptosystem architecture in the 180nm Static digital CMOS library. Table I gives the count of the combinational and non-combinational cells implemented in the system. Table II enlists the area requirements of various design units the system. Table III gives the timing requirements of the core RSA module. E0 and E1 represent the number of 0's and 1's in exponent bit-word and N is the key length of the RSA. Table IV compares the area and cells required for optimized design to that for DIMS-3 compliant DPA resistant RSA datapath. Further, this work presents the results of the RSA datapath transformed into Differential Power Balanced DIMS-3 DPA resistance compliant design. The results of both the pre-transformed and post-transformed designs are presented for comparison.

Table I. 8-bit RSA cryptosystem cell count

CELL TYPE	CELL COUNT
combinational cells	1191
non-combinational	316

Table II. Area report of modules for 8-bit RSA

DESIGN UNIT	AREA	AREA %
Rsa_System	4113	100.0
Sys_Controller	250	6.1
Sys_Datapath	3863	93.9
Sys_Datapath/Comparator_Unit	22.25	0.5
Sys_Datapath/Controller_Unit	616	15
Sys_Datapath/Counter_Unit	47	1.1
Sys_Datapath/Enc_Data_Reg	56	1.4
Sys_Datapath/Exponent_Unit	86.5	2.1
Sys_Datapath/Exp_Cntr_Unit	56	1.4
Sys_Datapath/Gcd_Inv_Unit	1488.25	36.2
Sys_Datapath/Multiplier_Unit	68	1.7
Sys_Datapath/Prng_Unit	106	2.6
Sys_Datapath/Rsa_Unit	856.5	20.8
Sys_Datapath/Unity_Unit	1.75	0.0

Table III. Timing requirements

MODULE	CLKs
RSA Module	$3+E_0(4+N)+E_1(8+2N)$

Table IV. Area Reports for pre-transformed and post-transformed RSA module designs

******************************* Report : area Design : RSA_DATAPATH (PRE-TRANSFORM) ******************************* Number of ports: 37 Number of nets: 191 Number of cells: 142 Number of combinational cells: 118 Number of sequential cells: 24 Number of macros: 0 Number of buf/inv: 28 Number of references: 14 Combinational area: 204.500000 Buf/Inv area: 14.000000 Non-combinational area: 126.000000 Net Interconnect area: 69.138399 Total cell area: 330.500000 Total area: 399.638399 ******************************	********************************* Report : area Design : RSA_DATAPATH (POST-TRANSFORM) ********************************* Number of ports: 40 Number of nets: 1520 Number of cells: 1497 Number of combinational cells: 1449 Number of sequential cells: 48 Number of macros: 0 Number of buf/inv: 95 Number of references: 13 Combinational area: 2316.500000 Buf/Inv area: 57.500000 Non-combinational area: 306.000000 Net Interconnect area: 1374.814115 Total cell area: 2622.500000 Total area: 3997.314115 ***********************************

REFERENCES

[1] Rivest R. L., Shamir A., and Adleman L., "A method for obtaining digital signatures and public-key cryptosystems," Communications of the ACM, 1978.

[2] Koc C. K., RSA Hardware Implementation, RSA Laboratories, Technical Report.

[3] Miller, Gray L., "Riemann's Hypothesis and tools for Primality", Journal of Computer and System Sciences, 300-317, 1976.

[4] Rabin, Micheal O., "Probabilistic algorithm for testing primality", Journal of Number Theory, 128-138, 1980.

[5] Hoffoss D., Notes on "The Rabin-Miller Primality Test", University of San Diego.

[6] Kleinberg B., Notes on "The Miller-Rabin Randomized Primality Test", Cornell University.

[7] Knuth D. E. , Seminumerical Algorithms, The Art of Computer Programming Vol-2, Addison-Wesley.

[8] Murphy J., "Standard Cell and Full Custom Power-balancing logic: ASIC implementation", Technical Report Series, New Castle University.

[9] Rahman M., Rokon I. R., Rahman M., "Efficient Hardware Implementation of RSA Cryptography", 3rd International Conference on Anti-Counterfeiting, Security, and Identification in Communications, 2009.

[10] Shams R., Khan F. H., Umair M., "Cryptosystem an Implementation of RSA using Verilog", International Journal of Computer Networks and Communications Security, Vol.-1, No. 3, August 2013, p102-109.

[11] Vishak M, Shankaraiah N., "Implementation of RSA key generation based on RNS using Verilog", International Journal of Communication Network Security, Vol.-1, Issue-4, 2012.

[12] Garg V., Arunachalam V., "Architectural Analysis of RSA cryptosystem on FPGA", International Journal of Computer Applications, Vol-26, No.-8, July 2011.

AUTHORS

Nehru Varun received B.E. (Hons.) from Panjab University in 2010 and joined Semi-Conductor Laboratory. Since joining he has been working in VLSI design division and has been involved in digital designs.

Jattana H.S. received his engineering education from BITS Pilani and joined SCL as ATE engineer. He worked on test programs development /characterization for pulse dialler/tone ringer, Audio codec, IIR filter, signal processor for sonar applications and many ASICs. For the last over ten years he has been involved in design of VLSI products, and have contributed in many design projects like range of transceivers (400Mbps to 1.2 Gbps), power management chips, converters (12-bit pipeline ADC, 12-bit current steering DAC, 16-bit sigma-delta), CMOS Imaging sensor, cold sparing pads, read-hard tolerant digital cells and memory cell, and many ASICs.

He has worked at Rockwell Semiconductor, Newport Beach, USA for characterization of R65 series of devices and at AMS Austria for porting of 2 um and 1.2 um processes and ATE testing/characterization of products fabricated in these processes.

INTENTIONAL BLANK

DYNAMIC SELECTION OF SYMMETRIC KEY CRYPTOGRAPHIC ALGORITHMS FOR SECURING DATA BASED ON VARIOUS PARAMETERS

Ranjeet Masram[1], Vivek Shahare[1], Jibi Abraham[1], Rajni Moona[1], Pradeep Sinha[2], Gaur Sunder[2], Prashant Bendale[2] and Sayali Pophalkar[2]

[1]Department of Computer Engineering and Information Technology,
COEP, India
masram.ranjeet@gmail.com, vivek.shahare27@gmail.com,
ja.comp@coep.ac.in, rajnimoona@yahoo.com
[2]C-DAC Pune, India
psinha@cdac.in, gaurs@cdac.in, prashantb@cdac.in,
psayali@cdac.in

ABSTRACT

Most of the information is in the form of electronic data. A lot of electronic data exchanged takes place through computer applications. Therefore information exchange through these applications needs to be secure. Different cryptographic algorithms are usually used to address these security concerns. However, along with security there are other factors that need to be considered for practical implementation of different cryptographic algorithms like implementation cost and performance. This paper provides comparative analysis of time taken for encryption by seven symmetric key cryptographic algorithms (AES, DES, Triple DES, RC2, Skipjack, Blowfish and RC4) with variation of parameters like different data types, data density, data size and key sizes.

KEYWORDS

AES, DES, Triple DES, RC2, Skipjack, Blowfish, RC4, data type, data size, data density and encryption time

1. INTRODUCTION

Cryptography is a science of "secret messages" that is being used by man from thousands of years [9]. In cryptography original message is basically encoded in some non readable format. This process is called encryption. The only person who knows how to decode the message can get the original information. This process is called decryption. On the basis of key used, cipher algorithms are classified as asymmetric key algorithms, in which encryption and decryption is done by two different keys and symmetric key algorithms, where the same key is used for encryption and decryption [8]. On the basis of the input data, cipher algorithms are classified as block ciphers, in which the size of the block is of fixed size for encryption and stream ciphers in which a continuous stream is passed for encryption and decryption [9].

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 185–192, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4519

A data file formats represents the standard for encoding the information to be stored in computer file. There are file formats like textual, image, audio and video data file formats. Textual data formats are ANSII, UNICODE (16 & 32 bit little and big Endian and UTF-8). ANSII is encoding scheme for 128 characters basically made for English alphabets. It contains alphabets a-z and A-Z, numbers 0-9 and some special characters. In Unicode standard unique numbers are provided for every character independent of platform. Image file formats are JPEG, TIFF, BMP, GIF and PNG [10]. JPEG is image file format for digital images that uses lossy compression method. TIFF and BMP are image file format that are used to store images of raster graphics. GIF image is similar to image format of bitmap images. GIF uses LZW (Lempel-Ziv-Welch) technique of compression and for each image it can support up to 8 bits/pixel. PNG is alternative to GIF image file format and allows more compression than GIF. Audio file formats are WAV, AIFF, M4A, MP3 and WMA. WAV and AIFF are usually uncompressed audio file format. M4A (audio) uses Apple Lossless compression format but often it is compressed with Advance audio coding (lossy). MP3 and WMA are lossy compression audio formats. Video file formats are AVI, M4V, MPEG and WMV etc. AVI format contains the data (audio and video data) file container; which allows audio-with-video playback synchronously. M4V and MP4 are very similar format, but M4v can be protected by Digital Rights Management copy protection. MPEG contains compressed audio and visual digital data. WMV is compressed video format developed by Microsoft.

Density of data represents the amount of different information present in the data file [10]. File is said to be dense file if file size is less and content is more. For example if there are two file X and Y both containing 2000 words and having sizes 50kb and 200kb respectively, then file X is denser. The more the information, the dense is the data and lesser the information, sparse is the data. Sparse file is a file that contains most of the empty spaces and attempts to use the computer space more effectively.

Data size is space occupied by a file on a disk. Audio, video takes more space on disk than textual files as they contain multimedia information. Key size in cryptography represents the size of key file in bits. For example AES is having key sizes 128, 192 and 256 bits.

The main objective of this paper is to analyze time taken for encryption by various cryptographic algorithms for parameters like data type, data size, data density and key size.

2. CRYPTOGRAPHIC ALGORITHMS

This section provides information about the various symmetric key cryptographic algorithms to be analyzed for performance evaluation, to select the best algorithm with appropriate parameter suitable to provide security for data. The various features of the cryptographic algorithm are listed in Table 1.

Table 1. Cryptographic Algorithms Information.

Algorithm Name	Structure	Cipher Type	Rounds	Key Size(In bits)
AES	Substitution-permutation network	Block	10, 12, 14	128, 192, 256
DES	Balanced Feistel network	Block	16	56
Triple DES	Feistel network	Block	48	112, 168

RC2	Source-heavy Feistel network	Block	18	40 to 1024
Blowfish	Feistel network	Block	16	32 to 448
Skipjack	Unbalanced Feistel network	Block	32	80
RC4	----	Stream	256	40 to 2048

3. Related Work

This section provides the information and results which are obtained from the numerous sources. Cryptographic algorithms have been compared with each other for performance evaluation on basis of throughput, CPU Memory utilization, energy consumption, attacks, Encryption time, Decryption time etc.

In [3] the author compared AES and RC4 algorithm and the performance metrics were encryption throughput, CPU work load, memory utilization, and key size variation and encryption and decryption time. Results show that the RC4 is fast and energy saving for encryption and decryption. RC4 proved to be better than AES for larger size data. In [2] author compared AES and DES algorithms on image file, MATLAB software platform was used for implementation of these two cipher algorithms. AES took less encryption and decryption time than DES. In [4] the author compared cipher algorithms (AES, DES, Blowfish) for different cipher block modes (ECB, CBC, CFB, OFB) on different file sizes varying from 3kb to 203kb. Blowfish algorithm yield better performance for all block cipher modes that were tested and OFB block mode gives better performance than other block modes. In [7] the author talks about comparison between three algorithms (DES, Triple DES, Blowfish) on processing time. They found, that the key generation time for all these three algorithms is almost same but there is a difference in time taken by CPU for encryption. On SunOS platform Blowfish seem to be fastest, followed by DES and Triple DES respectively. They analyzed CPU execution time for generating the secret key, encryption and decryption time on 10MB file. In [6] the author compared cipher algorithms (AES, DES, 3-DES and Blowfish) for varying file size and compared the encryption time on two different machines Pentium-4, 2.4 GHz and Pentium-II 266 MHz in EBC and CFB Mode. The author concluded that Blowfish is fastest followed by DES and Triple DES and CFB takes more time than ECB cipher block mode.

4. Proposed Work

From the related works, it is realized that none of the work did a very detailed analysis of the performance of various symmetric algorithms on various parameters on different type of files, especially the files which are used for medical health related data.

The main objective of this paper is to analyze the time taken for encryption by various cryptographic algorithms for parameters like data type, data size, data density and key size in order to select the most suitable cryptographic algorithm for encryption.

5. Experimental Setup and Testing

The execution results are taken on machine having Intel® Core™ i7-2600 (3.40 GHz) processor with Intel® Q65 Express 4 GB 1333 MHz DDR3 (RAM) and Ubuntu 12.04 LTS operating System. The java platform (openjdk1.6.0_14) is used for implementation. JCA (Java Cryptography Architecture) and JCE (Java Cryptography Extension) are used for cipher algorithm implementation. The JCA is a major platform that contains ”provider" architecture and the set of APIs for encryption (symmetric ciphers, asymmetric ciphers, block ciphers, stream ciphers), message digests (hash), digital signatures, certificates and certificate validation, key generation and secure random number generation. Here we have used sun and Bouncy Castel provider for implementing cryptographic algorithms.

The brief analysis of different symmetric key cryptographic algorithm for various parameters is as follows:

Case Study 1: Files with different Data types.

This case study has taken to check whether the encryption has dependency on type of data. Different data type files like audio, image, textual and video of nearly 50MB in size are chosen and encryption time of different cipher algorithms is calculated for these data types. For all executions of a specific cipher algorithm, varying parameter is data type and constant parameters are key size and block cipher mode. Key size and block mode are at kept at bare minimal parameters. The key size of AES, DES, 3-DES, RC2, Blowfish, Skipjack, and RC4 are kept at minimum values as 128, 56, 112, 40, 32, 80 and 40 bits respectively. Block cipher mode used is ECB with PKCS#5 padding scheme. Fig. 1 shows the execution time of the algorithms for different data type files.

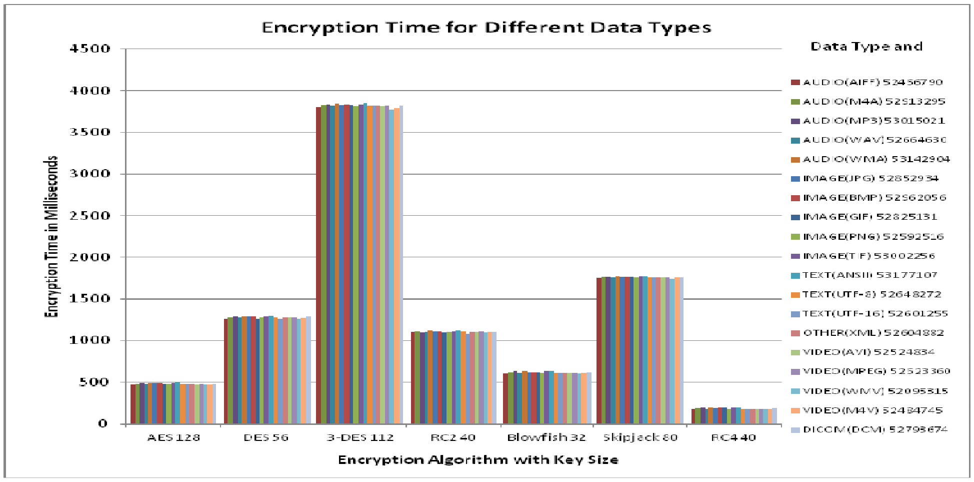


Figure 1. Encryption time Vs Cipher Algorithm for files of different data type

Observation: The result shows that the encryption time does not vary according to the type of the data. Encryption depends only on the number of bytes in the file and not on the type of file. AES works faster than other block ciphers. RC4 with key size 40 is fastest among the algorithms tested.

Case Study 2: Data files of same type with different sizes.

This case study is taken to ensure once again the observations obtained in case study 1. Case study 1 revealed that encryption time depends on number of bytes in the file. To ensure this another study is made in which different files of same types but different sizes are given for encryption and estimated the encryption time. For all executions key size and block mode are kept at bare minimal parameters. Table 2 gives the details about the files used for all executions and Fig. 2 and 3 show the execution results.

Table 2. Execution Parameters for files of different size.

File Type	Varying Parameters (Data Size)	Constant Parameters
AIFF	10.7MB, 50MB, 100MB	Data Type, Key size
AVI	50MB, 100MB, 482MB	

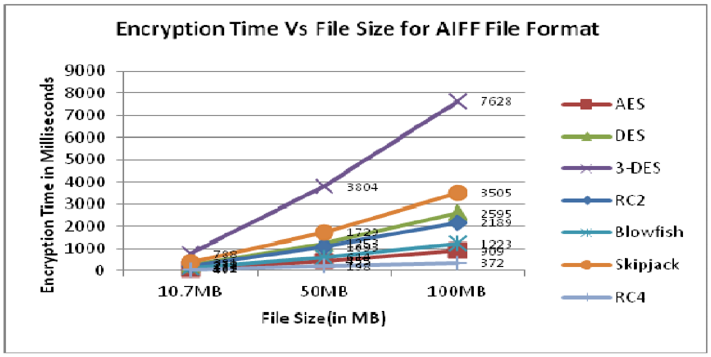


Figure 2. File size Vs Encryption time for AIFF file of different sizes.

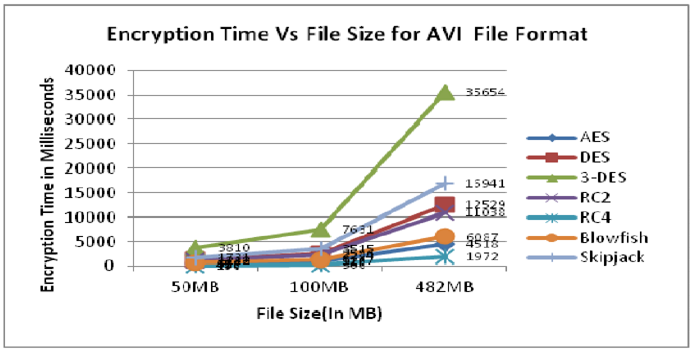


Figure 3. File size Vs Encryption time for AVI file of different sizes.

Table 3. Encryption time for files of different sizes

File Type	Size (In MB)	Encryption Time in Millisecond						
		AES	DES	3-DES	RC2	Blowfish	Skipjack	RC4
		128	56	112	40	32	80	40
AIFF	10.7	101	272	788	238	133	381	40
	50	455	1253	3804	1095	614	1729	198
	100	909	2595	7628	2189	1223	3505	372
AVI	50	456	1268	3810	1112	629	1731	196
	100	918	2586	7631	2224	1267	3515	360
	482	4518	12529	35654	11038	6087	16941	1972

Observation: From the results in Table 3 and Fig. 2 and 3 we can find that the result for different size of data varies proportional to the size of data file. Encryption time increases as file size increases in multiples of data size. For each encryption algorithm same parameters are used for files of different sizes.

Case Study 3: File with different data densities.

This case study is taken to check whether the encryption depends on density of data or not. Encryption rate is evaluated for the two different data density file; a sparse file of 69MB and a dense file of 58.5MB. For a cipher algorithm, key size and block mode are kept at bare minimal parameters. The results of execution are shown in Table 4.

Table 4. Execution rate for sparse and dense data file

Algorithm Name	Sparse (72000118 Bytes) AIFF file		Dense (61392454 Bytes) AIFF file	
	Encrypt Time(ms)	Encryption Rate(MB/s)	Encrypt Time(ms)	Encryption Rate(MB/s)
AES 128	634	108.28	540	108.40
DES 56	1801	38.11	1537	38.08
3-DES 112	5076	13.52	4365	13.41
RC2 128	1520	45.16	1285	45.55
Blowfish 128	854	80.38	723	80.96
Skipjack 128	2386	28.77	2042	28.66
RC4 128	253	271.35	216	271.01

Observation: Encryption rate for sparse and dense file has been calculated. The Table 4 shows that the encryption time is not affected by density of data in a file. The variation in time with respect to different algorithms follows the same pattern for both sparse and dense files. The encryption rate for a particular cipher algorithm remains the same, even if the file is sparse or dense. It depends on only the number of bytes in the file.

Case Study 4: Encryption Algorithms with different key sizes

This case study is to analyze the effect of changing the size of encryption key on encryption time. BMP file of 50.5MB is taken and different cipher algorithms are executed for different size of keys supported by them in ECB mode with PKCS#5 padding scheme. The various key sizes mentioned in Table 1 are used during experimentation. Fig. 4 shows the result of execution for key size variation.

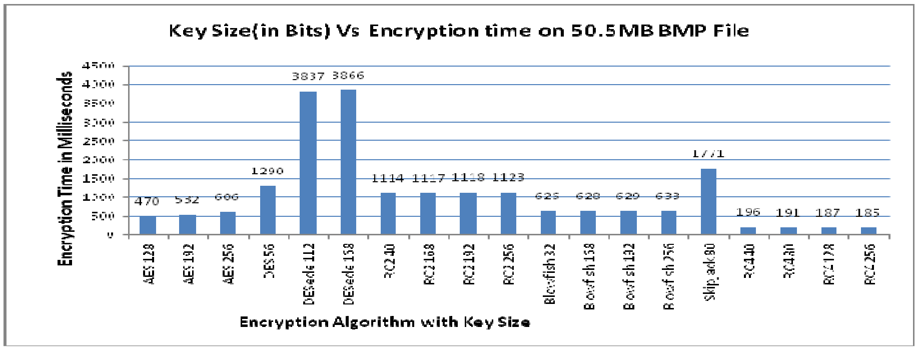


Figure 4. Variation of key sizes for different cipher Algorithms

Obsevation: The execution results show that for all ciphers algorithms, the encryption time varies with the change in the size of the of the key. Encryption time increases with increase in key size for block ciphers. The variation in time is very small. AES dominates in the block cipher. RC4 is fastest among all algorithms tested.

6. CONCLUSION

In this paper different symmetric key algorithm have been analyzed for various parameters like different data type, data size, data density, key size, cipher block modes and tested how the encryption time varies for different algorithms. From the execution results it is concluded that encryption time is independent of data type and date density. The research shown that, encryption time only depends upon the number of bytes of the file. It also reveled that encryption time varies proportionally according to the size of data. For all block cipher algorithms that are analyzed, with increase in key size, encryption time also increases, but reduces with increase in key size for RC4. AES is fastest block cipher, but RC4 appears to be fastest among all analyzed ciphers.

ACKNOWLEDGEMENT

We would like to thanks C-DAC, Pune for all their support during the project.

REFERENCES

[1] AL.Jeeva1, Dr.V.Palanisamy and K.Kanagaram, “Comparative Analysis of Performance Efficiency and Security Measures of Some Encryption Algorithms”, International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 3, May-Jun 2012, pp.3033-3037.

[2] S. Soni, H. Agrawal, M. Sharma, “Analysis and comparison between AES and DES Cryptographic Algorithm”, International Journal of Engineering and Innovative Technology, Vol 2, Issue 6, December 2012, pp.362-365.

[3] Nidhi Singhal and J.P.S.Raina, “Comparative Analysis of AES and RC4 Algorithms for Better Utilization”, International Journal of Computer Trends and Technology, Vol 2, Issue 6, July-Aug 2011, pp.177-181.

[4] Jawahar Thakur and Nagesh Kumar, “DES, AES and Blowfish: Symmetric Key Cryptography Algorithms Simulation Based Performance Analysis”, International Journal of Emerging Technology and Advanced Engineering, Vol 1, Issue 2, December 2011, pp.6-12.

[5] Allam Mousa and Ahmad Hamad, “Evaluation of the RC4 Algorithm for Data Encryption”, International Journal of Computer Science & Applications, Vol 3,Issue 2, June 2006, pp.44-56.

[6] Aamer Nadeem, Dr M. Younus Javed, “A Performance Comparison of Data Encryption Algorithms”, First International Conference on IEEE Information and Communication Technologies (ICICT), Vol 1, Issue 6, 27-28 Aug. 2005, pp 84-89.

[7] Kofahi, N.A, Turki Al-Somani, Khalid Al-Zamil, “Performance evaluation of three Encryption/Decryption Algorithms”, IEEE 46th Midwest Symposium on Circuits and Systems, Vol 2, Issue 1, 30-30 Dec. 2003, pp. 790-793.

[8] Jonathan Knudsen, Java Cryptography, 2nd Edition, O’Reilly, 2011.

[9] Behrouz A. Forouzan, Debdeep Mukhopadhyay, Cryptography and Network Security, 2nd Edition, Tata McGraw Hill, 2012. [10]John Miano, Compressed Image File Formats, 1st Edition, Addison Wesley Longman, Inc,1999.

AUTHORS

Ranjeet Masram is M tech Student in Computer Engineering from College of Engineering, Pune (India). He is appointed as JRF for Joint project on Medical data Security between C-DAC,Pune and College of Engineering, Pune for a period of one year.

Vivek Shahare received his Bachelor Degree in Computer Science and Engineering from Government College of Engineering, Amravati(India). He is appointed as JRF for Joint project on Medical data Security between C-DAC, Pune and College of Engineering, Pune for a period of one year.

Dr. Jibi Abraham is Professor at College of Engineering, Pune. She received her Doctor of Philosophy (PhD) in Computer Engineering from Visvesvaraya Technological University. She is the Principal Investigator from COEP for Joint project on Medical data Security between C-DAC, Pune and College of Engineering.

Dr. Rajni Moona was project engineer at IIT Kanpur. She was Visiting faculty at International Institute of Information Technology. She is the co-investigator for Joint project on Medical data Security between C-DAC, Pune and College of Engineering.

Dr.Pradeep K.Sinha is Senior Director, C-DAC,Pune. Dr. P.K. Sinha, Programme Coordinator, High Performance Computing and Communication (HPCC) Group, was included in the Sixteenth Edition of the "MARQUIS Who's who in the World," which is a prestigious international registry of outstanding men and women in a wide range of professions and geographical locations in the world. He is a visiting faculty at college of Engineering Pune. He is th first Indian to be conferred the Distinguished Engineer '09 honour.

Mr. Gaur Sunder is the Coordinator and Head of the Medical Informatics Group (MIG) at C-DAC, Pune. He is Computer Scientist working in HPC, Distributed Systems, Cloud Computing & Virtualization, Cluster & Grid Computing, Data Repository (Big-data), Imaging, Networking, Platform (Win/Lin) and Web Technologies, and allied areas.

Mr. Prashant Bendale is Senior Technical Officer at C-DAC, Pune. His research areas include medical informatics standards, distributed / cloud computing technologies. He led the software development kits activities for medical informatics standard like DICOM & HL7.

Mrs. Sayali Pophalkar is a project Engineer at C-DAC, Pune.

A MULTI-LAYER ARCHITECTURE FOR SPAM-DETECTION SYSTEM

Vivek Shandilya, Fahad Polash and Sajjan Shiva

Department of Computer Science, University of Memphis, Memphis, TN, USA
vmshndly,fpolash,sshiva@memphis.edu

ABSTRACT

As the email is becoming a prominent mode of communication so are the attempts to misuse it to take undue advantage of its low cost and high reachability. However, as email communication is very cheap, spammers are taking advantage of it for advertising their products, for committing cybercrimes. So, researchers are working hard to combat with the spammers. Many spam detections techniques and systems are built to fight spammers. But the spammers are continuously finding new ways to defeat the existing filters. This paper describes the existing spam filters techniques and proposes a multi-level architecture for spam email detection. We present the analysis of the architecture to prove the effectiveness of the architecture.

KEYWORDS

Email, Spam, Spam detection, Filters, Multi-Layer Architecture

1. INTRODUCTION

In general, unsolicited emails are regarded as spam email. But according to Mail Abuse Prevention System, L.C.C. [2], the three conditions to consider an email as spam are: 1) The recipient's personal identity and context are irrelevant because the message is equally applicable to many other potential recipients, 2) The recipient has not verifiably granted deliberate, explicit, and still revocable permission for it to be sent, and 3) The transmission and reception of the message appears to the recipient to give a disproportionate benefit to the sender. Spam affects the users in several ways. The user will lose productive time by looking into the spam emails. The user mailbox is overburdened by spam emails. Spam emails consume network bandwidth. Radicati Research Group Inc., a Palo Alto, CA, based research firm, estimates that spam costs businesses $20.5 billion annually in decreased productivity as well as in technical expenses. Nucleus Research estimates that the average loss per employee annually because of spam is approximately $1934[3]. The main success of the spammer is to sell a product advertised in the spam email. Though most of the users ignore the advertisement, even if some order the advertised product, it is profitable for the spammer, as it costs very less to send millions of spam. An internet connection and a single click is enough for the spammer to send a spam email to many email users. According to the industry figures, 1 out of the 12.500.000 spam messages that are sent, lead to a sale [4]. Spammers get a high percentage profit share for each of the sale generated.

The damage due to spam has met with many attempts to detect and stop them. Many commercial spam email filters are available in the market [5]. For example some of the client side filters are: ASB AntiSpam, Outlook Spam Filter, Spam Alarm, SpamButcher, Qurb Spam, Spam Arrest, Spam Bully, MailWasher Pro, McAfee SpamKiller, Feox for Outlook/OE, Edovia AntiSpam, SAproxy Pro, Dewqs' NMS for Outlook, AntiSpamWare and LashBack. Some of the server side

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 193–200, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4520

spam filters are: GFI Anti Spam Filter,M-Switch Anti-Spam, Astaro Security Gateway, Hexamail Guard, Symantec AntiSpam for SMTP, Accessio Server, SpamSentinel for Domino Server and Kaspersky Anti-Spam Enterprise Edition by Alligate. In spite of active countermeasures spamming is thriving. In November 2013, the percentage of spam email was 72.5% out of all email traffic [1]. This calls for continuous efforts to discourage the spammers.

In this paper, we propose a multi-level architecture which will combine the existing spam filters in different layers. This architecture could be used as a generic framework for spam email detection. Existing techniques employed at each layer are also described. Our main contribution can be summarized as,

1. A multi-layer architecture using the present day state of art spam detection technologies.
2. Analysis to prove the advantages of our novel method of having two thresholds over the traditional single threshold based classification, in improving the accuracy and reducing the false positives while keeping the computational load for filtering low when using each of the filtering features.

The rest of the paper is organized as follows: Section 2 describes related works regarding multi-level spam detection approaches, Section 3 describes the proposed architecture, Section 4 presents the performance evaluation measurements of spam detection and we conclude in Section 5.

2. RELATED WORKS

Many researchers have already given efforts to fight with the spammers. Some of the works are related to our proposed multi-level architecture for spam detection. However, our proposed model is different from these salient works. Jianying et al. [6] describe a multi layer framework for spam detection. They divide the spam detection techniques between server and client side deployments. Our proposed model does not differentiate between server and client. It can be equally applied to both server and client side anti-spam countermeasures. Rafiul et al. [7] proposes a multi-tier classification for phishing email. The classification result in the first tier is given to a second tier classifier. If the classification of the second tier and first tier matches, then the result is considered as the right output. But if the results differ, then a third tier classifier is used to classify the email. The output of the third classifier is considered the correct classification of the email. Thus, best of three classifiers are used to get the classification of an email. But in our proposed model, if any layer can classify an email with the confidence above the threshold, then the lower level is not invoked. And our proposed model can handle more than three levels of classifier to reduce the false positive rate as much as possible. In [8], Xiao et al. proposed a hierarchical framework for spam email detection. The first layer in their framework is a text classifier. But in our case, we have considered other behavioural features of spam email like blacklisting sender, sender reputation etc. Again, we have included negative selection based detection in the last layer which will be more effective against new spam emails. Zhe et al. [9] presents an approach targeting mainly at image spam email. The architecture presented by them is two layered. The first layer classifies non image spam and the second layer classifies image spam. The second layer involves multiple spam filters which will take longer time for training the detectors. Our proposed multi-level architecture is presented in Figure 1. The purpose of this model is to put the existing techniques in an organized way so that the detection of the spam would be faster and the false positive rate would be lower. If the detection of the spam could be possible in the upper layer, then the lower layer would not be invoked, and thus it will reduce the computational load. And in each layer, we can increase the threshold value so that the rate of the false positive would be minimum. As a result the overall performance of the spam filter detection process would be better.

This model comprises of the following layers: 1) Blacklist/Whitelist layer, 2) Content based filter, 3) Image based filter, 4) Negative Selection of Unknown Spam Email, and 5) Recipient Decision. The description of each layer is given below.

3. MULTI-LAYERED ARCHITECTURE

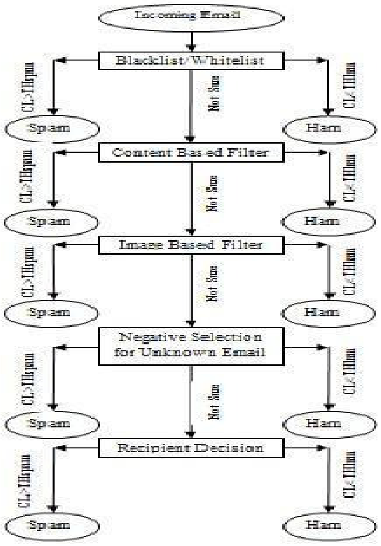


Figure 1. Multi-Layer spam detection Architecture

Our proposed multi-level architecture is presented in Figure 1. This model organizes the existing techniques for better detection of spam with lower false positive rate. If the detection of the spam could be possible in the upper layer, then the lower layer would not be invoked, and hence reduces computational expenditure. And in each layer, we can increase the threshold value so that the rate of the false positive would be minimum. As a result the overall performance of the spam filter detection process would be better. This model comprises of the following layers: 10 Black list/White list Layer, 20. Content based Filter, 3). Image Based Filter, 4). Negative Selection of Spam email and 5). Recipient decision. The description of each layer is given below.

3.1. Blacklist / Whitelist

Blacklist and Whitelist filters can classify the emails without reading the messages. Based on the senders reputations, blacklist and whitelist are prepared. Blacklisted senders' emails are classified as spam whereas whitelisted senders' emails are classified as ham emails. Any user can add email sender's email address(and in advanced cases IP addresses) in the blacklist or whitelist. The advantage of this approach is that the classification is very fast as it does not require to go through the messages. However, one disadvantage of this approach is that it requires the blacklist/whitelist to be updated regularly. Otherwise the false positive rate would be higher.

Duncan et al. [10] proposes a mechanism by which the blacklist will be updated dynamically. By checking the log files of a particular IP address and then other suspicious activities of the sender, the sender’s IP address is added in the blacklist. Anirudh et al. [11] proposes a technique with which the blacklist is updated based on how the sender is sending email, rather than not relying on IP address, the sending pattern is observed and the IP address of the sender is added in the blacklist.

3.2. Content Based Filter

Content based filter requires the whole message to be read before taking the decision. As a result it will take computationally more time than the Black list /White List layer. Machine learning and fingerprint based filters are popular among the content based filters.

Fingerprint filters generate unique fingerprints for known spam messages and store in a database. It compares the fingerprint of the incoming email to that of stored spam messages. If a match is found for the spam messages' fingerprint, the email is classified as a spam. However the matching should be above a certain threshold level. Damiani et al. [12] proposes a robust way to generate the fingerprint of an email.

There are several methodologies in machine learning techniques. Statistical and artificial immune systems are notable among them. Support Vector Machine (SVM) is a famous statistical tool. In SVM, each email is considered as an n-dimensional vector. Each dimension could be the frequency of a certain word. Harris et al. [13] has compared among different algorithms for statistical filtering. The results of their experiments proves that SVM is better than Ripper, Rocchio, and Boosting Decision Trees algorithm. Mehran et al. [14] presented bayesian classification for spam email where some domain specific features like the domain of the sender(.edu or .gov), the time of sending the email, whether the email contains attachment are taken into consideration. Their experiments show that bayesian classifier works better when the classification is done with additional domain specific features along with the contents of the message. Dat et al. [15] have proposed an approach to detect misspelled spam words in spam email. For example, the word 'viagra' is commonly used in spam emails. So, this word is a blacklisted word. To defeat the spam filter, spammer can use the word 'viaaagra'. The possibility theory will calculate all the possibilities of misspelling of spam keyword and thus can classify the email accurately. Clotilde et al. [16] presents a symbiotic filtering approach where trained filters are exchanged among the users. As it exchange filters, not emails, so the communication and computational cost is minimum in this approach while it achieves better performance.

Artificial Immune System uses machine learning methods inspired by the human immune systems for fighting the spam. Human immune system distinguishes between self and non-self, and artificial immune system distinguishes between a self of legitimate email and a non-self of spam email. The heart of artificial immune system is detectors which are randomly generated from a set of gene library. Oda et al. [17] proposes the approach of artificial immune system in spam detection successfully.

3.3. Image Based Filter

As text based filters can classify emails successfully, spammers are taking resort to image spam emails in order to defeat the existing text based filters. Initially, the optical character recognition(OCR) was being used to detect the text embedded in the image. However, spammers use randomization in creating image spam to defeat OCR. For example, spammers introduce additional dots, frames, bars in the image. They can change the font type of the text included in the image. Zhe et al. [9] proposes a technique which is effective in image spam detection. They have involved three different types of image filters: Color Histogram Filter, Haar Wavelet Filter and Orientation Histogram Feature. Each of the filters works better in different types of randomization detection. After combining the output from three different filters, decision is taken to classify the email. Their experiments have shown less than 0.001% false positive rate in image spam detection. Uemera et al. [18] proposes an image filtering technique based on Bayesian filter. The filter will consider the image file size, file name, compressibility technique and area of the image to classify the spam image email. Their experiment also exhibited low false positive rate.

3.4. Negative Selection of Unknown Spam Email

Negative selection is a part of Artificial Immune System. Dat et al. [19] proposed a novel technique for spam email detection based on negative selection. The difference between this filter and others is that negative selection does not require any prior knowledge of spam emails. It does not require any prior training. As a result, this filter can be used readily. And as such the unknown spam emails could be classified by the technique proposed by Dat.

3.5. Recipient Decision

If the above layers cannot classify the incoming email either as ham or spam with confidence level above the required threshold, then the email could be tagged with a probability number of being a ham or spam, but not classified and let into the inbox. The user can decide whether the email should be forwarded to spam or inbox folder. Based on the decision of the user, the filters of the upper layer can learn and use the knowledge for future classification of this sort of email. Elena et al. [20] proposed a spam filtering technique based on the reputation of the reporters. Whenever any of the users report any email as spam, the system maintains the trustworthiness of the reporter and use the feedback to classify emails.

4. PERFORMANCE ANALYSIS

In spam email detection, if a spam is detected correctly, it is called true positive (TP), if a legitimate email is classified correctly, it is called true negative (TN). Similarly, the misclassification of legitimate email into spam email is called false positive (FP) and the misclassification of spam email as a legitimate email is called false negative (FN). The goal of the spam detection is to classify as many as possible emails correctly and at the same time to reduce the false positive rate. Because if any legitimate email is classified as spam and the user overlooks that email, he/she might miss valuable information. As a result the cost of a false positive classification is very high. Description of various parameters [21] for spam detection are given below:

a) Recall = TP/ (TP + FN). It explains how good a test is at detecting the positives. i.e. predicting positive observations as positive. A high recall is desired for a good model. Recall is also known as sensitivity or TP Rate.

b) FP Rate = FP/ (FP+ TN). It explains how good a model is at detecting the negatives. A low value is desirable.

c) Precision = TP/ (TP + FP). It determines how many of the positively classified are relevant. It is the percentage of positive classifications being correct. A high precision is desirable.

d) Accuracy = TP + TN/ (TP+TN+FP+FN). It tells how well a binary classification test correctly i.e. what percentage of predictions that are correct. Accuracy alone is not a good indicator, as it does not tell how well the model is in detecting positives or negatives separately.

In our proposed model the classification of the emails would be done as follows:

Each filter calculates the correlation factor of the incoming mail based on the known characteristics of the spam based on the features of the filter. Then the correlation factor is compared to two threshold values, which are calibrated to decide if the email is a spam, a ham or not decidable with the information at hand. If the email is classified into the third category, it is sent to another more rigorous and computationally expensive filtering layer. This architecture leads us to have a classification which is having an acceptable accuracy and an acceptable false-positive rate, while considering each of the features when situation requires. As we show in the analysis below with the given technologies at hand this architecture provides an improvement over using the filters individually as some of the features considered by one filter will not be considered by the other filter.

Let A= Accuracy, T = Percentage of Correctly Classified emails, F= Percentage of Incorrectly classified emails = False positives + False Negatives. Then by definition we have, A = 1/(1+(F/T)). The effectiveness of the filtering depends on the nature of the incoming traffic and the filter's response to it. More precisely, when a filter can expect that, in the incoming set of emails, if there are spams, and those spams have a particular feature, then it can check for high correlation for that feature and classify successfully the spam emails and keep them out of inboxes of the end users. The only thing that the filters can control is the threshold of correlation factor to classify the incoming email as spam or ham. If the threshold for classification as spam is high then many spam end up in the inbox. If it is low, then many hams may be misclassified as spam, increasing the false positive. It is learnt from experience that false positive should be as less as possible even if that allows some spams to enter the inbox.

Thus let us have $\alpha_{1,}$ $\alpha_{2,}$ α_3 and α_4 as the percentage of incoming emails expected to be classified as spam emails and $\beta_{1,}$ $\beta_{2,}$ β_3 and β_4 be the expected to be classified as ham emails respectively at each layer. Let and γ_1, $\gamma_{2,}$ γ_3 and γ_4 be the percentage of false positives and δ_1 , $\delta_{2,}$ δ_3 and δ_4 be the false negatives of each layer correspondingly from top layer to the bottom layer in Figure 1. Then we get the total False Positive = $\alpha_1\gamma_1 + (1-(\alpha_1 + \beta_1))\, \alpha_2\, \gamma_2 + (1-(\alpha_1+ \alpha_2+ \beta_1+ \beta_2))\, \alpha_3\, \gamma_3 + (1-(\alpha_1+ \alpha_2+ \alpha_3 + \beta_1+ \beta_2 + \beta_3)\, \alpha_4\, \gamma_4$. Similarly we have the total False Negative = $\beta_1\, \delta_1 + (1-(\alpha_1 + \beta_1))\, \beta_2\, \delta_2 + (1-(\alpha_1+ \alpha_2+ \beta_1+ \beta_2))\, \beta_3\, \delta_3 + (1-(\alpha_1+ \alpha_2+ \alpha_3 + \beta_1+ \beta_2 + \beta_3)\, \beta_4\, \delta_4$. On simplification we can verify that the percentage of emails the effective false positives = $(\alpha_1\gamma_1 + \alpha_2\, \gamma_2 + \alpha_3\, \gamma_3 + \alpha_4\, \gamma_4)$ – (non-negative term) = Sum of False positives of individual filters – positive term. So, we know that collectively the false positive and by symmetry false negative are better off than if we had considered individual filters separately.

5. CONCLUSIONS

In this paper we have proposed a multi-layer architecture which provides a layered approach for spam detection process using the existing techniques. As the spammers are coming up with new ways to defeat the existing filters, continuous efforts are required to improve the filters in each layer. Detecting spam email closer to the source will avoid wasting bandwidth and traffic processing. With our analysis we show that our architecture yields more correct classifications for the same given thresholds of spam without adversely affecting false positives and the threshold of ham affecting the false negatives. Further research into the exact machine learning algorithms to detect spam incorporating this overall architecture for the layers would lead to better preparedness to fight the increasing spam traffic.

REFERENCES

[1] http://www.securelist.com/en/analysis/204792321/Spam_in_November_2013

[2] http://www.sans.org/reading-room/whitepapers/email/is-affect-us-deal-spam-1111

[3] http://www.spamlaws.com/spam-stats.html

[4] http://www.spamexperts.com/en/news/motivation-spammers

[5] http://www.spamhelp.org/software/

[6] Jianying Zhou, Wee-Yung Chin, Rodrigo Roman, and Javier Lopez,(2007) "An Effective Multi-Layered Defense Framework against Spam", Information Security Technical Report 01/2007.

[7] Rafiqul Islam,JemalAbawajy,"A multi-tier phishing detection and filtering approach (2013)",Journal of Network and Computer Applications,Volume 36, Issue 1, January, pp. 324–335.

[8] Xiao Mang Li, Ung Mo Kim,(2012)"A hierarchical framework for content-based image spam filtering", 8th International Conference on Information Science and Digital Content Technology (ICIDT), Jeju,June , pp. 149-155.

[9] Z. Wang, W. Josephson, Q. Lv, M. Charikar and K. Li.(2007) Filtering Image Spam with near-Duplicate Detection, in Proceedings of the 4th Conference on Email and Anti-Spam CEAS.

[10] Duncan Cook, Jacky Hartnett, Kevin Manderson and Joel Scanlan(2006),"Catching spam before it arrives: domain specific dynamic blacklists", ACSW Frontiers '06 Proceedings of the 2006 Australasian workshops on Grid computing and e-research - Volume 54, January , pp. 193-202.

[11] Anirudh Ramachandran, Nick Feamster, and Santosh Vempala,(2007)"Filtering spam with behavioral blacklisting", Proceedings of the 14th ACM conference on Computer and communications security CCS'07, NY, USA, pp. 342-351.

[12] Damiani E., Vimercati S. D. C. d. et al.,(2004) "An Open Digest-based Technique for Spam Detection", San Francisco, CA, USA, pp. 1-6.

[13] Harris Drucker,Donghui Wu, and Vladimir N. Vapnik,(1999)"Support Vector Machines for Spam Categorization", IEEE Transactions On Neural Networks, Vol. 10, No. 5, September.

[14] Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz(1998),"A Bayesian Approach to Filtering Junk E-Mail",Learning for Text Categorization: Papers from the 1998 Workshop, Madison, Wisconsin, AAAI Technical Report WS-98-05.

[15] Dat Tran, Wanli Ma, Dharmendra Sharma, and Thien Nguyen,(2007)"Possibility Theory-Based Approach to Spam Email Detection",IEEE International Conference on Granular Computing.

[16] Clotilde Lopes, Paulo Cortez, Pedro Sousa, Miguel Rocha, and Miguel Rio,(2011) "Symbiotic filtering for spam email detection", Expert Systems with Applications: An International Journal, Volume 38 Issue 8, August, pp.9365-9372.

[17] Oda, T. and T. White.(2005) Immunity from Spam: An Analysis of an Artificial Immune System for Junk Email Detection. in 4th International Conference on Artificial Immune Systems (ICARIS).

[18] M. Uemura and T. Tabata,(2008)"Design and Evaluation of a Bayesian-filter based Image Spam Filtering Technique", in Proceedings of the International Conference on Information Security and Assurance(ISA).

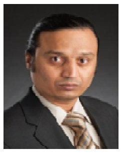
[19] Dat Tran, Wanli Ma, and Dharmendra Sharma,(2009) "A Novel Spam Email Detection System Based on Negative Selection",In Proceedings of the 4th International Conference on Computer Sciences and Convergence Information Technology (ICCIT'09). Los Alamitos, CA, 2009,pp. 987-992.

[20] Elena Zheleva, Aleksander Kolcz and Lise Getoor,(2008) "Trusting spam reporters: A reporter-based reputation system for email filtering", ACM Transactions on Information Systems (TOIS),Volume 27 Issue 1, December.

[21] P.K Panigrahi,(2012)"A Comparative Study of Supervised Machine Learning Techniques for Spam E-mail Filtering", Fourth International Conference on Computational Intelligence and Communication Networks (CICN), Mathura, Nov,pp.506-512.

AUTHORS

Vivek Shandilya holds a BE in Electronics and Communication Engineering from Bangalore university, MS in Computer Science and is a PhD candidate in Computer Science at University of Memphis. His research areas are optimization and security of stochastic systems.



Fahad Polash is a PhD student at Department of Computer Science at university of Memphis. He has a bachelor's degree in computer science and worked in telecom industry before starting his graduate studies. His research areas are computer networking, network security and forensics.

Sajjan Shiva is the chair and a professor at Department of Computer Science at university of Memphis. He was formerly chair of the Department of Computer Science at University of Alabama, Huntsville. His research areas are computer organization and architecture, parallel processing, software engineering, security systems and cloud computing. He is a fellow of IEEE

IMPROVED SECURE ADDRESS RESOLUTION PROTOCOL

Abhishek Samvedi[1], Sparsh Owlak[2] and Vijay Kumar Chaurasia[3]

Indian Institute of Information Technology, Allahabad, India
[1]abhisheksmvd@gmail.com
[2]maadhav.owlak@gmail.com
[3]vijayk@iiita.ac.in

ABSTRACT

In this paper, an improved secure address resolution protocol is presented where ARP spoofing attack is prevented. The proposed methodology is a centralised methodology for preventing ARP spoofing attack. In the proposed model there is a central server on a network or subnet which prevents ARP spoofing attack.

KEYWORDS

ARP, ARP poisoning, Central Server, DHCP server

1. INTRODUCTION

ARP spoofing has become a major problem in the present scenario. ARP spoofing can lead to many other attacks like man in the middle attack in secure socket layer. Thus steps must be taken to prevent this type of attack. In this paper a scheme is proposed to prevent ARP spoofing attack.

Section I gives a brief introduction of the situation. Section II discusses the proposed solution for preventing ARP spoofing attack in detail. Section III discusses the message formats used in proposed scheme. Section IV discusses the performance evaluation of proposed scheme against standard ARP protocol. Section V summarizes and concludes the paper.

1.1 Address Resolution Protocol

The Address Resolution Protocol (ARP) is used by Internet Protocol [IP], defined in [RFC826], to bind the IP addresses to the MAC addresses, which is stored in ARP cache of each client machine. This protocol works on network layer. In IPv6 this functionality is provided by Neighbour Discovery Protocol (NDP).

1.2 ARP Poisoning

ARP Poisoning is done by sending ARP Reply packet to victim node with sender’s (attacker) IP address and MAC address as destination IP and MAC address respectively, as shown in figure 1. The victim when processes the ARP Reply packet, this brings change in its ARP table for destination IP address with attacker’s MAC address, which causes the victim to send all packets destined to target host to attacker. The attacker then can read and modify packets flowing between target node and victim node.

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 201–211, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4521

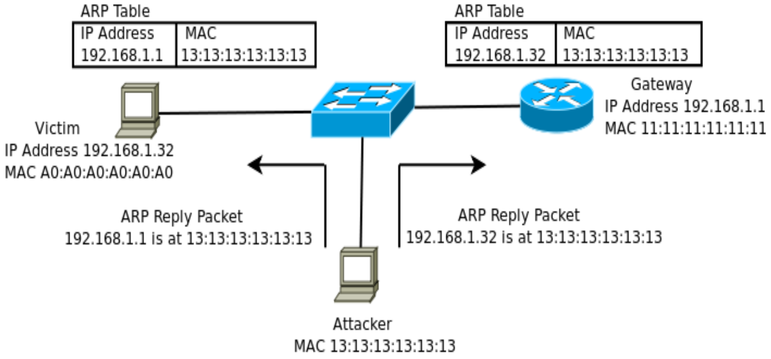


Figure 1. ARP poisoning

1.3 Dynamic Host Configuration Protocol

The Dynamic Host Configuration Protocol (DHCP) is a protocol, which provides IP address to client from IP address pool, which in return results into the connectivity of client with rest of the network. It involves clients and a server operating in a client-server model. When the DHCP server receives a request from a client, server determines the network to which the client is connected, and allocates an IP address or prefix and configuration information to the client. DHCP server grants IP addresses to clients for a limited period and clients are responsible for renewing their IP address before that time expires. DHCP is used in IPv4 and IPv6. This protocol is defined in [RFC2131].

2. RELATED WORK

Previous work relating ARP security are S-ARP[1], Ticket based ARP[2], Enhanced ARP[3], securing unicast protocol by extending DHCP[4] and a centralised detection and prevention technique against ARP poisoning[5]. In S-ARP each host has a public private key pair and a digital certificate obtained from a central certificate authority(CA) which is local to the organisation. Each host has to register its IP address, and public key (contained in certificate) to the authoritative key distributer (AKD), as every host has to get a certificate the total overhead incurred by the local CA will be high to verify every new host. In this scheme as every host has its certificate the chances of such attack are high where an attacker steals the certificate of a host and pretends to be that victim host on the network and performs malicious activities. In S-ARP for the case of dynamic networks the DHCP server has to first consult the AKD server before providing the new host an IP address. This means that a new host has to register an entry with AKD server first which has to be done manually as the new host does not have an IP address at this stage. Also the communication between new host and local CA will be done manually in dynamic network situation. In Ticket based Address Resolution Protocol [2] research paper there is a local Local Ticket Agent (LTA) which issues tickets to new host. Thereafter the communications in ARP protocol are done using the ticket concept. Here as every new host requires a ticket, overhead incurred will be high. Another possible attack here is that an attacker makes a copy of a ticket of genuine host and then uses that ticket for malicious activities on the network. In Enhanced ARP[3] research paper each host maintains a long term IP-MAC mapping table apart from ARP cache. This IP-MAC mapping table contains IP-MAC mapping of all hosts on its network. This scheme involves too much memory usage and thus cost incurred will also be high. Also the updating of IP-MAC mapping table for each host will be difficult. In dynamic networks the IP addresses provided by DHCP server to various hosts have to be renewed after some time thus frequent updating of IP-MAC table will be required which will require enough overhead in this scheme. In the research paper securing unicast protocol by extending DHCP [4], the DHCP man in the middle attack is possible. In the research paper a centralised detection and prevention technique against ARP poisoning [5] the ACS (ARP central server) acts very much similar to DHCP server and is vulnerable to DHCP man in the middle attack.

3. PROPOSED SOLUTION

The proposed scheme which removes ARP spoofing attack is for dynamic networks, where there are hosts, and a new entity called Central Server. The Central Server maintains an IP-mac table for the subnet or the network it is present in. This IP-mac table contains information of IP-mac binding information of all hosts on the subnet or network the central server is present in and which have been allocated IP address by the DHCP server nearest to this central server. Relay Agents can be used in case of large networks. Whenever a host on a subnet or network is allocated an IP address by the DHCP server, it also informs the central server on that subnet or network through IP_send message. This message is sent on data link layer. It is signed by the secret key shared between DHCP server and central server. The central server then sends a message IP_reply to the DHCP server to show acknowledgement of IP_send message. This message is sent on data link layer and is signed by the central server using the symmetric key shared between central server and DHCP server. In a network all the ARP request and ARP reply messages will be sent to this Central Server. The client or hosts will not communicate the ARP request and ARP reply messages to each other.
The scenario of a network can be shown as:-

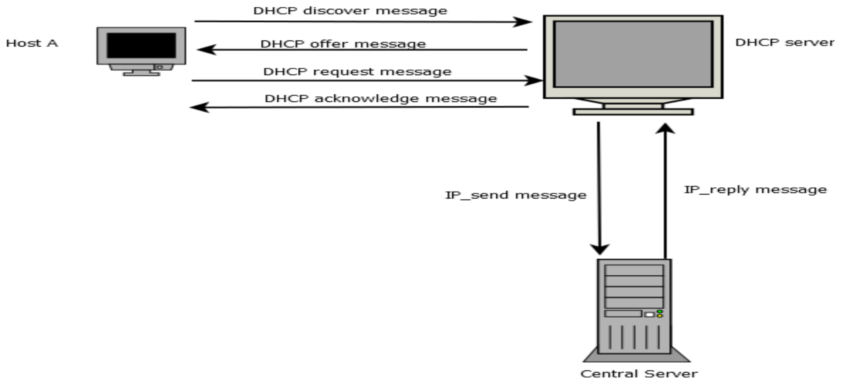


Figure 2. Proposed subnet or network setup

The procedure for entry of a new host in a subnet or network will be as follows:-

1. The new host broadcasts the DHCP discover message containing the MAC address of the host.

2. The DHCP server allocates an IP address to this new host following the standard DHCP protocol.

3. The DHCP server then sends an IP_send message to the central server. This message is sent on data link layer. It is digitally signed by the symmetric secret key shared between DHCP server and central server.

4. The central server after receiving this message will update its IP-mac table. The frame format of IP_send message is discussed in the following section.

5. After this the central server will send IP_reply message to the DHCP server showing acknowledgement of IP_send message. This message is sent on data link layer and is digitally signed by symmetric key shared between central server and DHCP server. The frame format of IP_reply message is discussed in the following section.

3.1.Prevention of ARP spoofing attack

In this scheme all the ARP reply and ARP request messages are send to Central Server. When the Central Server receives an ARP request message it provides the requesting host with the MAC address of corresponding IP address it wants by sending ARP reply message which is again digitally signed by the central server. In this case asymmetric cryptography is used. The digital certificate of central server is also attached in this ARP reply message. This digital certificate can be obtained from public certificate authority like VeriSign. One digital certificate is required and it can be distributed to all central servers. When the Central Server receives an ARP reply message which is sent by a host which wants to get its IP-mac combination information changed due to change in its MAC address then the Central Server checks it's IP-mac table and sends 50 ARP_Check messages to the previous MAC address stored as a combination with the IP address provided. These ARP_Check messages are again digitally signed by the Central Server. The Central Server also attaches its digital certificate with these messages. These ARP_Check messages are sent on data link layer and its frame format is discussed in the following section. If the Central Server gets a reply even to any one of these messages then it keeps the previous IP-mac combination in IP-mac table otherwise it changes the previous entry with new entry. In case the Central Server gets an ARP reply message from previous MAC address it sends an ARP_NoChange message to the MAC address which initiated this procedure. This ARP_NoChange is digitally signed by central server with digital certificate attached with this message. In case of Denial of Service (DOS) attack on the system with MAC address which has to be changed in the IP-mac table of Central Server, at least one ARP reply will be returned by it to the Central Server with a probability of 99.5% (=1-0.950). Thus sending 50 ARP_Check messages just increases the probability that we will get some reply from a host which has the previous MAC address which is to be changed in the case of ARP spoofing attack. On the other hand if the Central Server does not gets an ARP reply message from previous MAC address it sends an ARP_Ack message to theMAC address which initiated this procedure. This message indicates the client that appropriate modification in IP-MAC table in central server has been done. This message is again digitally signed by central server with digital certificate attached.

In this situation two types of ARP spoofing attack can be thought of:-

3.1.1. The attacker sends a fake ARP reply message to Central Server asking it to change its IP-mac combination in IP-mac table. In this situation the Central Server uses the checking scheme of sending 50 ARP_Check messages to previous MAC address as already discussed. This checking scheme prevents false entry in IP-mac table of Central Server thus preventing ARP spoofing attack.

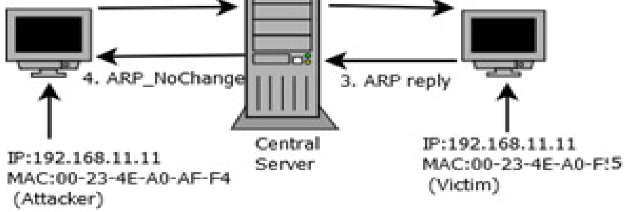


Figure 3: first possible type of ARP spoof attack

3.1.2. In this case a genuine host wants to get its IP-mac combination stored in IP-mac table in Central Server to get changed because of change in its MAC address, thus it will send the ARP reply message to Central Server for the same. Here in this case the Central Server will again follow its checking procedure. At this step the attacker knowing the previous MAC address of the genuine host who wants to get its IP-mac combination changed will falsely send an ARP reply message to the Central Server using the previous MAC address of genuine host. Thus the request of genuine host will get cancelled. Still In this situation the ARP spoofing attack is prevented because the Central Server will send a digitally signed ARP_NoChange message to the host which initiated this procedure. Thus this genuine host can again request a new IP address from Secure-DHCP server.

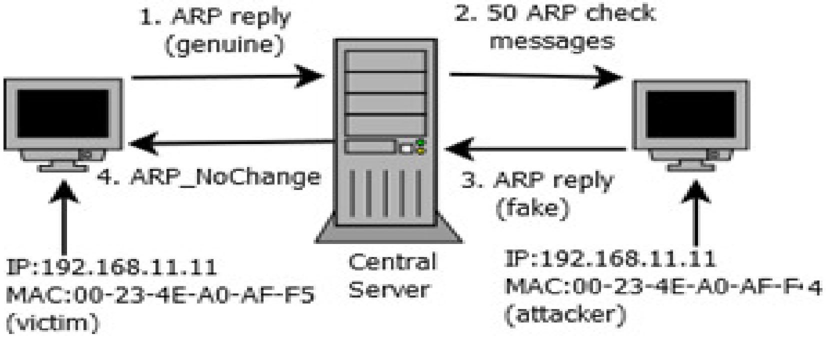


Figure 4. Second possible type of ARP spoof attack

Other possible attack in this scenario is the DOS attack on central server by a number of fake ARP reply messages. This DOS attack can be prevented by monitoring mechanism using Intrusion Detection System (IDS).

Also a situation arises that an attacker is performing a DOS attack on central server by sending a number of fake ARP reply messages to central server. Now at the same time a genuine user sends ARP reply message to central server then this message will also be discarded by the central server in order to mitigate the DOS attack. However since the client will not receive ARP_Ack message from central server it can conclude that its request has not been accepted so it can again send ARP_reply message to central server after some time.

4. MESSAGE FORMAT

The messages IP_send, IP_reply, can be send on data link layer with adequate changes in frame format of standard protocols used like Ethernet II to accommodate the required information. The messages IP_send, IP_reply, are messages communicated in a network amongst DHCP server, Central Server so these messages can be communicated on data link layer. Relay agents can be used in case of large networks. The messages ARP_Check,ARP_NoChange, ARP_Ack can be also be send on data link layer. ARP request and ARP reply follow the standard ARP protocol frame format. To make the scheme compatible with networks following the normal standard protocols the gateways should be modified with extra capability to identify traffic flowing from outside network or inside network so that the gateway can follow the required network protocols to forward the traffic. The possible frame format for these messages is as follows:

4.1.IP_send message

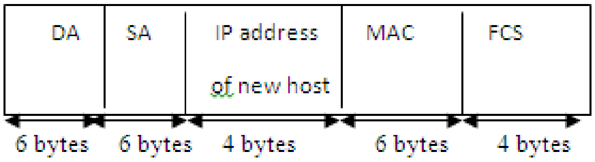


Figure 5. Frame Format of IP_send message

The fields in this frame format are:

4.1.1 Destination Address (6 bytes): This is the MAC address of destination system (CentralServer).

4.1.2 Source Address (6 bytes): this is the MAC address of source system (Secure DHCP).

4.1.3IP address (4 bytes): This is the IPv4 address of the new host.

4.1.4MAC address of new host (6 bytes): This is the MAC address of the new host.

4.1.5 FCS (4 bytes): This field stands for frame check sequence.

4.2.IP_reply message

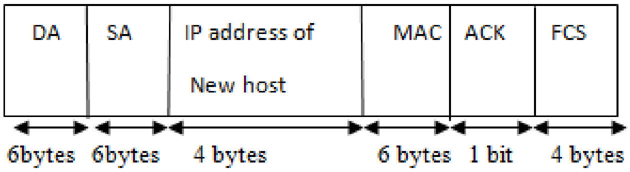


Figure 6. Frame Format of IP_reply messages

The fields in this frame format are:

4.2.1 Destination Address (6 bytes): This is the MAC address of destination system (SecureDHCP).

4.2.2 Source Address (6 bytes): This is the MAC address of source system (CentralServer).

4.2.3IP address (4 bytes):This is the IPv4 address of the new host.

4.2.4 MAC address (6 bytes): This is the MAC address of new host.

4.2.5ACK (1 bit): This is a 1 bit field showing the acknowledgement of IP address by new host. This bit is set to 1 to show acknowledgement of IP address.

4.2.6 FCS (4 bytes): This field stands for frame check sequence.

4.3 ARP_CHECK message

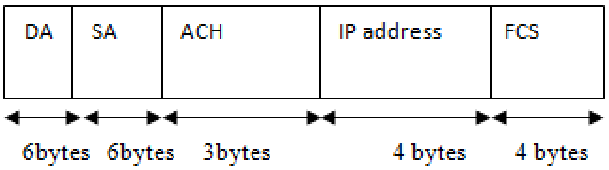


Figure 7. Frame Format of ARP_CHECK messages

The fields in this frame format are:

4.3.1 Destination Address (6 bytes): This is the MAC address of destination system (The host (if exists) with the MAC address stored in IP-mac table which has to be replaced with new MAC address).
4.3.2 Source Address (6 bytes): This is the MAC address of source system (CentralServer).
4.3.3ACH field (3 bytes):Thisfield will store the string ACH which indicates ARP_CHECK message.
4.3.4 IP address (4 bytes): This is the IPv4 address of the host whose MAC address has to be changed.
4.3.5 FCS (4 bytes): This field stands for frame check sequence.

4.4.ARP_NoChange message

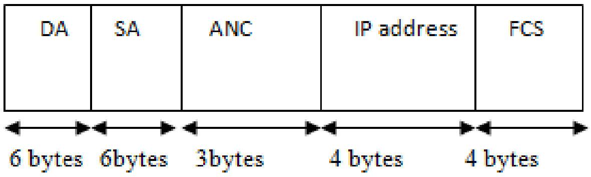


Figure 8. Frame Format of ARP_NoChange

The fields in this frame format are:

4.4.1 Destination Address(6 bytes):This is the MAC address of destination system (The host which initiated the ARP reply message to Central Server with the need to make changes in the IP-mac table and update it with new MAC address of this host).
4.4.2 Source Address (6 bytes):This is the MAC address of source system (CentralServer).
4.4.3ANC field (3 bytes): This field will store the string ANC which indicates ARP_No Change message.
4.4.4IP address (4 bytes):This is the IPv4 address of the host which initiated the request to update the IP-mac table.
4.4.5FCS (4 bytes): This field stands for frame check sequence.

4.5.ARP_Ack message

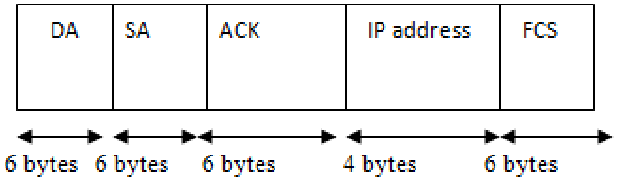


Fig 9. Frame Format of ARP_Ack

The fields in this frame format are:

4.5.1 Destination Address(6 bytes): This is the MAC address of destination system (The host which
4.5.2 Initiated the ARP reply message to Central Server with the need to make changes in the IP-mac table and update it with new MAC address of this host).
4.5.3 Source Address (6 bytes): This is the MAC address of source system (CentralServer).
4.5.4ACK field (3 bytes): This field will store the string ACK which indicates ARP_Ackmessage.

4.5.5IP address (4 bytes): This is the IPv4 address of the host which initiated the request to update the IP-mac table.
4.5.6FCS (4 bytes): This field stands for frame check sequence

5. PERFORMANCE EVALUATION

We will evaluate the performance of proposed scheme against standard ARP protocol and standard DHCP protocol.

5.1. Performance Evaluation of the proposed scheme against DHCP protocol

The figure below shows the scenario where IP address is provided to new host using standard DHCP protocol. This scenario involves just one DHCP server.

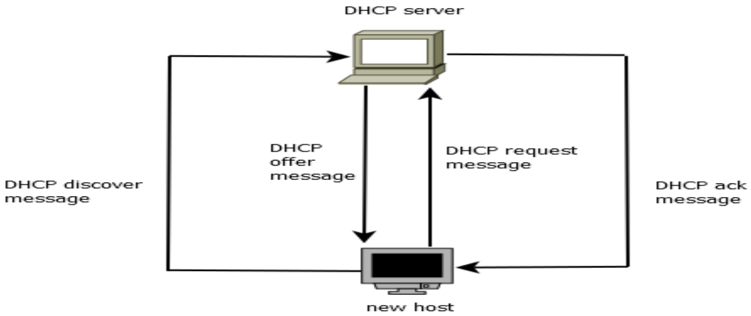


Figure 10. Performance evaluation of DHCP protocol

As is seen from Figure 10 the total transaction of messages involved in this process is 4.

Now we evaluate our proposed scheme. The figure below shows the situation:

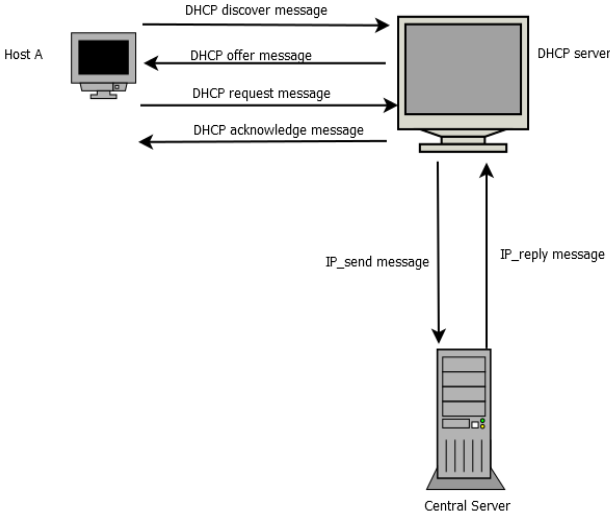


Figure 11. Performance evaluation of proposed scheme

As is seen from figure 11 that total transaction of messages involved is 6, Thus if we compare the performance of standard DHCP protocol and the proposed scheme, we find that the standard DHCP protocol is slightly better than our proposed scheme in terms of cost involved in transaction of messages but our proposed scheme also removes the ARP spoofing attack thus making the network secure.

5.2 Performance Evaluation of proposed scheme against ARP protocol

To evaluate performance of proposed scheme against standard ARP protocol we consider two situations:

5.2.1 The host A wants to know MAC address of host B: In the case of standard ARP protocol the host A will broadcast ARP request message on the network and will wait for ARP reply from host B. The situation is shown in figure below:

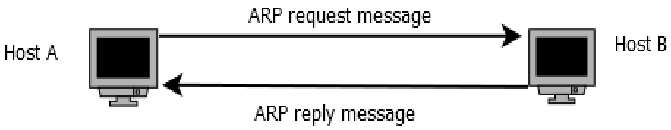


Figure 12. Performance evaluation of ARP protocol

As is seen from figure 12 the total transaction of messages involved in this case is 2.
If we consider the same situation in a subnet or network following our proposed scheme then the host A will send the ARP request message to the Central Server. The Central Server will reply with ARP reply message to host A providing MAC address of host B. The situation is shown below:

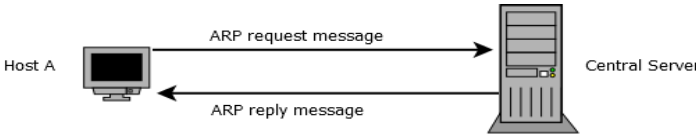


Figure 13. Performance evaluation of proposed scheme

As is seen from figure 13 the total transaction of messages involved is 2.Thus in this situation we see that performance of proposed scheme is equivalent to standard ARP protocol in terms of cost involved in transaction of messages.

5.2.2 Host A has got its MAC address changed and it wants to inform other hosts on the network:In this situationif the network follows standard ARP protocol then host A broadcasts ARP reply message to all other hosts on the network containing the information of its changed MAC address. The situation is shown in figure below.

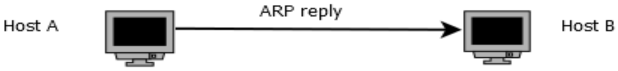


Fig 14. Performance evaluation of ARP protocol

In this situation in the case of standard ARP protocol the total transaction of messages is just 1.

If such a situation occurs in a network following the proposed scheme then the situation can be shown by the figure:

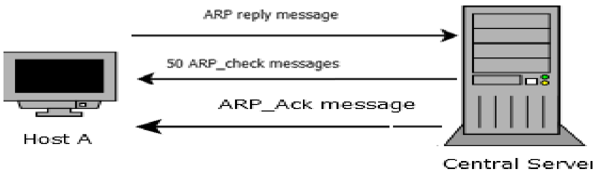


Figure 15. Performance evaluation of proposed scheme

In this situation we see that the host Asents ARP reply message to Central Server for updationof IP-mac table. In response to this the Central Server first sends 50 ARP_check messages to the previous MAC address. If it gets no reply to any of these messages then it updates the IP-mac table and finally the ARP_Ack message is sent from central server to the client as an acknowledgement for the change.

Total Transaction of messages involved in this situation is 52 messages.

Here we observe that the total cost involved in transaction of messages in the case of our proposed scheme is more than the standard ARP protocol , however the proposed scheme makes a network secure and ARP spoofing attack is not possible in it.

5. CONCLUSION

The proposed scheme removes ARP spoofing attack in a dynamic network. The scheme is highly secure and good for dynamic networks. The scheme is also compatible with existent networks following standard network protocols with some modifications in gateways. DHCP denial of service attack can also be removed if we follow the monitoring mechanism which is generally followed to prevent such attacks. Thus the proposed scheme is highly secure.

REFERENCES

[1] D. Bruschi, A. Ornaghi, and E. Rosti, “S-arp: a secure address resolution protocol,” in Computer Security Applications Conference, 2003. Proceedings.19th Annual. IEEE, 2003, pp. 66–74.

[2] W. Lootah, W. Enck, and P. McDaniel, “Tarp: Ticket-based address resolution protocol,” vol. 51, no. 15. Elsevier, 2007, pp. 4322–4337.

[3] S. Nam, D. Kim, and J. Kim, “Enhanced arp: preventing arp poisoning based man-in-the-middle attacks,”Communications Letters, IEEE, vol. 14, no. 2, pp. 187–189, 2010.

[4] B. Issac and L. Mohammed, “Secure unicast address resolution protocol (s-uarp) by extending dhcp,” in Networks, 2005. Jointly held with the 2005 IEEE 7th Malaysia International Conference on Communication. 2005 13th IEEE International Conference on, vol. 1.IEEE, 2005, pp. 6–pp.

[5] Sumit Kumar and ShashikalaTapaswi, “A centralized detection and prevention technique against ARP poisoning.”,IEEE,pp. 259-264.

[6] Internet Engineering Task Force. (2004) Dynamic Host Configuration Protocol [Online]. Available from: http://tools.ietf.org/html/rfc2131 [Accessed 24rd September,2013].

AUTHORS

AbhishekSamvedi is a student of Master of Science. His domain of M.S. program is CLIS (Cyber law and information security) from IIIT-Allahabad.

SparshOwlakis a student of Master of Science. His domain of M.S. program is CLIS (Cyber law and information security) from IIIT-Allahabad.

Vijay Kumar Chaurasiya is Doctor of philosophy from IIIT-Allahabad. His area of Specialization Is in Wireless and mobile network.

INTENTIONAL BLANK

TAXONOMY: MOBILE MALWARE THREATS AND DETECTION TECHNIQUES

Lovi Dua and Divya Bansal

Computer Science Department, PEC University of Technology,
Sector 12, Chandigarh 160012, India
dualovi@gmail.com, divya@pec.ac.in

ABSTRACT

Since last-decade, smart-phones have gained widespread usage. Mobile devices store personal details such as contacts and text messages. Due to this extensive growth, smart-phones are attracted towards cyber-criminals. In this research work, we have done a systematic review of the terms related to malware detection algorithms and have also summarized behavioral description of some known mobile malwares in tabular form. After careful solicitation of all the possible methods and algorithms for detection of mobile-based malwares, we give some recommendations for designing future malware detection algorithm by considering computational complexity and detection ration of mobile malwares.

KEYWORDS

Smart-phones, Malware, Attacks, Static analysis, Dynamic analysis

1. INTRODUCTION

Now, there is a thin line difference between Smart-phones, PCs(Personal Computers) and other newly emerged devices like tabs, notebooks and laptops as all are now connected technologies. Due to various services like social networking and gaming provided by smart-phones with the help of applications, these are exposed to gain some confidential information from mobile-devices. Smart-phone OSs includes symbian, android, palmOS and embedded Linux etc. Android is the popular platform for smart-phone based malware authors as any third-party vendor can create applications for android phones and deploy it on android market. Sometimes, even trusted applications are able to leak user's location and phone's identity and share it on server without its consent. Due to this growing skill-set of cyber-criminals who device their algorithms for breaching privacy, embarrassing service-provider and bring inconvenience to the users. So, it requires special care to secure these networked devices from malwares with the help of anti-developed techniques and algorithms for detection. This paper focuses on describing mobile-based threats and its counter detection techniques.

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 213–221, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4522

1.1 Current State of Study

This section discusses some current malwares reported by security researcher groups. In 2010,different types of mobile malwares are found including DroidDream, Geinimi, GGTracker, Plankton Tonclank and HongTouTou. These malwares are much like original Cabir worm. LookOut security firm reported that over one million of android devices are affected in first half of 2011[21]. In 2012, it is reported by Homeland security department that 79 percent of the mobile threats were targeted to Android operating systems. In January 2012, Symantec identified Trojan horse named AndroidCounterclank for stealing information [3]. Security firm Kaspersky found in 2013 that 98 percent of malware was directed at android platform.

1.2 Organization of paper

In this paper section 2 will discuss mobile device attack vectors and types of detection techniques for mobile malwares. Section3 will discuss detection techniques and algorithms proposed by various researchers and section 4 will give conclusion by analyzing various techniques proposed by different researchers followed by some future recommendations.

2. MOBILE MALWARES

Malware exhibits malicious behavior which targets to mobile phones without user's consent by adding or changing malicious code into software system. Malware is employed intentionally to cause harm to system by gaining confidential information from the device and modifying file contents etc [4]. Malicious executables are further classified into following categories: virus, worm, trojan-horse and botnets. Virus injects malicious code into existing programs to infect them which in turn infects other programs when gets executed. On the other hand, worms spread over the network by exploiting vulnerabilities on the computers connected to network. Trojan appears as benign program but do some malicious activities and botnet gives the attacker ability to remotely control set of user's devices [21].

2.1 Mobile Device Threats

Numerous attack vectors exist which compromises security of mobile devices [5]. Three main categories of attacks could be carried over mobile devices which includes- malware attacks, grayware attacks and spyware attacks described as:-

2.1.1 Malware- These kind of attacks steal personal data from mobile devices and damage devices [22]. With device vulnerabilities and luring user to install additional apps, attacker can gain unauthorized root access to devices. Some of the malware attacks are listed as:-

- Bluetooth attacks: With Bluetooth attacks, attacker could insert contacts or SMS messages, steals victim's data from their devices and can track user's mobile location. Blue-bugging is kind of blue-tooth attack through which attacker could listen conversations by activating software including malicious activities [22].

- SMS attacks: Through SMS attacks, attacker can advertise and spread phishing links. SMS messages can also be used by attackers to exploit vulnerabilities [22].

- GPS/Location attacks: User's current location and movement can be accessed with global positioning system (GPS) hardware and then information can be sold to other companies involved in advertising[22].

- Phone jail-breaking: With jail-breaking, an attacker can remove security implications of operating system like it allows OS to install additional and unsigned applications. Users are attracted to install them as they could get additional functionality [22].

- Premium rate attacks: They posed serious security concerns because premium rate SMS messages could go unnoticed until attacker faces thousands or dollars of bill on his device as they don't need permissions to send SMS on premium rated numbers [22].

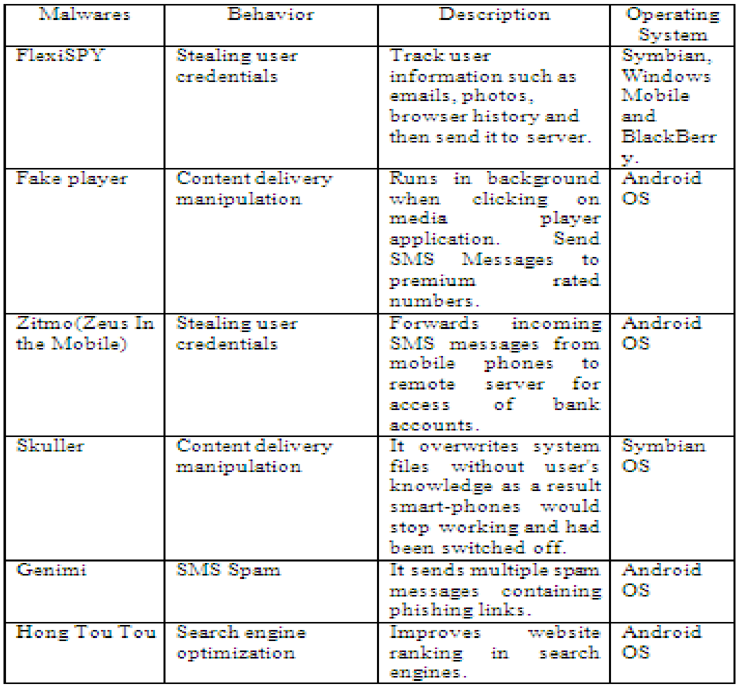
2.1.2 Grayware: Grayware include applications which collects the data from mobile devices for marketing purposes. Their intention is make no harm to users but annoy them.

2.1.3 Spyware: Spyware collects personal information from user's phone such as contacts, call history and location. Personal spyware are able to gain physical access of the device by installing software without user's consent. By collecting information about victim's phone, they send it to attacker who installed the app rather than the author of the application.

2.2 Behavioral Classification

Malware may also be classified on the basis of their behavior. Table 1 depicts behavioral classification of some known malwares as shown below:-

Table 1: Malware Behavioral classification



Malwares	Behavior	Description	Operating System
FlexiSPY	Stealing user credentials	Track user information such as emails, photos, browser history and then send it to server.	Symbian, Windows Mobile and BlackBerry.
Fake player	Content delivery manipulation	Runs in background when clicking on media player application. Send SMS Messages to premium rated numbers.	Android OS
Zitmo(Zeus In the Mobile)	Stealing user credentials	Forwards incoming SMS messages from mobile phones to remote server for access of bank accounts.	Android OS
Skuller	Content delivery manipulation	It overwrites system files without user's knowledge as a result smart-phones would stop working and had been switched off.	Symbian OS
Genimi	SMS Spam	It sends multiple spam messages containing phishing links.	Android OS
Hong Tou Tou	Search engine optimization	Improves website ranking in search engines.	Android OS

2.2 Malware detection techniques

Malware can be analyzed with the help of detection techniques. Malware analysis is the process of studying code, behavior and functionality of malware so that severity of attack can be measured. Detection techniques are broadly categorized into three types- static analysis, dynamic analysis and permission-based analysis as shown in Fig 3.1. Figure depicts that static analysis can be done with parameters-static code analysis, taint tracing and control flow dependencies. Dynamic analysis considers parameters including-network traffic, native code and user interaction. Permission-based analysis can be done with the help of permissions specified in manifest file. In literature, various techniques exist for detection of mobile malware.

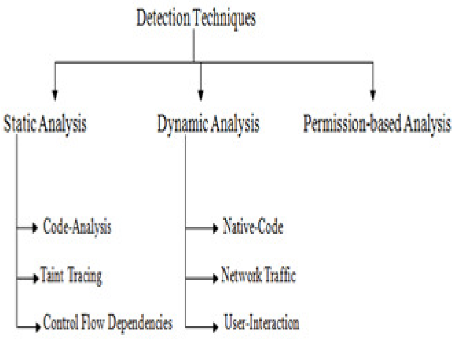


Figure 3.1 Malware detection techniques

2.2.1 Static analysis

Static analysis investigates downloaded app by inspecting its software properties and source code. It is an inexpensive way to find malicious activities in code segments without executing application and notifying its behavior. Many techniques can be used for static analysis: de-compilation, decryption, pattern matching and static system call analysis etc. However, obfuscation and encryption techniques embedded in software makes static analysis difficult. Static analysis is further categorized into two categories- misuse detection and anomaly detection traditionally used by anti-viruses.

2.2.1.1 Misuse detection: Misuse detection uses signature-based approach for detection of malware based on security policies and rule-sets by matching of signatures. In static analysis, data flow dependency and control flow dependencies in source code that would help to understand the behavior of apps.

2.2.1.2 Anomaly detection: Anomaly detection uses machine learning algorithms for learning of known malwares and predicting unknown malware. This approach is suitable for identifying action of malware rather than pattern. Here, methods are used to construct suspicious behavior of applications and then observed signatures are matched against database of normal behavior applications. It is able to distinguish between malicious and normal behavior by training network with classifier such as support vector machine (SVM).

2.2.2 Dynamic analysis

Dynamic analysis involves execution of application in isolated environment to track its execution behavior. In contrast to static analysis, dynamic analysis enables to disclose natural behavior of malware as executed code is analyzed, therefore immune to obfuscation attempts. Various heuristics are considered for monitoring dynamic behavior which includes-monitoring network activity, file changes and system call traces. Android applications can run in an Android SDK, a mobile device emulator running on desktop computer for emulation of software and hardware features except generating phone calls. For testing purposes emulator supports Android Virtual Device(AVD) configurations. When applications start running on the emulator, it can use all services like to invoke other applications, accessing network state, play audio and video, store and retrieve data. Console output is used for debugging, logging of simulated events such as generating phone calls and receiving SMS messages and kernel logs can also be obtained.

2.2.3 Permission-based analysis

Permissions play key role while analyzing android applications .They are listed in Manifest.xml file while each application is installed. Install time permissions limits application behavior with control over privacy and reduces bugs and vulnerabilities [2]. Users have right to allow or deny the installation of applications but he cannot go for the selection of individual permissions. These permissions are required in android applications because the use of resources in android phones is based on these permission set. Some researchers are able to detect malicious behavior of android applications on the basis of permissions specified in Manifest.xml.

3. RELATED WORK

3.1 Static analysis

Kim *et al*. [11] proposed framework for detection and monitoring of energy greedy threats by building power consumption from the collected samples. After generating power signatures, data analyzer compares them with signatures present in a database. Batyuk *et al.*[18] proposed system for static analysis of android applications . First, they provide in-depth static analysis of applications and present readable reports to user for assessment and taking security relevant decisions-to install or not to install an application. Then the method is developed to overcome security threats introduced by the applications by disabling malicious features from them. Ontang *et al.*[19] proposed Secure application Interaction Framework (Saint) by extending android security architecture for protection of interfaces and enhancing interaction policies between calling and callee applications.

Wei *et al.*[15] proposed a static feature-based approach and develop system named Droid Mat able to detect and distinguish android malware . Their mechanism considers the static information including permissions, intents and regarding components to characterize android malware , clustering algorithm is applied to enhance malware modeling capability .K-Nearest Neighbor algorithm classify applications as benign and malicious applications. Finally their results are compared with well known tool Androguard, published in Blackhat 2011 and it is found that DroidMat is efficient as it takes only half time than Androguard to predict 1738 applications.

Bose *et al.* [12] present behavioral detection framework for representation of malware behavior by observing logical ordering of applications actions. Malicious behavior is discriminated from normal behavior by training SVM. System is evaluated for both real-world and simulated mobile malwares with 96% accuracy.

Schmidt *et al.*[10] describes a method for symbianOS malware analysis called centroid based on static function call analysis by extracting features from binaries and clustering is applied for detection of unknown malwares. VirusMeter [9] is proposed to detect anomalous behavior on mobile devices by catching malwares which are consuming abnormal power .Machine learning algorithms helped to improve its detection accuracy. pBMDS [20] an approach through which user-behavior is analyzed by collecting data through logs of key-board operations and LCD displays and then correlated with system calls to detect anomalous activities. Hidden markov model(HMM) is leveraged to learn user-behavior and malware behavior for discrimination of differences between them.

3.2 Dynamic analysis

Batyuk *et al.* [8] proposed an android application sandbox (AA Sandbox) system for analysis of android applications consists of fast static pre-check facility and kernel space sand-box. For suspicious application detection, both static and dynamic analysis is performed on android applications. AASandbox takes APK file and list out following files by decompressing them-Androidmanifest.xml, res/, classes.dex. Manifest file holds security permissions and description of application. Res/ folder defines layout, graphical user interface (GUI) elements and language of application. Classes.dex file contains executable code for execution on dalvik virtual machine which is then de-compiled to java files with baksmali and then code is searched for suspicious patterns. Monkey program designed for stress testing of applications generates pseudo random sequences of user-events such as touches and mouse-clicks. It is used to hijack system calls for logging operation and helpful to get the logging behavior of application at system level. Around 150 applications are collected for testing and evaluation.

Min *et al.* [13] proposed run-time based behavior dynamic analysis system for android applications. Proposed system consists of event detector, log monitor and parser. Event trigger is able to simulate the user's action with static analysis. Static analyzer generates manifest.xml and java code with the help of application .apk file. Semantic analysis find list of risk based permissions, activities and services including other information such as hash code and package name. Data flow analysis creates control flow graph (CFG) of the application by mapping of user-defined methods and API calling. By running application in a customized emulator with loadable LKM, sensitive information about application can be captured such as sent SMS , call log and network data for entry address of system calls. Logs recorded with debugging tool logcat for sensitive behavior sent to Log parser. Log monitor gathers log data as the application runs and parser analyzes log data by picking sensitive information and filtering out unnecessary information. By collecting 350 apps from the Amazon Android Market, results found that about 82 applications leak private data.

Enack *et al.* [14] proposed Apps-playground framework for automatic dynamic analysis of android applications. Designed approach is able to analyze malicious applications in addition to applications leaking private data from smart-phones without the user's consent. Dynamic analysis should possess detection techniques including ability to explore application code as much as possible and the environment should be as much real that malicious application could not obfuscate. Automatic analysis code integrates the detection, exploration and disguise techniques to explore android applications effectively. Detection techniques detect the malicious functionality while app is being executed .It includes taint tracing which monitor sensitive APIs with TaintDroid such as SMS APIs and kernel level monitoring for tracing of root exploits.Automatic exploration techniques are helpful for code coverage of applications by simulating events such as location changes and received SMS so that all application code is covered. Fuzzy testing and intelligent black box execution testing is used for automatic exploration of android applications. Disguise techniques create realistic environment by providing data such as International mobile equipment identity(IMEI), contacts, SMS, GPS coordinates etc.

Enck *et al.* [7] proposed TaintDroid for dynamic analysis. First dynamic analysis tool used for system wide analysis of android applications by tracking flow of sensitive information through third-party applications. TaintDroid integrates multiple granularities at object level i.e, variable, method, message and file level. It is able to monitor how the sensitive data are used by applications and then taints are labeled. TaintDroid is tested on around 30 applications and it is found that 15 of them uses personal information.

3.3 Permission-based analysis

Johnson *et. al.* [16] proposed architecture for automatic downloading of android applications from the android market. Different algorithms employed for searching of applications such as downloading applications by application category. With static analysis, required permissions can be obtained based on its functionality. Permission names are searched in android source code and then mapped with API calls to know that whether requested permissions are correct or not. Program examines all smali files of application to obtain list of method calls used in an application. Each method call is then compared with method call listed in permission protected android API calls to know exact permissions. Restricted permission set is compared with all the permissions specified in AndroidManifest.xml file to find out extra permissions, lacking of permissions and exact permission set required for its functionality.

Zhou *et al.* [17] proposed DroidRanger for systematic study on overall health of both official and unofficial Android Markets with the focus on the detection of malicious apps. DroidRanger leverages a crawler for collection of apps from the Android Market and saved into local repository. Features extracted from collected apps include requested permissions and author information. Two different detection engines are used for detection of known and unknown malwares. First detection engine is permission-based behavioral foot-printing scheme able to distil apps requiring dangerous permissions such as SEND_SMS and RECEIVE_SMS permissions. Therefore, number of apps to be processed for second detection engine is reduced. In second step, multiple dimensions for behavioral foot-printing scheme chosen for listening of all system-wide broadcast messages if they contains receiver named android.provider.Telephony.SMS_RECEIVED. Obtained callgraph associates API calls to specific components specified in a rule. For example- by calling abortBroadCast function with specific rule, a method is obtained to detect apps monitoring incoming SMS messages. Second detection engine includes some heuristics to detect suspicious apps and zero-day malwares. Heuristics attempts to dynamically fetch and run code from untrusted websites which is further monitored during run-time execution to confirm whether it is truly malicious or not.

4. CONCLUSION

Smart-phones are becoming popular in terms of power, sensor and communication. Modern, smart-phones provide lots of services such as messaging, browsing internet, emailing, playing games in addition to traditional voice services. Due to its multi-functionality, new security threats are emerged for mobile devices. In this paper, we presented survey on various techniques for detection of mobile malware. We have categorized various mobile malware detection techniques based on features extracted from them and monitoring system calls as they provide us low level information. We have analyzed that information-flow tracking, API call monitoring and network analysis provide more deeper analysis and useful information for detection of mobile malware.

5. RECOMMENDATIONS FOR FUTURE

Following are some recommendations for designing algorithm to detect mobile-based applications containing malwares.

1. Multiple sources for feature extraction should be used for building feature-set to detect mobile malwares.
2. There should be national or international database for reporting malware incidents so that developers are aware of distinct vulnerabilities related to mobile malwares.
3. Artificial intelligence algorithms(neural network-based) to improve detection ratio.
4. Machine to machine communication and authentications tools must be used in between multiple device platforms

REFERENCES

[1] F-Secure. Trojan:symbos/yxe, http://www.virus.fi/v-descs/trojan_symbos_yxe.shtml.

[2] Manifest.permission,Androiddeveloper, http://developer.android.com/reference/android/Manifest.permission.html

[3] Android.Counterclank Found in Official Android Market, http://www.symantec.com/connect/fr/blogs/androidcounter

[4] M.L.Polla ,F. Martinelli, D.Sgandurra: A Survey on Security for Mobile Devices: Communications Surveys and Tutorials, pp.446-471.IEEE(2013)

[5] McAfee Labs Q3 2011 Threats Report Press Release, 2011, http://www.mcafee. com/us/about/news/2011/q4/20111121-01.aspx

[6] M. Chandramoha, H.Tan: Detection of Mobile Malware in the Wild.:Computer (Volume:45 , Issue: 9) ,pp.65-71(2012)

[7] W.Enck, P. Gilbert, B.G. Chun, L.P.Cox, J.Jung, P.McDaniel, A.P.Sheth: TaintDroid: an information-ow tracking systemfor realtime privacy monitoring on smart-phones.:In OSDI'10 Proceedings of the 9th USENIX conference on Operating systems design and implementation,pp.1-6 ,USENIX Association Berkeley, CA,USA (2010)

[8] T.Blasing, L.Batyuk, A.D.Schimdt, S.H.Camtepe, S.Albayrak,:An Android Application Sandbox System for Suspicious Software Detection.

[9] L.Liu,G.Yan, X.Zhang, S.Chen,: VirusMeter: Preventing Your Cell phone from Spies.: Proceedings of the 12th International Symposium on Recent Advances in Intrusion Detection.,pp.244-264, Springer-Verlag,Berlin, Heidelberg(2009).

[10] A.D.Schmidt, J.H.Clausen,S.H.Camtepe, S.Albayrak: Detecting Symbian OS Malware through Static Function Call Analysis: In Proceedings of the 4th IEEE International Conference on Malicious and Unwanted Software,pp.15-22.IEEE(2009).

[11] H.Kim, J.Smith, K.G.Shin,:Detecting energy-greedy anomalies and mobile malware variants: In MobiSys 08: Proceeding of the 6th international conference on Mobile systems, applications, and services,pp.239-252.ACM,NewYork(2008).

[12] A. Bose,X.Hu, K.G.Shin, T.Park: Behavioral detection of malware on mobile handsets:In MobiSys 08: Proceeding of the 6th international conference on Mobile systems, applications, and services,pp.225-238.,ACM,NewYork(2008).

[13] L.Min,Q.Cao: Runtime-based Behavior Dynamic Analysis System for Android Malware Detection: Advanced Materials Research,pp.2220-2225.

[14] V.Rastogi, Y.Chen, W.Enck: AppsPlayground: Automatic Security Analysis of Smartphone Applications:In CODASPY'13 Proceedings of the third ACM conference on Data and application security and privacy,pp.209-220.ACM,NewYork(2013)

[15] D.J.Wu,C.H.Mao,T.E.Wei,H.M.Lee,K.P.Wu: DroidMat: Android Malware Detection through Manifest and API Calls Tracing.: In Information Security (AsiaJCIS), 2012 Seventh Asia Joint Conference ,pp.62-69.IEEE,Tokyo(2012)

[16] R.Jhonson, Z.Wang, C.Gagnon, A.Stavrou,: Analysis of android applications' permissions.:In Software Security and Reliability Companion (SERE-C) Sixth Inter-national Conference,pp.45-46.IEEE(2012)

[17] Y.Zhou,, Z.Wang, W.Zhou,X.Jiang: Hey, You, Get o_ of My Market: Detecting Malicious Apps in O_cial and Alternative Android Markets: In Proceedings of the 19th Network and Distributed System Security Symposium,San Diego,CA(2012).

[18] L.Batyuk,M.Herpich,S.A.Camtepe,K.Raddatz,A.D.Schmidt,S.Albayrak:Using static analysis for automatic assessment and mitigation of unwanted and malicious activities within Android applications.: In 6th International Conference on Malicious and Unwanted Software,pp.66-72.IEEE Computer Society(2011)

[19] M.Ongtang,S.E.McLaughlin,W.Enck,P.D.McDaniel,:Semantically rich application-centric security in android:In Proceedings of the 25th Annual Computer Security Application Conference (ACSAC),pp.340-349(2009)

[20] L.Xie, X.Zhang, J.P.Siefert, S.Zhu: pBMDS: a behavior-based malware detection system for cellphone devices.:In Wisec'10 Proceedings of the third ACM conference on Wireless network security,Hoboken,pp.37-48.ACM,USA(2010)

[21] A.P.Felt ,M.Finifter,E.Chin,S.Hanna,D.Wagner:A survey of mobile malware in the wild.:In Proceedings of the 1st ACM workshop on Security and privacy in smart phones and mobile devices,pp.3-14.ACM,NewYork(2011)

[22] D.Stites, A.Tadimla :A Survey Of Mobile Device Security: Threats, Vulnerabilities and Defenses./urlhttp://afewguyscoding.com/2011/12/survey-mobile-devicesecurity-threats-vulnerabilities-defenses

INTENTIONAL BLANK

SURVEY ON CLASSIFICATION TECHNIQUES FOR INTRUSION DETECTION

Pritam Sapate[1] and Shital A. Raut[2]

[1,2]Department of Computer Science and Engineering, VNIT, Nagpur, India
pritamsapate@gmail.com
saraut@cse.vnit.ac.in

ABSTRACT

Intrusion detection is the most essential component in network security. Traditional Intrusion Detection methods are based on extensive knowledge of signatures of known attacks. Signature-based methods require manual encoding of attacks by human experts. Data mining is one of the techniques applied to Intrusion Detection that provides higher automation capabilities than signature-based methods. Data mining techniques such as classification, clustering and association rules are used in intrusion detection. In this paper, we present an overview of intrusion detection, KDD Cup 1999 dataset and detailed analysis of different classification techniques namely Support vector Machine, Decision tree, Naïve Bayes and Neural Networks used in intrusion detection.

KEYWORDS

Intrusion Detection, Data Mining, KDD Cup 1999, Classification.

1. INTRODUCTION

Internet plays vital role in today's world. It is used in business, education, shopping, social networking etc. This has increased risk of computer systems connected to the internet becoming targets of intrusions by cyber criminals. Cyber criminals attack systems to gain unauthorized access to information, misuse information or to reduce the availability of information to authorized users. This results in huge financial losses to companies besides losing their goodwill to customers. Intrusion prevention techniques such as user authentication (e.g. using password or biometrics), information protection (e.g. encryption), avoiding programming errors and firewalls have been used to protect computer systems. But, unfortunately these intrusion prevention techniques alone are not sufficient. There will always be unknown exploitable weaknesses in the system due to design and programming flaws in application programs, protocols and operating systems. Therefore, we need mechanism to detect intrusions as soon as possible and take appropriate actions [1].

Intrusion detection system monitors data coming from the network and various system logs and analyses them to detect potential attacks. Traditional intrusion detection methods are based on extensive knowledge of signatures of known attacks. The signatures describing attacks have to be hand-coded by human experts. Newly captured events are then matched against the available signatures of attacks to detect intrusion. Whenever new type of intrusion is discovered, the

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 223–231, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4523

signature database has to be manually revised by human expert. In other words, signature-based approach has failed to provide required level of automation. Other techniques including statistical methods, machine learning and data mining methods have been proposed as a way of dealing with limitations of signature-based approaches. These techniques provide higher automation in intrusion detection process along with good detection rate. Currently many researchers have shown an increasing interest in intrusion detection techniques based on data mining techniques [2] [3].

Data mining based intrusion detection techniques can be classified into two categories: misuse detection and anomaly detection. In misuse detection technique, each instance in a dataset is labelled either as 'normal' or 'intrusion' and learning algorithm is trained over labelled data to build model. Whenever a new type of attack is discovered, learning algorithm can be retrained with new dataset that includes labelled instances of new attack. In this way, models of misuse detection are created automatically and can be more precise than manually created signatures. In anomaly detection technique, models are built on normal behaviour and any deviation from normal behaviour is identified as intrusion [2].

This paper is organized as follows: Section 2 describes attack types, intrusion detection and general working of intrusion detection systems. Section 3 gives details of KDD Cup 1999 benchmark intrusion detection dataset. Data mining and intrusion detection are discussed in Section 4. Section 5 presents detailed analysis of different classification techniques used for intrusion detection. Finally conclusion is mentioned in section 6.

2. BACKGROUND

2.1. Attack Types

According to taxonomy proposed by kendall [4], attacks can be classified into following four categories:

2.1.1. Denial of Service (DoS)

A denial-of-service (DoS) or distributed denial-of-service (DDoS) attack is an attack in which the attacker tries to make computer resource too busy or too full to respond to its intended users. Examples of such attacks include Smurf, Teardrop, Back, Ping of death, Neptune, Land etc.

2.1.2. User to Root

A User to Root is an attack that aims to gain super user access to the system. Attacker gain super user access by exploiting vulnerability in operating system or application software. The attacker starts out with access to a normal user account on the system (perhaps gained by sniffing password, a dictionary attack or social engineering) and is able to exploit some vulnerability to gain root access to the system. Most common attack in this class of attack is buffer overflow attack. Other attacks include Loadmodule, Perl, Ps, Xterm etc.

2.1.3. Remote to User

A Remote to User is an attack in which the attacker tries to gain unauthorized access from a remote machine into super user account of the target system. In this type of attack, attacker sends packets to a machine over a network and then exploits some vulnerability to gain local access as a user of that machine. Examples of remote to user attack are Dictionary, Ftp_write, Guest, Imap, Phf etc.

2.1.4. Probing

Probing is an attack in which the attacker scans a network of computers to gather information or find known vulnerabilities. An attacker who knows which machines and services are available on network can use this information to look for weak points. He will use this information to plan future attacks. There are many tools available for probe attack which can be used by even a very unskilled attacker. Examples of probing attack are Ipsweep, Mscan, Nmap, Saint, Satan etc.

2.2. Intrusion detection

Intrusion detection is the act of detecting actions that tries to compromise the confidentiality, integrity and availability of a resource. Based on analysis strategy intrusion detection techniques can be divided into [1] [24]:

Anomaly Detection. Anomaly detection tries to determine whether deviation from normal usage pattern can be flagged as intrusion. It establishes normal usage patterns using statistical measures on system audit data and network data. The major limitation of this technique is high false alarm rate.

Misuse Detection. Misuse detection uses patterns of well known attacks to identify intrusions. It is very good at detecting known attacks. The main disadvantage of such system is it is unable to detect any future (unknown) intrusions that don't have matched pattern stored in the system.
Based on the source of audit data Intrusion detection techniques can be divided as Host based and network based.

Host-Based IDS. Data coming from various host activities including audit records of operating system, system logs and process activities is used for analysis.

Network-Based IDS. Data coming from network traffic is collected for analysis using sniffing software like TCPDUMP.

2.3. Working of intrusion detection systems

Following four steps are proposed for generalized working of IDS by authors of [6].

2.3.1. Data Collection

Data useful for detecting intrusion is collected in this step. For network-based intrusion detection network traffic is collected using sniffer software like TCPDUMP. For host-based intrusion detection data such as process activity, disk usage, memory usage and system calls are collected. Commands such as netstas, ps and strace are used for this purpose.

2.3.2. Feature selection

The collected data is substantially large and cannot be used as it is, so subset of this data is selected by creating feature vectors that contain only necessary information needed for intrusion detection. In network based intrusion detection, it can be IP packet header information which includes source and destination IP addresses packet length, layer four protocol type and other flags. In host-based intrusion detection it includes user name, login time and date, duration of session and number of opened files.

2.3.3. Analysis

The collected data is analyzed in this step to determine whether the data is anomalous or not. This is the main research area where many methods have been proposed and used to detect intrusion.

2.3.4. Action

IDS alerts the system administrator that an attack has happened using several methods like e-mail, alarm icons and visualization techniques. IDS can also stop or control attack by closing network ports or killing processes.

3. INTRUSION DETECTION DATASET

In this section, brief description of KDD Cup 1999 dataset [4][16] which was derived from the 1998 DARPA intrusion detection evaluation program is provided. It is the most widespread dataset collected over a period of nine weeks for a LAN simulating a typical U.S. Air Force LAN. The dataset contains a collection of simulated raw TCP dump data, where multiple intrusion attacks were introduced and widely used in the research community. From seven weeks of network traffic, four gigabytes of compressed binary TCP dump training data was processed into five million connection records. Similarly, two weeks of test data yielded about two million connection records. The dataset contains 4,898,430 labelled and 311,029 unlabeled connection records. The labelled connection records consist of 41 features. Features characterizing each connection are divided into:

- basic features of individual TCP connections,
- content features within a connection suggested by domain knowledge,
- time based features computed using a two second time window and
- host based features computed using a window of 100 connections used to characterize attacks that scan the hosts (or ports) using much larger time interval than two seconds.

In network data of KDD99 dataset, each instance represents feature values of a class, where each class is categorized either as normal or attack. The classes in dataset are divided into one normal class and four main intrusion classes: Denial of Service (DoS), Probe, User-to-Root (U2R), Remote-to-Login (R2L).

4. DATA MINING AND INTRUSION DETECTION

Data mining is used in applications that require data analysis. In recent years, data mining techniques have been highly researched in intrusion detection domain. Different data mining techniques such as classification, clustering, and association rules are used to acquire information about intrusions by analysing system audit data and network data [1][9]. The main approach of data mining is classification, which maps a data item into one of several predefined categories. Here we present a review of different classification techniques used for detecting intrusions.

5. CLASSIFICATION TECHNIQUES FOR INTRUSION DETECTION

Classification is the process of assigning each data instance to one of the predefined categories. Data classification is a two step process: Learning and classification. In first step, classifier is built by analysing a training set made up of data instances and their associated class labels. Because the class label of each training instance is provided, this is known as supervised learning. In second step, built classifier is used to predict the class for unlabelled data instance. Different types of classification techniques are decision trees, neural networks, bayesian classification, support vector machines, nearest neighbour classification, genetic algorithm and fuzzy logic [10].

Intrusion detection can be thought of as a classification problem. We can gather sufficient audit data in which each data instance will be labelled as either “normal” or “abnormal”. We then use classification algorithm on audit data to build classifier. This classifier will then predict class of new unseen audit data as “normal” or “abnormal”. Classification approach can be used for both misuse detection and anomaly detection but it is mostly used for misuse detection [1]. In this section, we present an overview of different classification techniques used for intrusion detection.

5.1. Support Vector Machine

Support vector Machine (SVM), a promising pattern classification technique, proposed by Vapnik [19]. SVMs are supervised learning models with associated learning algorithms that have been applied increasingly to misuse detection in the last decade. SVM maps the input vector into a higher dimensional feature space and obtain the optimal separating hyper-plane in the higher dimensional feature space.

Srinivas Mukkamala and Guadalupa Janoski [20] proposed Support Vector Machine (SVM) and Neural Networks (NN) for intrusion detection system. Two main reasons for using SVM for intrusion detection are: speed and scalability. The experiments were carried using DARPA 1998 dataset. The training time for SVMs is significantly shorter (17.77 sec) than that for neural networks (18 min). This becomes an important advantage in situations where retraining needs to be done quickly. The performance of SVM showed that SVM IDS have slightly higher rate of making the correct detection than neural networks. However, SVMs can make only binary classifications which will be disadvantage when IDS requires multiple-class identifications.

Chen R. C. et al. [25] proposed use of Rough Set Theory (RST) and Support Vector Machine (SVM) for intrusion detection. They used KDDCUP99 dataset for experiment. RST is used to pre-process the data and to reduce the number of features. The features selected by RST are used to learn the SVM model and to test the model respectively. Using all 41 features accuracy was 86.79% and false positive rate was 29.97%. While with 29 features selected using RST accuracy was 89.13% and false positive rate was reduced to 13.27%. This shows that method is effective in increasing accuracy and reducing false positive rate.

Wang Hui et al. [26] proposed an intrusion detection method based on improved SVM by combining Principal Component Analysis (PCA) and Particle Swarm Optimization (PSO). KDDCUP99 dataset was used for experiment. PCA is an effective data mining technique used to reduce dimensionality of dataset. Then PSO was used to elect punishment factor C and kernel parameters σ in SVM. The intrusion detection rate (97.752%) of improved SVM by combining PCA and PSO was higher than those of PSO-SVM (95.635%) and that of standard SVM (90.476%).

5.2. Decision tree

Quinlan [13] proposed a decision tree classifier which is one of the most known machine learning techniques. A decision tree composed of three basic elements [14]:

- A decision node representing test or condition on data item.
- An edge or a branch which corresponds to the one of the possible attribute values which means one of the test attribute outcomes.
- A leaf which determines the class to which the object belongs.

To classify an object, one starts at the root of the decision tree and follows the branch indicated by the outcome of each test until a leaf node is reached. The name of the class at the leaf node is the class of an unknown object. The best attribute to divide the subset at each stage is selected using the information gain of the attributes.

Ben Amor et al. [14] performed experiment on KDDCUP99 intrusion data set for comparative analysis of naïve bayes versus decision tree. They found that decision tree gives slightly better results than naïve bayes. However, from computational point of view, construction of decision tree is slower than naïve bayes. The decision tree selects the best features for each decision node during the construction of the tree based on some well defined criteria. Decision trees generally have very high speed of operation and high attack detection accuracy. The Naïve Bayes classifiers make strict independence assumption between the features in an observation that result in lower attack detection accuracy when the features are correlated.

In [14] they used all 41 features in KDDCUP99 dataset. However, Gary Stein et al. [15] suggest that not all 41 features are required for classification of four categories of attack: Probe, DOS, U2R and R2L. In their work they used Genetic Algorithm to select relevant features for decision tree, with a goal of increasing detection rate and decreasing false alarm rate. They performed experiment for each of the above four categories of attack separately. The GA made drastic improvements in some of the categories like performance gain on Probe is 23% on the average. However, Performance improvement on R2L and U2R are limited. This may be because the proportions of R2L and U2R are very low in the training data, but much higher in the testing data.

S. Sheen and R. Rajesh [23] used three different approaches for feature selection namely Chi square, Information Gain and ReliefF and compared the performance of these three approaches using decision tree classifier. The KDDCUP99 dataset is used for experiment. They found that Chi square and information gain had similar performance while ReliefF was giving a lower performance.

5.3. Naïve Bayes

Naïve Bayes can be considered as an upgraded version of Bayes Theorem as it assumes strong independence among attributes. Bayesian classifier encodes probabilistic relationships among variables of interest. This means that the probability of one attribute does not affect the probability of the other.

Mrutyunjaya Panda and Manas Ranjan Patra [17] proposed a framework of network intrusion detection system based on naïve bayes algorithm. They performed experiment on 10% KDDCUP99 dataset and evaluated system using 10-fold cross validation. Their approach achieved higher detection rate than neural network based approach. The detection rate was 95%, with an error rate of 5%. Moreover, it performed faster and was cost effective. However, it generates somewhat more false positives.

Dewan Md. Farid et al. [18] proposed a new hybrid learning algorithm for adaptive network intrusion detection using naive Bayesian classifier and ID3 algorithm. They evaluated the performance of proposed algorithm for network intrusion detection using 10% of KDDCUP99 dataset. The attacks of KDD99 dataset were detected with 99% accuracy and minimized false positives.

In [29] Z. Muda et al. proposed use of a hybrid learning approach through combination of K-means clustering and naïve bayes classification. An experiment is carried out using KDDCUP99 dataset to evaluate the performance. In first stage, they grouped similar data instances based on their behaviours by utilizing a K-Means clustering. In second stage, they used Naïve Bayes classifier to classify resulting clusters into attack classes. This approach detected better percentage of attacks with above 99% of accuracy and detection rate and below 0.5% of false alarm.

5.4. Neural Networks

A neural network consists of a collection of processing elements that are highly interconnected and transform a set of inputs to a set of desired outputs. The result is determined by the characteristics of the elements and the weights associated with the interconnections between them. By modifying the connections between the nodes, the network can adapt to the desired outputs. Neural networks have been used in both anomaly detection and misuse detection. For anomaly detection, neural networks were modelled to learn the typical characteristics of system users and identify significant variations from the user's established behaviour as anomaly. In misuse detection, the neural network would receive data from the network stream and analyze the information for instances of misuse [22].

Ryan et al. in [21] performed first works to intrusion detection using NN. They trained and tested a back propagation neural network called NNID (Neural Network Intrusion Detector) on a system of ten users. The data source for training and testing was operating system logs in UNIX environment. The system showed 96% accuracy in detecting unusual activity with 7% false alarm rate.

Jirapummin et al. [27] presented a methodology for both visualizing intrusions by using SOM and classifying intrusions by using Resilient Propagation. They selected Neptune attack (SYN flooding), Portsweep and Satan attacks (port scanning) from KDD Cup 1999 dataset. For Resilient Propagation algorithm (RPROP), they utilized 3-layer NN with 70 nodes in first hidden layer, 12 neurons in second hidden layer and 4 neurons in the output layer. The transfer functions for the first hidden layer, second hidden layer and the output layer of RPROP were tan-sigmoidal, log-sigmoidal and log-sigmoidal respectively. They achieved more than 90 % detection rate and less than 5 % false alarm rate in three selected attacks.

Iftikhar Ahmad, et al. [28] performed comparison between three back propagation algorithms used in intrusion detection. These three algorithms were:

a. The basic On-Line BackProp algorithm,
b. The Batch BackProp algorithm and
c. The Resilient BackProp algorithm.

They performed experiment on KDDCUP99 dataset and found that the Resilient BackProp algorithm give better performance than online and batch.

6. CONCLUSIONS

Data mining techniques have been highly researched in the domain of intrusion detection in order to reduce the hassle of manually analysing huge volumes of audit data. In this paper, we reviewed different classification approaches used by researchers for detecting intrusion. The challenge is to achieve high detection rate and reduce false alarm rate. Any one classifier alone is not sufficient to achieve this. More than one classifier can be combined to remove disadvantages of one another. Combining classifiers lead to a better performance than any single classifier.

REFERENCES

[1] Lee, W., & Stolfo, S. (1998), “Data mining approaches for intrusion detection,” In Paper presented at the proceedings of the seventh USENIX security symposium (SECURITY’98). San Antonio, TX.

[2] Paul Dokas , Levent Ertoz, V Kumar, Lazarevic, Srivastava & Pang-Nig Tan, “Data Mining for Network Intrusion Detection,” In Proc. 2002 NSF Workshop on Data Mining, pp. 21-30.

[3] C. A.Catania and C. G.Garino, “Automatic Network Intrusion Detection: Current Techniques and Open Issues,” Computer & Electrical Engineering 5 (2012), pp. 1062-1072.

[4] K. Kendall, “A Database of Computer Attacks for the Evaluation of Intrusion Detection Systems,” Massachusetts Institute of Technology Master's Thesis, 1998.

[5] LI Min, An Yang Institute of Technology,.”Application of DataMining Techniques in Intrusion Detection,” 2005.

[6] Khaled Labib, "Computer Security and Intrusion Detection," from Crossroads The ACM students magazine.

[7] Chang-Tien Lu,Arnold P.Boedihardjo,Prajwal Manalwar, "Exploiting efficient data mining techniques to enhance Intrusion Detection Systems," 0-7803-9093-8/05/$20.00 2005 IEEE, pp. 512-517.

[8] Brugger S. T, "Data mining methods for network intrusion detection," Technique Report, UC davis, 2004.

[9] Portnoy, L., Eskin, E., and Stolfo, S. 2001, “Intrusion detection with unlabeled data using clustering,” In Proceedings of the ACM Workshop on Data Mining Applied to Security.

[10] Data mining: concepts and techniques by jiawei Han, Michelle Kamber.

[11] Wang, H., Zhang, G., Chan, H. & Jiang, X. 2009, “Mining Association Rules for Intrusion Detection,” International Conference on Frontier of Computer Science and Technology.

[12] Reema Patel, AmitThakkar, Amit Ganatra, “A Survey and Comparative Analysis of Data Mining Techniques for Network Intrusion Detection Systems,” International Journal of Soft Computing and Engineering (IJSCE), Vol-2, 2012.

[13] Quinlan, C4.5: Programs for Machine Learning, 1993, Morgan Kaufmann Publishers, San Mateo, CA.

[14] Ben Amor, Benferhat, Elouedi, “Naive Bayes vs. Decision Trees in Intrusion Detection Systems,” Proc. of the 2004 ACM symposium on applied computing, 2004, pp. 420–424.

[15] Stein G, Chen B, Wu AS, Hua KA (2005), “Decision tree classifier for network intrusion detection with GA-based feature selection,” In: Proceedings of the 43rd annual southeast regional conference ACM vol 2, pp 136–141.

[16] KDD. http://kdd.ics.uci.edu/databases/kddcup99.

[17] Mrutyunjaya Panda, Manas Ranjan Patra, “Network Intrusion Detection Using Naïve Bayes,” International Journal of Computer Science and Network Security,vol.7 no.12, 2007, pp.258-262.

[18] Dewan Md. Farid, Nouria Harbi, Mohammad Zahidur Rahman, “Combining Naive Bayes and Decision Tree for daptive Intrusion Detection,” Proc. Of Intl. Journal of Network Security & Its Applications (IJNSA), Vol. 2, No. 2, 2010, pp.12-25.

[19] Cortes, Vapnik, Support-vector networks, Machine Learning, vol.20, 1995, pp.273–297.

[20] Srinivas Mukkamala, Guadalupe Janoski, Andrew Sung, “Intrusion Detection: Support VectorMachines and Neural Networks,” In Proceedings of the IEEE International Joint Conference on Neural Networks, 2002, pp. 1702-1707

[21] J. Ryan, M. -J. Lin, R. Miikkulainen, ”Intrusion detection with neural networks”, in Proceedings of AAAI -97 Workshop on AI Approaches to Fraud Detection and Task Management, 1997, pp. 92–97.

[22] J. Cannady, “Artificial Neural Networks for Misuse Detection,” National Information Systems Security Conference, 1998.

[23] S. Sheen and R. Rajesh, “Network Intrusion Detection using Feature Selection and Decision tree classifier,” IEEE Region 10 Conference, TENCON08 (2008), pp. 1–4.

[24] A. Lazarevic, V. Kumar and J. Srivastava, "Intrusion detection: A survey,” Managing Cyber Threats, pp.19 -78, 2005.

[25] Chen R. C., Cheng K. F., and Hsieh C. F., “Using rough set and support vector machine for network intrusion detection,” International Journal of Network Security & Its Applications (IJNSA), Vol. 1, No. 1, 2009, pp. 1–13.

[26] WANG Hui, ZHANG Guiling, E Mingjie, SUN Na, "A Novel Intrusion Detection Method Based on Improved SVM by Combining PCA and PSO," Wuhan University Journal of Natural Sciences, 2011, vol. 16, No. 5, pp. 409-413.

[27] Jirapummin, C., Wattanapongsakorn, N. and Kanthamanon P, "Hybrid Neural Networks for Intrusion Detection System," The 2002 International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC 2002), pp. 928-931, Phuket, Thailand.

[28] Iftikhar Ahmad, Dr. M.A Ansari, Dr. Sajjad Mohsin,"Performance Comparison between Backpropagation Algorithms Applied to Intrusion Detection in Computer Network Systems," Proceedings of the 7th WSEAS International Conference on Applied Computer and Applied Computational Science as ACM guide, pp. 47-52, 2008.

[29] Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir. "Intrusion Detection based on K-Means Clustering and Naïve Bayes Classification," 7th International Conference on IT in Asia (CITA), 2011.

AUTHORS

Mr. Pritam Sapate is currently an M.Tech student at Visvesvaraya National Institute of Technology, Nagpur. He has received his B.Tech degree in Information Technology from S.G.G.S.I.E. & T, Nanded (M.H).

Mrs. Shital A. Raut is an Assistant Professor at Visvesvaraya National Institute of Technology, Nagpur. Her areas of interest include Data Mining and Warehousing and Business Information Systems.

INTENTIONAL BLANK

EFFICIENT ALGORITHM FOR RSA TEXT ENCRYPTION USING CUDA-C

Sonam Mahajan[1] and Maninder Singh[2]

[1,2]Department of Computer Science Engineering, Thapar University, Patiala, India
sonam_mahajan1990@yahoo.in
msingh@thapar.edu

ABSTRACT

Modern-day computer security relies heavily on cryptography as a means to protect the data that we have become increasingly reliant on. The main research in computer security domain is how to enhance the speed of RSA algorithm. The computing capability of Graphic Processing Unit as a co-processor of the CPU can leverage massive-parallelism. This paper presents a novel algorithm for calculating modulo value that can process large power of numbers which otherwise are not supported by built-in data types. First the traditional algorithm is studied. Secondly, the parallelized RSA algorithm is designed using CUDA framework. Thirdly, the designed algorithm is realized for small prime numbers and large prime number . As a result the main fundamental problem of RSA algorithm such as speed and use of poor or small prime numbers that has led to significant security holes, despite the RSA algorithm's mathematical soundness can be alleviated by this algorithm.

KEYWORDS

CPU, GPU, CUDA, RSA, Cryptographic Algorithm.

1. INTRODUCTION

RSA (named for its inventors, Ron Rivest, Adi Shamir, and Leonard Adleman [1]) is a public key encryption scheme. This algorithm relies on the difficulty of factoring large numbers which has seriously affected its performance and so restricts its use in wider applications. Therefore, the rapid realization and parallelism of RSA encryption algorithm has been a prevalent research focus. With the advent of CUDA technology, it is now possible to perform general-purpose computation on GPU [2]. The primary goal of our work is to speed up the most computationally intensive part of their process by implementing the GCD comparisons of RSA keys using NVIDIA's CUDA platform.

The reminder of this paper is organized as follows. In section 2, we study the traditional RSA algorithm. In section 3, we explained our system hardware. In section 4, we explained the design and implementation of parallelized algorithm. Section 5 gives the result of our parallelized algorithm and section 6 concludes the paper.

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 233–241, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4524

2. TRADITIONAL RSA ALGORITHM[1]

RSA is an algorithm for public-key cryptography [1] and is considered as one of the great advances in the field of public key cryptography. It is suitable for both signing and encryption. Electronic commerce protocols mostly rely on RSA for security. Sufficiently long keys and up-to-date implementation of RSA is considered more secure to use.

RSA is an asymmetric key encryption scheme which makes use of two different keys for encryption and decryption. The public key that is known to everyone is used for encryption. The messages encrypted using the public key can only be decrypted by using private key. The key generation process of RSA algorithm is as follows:

The public key is comprised of a modulus n of specified length (the product of primes p and q), and an exponent e. The length of n is given in terms of bits, thus the term “8-bit RSA key" refers to the number of bits which make up this value. The associated private key uses the same n, and another value d such that d*e = 1 mod φ (n) where φ (n) = (p - 1)*(q - 1) [3]. For a plaintext M and cipher text C, the encryption and decryption is done as follows:

$$C = M^{e} \bmod n,\ M = C^{d} \bmod n.$$

For example, the public key (e, n) is (131,17947), the private key (d, n) is (137,17947), and let suppose the plaintext M to be sent is: ***parallel encryption***.

- Firstly, the sender will partition the plaintext into packets as: pa ra ll el en cr yp ti on. We suppose a is 00, b is 01, c is 02, ... z is 25.
- Then further digitalize the plaintext packets as: 1500 1700 1111 0411 0413 0217 2415 1908 1413.
- After that using the encryption and decryption transformation given above calculate the cipher text and the plaintext in digitalized form.
- Convert the plaintext into alphabets, which is the original: ***parallel encryption***.

3. ARCHITECTURE OVERVIEW[4]

NVIDIA's Compute Unified Device Architecture (CUDA) platform provides a set of tools to write programs that make use of NVIDIA's GPUs [3]. These massively-parallel hardware devices process large amounts of data simultaneously and allow significant speedups in programs with sections of parallelizable code making use of the Simultaneous Program, Multiple Data (SPMD) model.

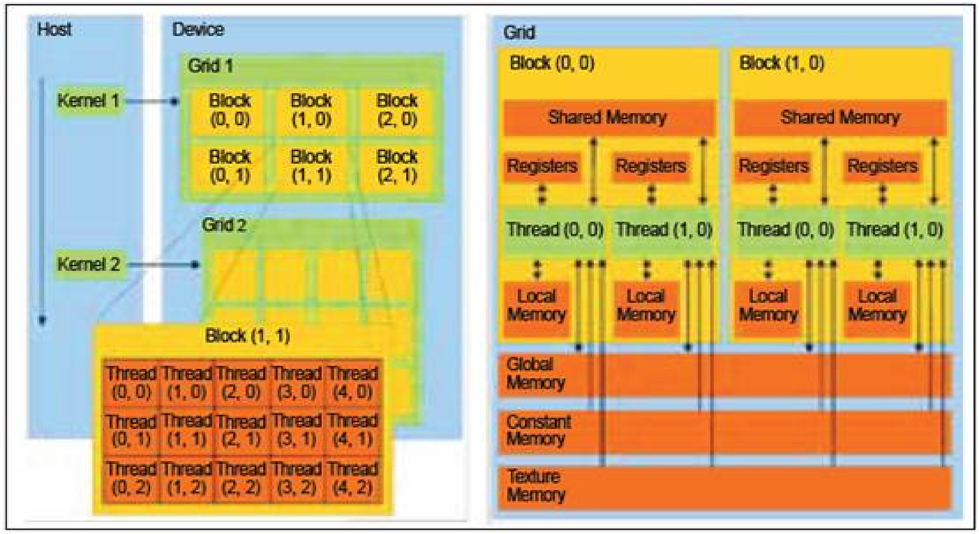


Figure 1. CUDA System Model

The platform allows various arrangements of threads to perform work, according to the developer's problem decomposition. In general, individual threads are grouped into up-to 3-dimensional blocks to allow sharing of common memory between threads. These blocks can then further be organized into a 2-dimensional grid. The GPU breaks the total number of threads into groups called warps, which, on current GPU hardware, consist of 32 threads that will be executed simultaneously on a single streaming multiprocessor (SM). The GPU consists of several SMs which are each capable of executing a warp. Blocks are scheduled to SMs until all allocated threads have been executed. There is also a memory hierarchy on the GPU. Three of the various types of memory are relevant to this work: global memory is the slowest and largest; shared memory is much faster, but also significantly smaller; and a limited number of registers that each SM has access to. Each thread in a block can access the same section of shared memory.

4. PARALLELIZATION

The algorithm used to parallelize the RSA modulo function works as follows:

- CPU accepts the values of the message and the key parameters.
- CPU allocates memory on the CUDA enabled device and copies the values on the device
- CPU invokes the CUDA kernel on the GPU
- GPU encrypts each message character with RSA algorithm with the number of threads equal to the message length.
- The control is transferred back to CPU
- CPU copies and displays the results from the GPU.

As per the flow given above the kernel is so built to calculate the cipher text $C = M^{e}$mod n. The kernel so designed works efficiently and uses the novel algorithm for calculating modulo value. The algorithm for calculating modulo value is implemented such that it can hold for very large power of numbers which are not supported by built in data types. The modulus value is calculated using the following principle:

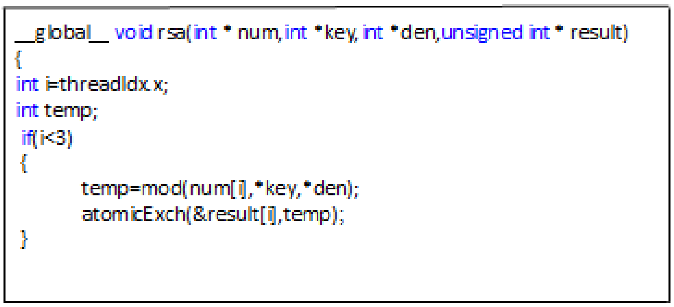
- ***$C = M^e \bmod n$***
- ***$C = (M^{e-x} \bmod n * M^x \bmod n) \bmod n$***

Hence iterating over a suitable value of x gives the desired result.

4.1. Kernel code

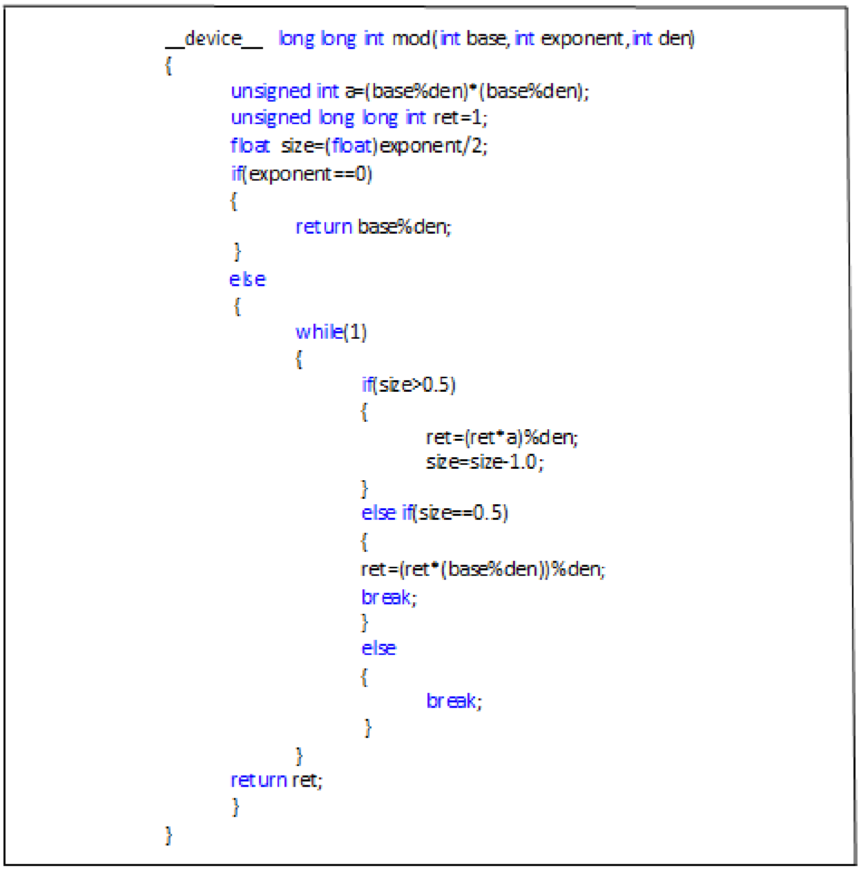
As introduced in section 2, RSA algorithm divides the plaintext or cipher text into packets of same length and then apply encryption or decryption transformation on each packet. A question is how does a thread know which elements are assigned to it and are supposed to process them? CUDA user can get the thread and block index of the thread call it in the function running on device. In this level, the CUDA Multi-threaded programming model will dramatically enhanced the speed of RSA algorithm. The experimental results will be showed in section 5.

The kernel code used in our experiment is shown below. First CUDA user assign the thread and block index, so as to let each thread know which elements they are supposed to process. It is shown in Figure 2. Then it call for another device function to calculate the most intense part of the RSA algorithm. Note in the below figure2 and figure3, it works for 3 elements.



```
__global__ void rsa(int * num,int *key,int *den,unsigned int * result)
{
int i=threadIdx.x;
int temp;
 if(i<3)
 {
        temp=mod(num[i],*key,*den);
        atomicExch(&result[i],temp);
 }
```

Figure 2. Kernel code



```
__device__  long long int mod(int base, int exponent,int den)
{
        unsigned int a=(base%den)*(base%den);
        unsigned long long int ret=1;
        float size=(float)exponent/2;
        if(exponent==0)
        {
                return base%den;
        }
        else
        {
                while(1)
                {
                        if(size>0.5)
                        {
                                ret=(ret*a)%den;
                                size=size-1.0;
                        }
                        else if(size==0.5)
                        {
                        ret=(ret*(base%den))%den;
                        break;
                        }
                        else
                        {
                                break;
                        }
                }
        return ret;
        }
}
```

Figure 3. Kernel’s Device code

5. VERIFICATION

In this section we setup the test environment and design three tests. At first test, we develop a program running in traditional mode for small prime numbers (only use CPU for computing). And at the second test, we use CUDA framework to run the RSA algorithm for small prime numbers in multiple-thread mode. Comparison is done between the two test cases and speed up is calculated. In the third test we run the GPU RSA for large prime numbers that is not supported by the built-in data types of CPU RSA. The test result will be showed in this section

5.1. Test environment

The code has been tested for :

- Values of message between 0 and 800 which can accommodate the complete ASCII table
- 8 bit Key Values

The computer we used for testing has an Intel(R) Core(TM) i3-2370M 2.4GHZ CPU, 4 GB RAM, Windows 7OS and a Nvidia GeForce GT 630M with 512MB memory, and a 2GHZ DDR2 memory. At the first stage, we use Visual Studio 2010 for developing and testing the traditional RSA algorithm using C language for small prime numbers. Series of input data used for testing and the result will be showed later.

At the second stage, we also use Visual Studio 2010 for developing and testing parallelized RSA developed using CUDA v5.5 for small prime numbers. After that the results of stage one and stage second are compared and hence calculating the respective speedup.

In the third test we run the GPU RSA for large prime numbers that is not supported by the built-in data types of CPU RSA. The test result will be showed in this section. At present the calculation of Cipher text using an 8-bit key has been implemented parallel on an array of integers.

5.2 Results

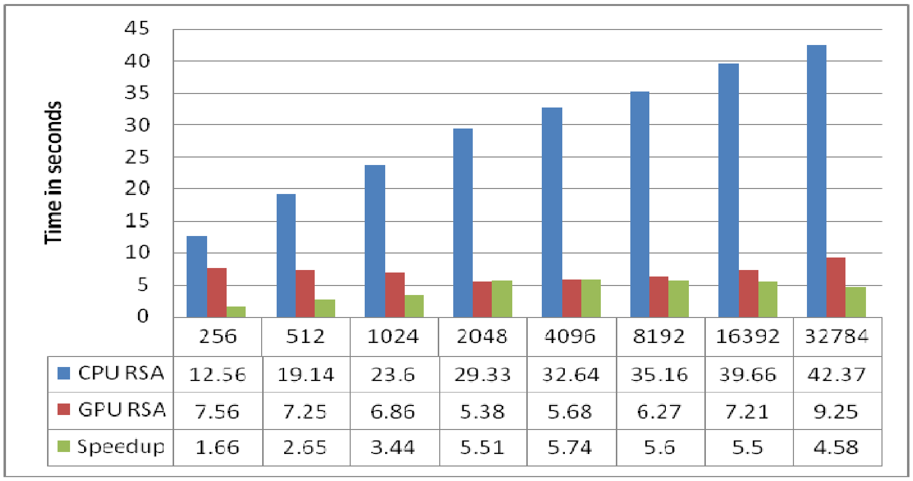
In this part, we show the experiment results for GPU RSA and CPU RSA for small value of n.

Table 1. Comparison of CPU RSA and GPU RSA for small prime numbers i.e (n=131*137)

Data Size(bytes)	No. of blocks	Threads per block	GPU RSA Time	CPU RSA Time	Speedup
256	4	64	7.56	12.56	1.66
512	8	64	7.25	19.14	2.65
1024	16	64	6.86	23.60	3.44
2048	32	64	5.38	29.33	5.51
4096	64	64	5.68	32.64	5.74
8192	128	64	6.27	35.16	5.60
16392	256	64	7.21	39.66	5.50
32784	512	64	9.25	42.37	4.58

Table 1 shows the relationship between the amount of data inputting to the RSA algorithm and the execution times (in seconds) in traditional mode and multiple thread mode. The first column shows the number of the data input to the algorithm, and the second column shows the number of blocks used to process the data input. In the above table 64 threads per block are used to execute RSA. The execution time is calculated in seconds. In the last column speed up is calculated. Above results are calculated by making average of the results so taken 20 times to have final more accurate and precise results.

The enhancement of the execution performance using CUDA framework can be visually demonstrated by Fig 4.



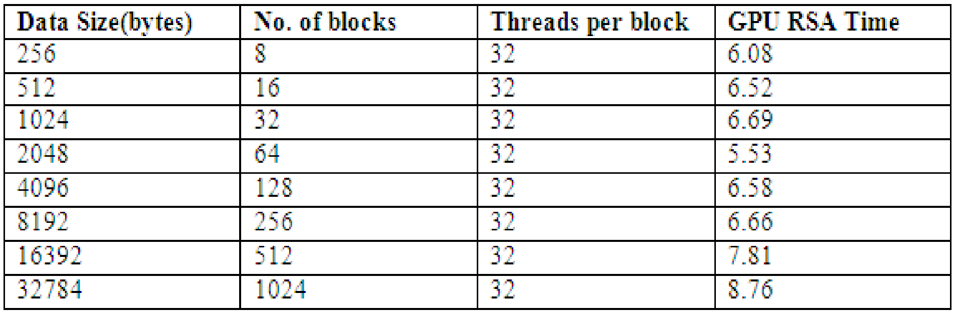
	256	512	1024	2048	4096	8192	16392	32784
🟦 CPU RSA	12.56	19.14	23.6	29.33	32.64	35.16	39.66	42.37
🟥 GPU RSA	7.56	7.25	6.86	5.38	5.68	6.27	7.21	9.25
🟩 Speedup	1.66	2.65	3.44	5.51	5.74	5.6	5.5	4.58

Figure 4. Graph showing effect of data input on CPU RSA and GPU RSA along with the Speedup

5.2.1 GPU RSA for large prime numbers

In this part, we show the experiment results for GPU RSA and CPU RSA for small value of n.

Table 2. GPU RSA for large prime numbers and large value of n (n = 1005 * 509)



Data Size(bytes)	No. of blocks	Threads per block	GPU RSA Time
256	8	32	6.08
512	16	32	6.52
1024	32	32	6.69
2048	64	32	5.53
4096	128	32	6.58
8192	256	32	6.66
16392	512	32	7.81
32784	1024	32	8.76

From Table 2, we can see the relationship between the execution time in seconds and the input data amount (data in bytes) is linear for certain amount of input. When we use 256 data size to execute the RSA algorithm, the execution time is very short as compared to traditional mode which is clearly proved in the above section where the comparison is made for CPU RSA and GPU RSA for small prime numbers and hence for small value of n. So we can say when the data size increases, the running time will be significantly reduced depending upon the number of threads used. Furthermore, we also find that when the data size increases from 1024 to 8192, the execution time of 7168 threads almost no increase, which just proves our point of view, the more the data size is, the higher the parallelism of the algorithm, and the shorter the time spent. Execution time varies according the number of threads and number of blocks used for data input. In the above table threads per block are constant i.e we use 32 threads per block and number of blocks used are adjusted according to the data input.

The enhancement of the execution performance of data input in bytes using the large value of prime numbers (n=1009 * 509) and hence large value of n on CUDA framework can be visually demonstrated by Figure 5.

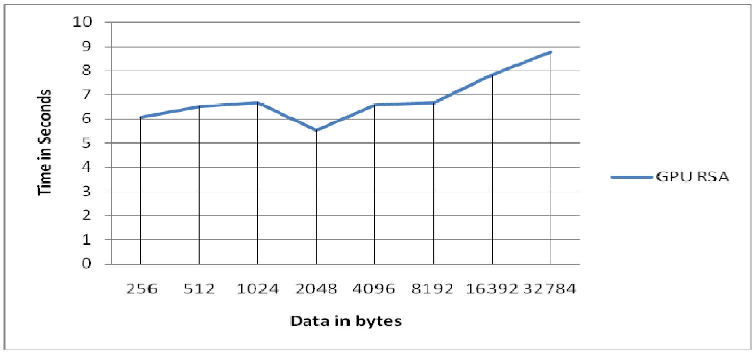


Figure 5. GPU RSA for large value of n (n=1009*509)

5.2.2. Execution time comparison of GPU RSA for large value of n (1009*509) with CPU RSA for small value of n(137*131)

In the third and final stage of test results analysis, we analyse our results between sequential RSA that is using small value of n (17947) and parallelized RSA that is making use of large prime numbers and large value of n (513581). The enhancement of the GPU execution performance of data input in bytes using the large value of prime numbers (n=1009 * 509) on CUDA framework and CPU RSA using small value of prime numbers (n=137*131) can be visually demonstrated by Figure 6. Hence, we can leverage massive-parallelism and the computational power that is granted by today's commodity hardware such as GPUs to make checks that would otherwise be impossible to perform, attainable.

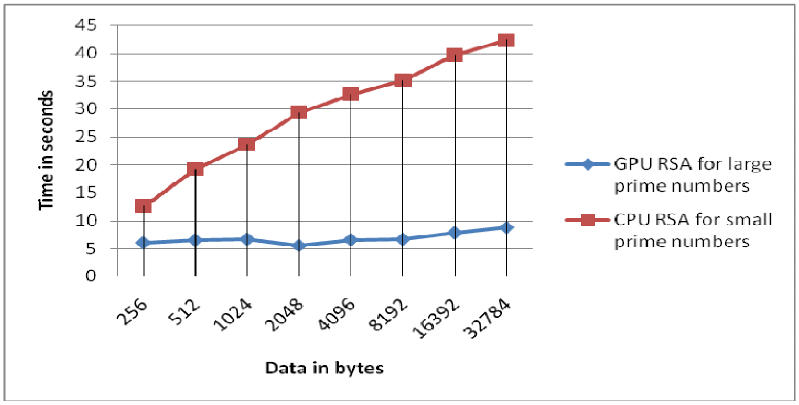


Figure 6. Comparison of CPU RSA for small prime numbers with GPU RSA for large prime numbers.

6. RSA Decryption Using CUDA-C

In this paper, we presented our experience of porting RSA encryption algorithm on to CUDA architecture. We analyzed the parallel RSA encryption algorithm. As explained above the encryption and decryption process is done as follows:

$$C = M^e \text{ mod n}, M = C^d \text{ mod n}.$$

The approach used for encryption process is same for decryption too. Same kernel code will work for decryption too. The only parameters that will change is the private key (d) and ciphertext in place of message bits used during encryption.

7. Conclusions

In this paper, we presented our experience of porting RSA algorithm on to CUDA architecture. We analyzed the parallel RSA algorithm. The bottleneck for RSA algorithm lies in the data size and key size i.e the use of large prime numbers. The use of small prime numbers make RSA vulnerable and the use of large prime numbers for calculating n makes it slower as computation expense increases. This paper design a method to computer the data bits parallel using the threads respectively based on CUDA. This is in order to realize performance improvements which lead to optimized results.

In the next work, we encourage ourselves to focus on implementation of GPU RSA for large key size including modular exponentiation algorithms. As it will drastically increase the security in the public-key cryptography. GPU are becoming popular so deploying cryptography on new platforms will be very useful.

REFERENCES

[1] R. L. Rivest, A. Shamir, and L. Adleman. A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM, 21(2):120{126, 1978.

[2] J. Owens, D. Luebke, N. Govindaraju.. A survey of general-purpose computation on graphics hardware. Computer Graphics Forum, 26(1): 80{113 , March 2007.

[3] N. Heninger, Z. Durumeric, E. Wustrow, and J. A. Halderman. Mining your Ps and Qs: detection of widespread weak keys in network devices. In Proceedings of the 21st USENIX conference on Security symposium, pages 205{220. USENIX Association, 2012 .

[4] NVIDIA|CUDA documents |Programming Guide |CUDA v5.5.
http://docs.nvidia.com/cuda/pdf/CUDA_C_Programming_Guide.pdf

AUTHORS

Sonam Mahajan
Student , ME – Information Security
Computer Science and Engineering Department
Thapar University
Patiala-147004

Dr. Maninder Singh
Associate Professor
Computer Science and Engineering Department
Thapar University
Patiala-147004

INTENTIONAL BLANK

SURVEY ON INFORMATION HIDING TECHNIQUES USING QR BARCODE

Manoj S. Rewatkar[1] and Shital A. Raut[2]

[1,2]Department of Computer Science and Engineering,
Visvesvaraya National Institute of Technology,
Nagpur, India
manojrewatkar143@gmail.com
saraut@cse.vnit.ac.in

ABSTRACT

Nowadays, the information processing system plays crucial part in the internet. Online information security has become the top priority in all sectors. Failing to provide online information security may cause loss of critical information or someone may use or distribute such information for malicious purpose. Recently QR barcodes have been used as an effective way to securely share information. This paper presents the survey on information hiding techniques which can share high security information over network using QR barcode.

KEYWORDS

QR Barcode, Information Hiding, Online information Security.

1. INTRODUCTION

Due to tremendous growth in communication technology, sharing the information through the communication network has never been so convenient. Nowadays information is processed electronically and conveyed through public networks. Such networks are unsecured and hence sensitive information needs to be protected by some means. Cryptography is the study of techniques that allows us to do this. In order to protect information from various computer attacks as well as network attacks various cryptographic protocols and firewalls are used. But no single measure can ensure complete security.

Nowadays, the use of internet and sharing information are growing increasingly across the globe, security becomes a vital issue for the society. Security attacks are classified as passive attacks and active attacks [11, 12]. In passive attacks, attacker monitors network traffic and looks for sensitive information but does not affect system resources. Passive attacks include traffic analysis, eavesdropping, Release of message contents [11, 12]. In active attack, attacker breaks protection features to gain unauthorized access to steal or modify information. Active attacks include masquerade, replay, modification of messages, and denial of service [11, 12].Therefore, security threats (such as eavesdropping, data modification, phishing, website leaks etc.) force us to develop new methods to counter them. Considering QR barcodes as an effective media of sharing information, many researchers have proposed information/data hiding methods [6,7, 8, 9.] as well as online transaction systems [1,2,3,4,5] using QR barcode. In this paper, we describe different information hiding schemes using QR barcode.

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 243–249, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4525

This paper is organized as follows: Section 2 gives details about QR barcode and their features. Section 3 gives details of different information hiding methods using QR barcodes and section 4 compares these methods. Section 5 presents our conclusion.

2. BACKGROUND

QR Code, also known as "Quick Response" [10] code, is a two dimensional matrix barcode that can store over 1800 characters of text information. QR Barcodes contain PDF 417 for its high data capacity, Data Matrix for its high density printing and MAXI Code for its high speed reading as shown in fig 1.

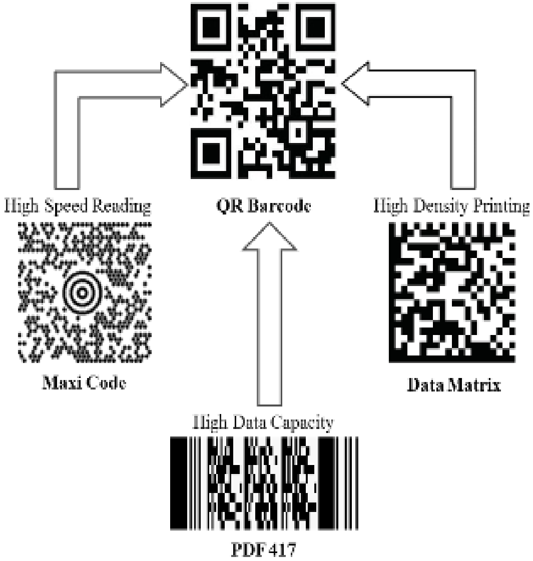


Fig.1.The formation of QR Code

QR Codes are capable of handling of data such as numbers, alphanumeric characters, Kanji, Kana, binary and control codes [10]. A QR code can store information [10] such as:

- Website URL
- SMS
- Text message
- Calendar event
- Contact Information
- Phone number
- Geographic location

2.1. Structure of QR Barcode

QR code consists of the functionality patterns for making it easily decodable. QR code has a position pattern for detecting the position of code, alignment pattern for correcting distortion, and timing pattern for identifying the central coordinate of each cell in the QR code. Quiet zone is the margin space for reading the QR code and the data area where the data is stored [10].

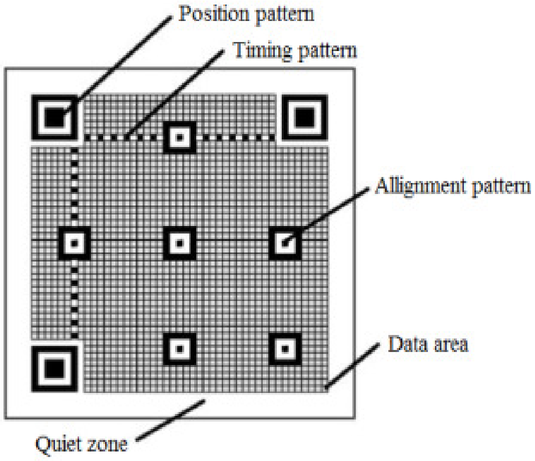


Fig.2. the internal pattern Structure of QR code

2.2. Features of QR Barcode

2.2.1. High Encoding Capacity

QR Barcode is capable of handling hundred times more data than conventional barcode. Conventional barcode has capacity to store maximum 20 digits [14]. While for QR code, up to 7,089(Numeric),4,296(Alphanumeric),2,953(Binary/byte),1,817(kanji/kana)characters can be encoded in one symbol.

2.2.2. Small Size

QR Barcode stores information in both horizontal and vertical fashion. QR Code is capable of storing the same amount of information in one-tenth the space of a conventional barcode [14].

2.2.3. Dirt and Damage resistant capability

QR Code has four different error correction levels, detailed as follows [14].

- L - Allows recovery of up to 7% damage.
- M - Allows recovery of up to 15% damage
- Q - Allows recovery of up to 25% damage
- H - Allows recovery of up to 30% damage

The error correction level can be selected by the user when he/she creates the symbol depending on how much damage the QR code is expected to suffer in its usage environment.

2.2.4. Structure linking functionality

QR Code has a structure appending functionality which will enable a single QR code to be represented in several symbols by dividing it as presented in fig 3. A single symbol can be divided into up to 16 symbols [14].

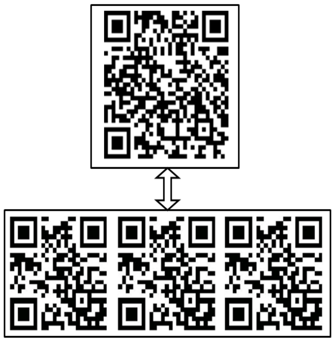


Fig 3.The association of the Symbols

2.2.5. The Confidentiality of the QR Code

The QR code can be easily encrypted and no one will be able to read the data until QR code is deciphered.

3. INFORMATION HIDING METHODS USING QR BARCODE

3.1. Using Hash function

Authors of [6] proposed an information hiding method using QR barcode. In this Method, information which is to be transmitted is first encrypted by using hash function, with a secret key K. The key K is known in advance to both sender as well as receiver. After the encryption process; QR code for encrypted information is created and sent over the network for the receiver. If an intruder were to try to extract the information from QR code, he/she would only be able to read the code with a QR code decoder but would not be able to get the secret information from QR code. Only the authorized user with secret key K can retrieve the secret information from QR code. The scheme is able to encode large amounts of secret information into a QR code based on selection of the QR version and the error correction level. The main disadvantage is that the whole secrecy of this scheme depends on key K. If someone gets the key, this scheme can reveal the secret information by simply decoding the QR code.

3.2. Using TTJSA symmetric key Algorithm

Authors of [7] proposed an encrypted information hiding mechanism using QR barcode. In this method, information which is to be transmitted is first encrypted using TTJSA symmetric key algorithm. For encrypted information, QR code is generated by using QR generator [15]. If an intruder tries to extract the information from QR code then he cannot do that because the cryptographic key is unknown to him. The decryption process is exactly reverse of the encryption process. TTJSA algorithm is free from attacks such as differential attacks, plain-text attacks or brute force attacks.

3.3. SD-EQR

Author of [8] presents a new technique using QR barcode to transfer information securely through public network. In this method, the password is entered along with the information. The secret key generated from the password which acts as the key for encryption process. The process of generating secret key is:

- Choose password of any size, but should consist of only ASCII characters (0-255).
- Find the length of the entered password denoted by “L”.
- Multiply ‘L2’ with the sum of the ASCII values of each letter of the word entered in the password to get S.
- Each digit of the S is added with each other. The ultimate sum is the secret key.

This secret key will be added to each character in the text entered in the information and complete the first phase of encryption process. After doing the first level of encryption, many other several encryption techniques are used to encrypt the message further to increase the level of security. At last final encrypted information is encoded into QR code. QR code efficiently handles the 1,264 characters of ASCII text in version 40 with Error correction level H. if encrypted information size is larger than capacity of QR code then other QR code is generated containing encrypted information after 1,264 characters. This method is continued until the whole encrypted information is converted into QR codes. Decryption is actually the reverse process of the encryption.

3.4. Using reversible data hiding

Authors of [9] propose a new algorithm in reversible data hiding, with the application associated with the QR code. Reversible data hiding is a new technique to hide data. During encoding process, data is hidden into original image. Hidden data and original image should be perfectly recovered during decoding process. The secret information which is to be conveyed is first encoded into QR code. At the lower portion of the original image, the pixels in this region are replaced by QR code. While decoding, the QR code is first removed from the image and original information can be recovered with reversible data hiding techniques from the rest of the image. During encoding process, the information in original image might be lost due to replacement of the corner portion of the original image with the QR code. The authors used reversible data hiding techniques to hide pixels in the corner portion of the original image into the rest of the original image in advance. The detailed process of information embedding and extraction by using reversible data hiding techniques is well explained in [10].

4. COMPARISON CHART

Table1. Comparison between Different Information Hiding Methods

Methods	**Using Hash function**	**Using TTJSA symmetric key Algorithm**	**SD-EQR**	**Using reversible data hiding**
Basic Application	Secret hiding	Secret hiding	Secret hiding	Image hiding
Computational Complexity	Low	Low	High	Low
Processing On QR code	No	No	No	No
Utilizing the error correction capability	Yes	Yes	Yes	No

Encryption on Data before embedding into QR code	Yes	Yes	Yes	No
Hiding Mechanism	Encrypted data embedded into QR Barcode			QR barcode of data embedded into cover image

5. CONCLUSION

This paper describes QR barcode and its use in different information hiding techniques. Such techniques employ traditional information hiding mechanisms like hash functions, image steganography, symmetric key algorithms, etc. in conjunction with QR barcodes. SD-EQR makes use of user entered password to formulate a private key and generates a QR barcode of the encrypted information. Finally the paper compares these techniques.

6. ACKNOWLEDGEMENTS

I take this opportunity to express my heartfelt gratitude to my guide Mrs Shital A Raut, Assistant Professor, VNIT Nagpur and my confederate Mr Viraj Jagtap for the time and efforts they have contributed to my work.

REFERENCES

[1] Kaushik S., “Strength of Quick Response Barcodes and Design of Secure Data Sharing System” International Journal on Advanced Computing & Science (IJACSA), Dec 2011.

[2] Kaushik S.; Puri S., “Online Transaction Processing using Sensitive Data Transfer Security Model”4th International Conference on Electronics Computer Technology (ICECT), IEEE, April. 2012.

[3] Suresh Gonaboina, Lakshmi Ramani Burra, Pravin Tumuluru,”Secure QR-Pay System With Ciphering Techniques In Mobile Devices” International Journal of Electronics and Computer Science Engineering.

[4] Jaesik Lee, Chang-Hyun Cho, Moon-Seog Jun,’’Secure Quick Response Payment(QR-Pay) System using Mobile Device”, Feb 2011.

[5] Sana Nseir, Nael Hirzallah, Musbah Aqel, ”A Secure Mobile Payment System using QR Code”, 5th International Conference on Computer Science and Information Technology (CSIT), 2013.

[6] Pei-Yu Lin, Yi-Hui Chen, Eric Jui-Lin Lu and Ping-Jung Chen ”Secret Hiding Mechanism Using QR Barcode”, International Conference on Signal-Image Technology & Internet-Based Systems, 2013.

[7] Somdip Dey, Asoke Nath, Shalabh Agarwal, ”Confidential Encrypted Data Hiding and Retrieval Using QR Authentication System”, International Conference on Communication Systems and Network Technologies, 2013.

[8] Somdip Dey,”SD-EQR: A New Technique To Use QR Codes in Cryptography” Use of QR Codes In Data Hiding and Securing.

[9] H. C. Huang, F. C. Chang and W. C. Fang, ”Reversible data hiding with histogram-based difference expansion for QR Code applications,” IEEE Transactions on Consumer Electronics, vol. 57, no. 2, pp. 779-787, 2011

[10] "QR Code, Wikipedia", http://en.wikipedia.org/wiki/QR_code [Online] .

[11] Cryptography & Network Security, Behrouz A. Forouzan, Tata McGraw Hill Book Company.

[12] Cryptography and Network Security, William Stallings, Prentice Hall of India.

[13] www.tldp.org/HOWTO/Secure-Programs-HOWTO/crypto.html.

[14] http://www.qrcode.com/en.

[15] http://zxing.appspot.com/generator.

AUTHORS

Mr. Manoj S. Rewatkar has received his **B.Tech** degree from Dr.B.A.T.U. Lonere Dist. Raigad (M.H.) He is currently pursuing **M.Tech** from Visvesvaraya National Institute of Technology in the department of computer science and engineering in Nagpur (M.H.) India.

Mrs Shital A. Raut is **Assistant Professor** at Visvesvaraya National Institute of Technology Nagpur. Her areas of interest include Data Mining and Warehousing, Business Information Systems.

INTENTIONAL BLANK

REALIZATION AND DESIGN OF A PILOT ASSIST DECISION-MAKING SYSTEM BASED ON SPEECH RECOGNITION

Jian ZHAO, Hengzhu LIU, Xucan CHEN and Zhengfa LIANG

School of Computer, National University of Defense Technology,
410073 Changsha, China
zhaojian9014@gmail.com

ABSTRACT

A system based on speech recognition is proposed for pilot assist decision-making. It is based on a HIL aircraft simulation platform and uses the microcontroller SPCE061A as the central processor to achieve better reliability and higher cost-effect performance. Technologies of LPCC (linear predictive cepstral coding) and DTW (Dynamic Time Warping) are applied for isolated-word speech recognition to gain a smaller amount of calculation and a better real-time performance. Besides, we adopt the PWM (Pulse Width Modulation) regulation technology to effectively regulate each control surface by speech, and thus to assist the pilot to make decisions. By trial and error, it is proved that we have a satisfactory accuracy rate of speech recognition and control effect. More importantly, our paper provides a creative idea for intelligent human-computer interaction and applications of speech recognition in the field of aviation control. Our system is also very easy to be extended and applied.

KEYWORDS

SPCE061A, Speech Recognition, DTW, Pilot Assist Decision-Making

1. INTRODUCTION

Speech recognition is a technology which is used to implement an appropriate control through correctly identifying and judging the speech characteristics and connotation [1]. In recent years, the applications of speech recognition technology in the fields like human-computer interaction have become more and more popular and challenging. As a very important technological progress, the pilot assist decision-making based on speech recognition can reduce burden on pilot, lower operating risk, and improve cockpit human-machine interface [2]. However, domestic application of speech recognition is still in a big blank at present. It's a great help to carry out pre-research in time, to understand and to master the technology, to overcome the application difficulties for improving the application level of our aviation control technologies.

Currently, DSP (Digital Signal Processor) chips are mainly applied to speech recognition. But they are generally more expensive, more complex and harder to be extended and applied [3]. The system proposed in our paper is realized with the HIL aircraft simulation platform and the 16-bit microcontroller SPCE061A. SPCE061A acts as the central processor for digital speech recognition to achieve better reliability and higher cost-effect performance. Technologies of LPCC and DTW are applied for isolated-word speech recognition to gain a smaller amount of

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 251–261, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4526

calculation and a better real-time performance. Besides, we adopt the PWM regulation technology to effectively regulate each control surface by speech, and thus to assist the pilot to make decisions.

The rest of the paper is organized as follows: algorithm of speech recognition is described in detail in the second part; hardware structure and software design of a pilot assist decision-making system based on speech recognition are respectively elaborated in Part III and Part IV; the performance of our whole system is evaluated in Part V; in the last part, we draw a summary and look forward to the future work.

2. ALGORITHM OF SPEECH RECOGNITION

As can be seen in Figure 1, speech recognition is essentially a kind of pattern recognition, which consists of basic units such as pretreatment, A/D conversion, endpoint detection, feature extraction and recognition judgment, [4] etc. According to the basic principle of pattern recognition, by comparing the pattern of the unknown speech with the reference pattern of the known, we can obtain the best matched reference pattern, namely, the result of recognition.

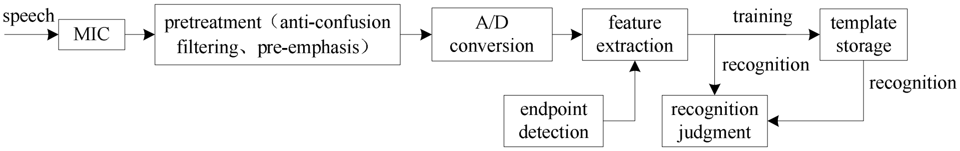


Figure 1. The basic structure of the speech recognition system

2.1. Endpoint Detection

Endpoint detection means using the digital processing technology to identify the start point and the end point among all kinds of paragraph (phonemes, morphemes, words, syllables, etc.) in the speech signal. The most commonly used method in endpoint detection is based on short-term energy and short-term zero-crossing rate [5, 6].

Short-term energy is defined as follows:

$$E_n = \sum_{m=-\infty}^{\infty} \left[x(m) \cdot w(n-m)\right]^2 \tag{1}$$

Here E_n reflects the law of the amplitude or energy of the voiced/unvoiced frames in the speech signal changing slowly over time [7]. According to the change of E_n, we can roughly judge the moment when voiced frames turn into unvoiced ones and the unvoiced frames turn into voiced ones. E_n is very sensitive to the high-level signal because the square of it was used when calculated formula (1). So in practice, we also use the following two types of definition:

$$E_n = \sum_{m=-\infty}^{\infty} \left|x(m) \cdot w(n-m)\right| \tag{2}$$

$$E_n = \sum_{m=-\infty}^{\infty} \log^2 \left[x(m) \cdot w(n-m)\right] \tag{3}$$

Short-term zero-crossing rate is defined as follows:

$$Z_n = \sum_{m=-\infty}^{\infty} \left| \mathrm{sgn}\left[x(n)\right] - \mathrm{sgn}\left[x(n-1)\right] \right| \cdot w(n-m) \tag{4}$$

sgn[•] is the sign function:

$$\mathrm{sgn}(n) = \begin{cases} 1, x(n) \geq 0 \\ 0, x(n) < 0 \end{cases} \tag{5}$$

$$w(n) = \begin{cases} \frac{1}{2N}, 0 \leq n \leq N-1 \\ 0, else \end{cases} \tag{6}$$

Z_n means the total number that the speech signal changes from positive to negative and from negative to positive per unit time [8]. According to Z_n, we can roughly obtain the spectral characteristics of the speech signal to distinguish the voiced/unvoiced frames and whether there's speech or not.

A two-stage judgment method is usually adopted in endpoint detection based on E_n - Z_n. As can be seen in Figure 2, firstly, select a relatively high threshold M1 according to the outline of E_n, in most cases, M1 is below E_n. In this way, we can do a rough judgment: the start point and the end point of the speech segment are located outside of the interval corresponding to the intersection points of envelops of M1 and E_n (namely, outside of segment AB). Then determine a relatively low threshold M2 based on the average energy of the background noise, and search from point A to the left, point B to the right, find out the first two intersection points C and D of the envelop of E_n and the threshold M2, so segment CD is the speech segment determined by the dual-threshold method according to E_n. From above we just finished the first stage of judgment, then turn to the second: use Z_n as the standard, and search from point C to the left, point D to the right, then find out the first two points E and F which are lower than threshold M3, so they are the start point and the end point of the speech segment.

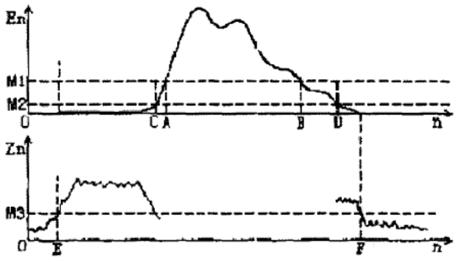


Figure 2. Endpoint detection based on E_n-Z_n

2.2. Feature Extraction

Feature extraction is a crucial step in speech recognition. If the speech features were effectively extracted, it is easier for us to distinguish among different categories in the feature space. Compared to the other speech features, linear predictive cepstral coding (LPCC) can effectively represent the speech feature, including channel characteristics and auditory features. LPCC has an excellent distinguishability, speed and accuracy. Besides, LPCC can effectively guarantee the real-time performance of speech recognition. LPCC is calculated based on linear prediction coefficient (LPC) characteristics:

$$\begin{cases} c(1) = a(1) \\ c(n) = \sum_{k=1}^{n-1} \left[1 - \frac{k}{n}\right] \cdot c(n-k) \cdot a(k) + a(n) \end{cases} \tag{7}$$

c(n) is the coefficient of LPCC, (n=1,2,…,p); p is the feature model order, most channel models of the speech signal can be sufficiently approximated when we take p=12; *a(k)* is the linear prediction coefficient (LPC) characteristics.

2.3. Recognition Judgment

We apply DTW algorithm, which is the commonly used identification method in speech recognition, to the recognition judgment part.

The basic idea of DTW is to find out the phonological characteristics of the speech signal and compare the distance， that is, to find out the differences (characteristic differences) between the frame characteristics in chronological order; And then accumulate characteristic differences included in phonological features and divided by the whole characteristic difference of the entire pronunciation. At last, we get the relative cumulative characteristic difference. Thus, despite the different articulation rate, the relative cumulative characteristic differences of the phonological characteristics are basically the same. Specific algorithm is as follows:

(I) normalize the feature data *L(i, j)* (the coefficient of LPCC) per frame, and get *S(i, j)*, the characteristic difference between two adjacent frames is:

$$t(j) = \sum_{i=1}^{c} \left| s(i,j) - s(i,j+1) \right| \tag{8}$$

The average characteristic difference is:

$$t = \frac{1}{N-1} \sum_{j=1}^{N-1} t(j) \tag{9}$$

N is the number of speech frames.

(II) Check *t(j)* from back to front, remove the ones that larger than the average characteristic difference, until it's less than the average characteristic difference, so as to remove the end part that contains less semanteme. The number of data frames reduced to *N'*. Assume the cumulative characteristic difference threshold is:

$$\Delta = \frac{1}{M} \sum_{j=1}^{N'} t(i) \qquad M \le N' - 1 \tag{10}$$

M is the number of reserved key frames. Usually we take *M*=8 for isolated character sound, and *M*=16 for double isolated words.

(III) Pick out the key frames: the first frame must be chosen, then plus *t(i)* in turn, the frame greater than △ is another key frame, until *M* key frames are picked out.

(IV) Piecewise linearization, and take the average of the characteristic differences between two key frames as the last speech eigenvector. In the process of training, save these feature vectors as templates. And in the process of recognition, match the speech signals and the templates, and calculate the distances, the minimum one within the scope of the distance threshold is the final recognition result.

3. HARDWARE STRUCTURE

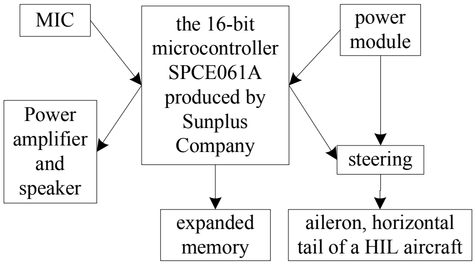

MIC
the 16-bit microcontroller SPCE061A produced by Sunplus Company
power module
Power amplifier and speaker
steering
expanded memory
aileron, horizontal tail of a HIL aircraft

Figure 3. Hardware block diagram of the system

As can be seen in Figure 3, the hardware structure of the pilot assist decision-making system based on speech recognition mainly includes a microprocessor circuit module based on a 16-bit microcontroller SPCE061A produced by Sunplus Company, expanded memory, audio circuit module, power module, steering and executive components.

SPCE061A contains multiple A/D converters, dual-channel 10-bit D/A converters and an online simulation circuit ICE interface. Besides, SPCE061A has the advantages of smaller size, higher integration, better reliability and cost-effective performance, easier to be expanded, stronger interrupt processing ability, more efficient instruction system and less power consumption, etc. than DSP chips [9]. In order to achieve real-time processing of speech signal, the whole hardware system is divided into the following several parts:

(I) Extraction, training, and judgment of speech features: we use speech processing and DSP functions of SPCE061A to pre-emphasis on the input speech digital signals, then cache and extract feature vectors, create templates under the function of training, and make judgment under the function of recognition.

(II) Acquisition of speech signals: SPCE061A has a microphone amplifier and single-channel speech A/D converters with the function of automatic gain control so that we could save much front-end processing hardware, simplify the circuit, and improve the stability. Connect the microphone to the anti-aliasing filter and access to the channel, and then complete sampling of the 10-bit 8 kHz signal.

(III) Expand memory: we need to expand a flash memory of 32 KB as the data memory because processing of speech signals requires a large amount of data storage. The storage space is divided into 4 parts: templates storage area is used to store isolated-word feature templates, and the number of stored templates (namely, the number of identifiable vocabulary) is determined by the size of the storage area; speech signal temporary storage area is used to store 62 frames of data of each speech signal to be identified; intermediate data storage area contains a 2 KB SRAM, and it’s used to store the intermediate computation, such as background noise characteristics and intermediate templates produced in the process of training, etc. the prompt speech information storage area is used to store function prompts speech and recognition response speech, etc. so as to facilitate human-computer interaction. Input of this part of speech signals can be achieved by the software wave_press provided by Sunplus Company.

(IV) Control input: consist of three keys, they are function switch key (recognition/study), function confirms and exit key, template revise and select key (select and modify a template). Through these three keys, we could realize the human-computer interaction of FS (Function Selection).

(V) Output: include a speaker output part and a control output part. The speaker is connected to a dual-channel 10-bit D/A converter with the function of audio output, and is used to output prompts speech and recognition response speech. The control output part is used to output a control signal through I/O interface, and then adjust corresponding steering, change flight attitudes, to realize assistant decision-making after having recognized speech instructions.

4. SOFTWARE DESIGN

Our system's software implement is developed in the integrated developing environment IDE3.0.4 of SPCE061A based on C language, which mainly includes three parts: main program, interrupt handling routine and function module subroutine. We will introduce the three parts in detail as follows.

4.1. Main Program

As can be seen in Figure 4, the processes of the main program are divided into initialization, training and recognition. The training and recognition of the speaker-dependent speech could be accomplished by calling related functions, and the corresponding operations could be performed according to the results of recognition.

(I) Initialization: The system collects 10 frames of background noise data at first after power on reset, extracts features of En and Zn after pre-emphasis, and then determines the threshold value as the basis to identify the start point and the end point.

(II) Training: enter by the function switch key, and prompt "now is the training function". And then prompt "now modify the first template", select the template to be modified by the template revise and select key, after per click it turns to the next template. And then prompt "speech input at the first time", you are asked to input 4 times here in order to ensure the accuracy of the template. Extract the feature vectors and temporary store the template. Only after 4 times all succeeded would it prompt "successfully modified". Otherwise, data won't be retained, and template won't be modified, neither. If one process lasted for more than 10s, it would prompt "Quit the training function" and do not make any changes.

(III) Recognition: The microcontroller constantly sample the outside signal, and save 10 frames of speech data to judge the start point; and then sample 52 frames of speech data to determine the end point. Handle the error if there's no end point. After that, calculate LPCC of each frame, and use LPCC and DTW to get the eigenvectors of isolated words in that speech segment. Compare them with the templates, if the distance is within the specified threshold, select the template with the minimum distance as the result. At the meantime, a corresponding response is made. However, if the distance is beyond that threshold, handle the error.

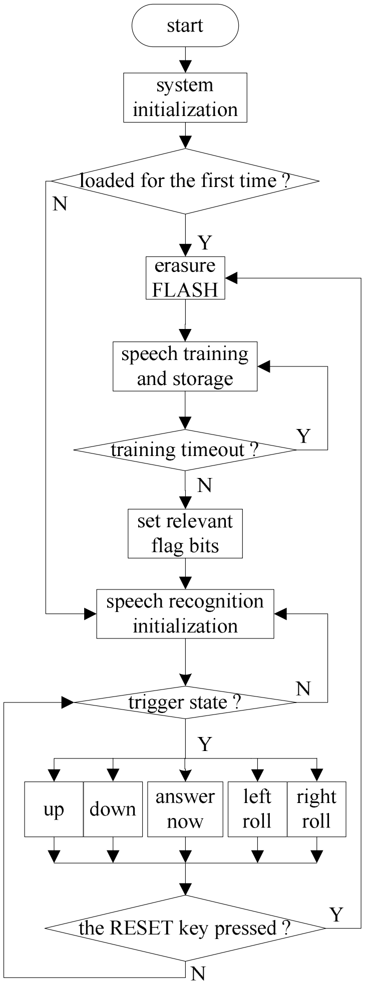


Figure 4. Software flow of the main program

4.2. Interrupt Handling Routine

As can be seen in Figure 5, A/D conversion results are read by the interrupt handling routine periodically and deposited in the buffer. The speech signal of MIC channel is the input of A/D. Interrupts are generated by speech recognition and playback TMA_FIQ interrupt sources, and judged by the flag bit whether it’s speech playback or speech recognition [10, 11]. Functions written in the process of speech recognition are: start point judgment function, end point judgment function, LPC function, LPCC function, characteristic differences piecewise linear dynamic distribution function, judgment function, error function, and the upper functions, feature extraction functions constructed by these sub-functions.

4.3. Function Module Subroutine

The function module subroutine includes the functions of up, down, left roll, right roll, reset and output of PWM wave, etc. In each function, corresponding operation is realized by configuring I/O output to provide related signal to the circuit of steering.

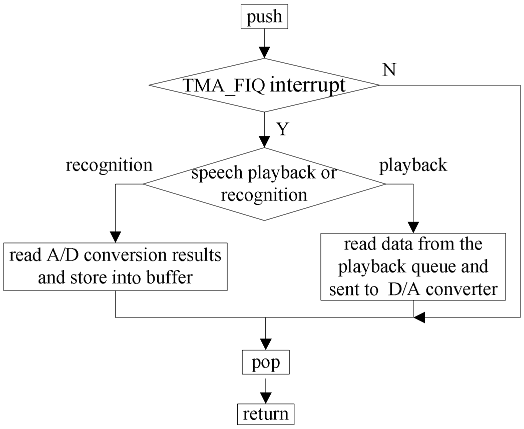

push
TMA_FIQ interrupt
N
Y
recognition
speech playback or recognition
playback
read A/D conversion results and store into buffer
read data from the playback queue and sent to D/A converter
pop
return

Figure 5. Software flow of the interrupt handling routine

The steering is a position servo actuator, and applies to those control systems in need of the ever-changing and maintaining angle [12]. The control of steering usually needs a time-base pulse around 20ms, and the high level part of the pulse is generally 0.5ms-2.5ms. Take the servo of 180 degrees angle for example, the corresponding control relationships are shown in Table 1.

Table 1. Control relationship between pulse and pivot angle of steering.

Variables	**Negative**	**Negative**	**Zero**	**Positive**	**Positive**
High level	0.5ms	1.0ms	1.5ms	2.0ms	2.5ms
Angle	0	45	90	135	180

In the drive program of steering, we take the angle of 90 degrees corresponding to 1.5ms as the initial position of the system, and realize the zero declination of control surface through the reset function. By changing the duty ratio of the output PWM wave in the functions of up, down, left roll and right roll, to control the positive and negative angle of steering, and thus, to control the aircraft flight attitudes. Figure 6 shows the output of PWM wave when the high level part is 1.5ms.

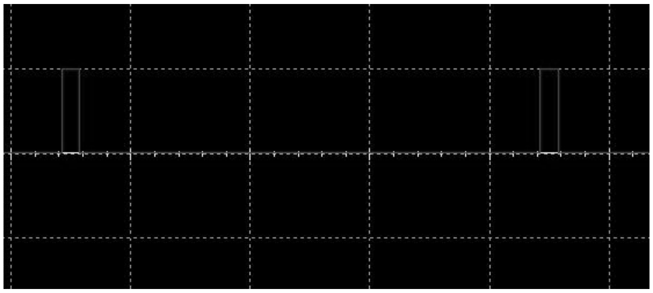


Figure 6. Output of PWM wave when the high level part is 1.5ms

5. System Performance Analysis

For the same training template, let the speaker dependent and the speaker independent respectively test the system based on a HIL aircraft simulation platform, 20 times test for each command and 100 times test for each group. The results show that: the recognition rate of speaker dependent has reached more than 95.3%, the recognition rate of the speaker-independent A is

81.2%, the recognition rate of the speaker-independent B is 85.7%; then select male speech as the template, and use female speech test the system, the recognition rate is 54.5%. From the results we know that the recognition rate of speaker dependent is higher than speaker independent. Besides, a higher recognition rate could be obtained if a more precise algorithm was taken during template matching.

In the aspect of aircraft flight attitudes control, we used SOLIDWORKS design and simulate the movements of the corresponding control surface. As can be seen in Figure 7, control parts move the front end of the aircraft horizontal tail upward after system recognized the instruction “down” of the pilot.

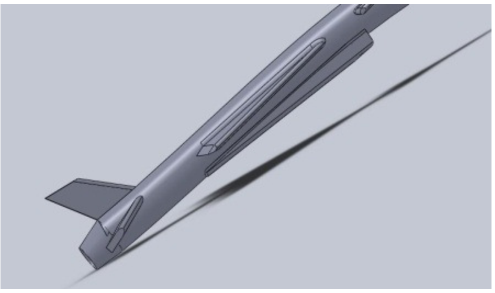


Figure 7. Aircraft down

As can be seen in Figure 8, control parts move the trailing end of the aircraft left aileron upward and the right aileron downward after system recognized the instruction “left roll” of the pilot.

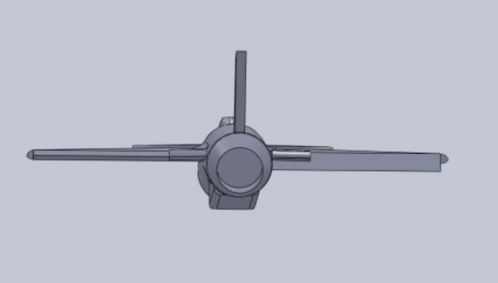


Figure 8. Aircraft left roll

By trial and error, the movements of each control surface under the corresponding speech instructions are in line with expectations and the overall performance is better than that gained by other methods.

6. CONCLUSIONS

As a very important technological progress, the pilot assist decision-making based on speech recognition can reduce burden on pilot, lower operating risk, and improve cockpit human-machine interface [13]. However, domestic application of speech recognition is still in a big blank at present. It’s a great help to carry out pre-research in time, to understand and to master the technology, to overcome the application difficulties for improving the application level of our aviation control technologies.

The system proposed in our paper is realized with the HIL aircraft simulation platform and the 16-bit microcontroller SPCE061A. SPCE061A acts as the central processor for digital speech recognition to achieve better reliability and higher cost-effect performance. And an artificial intelligence system is introduced in the control system of aircraft to achieve more flexible control and better human-computer interaction. Besides, speech recognition is optimized by certain mechanical structures and algorithms. Speech features and recognition methods fit for speaker-dependent isolated word are selected to achieve faster processing speed and higher recognition rate, so as to meet the needs of real-time speech recognition [14, 15]. Our system made the best advantages of speech control and realized a system for assisting the pilot to make decisions. By trial and error, it is proved that we have a satisfactory accuracy rate of speech recognition and control effect.

ACKNOWLEDGEMENTS

This work is supported by the Doctor Program Foundation of Education Ministry of China, under grant No. 20114307130003. We thank Haibin DUAN, Hongyu YAN and Chi MA from Beijing University of Areonautics and Astronautics in China for their valuable comments to improve this paper.

REFERENCES

[1] Lavner Y, Gath I, Rosenhouse J.: The effects of acoustic modificationson the identification of familiar voices speaking isolated vowels. Speech Communication (2000).

[2] Chu M K, Sohn Y S.: A user friendly interface operated by the improved DTW method. In: 10th IEEE International Confe-rence on Fuzzy Systems (2002).

[3] LEE C H.: On Automatic Speech Recognition at the Dawn of 21th Century. IEICE TRANS, INF & SYST (2003).

[4] Lee, D. Hyun, etc.: Optimizing feature extraction for speech recognition. IEEE Transactions on Speech and Audio Processing (2003).

[5] X. Huang, A. Acero, H. W. Hon.: Spoken language processing-a guide to theory, algorithm, and system development (2003).

[6] Zbancioc M, Costin M.: Using neural networks and LPCC to improve speech recognition signals. Proceedings of the International Symposium on Circuits and Systems (2003).

[7] SU JAY P, RH ISH IKESH L, SIDDH ARTH V.: On design and implementation of an embedded automatic speech recognition system. IEEE Transactions on VLSI Design (2004).

[8] Turner CW, Gantz BJ, Vidal C, et al.: Speech recognition in noise for cochlear implant listeners: benets of residual acoustic hearing. The Journal of the Acoustical Society of America (2004).

[9] Osamu Segawa, Kazuya Takeda, Fumitada Itakura.: Continuous speech recognition without end-point detection. Electrical Engineering (2006).

[10] Abu Shariah, Mohammad A. M. Ainon, Raja N. Zainuddin, Roziati. Khalifa, Othman O.: Human computer interaction using isolated-words speech recognition technology. 2007 International Conference on Intelligent and Advanced Systems, ICIAS 2007 (2007).

[11] J.H.L Hansen, L. M. Arslan.: Robust feature estimation and objective quality assessment for noisy speech recognition using the credit card corpus. IEEE Trans. Speech Audio Processing (2009).

[12] Je-Keun Oh, Giho Jang, Semin Oh, Jeong Ho Lee, Byung-Ju Yi, Young Shik Moon, Jong Seh Lee, Youngjin Choi.: Bridge inspection robot system with machine vision. Automation in Construction (2009).

[13] Rabiner. L. R.: An algorithm for determing the endpoints of isolated utterance. The Bell System Technical Journal (2010).

[14] Mayumi Beppu, Koichi Shinoda, Sadaoki Furui.: Noise Robust Speech Recognition based on Spectral Reduction Measure. APSIPA ASC (2011).

[15] Microsoft Speech SDK 5.1 Help. http://www.Microsoft.com.

AUTHORS

Jian ZHAO received the B.S. degree in Automatic Control and Information Technology from Beijing University of Aeronautics and Astronautics of China in 2012, and he is currently working toward the M.S. degree in Computer Science and Technology in National University of Defense Technology. His research interests are in the area of computer system architecture, software radio communication, and signal processing with huge data.

Hengzhu LIU received the Ph.D. degree in Computer Science and Technologyfrom National University of Defense Technology of China in 1999. Currently, he is a professor and vice leader with the Institute of Microelectronics and Micro-processors of computer school in National University of Defense Technology. His research interests include processor architecture, micro-architecture, memory architecture, system on chip (SoC), VLSI signal processing and high-performance digital signal. He is a member of IEEE, IEICE and China Computer Federation.

Xucan CHEN received the Ph.D. degree in Computer Science and Technologyfrom National University of Defense Technology of China in 1999. Currently, she is a professor of computer school in National University of Defense Technology. Her research interests include computer system architecture,micro-architecture, software radio communication, VLSI signal processing and high-performance digital signal.

Zhengfa LIANG received the M.S. degree in Electrical Science and Technology from National University of Defense Technology of China in 2012. He is currently working toward the Ph.D degree in Electrical Science and Technology in National University of Defense Technology. His research interests are in the area of wireless communication, parallel processing architectures, and circuits.

INTENTIONAL BLANK

A FUZZY INFERENCE SYSTEM FOR ASSESSMENT OF THE SEVERITY OF THE PEPTIC ULCERS

Kianaz Rezaei[1], Rahil Hosseini[1,2], and Mahdi Mazinani[1,2]

[1]Department of Computing, Kharazmi University, Tehran, Iran
[2]Faculty of Engineering, Islamic Azad University, Share-Qods Branch, Tehran, Iran
kianaz.rezae@gmail.com, rahilhosseini@gmail.com

ABSTRACT

Peptic ulcer disease is the most common ulcer of an area of the gastro- intestinal tract. The aim of this study is to utilize soft computing techniques to manage uncertainty and imprecision in measurements related to the size, shape of the abnormality. For this, we designed a fuzzy inference system (FIS) which emulates the process of human experts in detection and analysis of the peptic ulcer. The proposed approach models the vagueness and uncertainty associated to measurements of small objects in low resolution images In this study, for the first time, we applied soft computing technique based upon fuzzy inference system (FIS) for assessment of the severity of the peptic ulcer. Performance results reveal the FIS with maximum accuracy of 98.1%, which reveals superiority of the approach. The intelligent FIS system can help medical experts as a second reader for detection of the peptic ulcer in the decision making process and consequently, improves the treatment process.

KEYWORDS

Soft computing, Fuzzy inference system (FIS), Peptic ulcer.

1. INTRODUCTION

The second common cause of death from malignant disease is gastric cancer around the world. Detection and treatment of this painful disease has become one of the challenging medical problems.

Nowadays gastric ulcer is one of the most important concerns involves many factors especially widespread using of NSAIDs. Because of poorly understanding the pathophysiology of this disease [6], studies investigating new active compounds are needed. As well, various pharmaceutical products currently used for treatment of gastric ulcers are not completely efficient and cause many adverse side effects.

Peptic ulcer disease encompassing gastric and duodenal ulcer is the most prevalent gastrointestinal disorder [1]. They are caused by various factors such as drugs, stress or alcohol, due to an imbalance between offensive acid- pepsin secretion and defensive mucosal factors like mucin secretion and cell shedding [2]. Gastric ulcer therapy faces a major drawback due to the unpredictable side effects of the long-term use of commercially available drugs. It is shown that toxic oxygen radicals plays an important role in the etiopathogenesis of gastric damage

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 263–271, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4527

[3].Currently, focus on plant research has increased all over the world and a large source of evidence has been collected to show immense potential of medicinal plants used in various traditional systems [4].

One of the main group of problems in medical science is related to diagnosing diseases based on different tests on patients. However, the final diagnosis of an expert is associated with difficulties. This matter led the physicians to apply computer aided detection and diagnosing tools in the recent decades.

A prime target for such computerized tools is in the domain of cancer diagnosis. Specifically, where breast cancer is concerned, the treating physician is interested in ascertaining whether the patient under examination exhibits the symptoms of a benign case, or whether her case is a malignant one [16].

The uncertainty issues in decisions making and medical diagnosis are related to incompleteness of medical science. computer aided detection (CAD) tools are presented with the purpose of facilitating the diagnosis of different diseases and acceleration of the treatment process [18]-[20]. One of the current applications of the CAD systems is to analysis the severity diagnosis of peptic ulcer presented in [10].

This study is concerned with the severity diagnosis of peptic ulcer and uses fuzzy inference systems for automatic diagnosis of disease. The required medical knowledge are aided by fuzzy systems and achieved data are from tested stomachs. This method assorts patients according to the length of ulcer.

The present study has been undertaken with the aim to assess the peptic ulcer severity using a fuzzy system.

2. LITERATURE REVIEW

In recent years, a major class of problems in medical science involves the diagnosis of disease, based upon various tests performed upon the patient. When several tests are involved, the ultimate diagnosis may be difficult to obtain, even for a medical expert. This has given rise, over the past few decades, to computerized diagnostic tools, intended to aid the physician in making sense out of the welter of data [16].

Soft Computing techniques based on the concept of the fuzzy logic or artificial neural networks for control problems has grown into a popular research area [11]-[13]. The reason is that classical control theory usually requires a mathematical model for designing controllers. The inaccuracy of mathematical modeling of plants usually degrades the performance of the controllers, especially for nonlinear and complex control problems. Fuzzy logic has the ability to express the ambiguity of human thinking and translate expert knowledge into computable numerical data.

A fuzzy system consists of a set of fuzzy IF-THEN rules that describe the input-output mapping relationship of the networks. Obviously, it is difficult for human experts to examine all the input-output data from a complex system to find proper rules for a fuzzy system. To cope with this difficulty, several approaches that are used to generate the fuzzy IF-THEN rules from numerical data have been proposed [11]-[13].

Today, medical endoscopy is a widely used procedure to inspect the inner cavities of the human body. The advent of endoscopic imaging techniques allow the acquisition of images or videos created the possibility for the development of the whole new branch of computer-aided decision support systems. This section summarizes related works specifically targeted at computer-aided decision support in the gastrointestinal tract [10].

A symbiotic evolution-based fuzzy-neural diagnostic system for common acute abdominal pain presents a symbiotic evolution-based fuzzy-neural diagnostic system (SE-FNAAPDS) for diagnosis of common acute abdominal pain (AAP) without professional medical examination [11]. The computer-assisted diagnostic system is formatted a multiple-choice symptom questionnaire, with a prompt/help menu to assist user in obtaining accurate symptom data using nothing more technologically sophisticated than a medical-type thermometer and stethoscope. Compared to traditional methods, diagnostic decisions from SE-FNAAPDS shows 94% agreement with professional human medical diagnosis and less CPU time for system construction. The presented method is useful as a core module for more advanced computer-assisted diagnostic systems, and for direct application in AAP diagnosis [11].

Non-ulcer dyspepsia (NUD) has been attributed to gastritis and Helicobacter infection in A Quantitative analysis of symptoms of Non-Ulcer Dyspepsia as related to age, pathology, and Helicobacter Infection. The Sydney classification enables dyspepsia symptoms assessed quantitatively in relation to Helicobacter infection and topographic pathology in different gastric compartments. The method presented in this study for 348 patients with the NUD. It studied the unconfounded effects of age, pathology, and Helicobacter. It was concluded that age was the most important determinant of dyspeptic symptoms, but not pathology or Helicobacter [7].
Computer-aided capsule endoscopy images evaluation based on color. Rotation and texture features were used as an educational tool to physicians.

Wireless capsule endoscopy (WCE) is a revolutionary, patient-friendly imaging technique that enables non-invasive visual inspection of the patient's digestive tract and, especially, small intestine. Experimental results demonstrated promising classification accuracy (91.1%) exhibiting high potential towards a complete computer-aided diagnosis system that will not only reduce the Wireless capsule endoscopy (WCE) data reviewing time, but also serve as an assisting tool for the training of inexperienced physicians [9].

3. Materials And Methods

Fuzzy set A in universe of discourse X can be defined as a set of ordered pairs of element x in X and the grade of membership of x, μ_A (x), to fuzzy set A [15] as follows:

$$A = \{(x, \mu_A(x)) | x \in X\}$$

where the two dimensional membership function μ_A (x) is a crisp value between 0 and 1 for all $x \in X$. Linguistic terms are modelled using fuzzy sets. One of the parameters in the design of a fuzzy logic is the number of fuzzy sets associated to a linguistic term. Fuzzy inference system as a soft computing method mimics cognitive reasoning of the human mind based on linguistic terms for performing tasks in a natural environments.

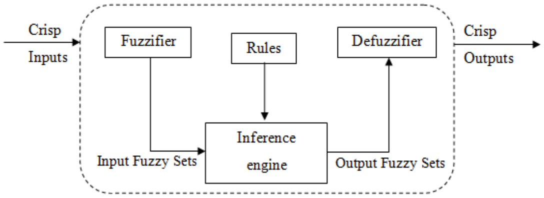


Figure 1. Architecture of a Fuzzy Inference System

The fuzzy inference system is a rule-based system that uses fuzzy logic, rather than Boolean logic, to reason about data. Its basic structure includes four main components, as depicted in Figure 1: (1) a fuzzifier, which translates crisp (real-valued) inputs into fuzzy values; (2) an inference engine that applies a fuzzy reasoning mechanism to obtain a fuzzy output; (3) a defuzzifier, which translates this latter output into a crisp value; and (4) a knowledge base, which contains both an ensemble of fuzzy rules, known as the rule base, and an ensemble of membership functions, known as the database[16].

The fuzzy inference system is a popular computing framework based on the concepts of fuzzy set theory, fuzzy if-then rules, and fuzzy reasoning. It has found successful applications in a wide variety of field, such as automatic control, data classification, decision analysis, expert systems, and pattern recognition.

This Mapping is accomplished by a number of fuzzy if-then rules, each of which describes the local behavior of the mapping. In particular, the antecedent of a rule defines a fuzzy region in the input space, while the consequent specifies the output in the fuzzy region.

Fuzzy logic models can be developed from expert knowledge or from process (patient) input-output data. In the first case, fuzzy models can be extracted from the expert knowledge of the process. The expert knowledge can be expressed in terms of linguistics, which is sometimes faulty and requires the model to be tuned. This process requires defining the model input variables and the determination of the fuzzy model system parameters.

Sugeno, or Takagi-Sugeno-Kang, method of fuzzy inference. Introduced in 1985, it is similar to the Mamdani method in many respects. The first two parts of the fuzzy inference process, fuzzifying the inputs and applying the fuzzy operator, are exactly the same. The main difference between Mamdani and Sugeno is that the Sugeno output membership functions are either linear or constant [21]. This study applies the Sugeno fuzzy inference model in order to present a measure of the sevirity of the peptic ulcer in the output of the FIS.

4. PEPTIC ULCER SPECIFICATION

This section explains the chemical process and animals used in laboratory experiments. The features extracted in the experiments were considered as the input of the FIS. These features are explained in details in this section.

The present study was tested on male Wistar rats for 15 days protected the gastric mucosa against the damage induced by indomethacin (25, 50 and 100 mg/kg) [17]. Male Wistar rats weighing 175 - 220 g were used in the study. The animals were in 6 separate groups consisting of 5 rats. The quantitative evaluation of experimentally induced gastric lesions is a problematic and error-prone task due to their predominantly multiple and irregularly shaped occurrence. The simplest type of lesion index for quantification of chemically induced ulcers were described as the cumulative length (mm) of all hemorrhagic erosions. The width of lesion has also been taken into account (ulcer index = length -width) [17]. Figure 2 shows Microscope views of a sample stomach of a rat with ulcer.

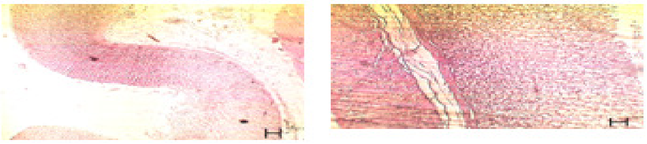


Figure 2. Microscope views of the rats stomach with ulcer

5. FUZZY MODELING OF THE FEATURE CHARACTERIZATION OF PEPTIC ULCER

In order to apply soft computing techniques based on the FIS for severity assessment of the peptic ulcer, we applied two methods as follows:

1) FCM (Fuzzy C-Means Clustering): for scatter partitioning of the input space and automatic generation of the membership functions
2) ANFIS (Adaptive Neuro-Fuzzy inference system) for learning the FIS rules and tuning of the membership functions

The rest of this section explains the abovementioned processes in details.

5.1. Scatter partitioning of the input space

Fuzzy c-means (FCM) is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. This technique was originally introduced by Jim Bezdek in 1981 as an improvement on earlier clustering methods. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters.

In this study we used Fuzzy Logic Toolbox in Matlab to implement the FIS. The FCM starts with an initial guess for the cluster centers, which are intended to mark the mean location of each cluster. The initial guess for these cluster centers is most likely incorrect. Additionally, fcm assigns every data point a membership grade for each cluster. By iteratively updating the cluster centers and the membership grades for each data point, FCM iteratively moves the cluster centers to the right location within a data set. This iteration is based on minimizing an objective function that represents the distance from any given data point to a cluster center weighted by that data point's membership grade Input features for assessment of the peptic ulcer are as Follows:

For each input and output variable of the FIS, three linguistic terms (Low, Medium and High) were considered. Table 1 shows all input variables of the peptic ulcer FIS.

Table 1. The FIS input variables

No.	Feature	Description
1	Score 1	Each fifth petechia was calculated as 1 mm
2	Score 2	lesion length between 1 and 2 mm
3	Score 3	lesion length between 2 and 4 mm
4	Score 4	lesion length between 4 and 6 mm
5	Score 5	lesion length more than 6 mm
6	Indomethacin	Explained in Section IV
7	Cimitidine	Explained in Section IV

The FIS output variable were considered ulcer index (UI) which represents severity of the peptic ulcer. The ulcer index (UI) was calculated using the following formula:

$$UI =(1*S1)+(2*S2)+(3*S3)+(4*S4)+(5*S5) \qquad (1)$$

where $S_{1,}$ $S_{2,}$ $S_{3,}$ $S_{4,}$ S_5 are related to the score 1 to score 5, respectively.

5.2. ANFIS (Adaptive Neuro-Fuzzy Inference System)

This syntax is the major training routine for Sugeno -type fuzzy inference systems. anfis uses a hybrid learning algorithm to identify parameters of Sugeno-type fuzzy inference systems. It applies a combination of the least-squares method and the backpropagation gradient descent method for training FIS membership function parameters to emulate a given training data set. ANFIS can also be invoked using an optional argument for model validation. We applied the ANFIS for learning rules in the FIS and tuning of the membership function parameters.

The flowchart of the approach applied for learning and tuning of the FIS parameters using the ANFIS approach is shown in Figure 3.

6. EXPERIMENTS RESULTS

In the process of the FIS parameter specification using the ANFIS model, we have a dataset including 30 real patients diagnosed with peptic ulcer information. We partitioned the dataset into two parts:

1) Training (70%)
2) Testing (30%)

Figures 4 to 9 represent the performance results on training and testing datasets in terms of the root mean square error (RMSE) and the histogram of the errors. The performance results are summarized in Table 2.

Table 2. System Performance on Train and Test datasets

	Accuracy
Average(train)	99.65%
Average(test)	97.74%

The rules and membership functions of the FIS for peptic ulcer risk assessment was designed using fuzzy c-means (FCM) clustering by extracting a set of rules that models the data behaviour using Fuzzy Logic Toolbox and ANFIS Toolbox in Matlab are shown in Figures 9 to 14. The rule extraction method first uses the FCM method to determine the number of rules and membership functions for the antecedents and consequents. Then ANFIS is applied to tune the FIS parameters. Table 3 shows the resulted FIS before training and after training process using the ANFIS approach. The RMSE was used as performance measure during evaluation process. The result of the RMSE and the histogram of the errors on train and test datasets are shown in Figures 3 to 8.

Table 3. Comparison of Results

Methods	Accuracy
FCM	94.9%
ANFIS	98.1%

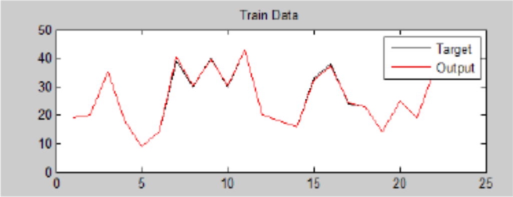


Figure 3. This section shows the Train data. Almost coincides target and Output.

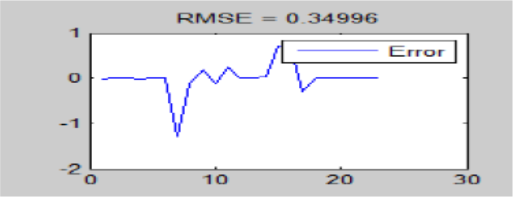


Figure 4. RMSE: Shows the maximum and minimum errors in the Train Data.

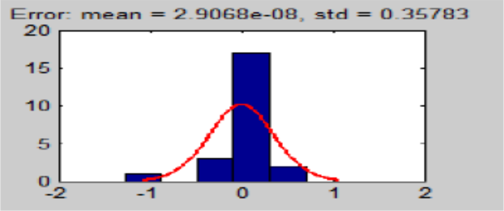


Figure 5. The third part is the histogram of the errors,and shows the mean and standard deviation of the error.

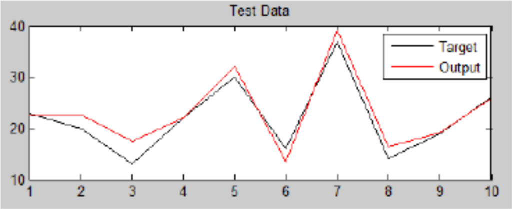


Figure 6. This section shows the Test data. There is between Output and Target.

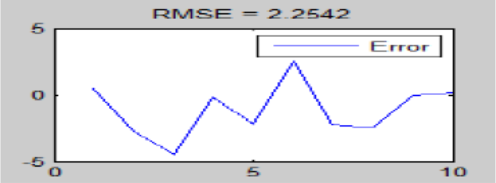


Figure 7. RMSE: Shows the maximum and minimum errors in the Test Data.

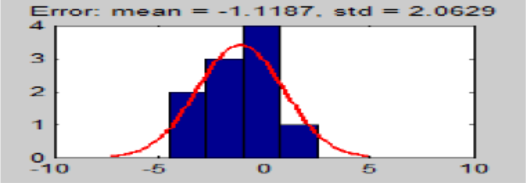


Figure 8. The third part is the histogram of the errors, and shows the mean and standard deviation of the error.

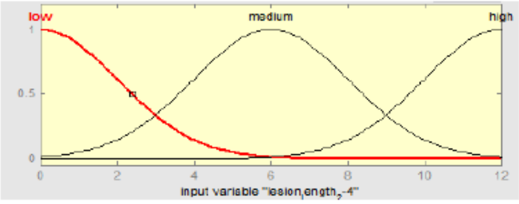


Figure 9. Membership functions related to the Score 3: lesion length between 2 and 4 mm

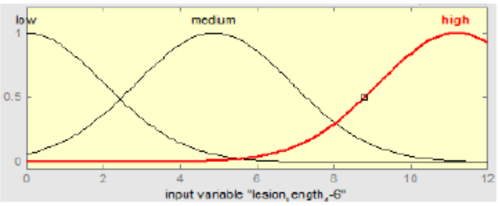


Figure 10. Membership functions related to the Score 4: lesion length between 4 and 6 mm

7. Conclusion

In this study, for the first time, a soft computing technique based upon fuzzy inference system (FIS) was proposed for the problem of peptic ulcer assessment. The FIS was generated using FCM and tuned using the ANFIS model. Performance results on a dataset including real patients reveal the FIS with maximum accuracy of 98.1%, which reveals superiority of the approach. The intelligent FIS system can help medical experts as a second reader for detection of the peptic ulcer in the decision making process and consequently, improves the treatment process.

REFERENCES

[1] Umashanker, M. and S. Shruti, Traditional Indian herbal medicine used as antipyretic, antiulcer, anti-diabetic and anticancer: A review. IJRPC, 2011. 1: 1152-1159.

[2] M. Shakeerabanu, K. Sujatha, C. Praveen-Rajneesh and A. Manimaran, "The Defensive Effect of Quercetin on In- domethacin Induced Gastric Damage in Rats," Advances in Biological Research, Vol. 5, No. 1, 2011, pp. 64-70.

[3] Cadirci, E., et al., Effects of< i> Onosma armeniacum</i> root extract on ethanol-induced oxidative stress in stomach tissue of rats. Chem Biol Int, 2007. 170(1): p. 40-48

[4] Anosike, C.A., et al., Anti-inflammatory and anti-ulcerogenic activity of the ethanol extract of ginger (Zingiber officinale). Af J Biochem Res, 2009. 3: 379-384.

[5] M. Khanavi, R. Ahmadi, A. Rajabi, S. Jabbari-Arfaee, G. Hassanzadeh, R. Khademi, A. Hadjiakhoondi and M. Sha- rifzadeh, "Pharmacological and Histological Effects of Centaurea bruguierana ssp. belangerana on Indometha- cin-Induced Peptic Ulcer in Rats," Journal of Natural Medicines, Vol. 66, No. 2, 2012, pp. 343-349.

[6] P. Malfertheiner, F. K. Chan and K. E. McColl, "Peptic Ulcer Disease," Lancet, Vol. 374, No. 9699, 2009, pp. 1449-1461.

[7] S. T.Lai , K.P.Fung, F.H. Ng. and K. C . Lee"A Quantitative Analvsis of Svmptoms of Non-Ulcer Dvspepsia as Related to Age pathology, and Helicobacter Infection" 1996,vol. 31,No.11,pages 1078-1082.

[8] G.Gopu, R.Neelaveni, K.Porkumaran "Noninvasive Technique for Acquiring and Analysis of Electrogastrogram" IJCA Special Issue on "Computer Aided Soft Computing Techniques for Imaging and Biomedical Applications" CASCT, 2010.

[9] Vasileios S. Charisis, Christina Katsimerou, Leontios J. Hadjileontiadis, Christos N. Liatsos, George D. Sergiadis "Computer-aided Capsule Endoscopy Images Evaluation based on Color".

[10] M. Liedlgrubera,_, A. Uhla "A Summary of Research Targeted at Computer-Aided Decision Support in Endoscopy of the Gastrointestinal Tract" Technical Report 2011-01.

[11] G. G. Towell and J. W. Shavlik, ―Extracting refined rules from knowledge-based neural networks,‖ Machine Learning, vol.13, pp.71-101, 1993.

[12] C. J. Lin and C. T. Lin, ―An ART-based fuzzy adaptive learning control network,‖ IEEE Transactions on Fuzzy Systems, vol.5, no.4, pp.477-496, 1997.

[13] C. F. Juang and C. T. Lin, ―An on-line self-constructing neural fuzzy inference network and its applications,‖ IEEE Transactions on Fuzzy Systems, vol.6, no.1, pp.12-31, 1998.

[14] Jyh-Shing Roger Jong, Chapter 4, Fuzzy Inference System (Jang, Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence.)

[15] Zadeh, Fuzzy Logic, Neural Networks, and Soft Computing-1994

[16] Carlos Andre´s Pen˜a-Reyes, Moshe Sipper, A fuzzy-genetic approach to breast cancer diagnosis, Artificial Intelligence in Medicine 17 (1999) 131–155.

[17] Ghazaleh Moghaddam1, Mohammad Sharifzadeh2, Gholamreza Hassanzadeh3, Mahnaz Khanavi4,5, MannanHajimahmoodi; Anti-Ulcerogenic Activity of the Pomegranate Peel (Punicagranatum) ,Published Online October 2013.

[18] M. Mazinani, J. Dehmeshki, R. Hosseini, S.D. Qanadli, "Afuzzy automatic method for measuring the volume of the stenosis by modeling thepartial volume problem in the coronary artery", USGG 2009, Lausanne, Switzerland(2009).

[19] RahilHosseini, J. Dehmeshki, S. A. Barman, MMazinani, S.D. Qanadli, "A Fuzzy Logic System for Classification ofthe Lung Nodule in Digital Images in Computer Aided Detection", In Proc.of Int. Conf. of Digital Society (ICDS), IEEE, September (2010), pp. 255- 259.

[20] RahilHosseini, S. D. Qanadli, S. Barman, M. Mazinani, T.Ellis, and J. Dehmeshki, "An Automatic Approach for Learning and Tuning Gaussian Interval Type-2 Fuzzy Membership Functions Applied to Lung CAD Classification System", IEEE Transactions on Fuzzy Systems, vol. 20, No. 2, April (2012), pp. 224-234.

[21] Sugeno, M., Industrial applications of fuzzy control, Elsevier Science Pub. Co., 1985.

AUTHORS

Kianaz Rezaei (born February, 1990), obtained her B.Sc. degree in Computers-Soft computing from Islamic Azad University, Share-e-Qods Branch, Tehran, Iran during 2012, and currently pursuing Masters Degree from Kharazmi University, Tehran, Iran.

Rahil Hosseini is assistant professor in software engineering and soft computing at Azad University. She is a member of the Digital Imaging Research Centre (DIRC) in the Faculty of Science, Engineering and Computing at Kingston University, London. She received her BSc at Teacher training University in Tehran followed by her MSc in software engineering. She commenced her PhD at Kingston University London. Her main area of interest in research is in the fields of soft computing, fuzzy modelling and pattern recognition in data mining problems and medical image analysis, specifically cancer diagnosis. She has professional work experience in the field of data mining, medical image analysis and fuzzy modelling for uncertain environments. She has published journal and conference papers in data mining, soft computing, distributed systems, fuzzy modelling, and pattern recognition for medical image analysis and cancer diagnosis.

Mahdi Mazinani is assistant professor in electronic engineering and image processing at Azad University. He is a member of the Quantitative Medical Imaging International Institutes (QMI3) and Digital Imaging Research Centre in the Faculty of Science, Engineering and Computing at Kingston University, London. After completing his BSc degree in Electronic engineering at Semnan University (securing the first rank) he completed his MSc degree in the Semnan University with first rank. He commenced his PhD. His main area of interest in research is in the field of medical image analysis. His work is focused on research into novel image analysis techniques to enable the recognition and quantification using CTA images.

He has published papers in the detection and quantification of the coronary arteries using fuzzy and medical image analysis techniques.

INTENTIONAL BLANK

FUZZY LOGIC MULTI-AGENT SYSTEM

Atef GHARBI[1] and Samir BEN AHMED[2]

[1]Department of Computer Engineering, INSAT, Tunis, Tunisia
atef.elgharbi@gmail.com
[2] Department of Computer Engineering, FST, Tunis, Tunisia
samir.benahmed@fst.rnu.tn

ABSTRACT

The paper deals with distributed planning in a Multi-Agent System (MAS) constituted by several intelligent agents each one has to interact with the other autonomous agents. The problem faced is how to ensure a distributed planning through the cooperation in our multi-agent system.

To do so, we propose the use of fuzzy logic to represent the response of the agent in case of interaction with the other. Finally, we use JADE platform to create agents and ensure the communication between them.

A Benchmark Production System is used as a running example to explain our contribution.

KEYWORDS

Multi-Agent System, Distributed Planning, Fuzzy Logic, JADE

1. INTRODUCTION

While Multi-Agent System (MAS) is a concept mainly used in research [23], by adapting it we must face various problems, some of which are serious enough to place the utility of MAS in the doubt. Since we wish to use the MAS in large scales, concurrent systems, and since we wish to address not very frequent, but demanding problems [24], MAS can become arbitrarily complex if MAS can not provide guarantees which help to order the system and ensure the progression of the total application.

We can not pretend the unicity nor the exactitude of an agent definition, however the most adapted one presented by [1] where an agent is defined as a physical or virtual entity (i) which is capable of acting in an environment; (ii) which can communicate directly with other agents; (iii) which is driven by a set of tendencies (in the form of individual objectives or of a satisfaction/survival function which it tries to optimize); (iv) which possesses resources of its own; (v) which is capable of perceiving its environment (but to a limited extent); (vi) which has only a partial representation of its environment (and perhaps none at all); (vii) which possesses skills and can offer services; (iix) which may be able to reproduce itself; (ix) whose behaviour tends towards satisfying its objectives, taking account of the resources and skills available to it and depending on its perception, its representation and the communications it receives.

In MAS, distributed planning is considered as a very complex task [3], [18]. In fact, distributed planning ensures how the agents should plan to work together, to decompose the problems into

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 273–290, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4528

subproblems, to assign these subproblems, to exchange the solutions of subproblem, and to synthesize the whole solution which itself is a problem that the agents must solve [19, 20, 4]. The actions of the other agents can induce a combinatorial explosion in the number of possibilities which the planner will have to consider, returning the space of research and the size of solution exponentially larger.

There are several techniques to reduce data-processing complexity of planning interactions with other agents including [22]: (i) dividing states in the classes of equivalence, (ii) reducing search space into states which are really required. (iii) planning on line, i.e., eliminating the possibilities which do not emerge during the execution of plan.

Our contribution in this research work is the use of another solution what is Fuzzy Logic Control. The Fuzzy Logic Control is a methodology considered as a bridge on the artificial intelligence and the traditional control theory [17]. This methodology is usually applied in the only cases when exactitude is not of the need or high importance [16]. Fuzzy Logic is a methodology for expressing operational laws of a system in linguistic terms instead of mathematical equations. Wide spread of the fuzzy control and high effectiveness of its applications in a great extend is determined by formalization opportunities of necessary behavior of a controller as a ”fuzzy” (flexible) representation [14]. This representation usually is formulated in the form of logical (fuzzy) rules under linguistic variables of a type ”If A then B” [12]. The Fuzzy Logic methodology comprises three phases: Fuzzyfication, Rule engine, Defuzzyfication [13].

This article is concerned with two important matters: how to define the MAS in a manner such that it has more utility to deploy it, and how to use such a MAS for the advanced software. The MAS must discover the action to be taken by supervising the application and its environment and analyzing the data obtained.

With MAS, we face two important matters: (i) the detection of a need for action. the need for action must be discovered by supervising the application and its environment and analyzing data obtained. (ii) the planning of the action. It consists to envisage the action (by proposing which modifications need to be made) and by programming it. In practice, the opposite dependency also requires consideration: Only those situations which can be repaired by an action taken which can really be planned should be considered during the analysis.

This paper introduces a simple Benchmark Production System that will be used throughout this article to illustrate our contribution which is developped as agent-based application. We implement the Benchmark Production System in a free platform which is JADE (JavaTM Agent DEvelopment) Framework. JADE is a platform to develop multi-agent systems in compliance with the FIPA specifications [5, 6, 2].

In the next section, we present the Benchmark Production System. The third section introduces the Fuzzy Multi-Agent System. We present in section 4 the creation of JADE agents.

2. BENCHMARK PRODUCTION SYSTEM

As much as possible, we will illustrate our contribution with a simple current example called RARM [11]. We begin with the description of it informally, but it will serve as an example for various formalism presented in this article. The benchmark production system RARM represented in the figure 1 is composed of two input and one output conveyors, a servicing robot and a processing-assembling center. Workpieces to be treated come irregularly one by one. The workpieces of type A are delivered via conveyor C1 and workpieces of the type B via the conveyor C2. Only one workpiece can be on the input conveyor. A robot R transfers workpieces one after another to the processing center. The next workpiece can be put on the input conveyor when it has been emptied by the robot. The technology of production requires that first one A-workpiece is inserted into the center M and treated, then a B-workpiece is added in the center, and last the two workpieces are assembled. Afterwards, the assembled product is taken by the robot and put above the C3 conveyer of output. the assembled product can be transferred on C3 only when the output conveyor is empty and ready to receive the next one produced.

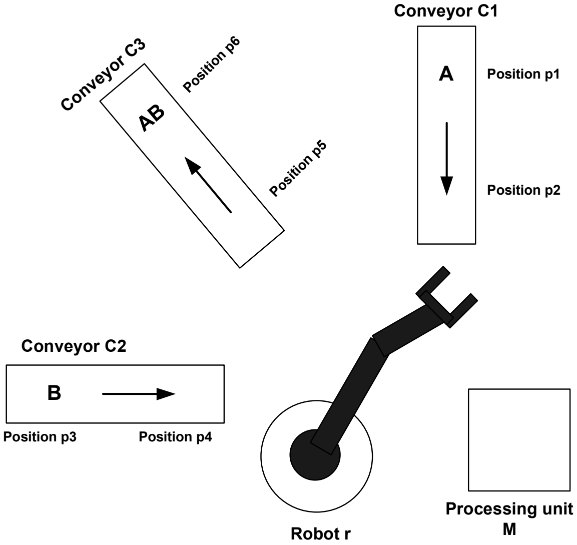


Figure 1. The benchmark production system RARM

Traditionally, the RARM systems are directly controlled by a central server. The server proposes the schedule for the system as a whole and dispatches commands to the robots. This results is reliable and predicable solutions. The central point of control also allows an easier diagnosis of the errors. However, a variation in user's needs leads to change the centralized architecture. Customers ask more and more for self-management system, i.e., systems that can adapt their behavior with changing circumstances in an autonomous way. Self-management with regard to the dynamics of system needs two specific quality requirements: flexibility and openess.

Flexibility refers to the capacity of the system to treat dynamic operating conditions. The openess refers to the capacity of the system to treat robots leaving and entering system.To treat these new quality requirements, a radically new architecture was conceived based on multi-agent systems (Figure 2).

Applying a situated multi-agent system opens perspective to improve the flexibility and the openess from the system: the robots can adapt to the current situation in their vicinity, order assignment is dynamic, the system can therefore treat in an autonomous way the robots leaving and reentring the system, etc.

However, a decentralized architecture can lead to a certain number of implications, in particular distributed planning can have an impact on the total efficiency of the system. In fact, this critical topic must be considered during the design and development of multi-agent system.

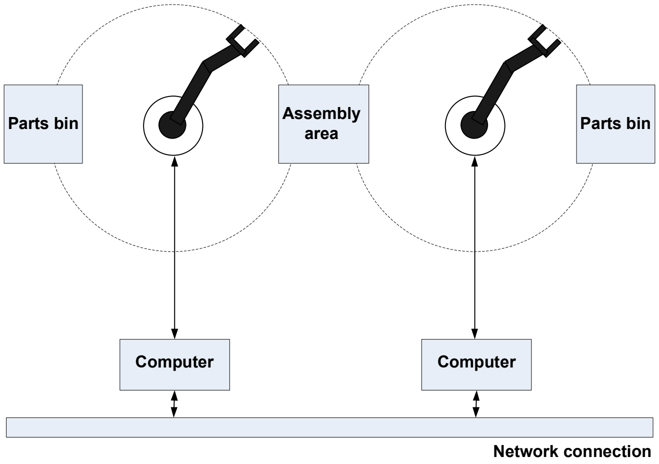


Figure 2. The distributed Production system

3. FUZZY MULTI-AGENT SYSTEM

Multi-agent planning problems can sometimes be translated into non deterministic single-agent planning problems by modifying the plan-execution agent's actions to incorporate the effects of the other agents' possible responses to those actions. For example, suppose an agent $RARM_1$ is going to reduce the production.

The another agent $RARM_2$ may either decrease the production (in which case the agents can cooperate together) or increase the production (in which case neither agent can cooperate). As shown in Figure 3, this two possible actions can be modeled as nondeterministic outcomes.

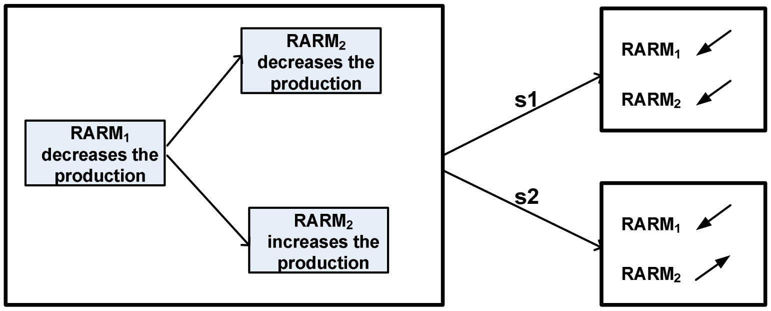

RARM₁ decreases the production
RARM₂ decreases the production
RARM₂ increases the production
s1
s2
RARM₁
RARM₂
RARM₁
RARM₂

Figure 3. Nondeterministic planning problem

The basic form of a fuzzy logic agent consists of: Input fuzzification, Fuzzy rule base, Inference engine and Output defuzzification (Figure 4).

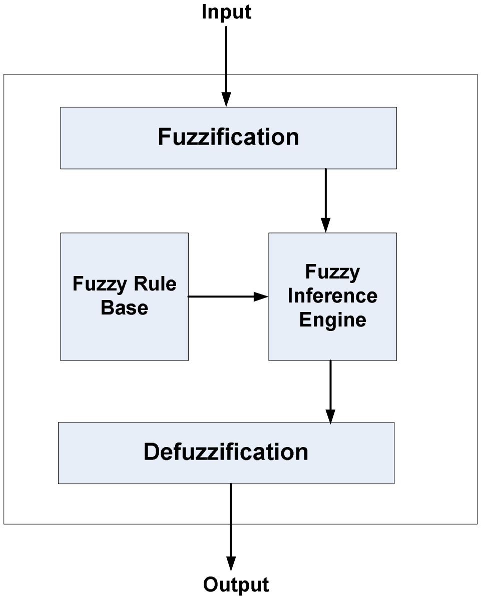


Figure 4. The Agent structure

3.1 Fuzzification

In the classical logic set, its characteristic function assigns a value of either 1 or 0 to each individual in the universal set, there by discriminating between members and non-members of the crisp set under consideration. However, a fuzzy set is a set containing elements that have varied degrees of membership in the set. The fuzzification can be defined as a conversion of a precise quantity to a fuzzy quantity.

Running example

The number of defected pieces is measured through a sensor related to the system. The range of number of defected pieces varies between 0 to 40, where zero indicates the rate of defected pieces of A that is null (each piece is well) and 40 indicates the rate of defected pieces of A is very high. Now assume that the following domain meta-data values for these variable, VF = very few, F = few, Md = medium, Mc = much, VMc = very much. Assume that the linguistic terms describing the meta-data for the attributes of entities are: VF = [0,..,10], F = [5,..,15], Md = [10,..,20], Mc = [15,..,25] and VMc = [20,..,40].

Based on the metadata value for each attribute the membership of that attribute to each data classification can be calculated. In the Figure 5 and 6, triangular and trapezoidal fuzzy set was used to represent the state of defected pieces from A classifications (i.e. state of defected pieces from A classification levels: VF , F, Md, Mc, VMc whereas state of defected pieces from B classification levels: F, Md, Mc).

In the figure 7, state of production system classification levels: Null, Low, Medium and High.

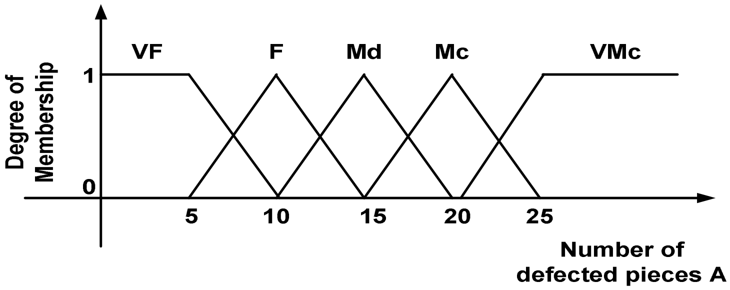


Figure 5. Fuzzy State of defected pieces from A

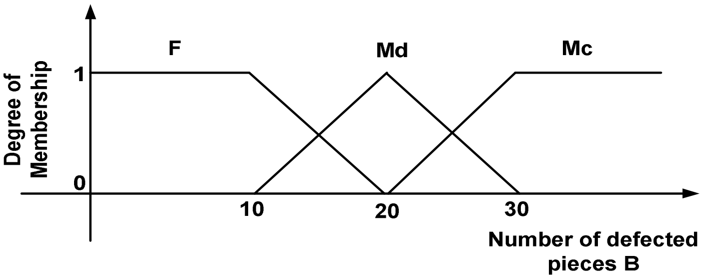


Figure 6. Fuzzy State of defected pieces from B

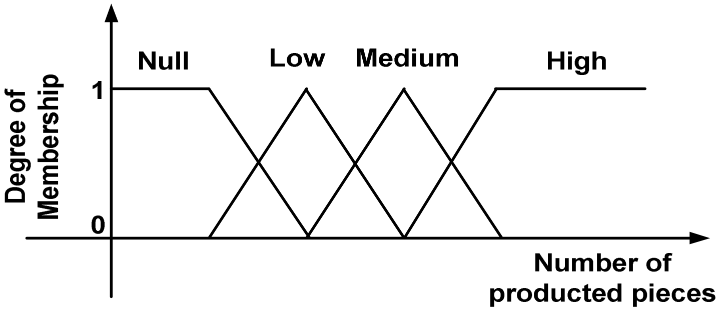


Figure 7. Fuzzy Production system

The membership value based on its meta-data can be calculated for all these classification using the formulas:

Formulas for calculation triangular fuzzy memberships

$$(1)\ \begin{cases} m_A(x) = 0 \text{ if } x < a_1, \\ m_A(x) = \dfrac{x-a_1}{a_2-a_1} \text{ if } a_1 \le x \le a_2, \\ m_A(x) = \dfrac{a_3-x}{a_3-a_2} \text{if } a_2 \le x \le a_3, \\ m_A(x) = 0 \text{ if } x > a_3 \end{cases}$$

Formulas for calculation trapezoidal fuzzy memberships

$$(2)\begin{cases} \mathrm{m_A}(\mathrm{x}) = 0 \text{ if } \mathrm{x} < \mathrm{a_1}, \\ \mathrm{m_A}(\mathrm{x}) = \dfrac{x-a_1}{a_2-a_1} \text{ if } \mathrm{a_1} \leq \mathrm{x} \leq \mathrm{a_2}, \\ \mathrm{m_A}(\mathrm{x}) = 1 \text{ if } \mathrm{a_2} \leq \mathrm{x} \leq \mathrm{a_3}, \\ \mathrm{m_A}(\mathrm{x}) = \dfrac{a_4-x}{a_4-a_3} \text{if } \mathrm{a_3} \leq \mathrm{x} \leq \mathrm{a_4}, \\ \mathrm{m_A}(\mathrm{x}) = 0 \text{ if } \mathrm{x} > \mathrm{a_4} \end{cases}$$

Running example

As an example, we consider the membership functions for the fuzzy variable defected pieces from A. Figure 5 shows various shapes on the universe of defected pieces from A. Each curve is a membership function corresponding to various fuzzy variables, such as very few, few, medium, much and very much (Figure 8).

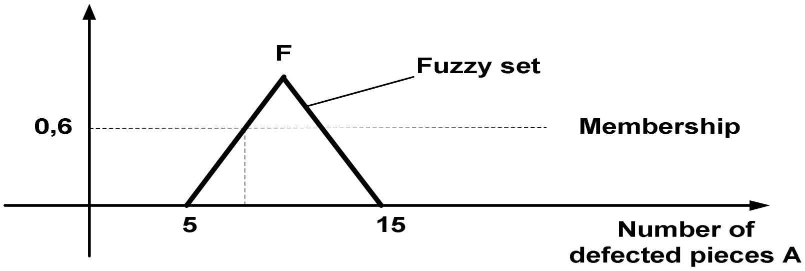


Figure 8. Membership function representing imprecision in number of defected pieces from A

3.2 Rule Engine

In the inference method we use knowledge to perform deductive reasoning. That is, we wish to deduce or infer a conclusion, given a body of facts and knowledge. Now that the data can be classified and categorized into fuzzy sets (with membership value), a process for determining precise actions to be applied must be developed. This task involves writing a rule set that provides an action for any data classification that could possibly exist. The formation of the rule set is comparable to that of an expert system. Thus, behaviors is synthesized as fuzzy rule base i.e. a collection of fuzzy if-then rules.

Each behavior is encoded with a distinct control policy goverened by fuzzy inference. We write fuzzy rules as antecedent-consequent pairs of If-Then statements (Figure 9).

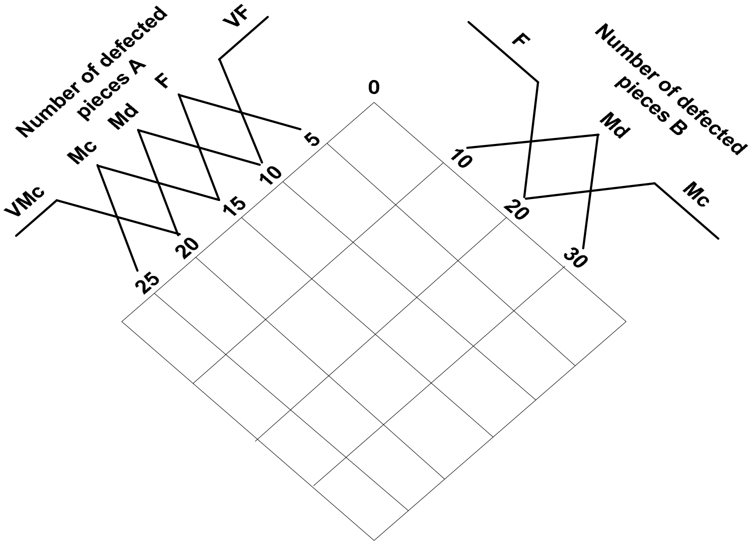


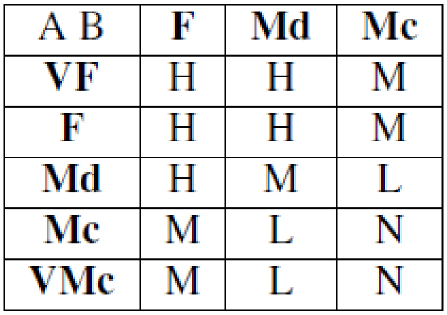
Figure 9. Fuzzy Rules of Production system

Running example

We take as example, the first column from the Table 1:

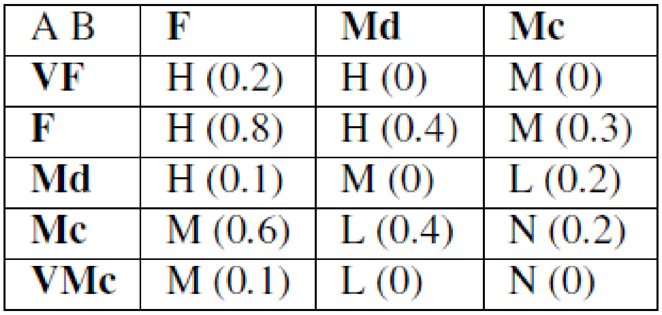
IF number of defected pieces from A is Very Few and number of defected pieces from B is Few Then Production is High.
IF number of defected pieces from A is Few and number of defected pieces from B is Few Then Production is High.
IF number of defected pieces from A is Medium and number of defected pieces from B is Few Then Production is High.
IF number of defected pieces from A is Much and number of defected pieces from B is Few Then Production is Medium.
IF number of defected pieces from A is Very Much and number of defected pieces from B is Few Then Production is Medium.

Table 1. Fuzzy Control rules for the Agent



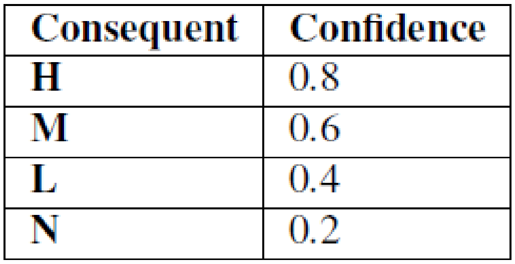
A B	**F**	**Md**	**Mc**
VF	H	H	M
F	H	H	M
Md	H	M	L
Mc	M	L	N
VMc	M	L	N

Table 2. Selection-based rules for the Agent



A B	F	Md	Mc
VF	H (0.2)	H (0)	M (0)
F	H (0.8)	H (0.4)	M (0.3)
Md	H (0.1)	M (0)	L (0.2)
Mc	M (0.6)	L (0.4)	N (0.2)
VMc	M (0.1)	L (0)	N (0)

Table 3. Final fuzzy values for the Agent



Consequent	Confidence
H	0.8
M	0.6
L	0.4
N	0.2

```
FzSet  AddLeftShoulderSet(std::string name,
                          double     minBound,
                          double     peak,
                          double     maxBound);

FzSet  AddRightShoulderSet(std::string name,
                           double     minBound,
                           double     peak,
                           double     maxBound);

FzSet  AddTriangularSet(std::string name,
                        double     minBound,
                        double     peak,
                        double     maxBound);

 FzSet  AddSingletonSet(std::string name,
                       double     minBound,
                       double     peak,
                       double     maxBound);

 //fuzzify a value by calculating its DOM in each of this variable's subsets
  void       Fuzzify(double val);

//defuzzify the variable using the MaxAv method
  double       DeFuzzifyMaxAv()const;

//defuzzify the variable using the centroid method
  double       DeFuzzifyCentroid(int NumSamples)const;
```

Running example

```
/* Add the rule set */
fm.AddRule(FzAND(A_VF, B_F), High);
        fm.AddRule(FzAND(A_VF, B_Md), High);
        fm.AddRule(FzAND(A_VF, B_Mc), Medium);
        fm.AddRule(FzAND(A_F, B_F), High);
fm.AddRule(FzAND(A_F, B_Md), Medium);
        fm.AddRule(FzAND(A_F, B_Mc), Medium);
        fm.AddRule(FzAND(A_Md, B_F), High);
        fm.AddRule(FzAND(A_Md, B_Md), Medium);
fm.AddRule(FzAND(A_Md, B_Mc), Low);
        fm.AddRule(FzAND(A_Mc, B_F), Medium);
        fm.AddRule(FzAND(A_Mc, B_Md), Low);
        fm.AddRule(FzAND(A_Mc, B_Mc), Null);
fm.AddRule(FzAND(A_VMc, B_VF), Medium);
        fm.AddRule(FzAND(A_VMc, B_VF), Low);
        fm.AddRule(FzAND(A_VMc, B_VF), Null);
```

3.3 Defuzzification

Fuzzy set is mapped to a real membered value in the interval 0 to 1.

If an element of universe, say x, is a member of fuzzy set A, then the mapping is given by $\mu A \in [0,1]

The output of a fuzzy process needs to be a single scalar quantity as opposed to a fuzzy set. Defuzzification is the conversion of a fuzzy quantity to a precise quantity. There are many methods to calculate it such as Max membership, Centroid method, Weighted average method, Mean max membership, Center of sums, Center of largest area and First (or last) of maxima. Obviously, the best defuzzification method is context-dependant [13].

4. CREATING JADE AGENTS

JADE is a Java tool and therefore creating a JADE-based multi-agent system requires creating Java classes. For more details, we refer to [7, 8, 9, 10].

Creating a JADE agent is very easy through defining a class that extends the jade.core.Agent class and implementing the setup() method. Each class introduced in the Figure 10 will be presented in the following paragraphs.

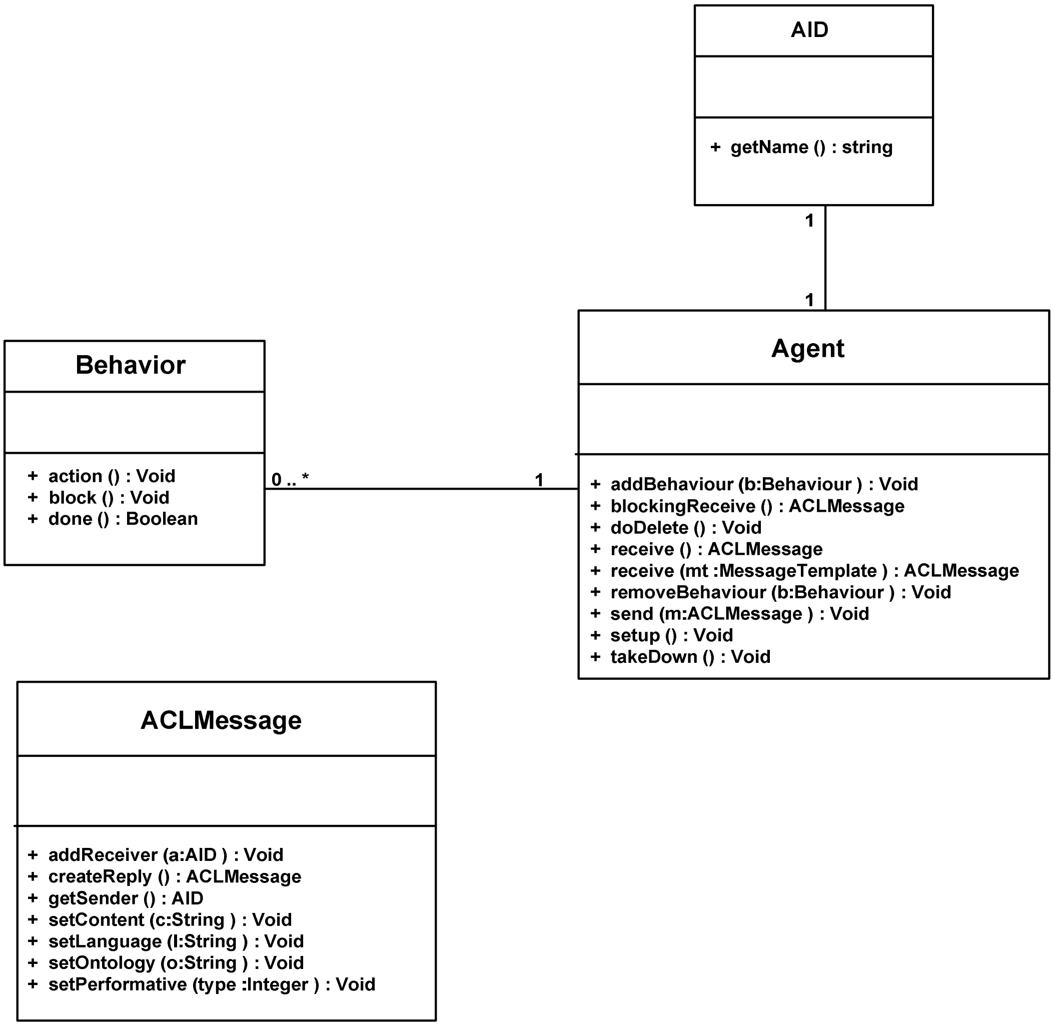

AID
+ getName () : string
1
1
Agent
+ addBehaviour (b:Behaviour) : Void
+ blockingReceive () : ACLMessage
+ doDelete () : Void
+ receive () : ACLMessage
+ receive (mt :MessageTemplate) : ACLMessage
+ removeBehaviour (b:Behaviour) : Void
+ send (m:ACLMessage) : Void
+ setup () : Void
+ takeDown () : Void
Behavior
+ action () : Void
+ block () : Void
+ done () : Boolean
0 .. *
1
ACLMessage
+ addReceiver (a:AID) : Void
+ createReply () : ACLMessage
+ getSender () : AID
+ setContent (c:String) : Void
+ setLanguage (l:String) : Void
+ setOntology (o:String) : Void
+ setPerformative (type :Integer) : Void

Figure 10. JADE agent

Running example

The setup() method is invoked when agent starts running and permits to initialize instance variables, register agent and attach one or more behaviors to the agent.

```
import jade.core.Agent;
public class Robot extends Agent {
            protected void setup() {
                  System.out.println("Hello everybody! I am an agent");
                }
}
```

4.1 Agent Identifier

Each agent is identified by an “agent identifier” represented as an instance of the jade.core.AID class. The getAID() method of the Agent class allows retrieving the agent identifier. An AID object includes a globally unique name plus a number of addresses. The name in JADE has the form <nickname>@<platform-name> so that an agent called Robot1 living on a platform called RARM will have Robot1@RARM as globally unique name. The addresses included in the AID are the addresses of the platform the agent lives in. These addresses are only used when an agent needs to communicate with another agent living on a different platform.

4.2 Agent discovery

The JADE platfrom allows the possibility to discover dynamically the available agents. To do so, a yellow pages service permits agents to describe one or more services they provide. An agent can register (publish) services and search to discover services.

Running example

In order to publish a service, an agent must create a proper description which is an instance of DFAgentDescription class and call the register() method of DFService class.

```
/// Register the Robot  in DFService
 DFAgentDescription dfd = new DFAgentDescription();
 dfd.setName(getAID());
 ServiceDescription sd = new ServiceDescription();
 sd.setType("Robot");
 sd.setName("Robot-executing");
 dfd.addServices(sd);
 try {
         DFService.register(this, dfd);
 }
 catch (FIPAException fe) {
         fe.printStackTrace();
 }
```

It is possible to search some agents, if the agent provides the DF with a template description. The result of the research is a list of all the descriptions matching the template.

Running example

The search() method of the DFService class ensures the result.

```
DFAgentDescription template = new DFAgentDescription();
       ServiceDescription   sd = new ServiceDescription();
         sd.setType("Robot");
         template.addServices(sd);
                     DFAgentDescription[] result ;
         try {
                      do
                      {
           result = DFService.search(myAgent, template);
           robotAgents = new AID[result.length];
          for (int i = 0; i < result.length; i++) {
           robotAgents[i] = result[i].getName();
                 }
                       }
                     while (result.length <= 0);
```

```
        }
        catch (FIPAException fe) {
          fe.printStackTrace();
        }
nbRobots=robotAgents.length;
```

4.3 Message exchanged between JADE Agents

Agents never interact through method calls but by exchanging asynchronous messages. Obviously, inter-agent interaction will be very difficult until all agents adopt the same communication language, and fortunately ACL standards ensure this requirement. All JADE agents communicate using messages that obey the FIPA ACL specification, which is described in : http//www.fipa.org.

This format comprises a number of fields and in particular: (1) the sender of the message, (2) the list of receivers, (3) the communicative intention (also called "performative") indicating what the sender intends to achieve by sending the message (for example the performative can be REQUEST, INFORM, QUERY_IF, CFP (call for proposal), PROPOSE, ACCEPT_PROPOSAL, REJECT_PROPOSAL, and so on). (4) The content i.e. the actual information included in the message which may be string in simple cases; otherwise we need a content language, a corresponding ontology, and a protocol. (5) The ontology i.e. the vocabulary of the symbols used in the content and their meaning (both the sender and the receiver must be able to encode expressions using the same symbols to be sure that the communication is effective)

4.3.1. Sending a message

Sending a message to another agent is as simple as filling the fields of an ACLMessage object and then call the send() method of the Agent class. The code below informs an agent whose nickname is Robot1 that the production must be decreased.

Running example

```
ACLMessage msg = new ACLMessage(ACLMessage.INFORM);
msg.addReceiver(new AID("Robot1", AID.ISLOCALNAME));
msg.setOntology("Production");
msg.setContent("We must decrease in the production");
send(msg);
```

4.3.2. Receiving a message

As mentioned above the JADE runtime automatically posts messages in the receiver's private message queue as soon as they arrive. An agent can pick up messages from its message queue by means of the receive() method.

This method returns the first message in the message queue (removing it) or null if the message queue is empty and immediately returns.

Running example

```
ACLMessage msg = receive();
if (msg != null) {
```

```
// Process the message
}
```

4.3.3. Blocking behavior waiting a message

Some behaviors must be continuously running and at each execution of their action() method, must check if a message is recceived and perform some action.

Running example

```
public void action() {
ACLMessage msg = myAgent.receive();
if (msg != null) {
// Message received. Process it
…
}
else {
block();
}
}
```

4.3.4. Selecting a message

When a template is specified, the receive() method returns the first message (if any) matching it, while ignores all non-matching messages. Such templates are implemented as instances of the jade.lang.acl.MessageTemplate class that provides a number of factory methods to create templates in a very simple and flexible way.

Running example

The action() method is modified so that the call to myAgent.receive() ignores all messages except those whose performative is REQUEST:

```
  public void action() {
MessageTemplate mt = MessageTemplate.MatchPerformative(ACLMessage.REQUEST);
ACLMessage msg = myAgent.receive(mt);
if (msg != null) {
// REQUEST Message received. Process it
...
}
else {
block();
}
}
```

4.4 Agent Behavior in JADE

A behavior is a kind of control thread for the agent where the method action() is similar to Thread.run(). New beahviors can be added at any time during the agent life. A behavior represents a task that an agent can carry out and is implemented as an object of a class that extends jade.core.behaviours.Behaviour. To make an agent execute the task implemeted by a behavior object, the behavior should be added to the agent by means of the addBehavior() method of the Agent class in the setup() method or inside other behavior (Figure 11).

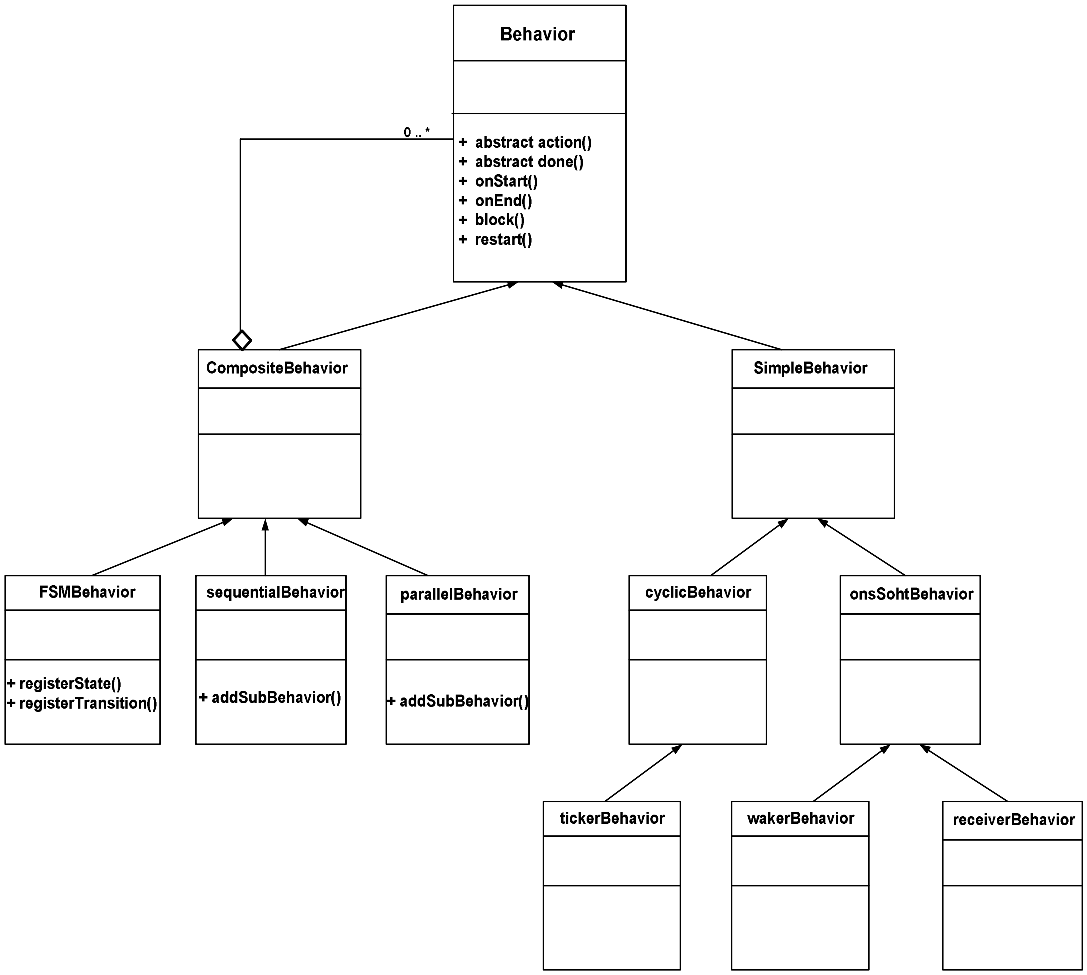


Figure 11. Behaviour class hierarchy in JADE

- ✓ class Behaviour : Each class extending the abstract class Behavior must implement two abstract methods. The action() method defines the operation to be performed when the behavior is in execution. The done() method returns a boolean value to indicate whether or not a behavior has completed. The Behaviour class also provides two methods, named onStart() and onEnd(). These methods can be overridden by user defined subclasses when some actions are to be executed before and after running behaviour execution. onEnd() returns an integer that represents a termination value for the behaviour. It should be noted that onEnd() is called after the behaviour has completed and has been removed from the pool of agent behaviours.
- ✓ class SimpleBehaviour: The SimpleBehaviour class is an abstract class modeling simple atomic behaviours. Its reset() method does nothing by default, but it can be overridden by user defined subclasses.
- ✓ class OneShotBehaviour: The OneShotBehaviour class models atomic behaviours that must be executed only once and cannot be blocked. So, its done() method always returns true. The class WakerBehaviour implements a one-shot task that must be executed only

once just after a given timeout is elapsed. The class TickerBehaviour implements a cyclic task that must be executed periodically.

✓ class CyclicBehaviour: The CyclicBehaviour class models atomic behaviours that must be executed forever. So its done() method always returns false. “Cyclic” behaviours that never complete and whose action() method executes the same operations each time it is called.

✓ class CompositeBehaviour: This abstract class models behaviours that are made up by composing a number of other behaviours (children). So the actual operations performed by executing this behaviour are not defined in the behaviour itself, but inside its children while the composite behaviour takes only care of children scheduling according to a given policy (sequentially for SequentialBehaviour class, concurrently for ParallelBehaviour class and finite state machine for FSMBehaviour class).

Running example

```
int   nbPositive = 0;

  protected void setup()
  {
  ACLMessage msg = newMsg( ACLMessage.QUERY\_REF );

   MessageTemplate template = MessageTemplate.and(
   MessageTemplate.MatchPerformative( ACLMessage.INFORM ),
   MessageTemplate.MatchConversationId( msg.getConversationId() ));

   SequentialBehaviour seq = new SequentialBehaviour();
   addBehaviour( seq );

   ParallelBehaviour par = new \textbf{ParallelBehaviour}( ParallelBehaviour.WHEN_ALL );
  seq.addSubBehaviour( par );

  for (int i = 1; i<= nbRobots; i++)
   {
            msg.addReceiver( new AID( "Robot" + i,  AID.ISLOCALNAME ));

            par.addSubBehaviour( new myReceiver( this, 1000, template)
              {
              public void handle( ACLMessage msg)
              {
                if (msg != null){
                         if (msg.getPerformative() == ACLMessage.ACCEPT) {
                             nbPositive = nbPositive+1;
                 } }  }
              });
    }
    seq.addSubBehaviour( new OneShotBehaviour()
     {
              public void action()
              {
              if (nbPositive = nbRobots)
                         System.out.println("All agents accept to change the production");
               else
```

```
                System.out.println("Some agents refuse to change the production");
        }
    });
```

5. CONCLUSION

Distributed planning is narrowly interlaced with the distributed resolution of the problems, being a problem in itself and means to solve a problem. The main aim of this paper is how to ensure a distributed planning in Multi-Agent System (MAS) composed of several intelligent autonomous agents able to take the initiative instead of simply reacting in response to its environment. Our solution to this problem is the use of fuzzy logic which is based on three steps: fuzzyfication, rule engine and defuzzyfication. We create the MAS through JADE platform and show the interaction between the different agents through exchanging messages. All our contributions are applied on the benchmark production system (RARM system).

REFERENCES

[1] Jacques Ferber, Multi-Agent System: An Introduction to Distributed Artificial, Intelligence, Harlow: Addison Wesley Longman, 1999, Paper: ISBN 0-201-36048-9.

[2] Bordini, R.H., and all. A Survey of Programming Languages and Platforms for Multi-agent Systems. Informatica, 30(1): pp. 33–44, 2006.

[3] David Jung, Alexander Zelinsky, An architecture for distributed cooperative planning in a behaviour-based multi-robot system Robotics and Autonomous Systems, Volume 26, Issues 2–3, 28 February 1999, Pages 149-174.

[4] Malik Ghallab, Dana Nau, Paolo Traverso, The actor's view of automated planning and acting: A position paper Artificial Intelligence, Volume 208, March 2014, Pages 1-17.

[5] Salvatore Vitabile, Vincenzo Conti, Carmelo Militello, Filippo Sorbello, An extended JADE-S based framework for developing secure Multi-Agent Systems Computer Standards \& Interfaces, Volume 31, Issue 5, September 2009, Pages 913-930.

[6] Chuan-Jun Su, Chia-Ying Wu, JADE implemented mobile multi-agent based, distributed information platform for pervasive health care monitoring, Applied Soft Computing, Volume 11, Issue 1, January 2011, Pages 315-325

[7] Fabio Bellifemine, Giovanni Caire, Tiziana Trucco, Giovanni Rimassa, Roland Mungenast, JADE ADMINISTRATOR'S GUIDE, 2010

[8] Giovanni Caire, JADE TUTORIAL : JADE PROGRAMMING FOR BEGINNERS , 2009

[9] Fabio Bellifemine, Giovanni Caire, Tiziana Trucco, Giovanni Rimassa, JADE PROGRAMMER'S GUIDE, 2010.

[10] Fabio Bellifemine, Giovanni Caire, Dominic Greenwood, Developing Multi-Agent Systems with JADE, 2004

[11] Branislav Hrúz, MengChu Zhou Modeling and Control of Discrete-event Dynamic Systems with Petri Nets and Other Tools 2007 p67)

[12] Kazem Sadegh-Zadeh, Advances in fuzzy theory, Artificial Intelligence in Medicine, Volume 15, Issue 3, March 1999, Pages 309-323

[13] Lotfi A. Zadeh, Is there a need for fuzzy logic? Information Sciences, Volume 178, Issue 13, 1 July 2008, Pages 2751-2779

[14] Belohlavek, R., Klir, G., Lewis, H., and Way, E. (2002) On the capability of fuzzy set theory to represent concepts. Int. J. Gen. Syst., 31, 569–585.

[15] Marijana Gorjanac Ranitović, Aleksandar Petojević, Lattice representations of interval-valued fuzzy sets, Fuzzy Sets and Systems, Volume 236, 1 February 2014, Pages 50-57.

[16] Jianhua Dai, Haowei Tian, Fuzzy rough set model for set-valued data, Fuzzy Sets and Systems, Volume 229, 16 October 2013, Pages 54-68

[17] Mária Kuková, Mirko Navara, Principles of inclusion and exclusion for fuzzy sets, Fuzzy Sets and Systems, Volume 232, 1 December 2013, Pages 98-109

[18] Oscar Sapena, Eva Onaindia, Antonio Garrido, Marlene Arangu, A distributed CSP approach for collaborative planning systems, Engineering Applications of Artificial Intelligence, Volume 21, Issue 5, August 2008, Pages 698-709

[19] Sergio Pajares Ferrando, Eva Onaindia, Context-Aware Multi-Agent Planning in intelligent environments, Information Sciences, Volume 227, 1 April 2013, Pages 22-42

[20] Pascal Forget, Sophie D'Amours, Jean-Marc Frayret, Multi-behavior agent model for planning in supply chains: An application to the lumber industry, Robotics and Computer-Integrated Manufacturing, Volume 24, Issue 5, October 2008, Pages 664-679

[21] Malik Ghallab, Dana Nau, Paolo Traverso, Automated Planning, 2004

[22] Tsz-Chiu Au, Ugur Kuter, and Dana Nau, Planning for Interactions among Autonomous Agents

[23] Chun-xia Dou, Da-wei Hao, Bao Jin, Wei-qian Wang, Na An, Multi-agent-system-based decentralized coordinated control for large power systems International Journal of Electrical Power & Energy Systems, Volume 58, June 2014, Pages 130-139

[24] Bo Liu, Housheng Su, Rong Li, Dehui Sun, Weina Hu, Switching controllability of discrete-time multi-agent systems with multiple leaders and time-delays, Applied Mathematics and Computation, Volume 228, 1 February 2014, Pages 571-588

AUTHORS

Atef Gharbi received his computer engineering Diploma from the National School in Computer Science (ENSI) of Tunisia, in 2005. After that, he received the Master degree from the National Institute of Applied Sciences and Technology (INSAT) of Tunisia in 2007. He obtained his Phd in 2013. He is currently related to LISI Research Laboratory in Tunisia. His research interests include specification of model, verification of properties related to functional safety, implementation of software solutions to ensure functional safety.

Samir Ben Ahmed is a Full Professor in Computer Science at Tunis-El Manar University, President of National Institute of Applied Sciences and Technology (INSAT), and Head of MOSIC Research Unit in Tunisia. He was Founder of ISI Institute of Computer Science, and Head of IT Department of Faculty of Science at Tunis-El Manar University in Tunisia. Prof. Ben Ahmed obtained his PhD Thesis in Automation and Computer Science at Paul Sabatier University in France. The Engineering Diploma was obtained before from National School of Electrical Engineering, Electronic, Computer Science and Hydraulic in Toulouse (ENSEEIHT). Prof. Ben Ahmed is strongly active in several National and International Projects and Collaborations.

REDUCT GENERATION FOR THE INCREMENTAL DATA USING ROUGH SET THEORY

Shampa sengupta[1], Asit Kumar Das[2]

[1]Department of Information Technology, MCKV Institute of Engineering,
Liluah, Howrah – 711 204, West Bengal, India
shampa2512@yahoo.co.in
[2]Department of Computer Science and Technology, Indian Institute of
Engineering, Science and Technology, Shibpur, Howrah – 711 103,
West Bengal, India
akdas@cs.becs.ac.in

ABSTRACT

In today's changing world huge amount of data is generated and transferred frequently. Although the data is sometimes static but most commonly it is dynamic and transactional. New data that is being generated is getting constantly added to the old/existing data. To discover the knowledge from this incremental data, one approach is to run the algorithm repeatedly for the modified data sets which is time consuming. The paper proposes a dimension reduction algorithm that can be applied in dynamic environment for generation of reduced attribute set as dynamic reduct. The method analyzes the new dataset, when it becomes available, and modifies the reduct accordingly to fit the entire dataset. The concepts of discernibility relation, attribute dependency and attribute significance of Rough Set Theory are integrated for the generation of dynamic reduct set, which not only reduces the complexity but also helps to achieve higher accuracy of the decision system. The proposed method has been applied on few benchmark dataset collected from the UCI repository and a dynamic reduct is computed. Experimental result shows the efficiency of the proposed method.

KEYWORDS

Dimension Reduction, Incremental Data, Dynamic Reduct, Rough Set Theory.

1. INTRODUCTION

In today's e-governance age, everything is being done through electronic media. So huge data is generated and collected from various areas for which proper data management is necessary. Retrieval of some interesting information from stored data as well as time variant data is also a very challenging task. Extraction of meaningful and useful data pattern from these large data is the main objective of data mining technique [1]. Data mining techniques basically uses the concept of database technology [2] and pattern recognition [3] principles. Feature selection [4] and reduct generation [5] are frequently used as a pre-processing step to data mining and knowledge discovery [6, 7]. For static data, it selects an optimal subset of features from the feature space according to a certain evaluation criterion. In recent years, dimension of datasets are growing rapidly in many applications which bring great difficulty to data mining and pattern recognition. As datasets changes with time, it is very time consuming or even infeasible to run

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 291–299, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4529

repeatedly a knowledge acquisition algorithm. Rough Set Theory (RST) [8, 9, and 10], a new mathematical approach to imperfect knowledge, helps to find the static as well as dynamic reduct. Dynamic reducts can put up better performance in very large datasets as well as enhance effectively the ability to accommodate noise data. The problem of attribute reduction for incremental data falls under the class of Online Algorithms and hence demands a dynamic solution to reduce re-computation. Liu [11] developed an algorithm for finding the smallest attribute set of dynamic reducts with increase data. Wang and Wang [12] proposed a distributed algorithm of attribute reduction based on discernibility matrix and function. Zheng et al. [13] presented an incremental algorithm based on positive region for generation of dynamic reduct. Deng [14] presented a method of attribute reduction by voting in a series of decision subsystems for generation of dynamic reduct. Jan G. Bazan et al. [15] presented the concept of dynamic reducts to solve the problem of large amount of data or incremental data.

In the proposed method, a novel heuristic approach is proposed to find out a dynamic reduct of the incremental dataset using the concept of Rough Set Theory. To understand the concepts of dynamic data, a sample dataset is divided into two sub sets considering one as old dataset and other as new dataset. Using the concept of discernibility matrix and attribute dependency of Rough Set Theory reduct is computed from old dataset. Then to handle the new data or incremental data, previously computed reduct is modified wherever changes are necessary and generates dynamic reduct for the entire system. The details of the algorithm are provided in subsequent section.

The rest of the paper is organized as follows: Basic Concepts of Rough Set Theory is described in section 2. Section 3 demonstrated the process of generation of dynamic reduct and Section 4 shows the experimental result of the proposed method. Finally conclusion of the paper is stated in section 5.

2. BASIC CONCEPTS OF ROUGH SET THEORY

The rough set theory is based on indiscernibility relations and approximations. Indiscernibility relation is usually assumed to be equivalence relation, interpreted so that two objects are equivalent if they are not distinguishable by their properties. Given a decision system DS = (U, A, C, D), where U is the universe of discourse and A is the total number of attributes, the system consists of two types of attributes namely conditional attributes (C) and decision attributes (D) so that A = C ∪ D. Let the universe U = {x_1, x_2... x_n}, then with any P ⊆ A, there is an associated P-indiscernibility relation IND(P) defined by equation (1).

$$IND(P) = \{(x, y) \in U^2 | \forall a \in P, a(x) = a(y)\} \qquad (1)$$

If (x, y) ∈ IND (P), then x and y are indiscernible with respect to attribute set P. These indistinguishable sets of objects, therefore define an indiscernibilty relation referred to as the P-indiscernibility relation and the class of objects are denoted by $[x]_P$.

The lower approximation of a target set X with respect to P is the set of all objects which certainly belongs to X, as defined by equation (2).

$$\underline{PX} = \{x | [x]_P \subseteq X\} \qquad (2)$$

The upper approximation of the target set X with respect to P is the set of all objects which can possibly belong to X, as defined by equation (3)

$$\overline{PX} = \{x | [x]_P \cap X \neq \emptyset\} \tag{3}$$

As rough set theory models dissimilarities of objects based on the notions of discernibility, a discernibility matrix is constructed to represent the family of discernibility relations. Each cell in a discernibility matrix consists of all the attributes on which the two objects have the different values. Two objects are discernible with respect to a set of attributes if the set is a subset of the corresponding cell of the discernibility matrix.

(a) Discernibility Matrix and Core

Given a decision system DS = (U, A, C, D), where U is the universe of discourse and A is the total number of attributes. The system consists of two types of attributes namely conditional attributes (C) and decision attributes (D) so that A = C ∪ D. Let the universe U = $\{x_1, x_2 ... x_n\}$, then discernibility matrix M = (m_{ij}) is a |U| × |U| matrix, in which the element m_{ij} for an object pair (x_i, x_j) is defined by (4).

$$m_{ij} = \{a \in C : a(x_i) \neq a(x_j) \wedge (d \in D, d(x_i) \neq d(x_j))\} \tag{4}$$

where, i, j = 1, 2, 3... n

Thus, each entry (i, j) in the matrix S contains the attributes which distinguish the objects i and j. So, if an entry contains a single attribute say, A_s, it implies that the attribute is self sufficient to distinguish two objects and thus it is considered as the most important attribute, or core attribute. But in reality, several entries may contain single attribute, union of which is known as core CR of the dataset, as defined in (5).

$$CR = \cup\{m_{ij} | m_{ij} \neq \emptyset \; and \; |m_{ij}| = 1, \forall\, i, j = 1, 2, \ldots, n\} \tag{5}$$

(b) Attribute Dependency and Reduct

One of the most important aspects of database analysis or data acquisition is the discovery of attribute dependencies; that establishes a relationship by finding which variables are strongly related to which other variables. In rough set theory, the notion of dependency is defined very simply. Assume two (disjoint) sets of attributes, P and Q, and inquire what degree of dependency is present between them. Each attribute set induces an (indiscernibility) equivalence class structure. Say, the equivalence classes induced by P is $[x]_P$, and the equivalence classes induced by Q is $[x]_Q$. Then, the dependency of attribute set Q on attribute set P is denoted by $\gamma_P(Q)$ and is given by equation (6).

$$\gamma_P(Q) = \frac{\sum_{i=1}^{N} |\underline{PX_i}|}{|U|} \tag{6}$$

Where, Q_i is a class of objects in $[x]_Q$; ∀ i = 1, 2, …, N.

A reduct can be thought of as a sufficient set of attributes to represent the category structure and the decision system. Projected on just these attributes, the decision system possesses the same equivalence class structure as that expressed by the full attribute set. Taking the partition induced by decision attribute D as the target class and R as the minimal attribute set, R is called the reduct if it satisfies (7). In other words, R is a reduct if the dependency of decision attribute D on R is exactly equal to that of D on whole conditional attribute set C.

$$\gamma_R(D) = \gamma_C(D) \qquad (7)$$

The reduct of an information system is not unique. There may be many subsets of attributes which preserve the equivalence-class structure (i.e., the knowledge) expressed in the decision system.

(c) Attribute Significance: Significance of an attribute a in a decision table *A= (U, CUD)* (with the decision set *D*) can be evaluated by measuring the effect of removing of an attribute a∈C from the attribute set C on the positive region. The number γ(C, D) expresses the degree of dependency between attributes C and D. If attribute 'a' is removed from the attribute set C then the value of (γ(C, D)) will be changed.

So the significance of an attribute a is defined as

$$\sigma^{a}_{(C,D)} = \frac{\gamma(C,D) - \gamma(C-\{a\},D)}{\gamma(C,D)} \qquad (8)$$

(d) Dynamic Reduct: The purpose of dynamic reducts is to get the stable reducts from decision subsystems. Dynamic reduct can be defined in the following direction.

Definition 1: If *DS* = (*U*, *A*, *d*) is a decision system, then any system *DT* = (*U'*, *A*, *d*) such that *U'* ⊆ *U* is called a subsystem of *DS*. By P (*DS*) we denote the set of all subsystems of *DS*. Let *DS* = (U, *A*, *d*) be a decision system and *F* ⊆ P (*DS*). By *DR* (*DS*, *F*) we denote the set *RED (DS)* $\cap \bigcap_{DT \in F}$ *RED (DT).*Any elements of *DR* (*DS*, *F*) are called an F-dynamic reduct of *DS*.
So from the definition of dynamic reducts it follows that a relative reduct of DS is dynamic if it is also a reduct of all sub tables from a given family of F.

Definition 2: Let *DS* = (U, *A*, *d*) be a decision system and *F* ⊆ P (*DS*). By *GDR* (*DS*, *F*) we denote the set

$$\bigcap_{DT \in F} RED\ (DT)$$

Any elements of *GDR* (*DS*, *F*) are called an F generalized dynamic reduct of *DS*. From the above definitions of generalized dynamic reduct it follows that any subset of *A* is a generalized dynamic reduct if it is also a reduct of all sub tables from a given family F.

Time complexity of computation of all reducts is NP-Complete. Also, the intersection of all reducts of subsystems may be empty. This idea can be sometimes too much restrictive, so more general notion of dynamic reducts are described. They are called (F, ε) dynamic reducts, where ε > 0. The set *DR (DS, F)* of all (F, ε) dynamic reducts is defined by

$$DR_{\varepsilon}^{(DS)} = \{C \in RED\ (DS, d): \quad \frac{card(DT \in F : C \in red(DT, d))}{Card\ (F)} \geq 1 - \varepsilon\}$$

3. DYNAMIC REDUCT GENERATION USING ROUGH SET THEORY

Various concepts of rough set theory like discernibility matrix, attribute significance and attribute dependency are applied together to compute dynamic reducts of a decision system. The term dynamic reduct is used in the sense that the method computes a set of reducts for the incremental data very quickly without unnecessarily increasing the complexity since they are sufficient to represent the system and subsystems of it. Based on the discernibility matrix M and the frequency value of the attributes, the attributes are divided [16] into the core set CR and noncore set NC for old subsystem DS_{old}. Next, highest ranked element of NC is added to the core CR in each iteration provided the dependency of the decision attribute D on the resultant set increases for the old subsystem ; otherwise it is ignored and next iteration with the remaining elements in NC is performed. The process terminates when the resultant set satisfies the condition of equation (7) for the old subsystem and is considered as an initial reduct RED_OLD. Then backward attribute removal process is applied for each noncore attribute x in the generated reduct RED_OLD, it is checked whether (7) is satisfied using RED_OLD – {x}, instead of R. Now if it is satisfied, then x is redundant and must be removed. Thus, all redundant attributes are removed and final reduct RED_OLD is obtained.

To generate the dynamic reduct, discernibility matrix is constructed for the new subsystem DS_{new} and frequency values of all conditional attributes are calculated. Now the previously computed reduct (RED_OLD) from the old dataset is applied to new dataset for checking whether it can preserve the positive region in the new data set i.e., whether the dependency value of the decision attribute on that reduct set is equal to that of the decision attribute on the whole conditional attribute set. If the condition is satisfied, then that reduct set is considered as dynamic reduct (DRED). Otherwise; according to the frequency values obtained using [16] of the conditional attributes, higher ranked attribute is added to the most important attribute set in each iteration provided attribute dependency of the resultant set increases and subsequently a reduct is formed after certain iteration when dependency of the decision attribute on the resultant set is equal to that of the decision attribute on the whole condition attribute set for the new subsystem. Then backward attribute removal process is applied for generation of final dynamic reduct of the system. In this process, significance value of each individual attribute is calculated using equation (8) except that most important attribute set in a reduct. If the significance value of a particular attribute is zero, then that attribute is deleted from the reduct. In this way, all redundant attributes are removed and finally dynamic reduct is generated by modifying the old reducts for the entire data.

The proposed method describes the attribute selection method for the computation of reducts from old data and dynamic reduct set DRED for entire data considering incremental data.

Algorithm1 generates initial reduct for the old decision system DS_{old} = (U, A, C, D) and Algorithm2 generates dynamic reduct for the entire data, by considering the old data as well as incremental data.

Algorithm1: Initial_Reduct_Formation (DS_{old}, CR, NC)

Input: DS_{old}, the decision system with C conditional attributes and D decisions with objects x, CR, the core and NC, the non-core attributes

Output: RED_OLD, initial reduct

```
Begin
        RED_OLD = CR   /* core is considered as initial reduct*/
        NC_OLD = NC /* take a copy of initial elements of NC*/
       /*Repeat-until below forward selection to give one reduct*/
        Repeat
             x = highest ranked element of NC_OLD
             If (x = $\phi$) break   /*if no element found in NC*/
             If ($\gamma_{RED\_OLD \cup \{x\}}$ (D) > $\gamma_{RED\_OLD}$ (D))
                 {
                   RED_OLD = RED_OLD $\cup$ {x}
                  NC_OLD = NC_OLD - {x}
                 }
        Until ($\gamma_{RED\_OLD}$(D) = $\gamma_C$(D))
           // apply backward removal
        For each x in (RED_OLD – CR)
               If ($\gamma_{RED\_OLD - \{x\}}$(D) = = $\gamma_C$(D))
                     RED_OLD = RED_OLD - {x}
       Return (RED_OLD);
End
```

Algorithm2: Dynamic_Reduct_Formation (DS, C, D)

//An algorithm for computation of dynamic reducts for incremental data

Input: DS = {DS_{new}}, the new decision system with C conditional attributes and D decisions attribute and reduct RED_OLD obtained from ‘Reduct Formation’ algorithm for the old dataset (DS_{old}).

Output: Dynamic reduct (DRED), reduct of $DS_{old} \cup DS_{new}$

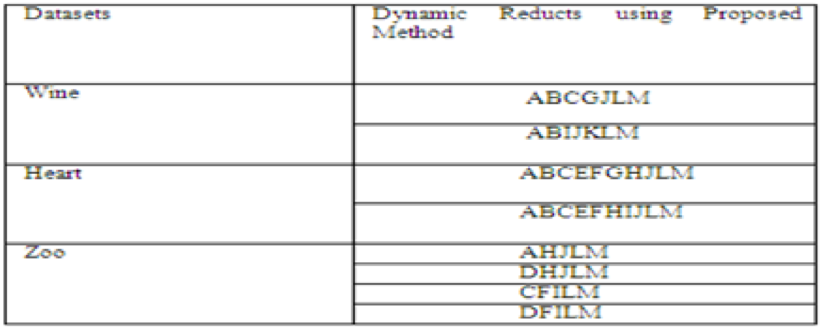
```
Begin
    If (($\gamma_{(RED\_OLD)}$ (D) = $\gamma_{(C)}$ (D))
        {
          DRED = RED_OLD
          Return DRED
        }

    Else {
          NC = C - RED_OLD
          CR = DRED  /*initial reduct is considered as core reduct of new system */
          Repeat
             DRED = RED_OLD   /* Old reduct is considered as core */
             x = highest frequency attribute of NC
```

```
  If ($\gamma_{DRED\cup\{x\}}(D) > \gamma_{DRED}(D)$)
           {
            DRED = DRED $\cup$ {x}
            NC = NC - {x}
           }

      Until ($\gamma_{DRED}(D) = \gamma_C(D)$)

      // apply backward removal

For each highest ranked attribute x in (DRED – CR) using (8)

       If ($\gamma_{DRED - \{x\}}(D) == \gamma_C(D)$)
          DRED = DRED - {x}
     Return (DRED);
  }  /* end of else*/
End
```

4. EXPERIMENTAL RESULTS

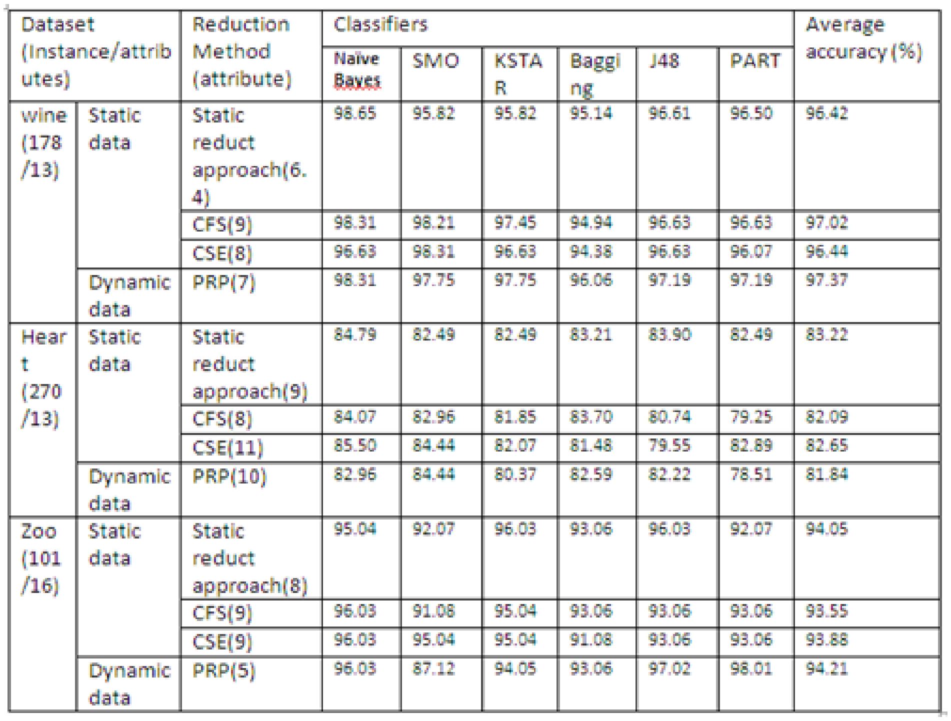
The method is applied on some benchmark datasets obtained from UCI repository ‘http: //www.ics.uci.edu/mlearn/MLRepository’. The wine dataset contains 178 instances and 13 conditional attributes. The attributes are abbreviated by letters A, B, and so on, starting from their column position in the dataset. In our method, for computation of dynamic reduct the wine dataset is divided into 2 sub tables considering randomly 80% of data as old data and other 20% of data is new data. Reduct is calculated for the old data using Algorithm1.Then based on previous reducts, the proposed algorithm worked on new data and generates two dynamic reducts {{ABCGJLM}, {ABIJKLM}} for the whole dataset. Similarly dynamic reducts are calculated for the heart and Zoo dataset. Reducts are also calculated for the modified data set using static data approach. All results are given in Table 1. Accuracies of the reduct of our proposed algorithm (PRP) are calculated and compared with existing attribute reduction techniques like ‘Correlation-based Feature Selection’ (CFS) and ‘Consistency-based Subset Evaluation’ (CSE), from the ‘weka’ tool [17] as shown in Table 2. The proposed method, on average, contains lesser number of attributes compared to CFS and CSE and at the same time achieves higher accuracy, which shows the effectiveness of the method.

Table 1. Dynamic reducts of datasets



Datasets	Dynamic Reducts using Proposed Method
Wine	ABCGJLM
	ABIJKLM
Heart	ABCEFGHJLM
	ABCEFHIJLM
Zoo	AHJLM
	DHJLM
	CFILM
	DFILM

Table 2. Classification accuracy of reducts obtained by proposed and existing method



Dataset (Instance/attributes)		Reduction Method (attribute)	Classifiers						Average accuracy (%)
			Naive Bayes	SMO	KSTAR	Bagging	J48	PART	
wine (178 /13)	Static data	Static reduct approach(6.4)	98.65	95.82	95.82	95.14	96.61	96.50	96.42
		CFS(9)	98.31	98.21	97.45	94.94	96.63	96.63	97.02
		CSE(8)	96.63	98.31	96.63	94.38	96.63	96.07	96.44
	Dynamic data	PRP(7)	98.31	97.75	97.75	96.06	97.19	97.19	97.37
Heart (270 /13)	Static data	Static reduct approach(9)	84.79	82.49	82.49	83.21	83.90	82.49	83.22
		CFS(8)	84.07	82.96	81.85	83.70	80.74	79.25	82.09
		CSE(11)	85.50	84.44	82.07	81.48	79.55	82.89	82.65
	Dynamic data	PRP(10)	82.96	84.44	80.37	82.59	82.22	78.51	81.84
Zoo (101 /16)	Static data	Static reduct approach(8)	95.04	92.07	96.03	93.06	96.03	92.07	94.05
		CFS(9)	96.03	91.08	95.04	93.06	93.06	93.06	93.55
		CSE(9)	96.03	95.04	95.04	91.08	93.06	93.06	93.88
	Dynamic data	PRP(5)	96.03	87.12	94.05	93.06	97.02	98.01	94.21

5. CONCLUSION

The paper describes a new method of attribute reduction for incremental data by using the concepts of Rough Set theory. Even if the data is not completely available at a time, i.e it keeps arriving or increasing, the algorithm can find the reduct of such data without recomputing the data that has already arrived. The proposed dimension reduction method used only the concepts of rough set theory which does not require any additional information except the decision system itself. Since, reduct generation is a NP-complete problem, so different researchers' use different heuristics to compute reducts used for developing classifiers. Dynamic reducts are very important for construction of a strong classifier. A future enhancement to this work is to formation of classifiers from dynamic reduct sets and finally ensemble them to generate an efficient classifier.

REFERENCES

[1] Han, and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001.

[2] Handbook of Research on Innovations in Database Technologies and Applications: Current and Future Trends Viviana E. Ferraggine , Jorge H. Doorn , Laura C. Rivero, ISBN-10: 1605662429 ISBN-13: 978-1605662428

[3] Devijver, P.A., and Kittler, J. (1982) Pattern Recognition: A Statistical Approach Englewood Cliffs, NJ: Prentice Hall.

[4] Della Pietra,S., Della Pietra, V., and Lafferty, J. (1997) Inducing features of random fields. IEEE transactions on pattern Analysis and Machine Intelligence, 19(4),pp. 380-393.

[5] R. Jensen, QiangShen, "Fuzzy-Rough Attribute Reduction with Application to Web Categorization, Fuzzy Sets and Systems, Vol.141, No.3, pp.469-485, 2004

[6] N.Zhong and A. Skowron, “A Rough Set-Based Knowledge Discovery Process”, Int. Journal of Applied Mathematics and Computer Science. 11(3), 603-619, 2001. BIME Journal, Volume (05), Issue (1), 2005

[7] Ethem Alpaydin Introduction to Machine Learning.PHI, 2010

[8] Pawlak, Z.: “Rough sets.: International journal of information and computer sciences,” Vol, 11, pp. 341-356 (1982)

[9] Pawlak, Z.: “Rough set theory and its applications to data analysis,” Cybernetics and systems 29 (1998) 661-688, (1998)

[10] K. Thangavel, A. Pethalakshmi. Dimensionality reduction based on rough set theory : A review, Journal of Applied Soft Computing, Volume 9, Issue 1, pages 1 -12, 2009.

[11] Z.T,Liu.: “An incremental arithmetic for the smallest reduction of attributes” Acta Electro nicasinicia, vol.27, no.11, pp.96—98,1999

[12] J.Wang and J.Wang,”Reduction algorithms based on discernibility matrx:The order attributes method.Journal of computer Science and Technology,vol.16.No.6,2001,pp.489-504

[13] G.Y.Wang,Z.Zheng and Y.Zhang”RIDAS-A rough set based intelligent data analysis system” Proceedigs of the 1st International conference on machine Learning and Cybernatics,Beiing,Vol2,Feb,2002,pp.646-649.

[14] D.Deng,D.Yan and J.Wang,”parallel Reducts based on Attribute significance”, LNAI6401, 2010, pp.336-343.

[15] G..Bazan ,”Dynamic reducts and statistical Inference” Proceedigs of the 6th International conference on Information Processing and Management of uncertainity in knowledge based system,July 125, Granada,Spain,(2),1996pp.1147-1152

[16] Asit Kumar Das, Saikat Chakrabarty, Shampa Sengupta “Formation of a Compact Reduct Set Based on Discernibility Relation and Attribute Dependency of Rough Set Theory” Proceedings of the Sixth International Conference on Information Processing – 2012 August 10 - 12, 2012, Bangalore, Wireless Network and Computational Intelligence Springer pp 253-261.

[17] WEKA: Machine Learning Software, http://www.cs.waikato.ac.nz/~ml/

INTENTIONAL BLANK

OPTIMAL RULE SET GENERATION USING PSO ALGORITHM

Shampa sengupta[1], Asit Kumar Das[2]

[1]Department of Information Technology, MCKV Institute of Engineering,
Liluah,
Howrah – 711 204, West Bengal, India
shampa2512@yahoo.co.in

[2]Department of Computer Science and Technology, Indian Institute of
Engineering, Science and Technology, Shibpur, Howrah – 711 103, West
Bengal, India
akdas@cs.becs.ac.in

ABSTRACT

Classification and Prediction is an important research area of data mining. Construction of classifier model for any decision system is an important job for many data mining applications. The objective of developing such a classifier is to classify unlabeled dataset into classes. Here we have applied a discrete Particle Swarm Optimization (PSO) algorithm for selecting optimal classification rule sets from huge number of rules possibly exist in a dataset. In the proposed DPSO algorithm, decision matrix approach was used for generation of initial possible classification rules from a dataset. Then the proposed algorithm discovers important or significant rules from all possible classification rules without sacrificing predictive accuracy. The proposed algorithm deals with discrete valued data, and its initial population of candidate solutions contains particles of different sizes. The experiment has been done on the task of optimal rule selection in the data sets collected from UCI repository. Experimental results show that the proposed algorithm can automatically evolve on average the small number of conditions per rule and a few rules per rule set, and achieved better classification performance of predictive accuracy for few classes.

KEYWORDS

Particle swarm optimization, Data Mining, Classifiers.

1. INTRODUCTION

Many particle swarm algorithms have been developed that deals only with continuous variables [1, 5, 10]. This is a significant limitation because many optimization problems are there which featuring discrete variables in the problem domain. Typical problems are there in the space which deals with the ordering, grouping or arranging of discrete variables such as scheduling or routing problems. Therefore, the developing of particle swarm algorithms that deals with discrete variables is important for such kind of problem. We propose a variant of Particle Swarm Optimization (PSO) algorithm applied to dicerete data for rule selection in Data Mining. We will refer to this algorithm as the discrete Particle Swarm Optimization (DPSO) algorithm. The DPSO deals with discrete valued data, and its population of candidate solutions contains particles of different sizes. Although the algorithm has been specifically designed for optimized rule selection task, it is by no means limited to this kind of application. The DPSO algorithm may be applied to

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 301–306, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4530

other optimization problems with little modification. Construction of classifier model for any decision system is an important job for many data mining applications. The objective of developing such a classifier is to classify unlabeled dataset (object belongs to test set) into classes. Here we propose a discrete Particle Swarm Optimization (PSO) algorithm for selection of optimal or near optimal classification rules from huge number of rules possibly exists in a dataset. Primarily, all exhaustive rules are generated from the dataset using decision matrix approach [11, 12] which is based on rough set theory. Then the DPSO method identifies important or significant rules from all possible classification rules without sacrificing predictive accuracy. For the larger dataset, if all decision rules are considered for data analysis then time complexity will be very high. For this reason a minimum set of rules are generated.

The paper is organized as follows. Section 2 briefly introduces PSO algorithm. Section 3 introduces the DPSO algorithm proposed in this paper for the task of optimal rule selection. Section 4 and 5 reports experimental methodology, experiments and results respectively and finally Section 6 presents conclusions of the work.

2. A Brief Introduction To Particle Swarm Optimization

PSO [5, 6, 7] is an evolutionary optimization algorithm proposed by Kennedy and Eberhart in 1995. In PSO, a population, called a swarm, of candidate solutions are encoded as particles in the search space. PSO starts with the random initialization of a population of particles. The whole swarm move in the search space to search for the best solution by updating the position of each particle based on the experience of its own and its neighbouring particles. In PSO a potential solution to a problem is represented by a particle X(i) = ($x_{(i,1)}$, $x_{(i,2),}$ …,$x_{(i,n)}$) in an n-dimensional search space. The coordinates $x_{(i,d)}$ of these particles have a rate of change(velocity) $v_{(i,d)}$, d = 1, 2, ..., n. Every particle keeps a record of the best position that it has ever visited. Such a record is called the particle's previous best position and denoted by B_i. The global best position attained by any particle so far is also recorded and stored in a particle denoted by G. Iteration comprises evaluation of each particle with the adjustment of $v_{(i, d)}$ in the direction of particle X(i)'s previous best position and the previous best position of any particle in the neighbourhood.

Generally speaking, the set of rules that govern PSO are: evaluate, compare and evolve. The evaluation phase measures how well each particle (candidate solution) solves the problem. The comparison phase identifies the best particles. The evolve phase produces new particles based on some of the best particles previously found. These three phases are repeated until a given stopping criterion is matched. The objective is to find the best particle which solves the target problem. Important concepts in PSO are velocity and neighbourhood topology. Each particle, X(i), is associated with a velocity vector. This velocity vector is updated at every generation. The updated velocity vector is then used to generate a new particle X(i). The neighbourhood topology defines how other particles in the swarm, such as B (i) and G, interact with X (i) to modify its respective velocity vector and, consequently, its position as well.

3. The Proposed Discrete PSO Algorithm

The algorithm presented here is based on DPSO algorithm [6]. The proposed algorithm deals with generation of optimized classification rules from discrete valued dataset, which is typically a rule mining problem. Primarily, all exhaustive rules are generated from the dataset using decision matrix approach [11, 12] which is based on rough set theory. Every rule has two parts, conditional part and decision part, conditional part comprises of some conditional attributes with their values and decision part has decision attribute with the corresponding decision value or class. In a rule, each conditional attribute with the corresponding value is termed as rule component. So a rule is formed by some rule components. In PSO, population of candidate solutions contains particles of different sizes. Here population of candidate solutions is generated from initial rule set. Each particle represents the antecedent part of a rule by considering the conclusion part is fixed at the beginning of the execution and represents a target attribute value. In this case to discover the optimal rule set, predicting different target/decision values, the algorithm has to run several times, one for each decision value. There are N (No of initial rules) particles in a swarm. The length of each particle may vary from 1 to n, where n is the number of unique rule component present in the initial rule set, which is made by considering only the antecedent part of each rule. Each particle R_i keeps a record of the best position it has ever attained. This information is kept in a distinguished particle labeled as B_i. The swarm also keeps the information of the global best position ever attained by any particle in the swarm. This information is also kept in a separated particle labeled G. G is equal to the best Bi present in the swarm.

3.1 ENCODING OF THE PARTICLES FOR THE PROPOSED DPSO ALGORITHM

Each rule is formed by some rule components and each rule component is identified by a unique positive integer number or index. These indices, vary from 1 to n, where n is the number of unique rule components present in the initial rule set. Each particle is subsets of indices without repetition. For example, corresponding to the rules R_1, R_2 and R_3 given below, the rule components are (A=1), (A=2), (B=3), (B=4) and (C=5) which are indexed as 1, 2, 3, 4 and 5 respectively.

R_1= {A=1, B=3, C=5}
R_2= {A=2, B=4}
R_3= {A=2, C=5}

Where, A, B, C are the conditional attributes, N (Number of initial rules) = 3. Here initial swarm representing candidate solution could looks as follows:

R_1= {1, 3, 5}
R_2= {2, 4}
R_3= {2, 5}.

3.2 THE INITIAL POPULATION FOR THE PROPOSED DPSO ALGORITHM

The initial population of particles is generated as follows.

Population of candidate solutions is generated from initial rule set. Each particle represents the antecedent part of a rule by considering the conclusion part as fixed at the beginning of the execution and represents a target attribute value. Rule encoding process is described in section 3.1. By considering the rule encoding process the particles are formed for each rule generated using decision matrix based classification method.

3.3 VELOCITIES

The DPSO algorithm does not use a vector of velocities. It works with proportional likelihoods. Basically, the idea of proportional likelihood used in the DPSO algorithm is almost similar with the idea of velocity used in the standard PSO algorithm. Each particle is associated with a 2 × n array of proportional likelihoods, where 2 and n represents number of rows and number of columns respectively. In this standard proportional likelihood array, each element in the first row of V_i represents the proportional likelihood based on which a rule component be selected. The second row of V_i has the indices of the rule components which is associated with the respective proportional likelihoods of the first row of the vector V_i. There is a one-to-one correspondence between the columns of this array. At the beginning, all elements in the first row of V_i are set to 1, e.g., V_i = {{1, 1, 1, 1, 1 }, {1,2,3,4,5}}.

After the initial population of particles is generated, this array is always updated before a new configuration for the particle associated to it is made. The updating process is based on R_i, B_i (particle's previous best position) and G (global best position) and works as follows.

In addition to R_i, Bi and G, three constant updating factors, namely, a, b and c are used to update the proportional likelihoods $v_{(i,d)}$. These factors determine the strength of the contribution of R_i, B_i and G to the adjustment of every coordinate $v_{(i,d)}$ € $V_{i.}$ Parameter values of a, b and c is chosen experimentally.

3.4 GENERATING NEW PARTICLES FOR THE PROPOSED DPSO ALGORITHM

The proportional likelihood array V_i is then used to sample a new configuration of particle R_i that is, the particle associated to it. First, for a particle with V_i, all indices present in Ri have their corresponding proportional likelihood increased by 'a'. Similarly, all indices present in B_i and G have their corresponding proportional likelihood increased by 'b' and 'c' respectively. Now each element of the first row of the array V_i is then multiplied by a uniform random number between 0 and 1. A new random number is generated for every single multiplication performed. The new particle is then defined by ranking the columns in V_i by the values in its first row. That is, the elements in the first row of the array are ranked in a decreasing order of value and the indices of the rule components (in the second row of V_i) follow their respective proportional likelihoods.

Thus for example, after all the steps if particle i has length 3, and the particle associated to the array

$$V_i = \begin{array}{ccccc} 0.74 & 0.57 & 0.50 & 0.42 & 0.20 \\ 5 & 4 & 3 & 1 & 2 \end{array}$$

Then first 3 indices from the second row of Vi would be selected to compose the new particle. That is, R_i = {*, *, *} the indices (rule components) 5, 4 and 3 would be selected to compose the new particle, i.e., R_i = {5, 4, 3}. Note that indices that have a higher proportional likelihood are, on average, more likely to be selected. If indices are the rule components with same attribute are selected, then higher confidence rule component is selected. In that case new particle size is also changes in the generation. The updating of R_i, B_i and G is identical to what is described earlier. In this way new particles are formed generation by generation.

4. EXPERIMENTAL METHODOLOGY

The purpose of this experiment was to evaluate the classification accuracy and comprehensibility by the number of rules in the data set, and the average number of rule conditions per rule. The fitness function f (Ri) of any particle i is computed as follows. Optimized rule selection process can be performed by DPSO as a multi objective problem to maximize the confidence (association rule mining concept) of a rule to achieve higher classification accuracy as well as minimizing the length of a rule. The goal is to see whether DPSO can select optimized set of rules to achieve a higher classification accuracy rate from initial set of rules.

In this regard following fitness function is considered.
$F_i = \alpha * rule_i\_confidence + (1-\alpha) * 1 / (rule_i\_length)$

$rule_i$ _confidence: A rule like, $R_i \rightarrow (C_{i1} = a)^\wedge(C_{i2} = b)\ldots^\wedge(C_{in} = c) \rightarrow (D = d_i)$ has the condition part C_i where conditional attributes are associated with value, and the decision part D has the decision values d_i. So here the rule maps $C_i \rightarrow d_i$.

Then the confidence of the rule is conf (R_i) = number of rows in *dataset* that match C_i and have class label ***d_i*/** number of rows in *dataset* that match only C_i.

Here relative importance of the rule confidence and the length of the rule are considered. Rule confidence is set larger than length of a rule because the classification performance is assumed more important than the number of conditional attributes present in a rule i.e. rule length. The objective is to find the fittest rules with which it is possible to classify the data set as belonging to one of the classes with an acceptable accuracy. Here α=0.8 is considered.

5. EXPERIMENTS AND RESULT

The computational experiments involved a 10-fold cross-validation method [13].To illustrates the method, wine dataset from UCI repository [14] of Machine Learning databases [10] is considered. First, the 178 records with 3 distinct classes (0, 1, 2) in the wine data set were divided into 10 almost equally sized folds. The folds were randomly generated but under the following regulation. The proportion of different classes in every single fold was similar to the one found in the original data set containing all the 178 records. Each of the 10 folds is used once as test set and the remaining of the data set is used as training set. The purpose of the experiment was to evaluate the generalization ability, measured as the classification accuracy on test set. Each training set was used for generation of initial classification rules. For each initial rule set, DPSO algorithm generates optimized set of classification rules according to their fitness which is used to classify the test set (each 10 test folds) accordingly. DPSO selects only the best particle in each run as the rule.

As the DPSO algorithm is stochastic algorithm, 20 independent runs for the algorithm were performed for every single fold and for every decision classes. The average number of rules by the rule selection algorithms has always been rounded to the nearest integer. The population size (initial rule set) used for the algorithm is on average 300 and the search stops in one run after 100 iterations. Other choices of parameter values were a = 0.10, b = 0.12 and c = 0.14. The results of experiment were as follows. For Wine on average 95 rules are selected as optimized rules from 300 rules. And using these 95 rules on test set 99.32% of average classification accuracy is achieved.Using fewer rules, DPSO algorithm obtained on average, a better predictive accuracy than the classification performed using all the initial classification rules for few classes. For the initial rule set average accuracy values on test set for 3 classes (0, 1, 2)were 99.05, 99.48, 99.77, while using optimized rule set corresponding accuracy values on test set were 98.38, 99.59, 100. The results also indicate that not only the predictive accuracy is good, but also the number of rule conditions or rule length is relatively short like 2- 4 rule components in a rule where the same was 3-5 in the initial rule set and there are a small number of rules in the rule set.

6. CONCLUSION

We proposed an efficient method for rule discovery using DPSO algorithm. The discovered rules are with of high accuracy and comprehensibility. Using fewer rules, DPSO algorithm obtained on average, a better predictive accuracy than the classification performed using all the initial classification rules for few classes. The results on the Wine data set show that our approach has good performance for rule discovery on discrete data. Few less coverage rules were also selected. Rule Coverage is an important parameter for selection of good quality rule. In the future by applying some rule pruning technique better quality rule set can be generated.

REFERENCES

[1] T. Blackwell and J. Branke. Multi-swarm optimization in dynamic environments. In Lecture Notes in Computer Science, volume 3005, pages 489{500. Springer-Verlag, 2004.

[2] E. S. Correa, M. T. Steiner, A. A. Freitas, and C. Carnieri. Using a genetic algorithm for solving a capacity p-median problem. Numerical Algorithms, 35:373{388, 2004.

[3] D. Freedman, R. Pisani, and R. Purves. Statistics. W. Norton & Company, 3rd edition, September 1997.

[4] A. A. Freitas. Data Mining and Knowledge Discovery with Evolutionary Algorithms. Springer-Verlag, October 2002.

[5] G. Kendall and Y. Su. A particle swarm optimisation approach in the construction of optimal risky portfolios. In Proceedings of the 23rd IASTED International Multi-Conference on Applied Informatics, pages 140{145, 2005. Artificial intelligence and applications

[6] J. Kennedy and R. C. Eberhart. A discrete binary version of the particle swarm algorithm. In Proceedings of the 1997 Conference on Systems, Man, and Cybernetics, pages 4104{4109, Piscataway, NJ, USA, 1997. IEEE.

[7] J. Kennedy and R. C. Eberhart. Swarm Intelligence. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001

[8] T. M. Mitchell. Machine Learning. McGraw-Hill, August 1997.

[9] R. Poli, C. D. Chio, and W. B. Langdon. Exploring extended particle swarms: a genetic programming approach. In GECCO'05: Proceedings of the 2005 Conference on Genetic and Evolutionary Computation, pages 169{176, New York, NY, USA, 2005. ACM Press.

[10] Y. Shi and R. C. Eberhart. Parameter selection in particle swarm optimization. In EP'98: Proceedings of the 7th International Conference on Evolutionary Programming, pages 591{600, London, UK, 1998. Springer-Verlag

[11] I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2nd edition, 2005

[12] N.Zhong and A. Skowron, "A Rough Set-Based Knowledge Discovery Process", Int. Journal of Applied Mathematics and Computer Science., 11(3), 603-619, 2001. BIME Journal, Volume (05),), 2005

[13] Ethem Alpaydin Introduction to Machine Learning.PHI, 2010.

[14] Murphy, P. and Aha, W.: UCI repository of machine learning databases (1996), http://www.ics.uci.edu/mlearn/MLRepository.html

AN UNSUPERVISED METHOD FOR REAL TIME VIDEO SHOT SEGMENTATION

Hrishikesh Bhaumik[1], Siddhartha Bhattacharyya[2] and Susanta Chakraborty[3]

[1,2]Department of Information Technology,RCC
Institute of Information Technology, Kolkata, India
hbhaumik@gmail.com, dr.siddhartha.bhattacharyya@gmail.com
[3]Department of Computer Science and Technology,
Bengal Engineering and Science University, Shibpur, Howrah, India
susanta.chak@gmail.com

ABSTRACT

Segmentation of a video into its constituent shots is a fundamental task for indexing and analysis in content based video retrieval systems. In this paper, a novel approach is presented for accurately detecting the shot boundaries in real time video streams, without any a priori knowledge about the content or type of the video. The edges of objects in a video frame are detected using a spatio-temporal fuzzy hostility index. These edges are treated as features of the frame. The correlation between the features is computed for successive incoming frames of the video. The mean and standard deviation of the correlation values obtained are updated as new video frames are streamed in. This is done to dynamically set the threshold value using the three-sigma rule for detecting the shot boundary (abrupt transition). A look back mechanism forms an important part of the proposed algorithm to detect any missed hard cuts, especially during the start of the video. The proposed method is shown to be applicable for online video analysis and summarization systems. In an experimental evaluation on a heterogeneous test set, consisting of videos from sports, movie songs and music albums, the proposed method achieves 99.24% recall and 99.35% precision on the average.

KEYWORDS

Real time video segmentation, spatio-temporal fuzzy hostility index, image correlation, three-sigma rule

1. INTRODUCTION

There has been a spectacular increase in the amount of multimedia content transmitted and shared over the internet. The number of users for video-on-demand and Internet Protocol Television (IP-TV) services are growing at an exponential rate. According to 2012 statistics, there were more than one million IP-TV streams per month on the Zatoo platform. Textual annotation i.e. associating a set of keywords for indexing multimedia content was performed to facilitate searching of relevant information in existing video repositories (e.g. YouTube, DailyMotion etc.). However, manual annotation of the millions of videos available in such repositories is a cumbersome task. Content based video analysis techniques [1, 3, 4, 5, 6, 7, 8, 9] do away with manual annotation of the video data and results in saving time and human effort, with increase in accuracy. Video segmentation is the preliminary step for analysis of digital video. These segmentation algorithms can be classified according to the features used, such as pixel-wise

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 307–318, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4531

difference [10], histograms [11], standard deviation of pixel intensities [12], edge change ratio [13] etc. A more daunting task arises for online analysis and indexing of live streaming videos such as telecast of matches or live performances. Generation of highlights or summarization of such events is a challenging task as it involves development of algorithms which have good performance and are adapted to work in real time. Although video edits can be classified into hard cuts, fades, wipes etc. [9], the live streams mainly contain hard cuts, which is the main focus of the proposed approach. Several methods for video segmentation [8, 10, 11, 13, 14] exist in the literature, which have been applied for non-real time videos. However, shot boundary detection for streaming videos is hard to find in the literature. The difficulties and problems related to setting thresholds and the necessity of automatic threshold have been discussed in [1]. Hence, the motivation behind this work was to develop a technique for video segmentation, which is able to detect with high accuracy, the hard cuts present in live streaming videos. Video segmentation is performed on the fly, as new video frames are streamed in. Also, in this work, the problem of setting an automatic threshold has been addressed. The threshold is set dynamically without any *a priori* knowledge about the type, content or length of the video. The algorithm incorporates a look back mechanism to detect any missed hard cuts particularly during the start of the video when the statistical measures used to set the threshold are unstable. The novelty of the proposed work lies in its applicability in video summarization tasks of real time events, such as producing highlights of sports videos.

The reminder of the paper is organized as follows. In section 2, the basic concepts and definitions are presented. The proposed method for real time video segmentation is described in section 3. The experimental results and analysis are presented in section 4. The comparison with other existing approaches is also given in the same section. Finally, the concluding remarks are mentioned in section 5.

2. BASIC CONCEPTS AND DEFINITIONS

2.1. Application of fuzzy set theory to image processing

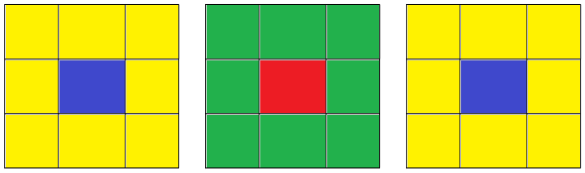
A colour image may be considered as a 3D matrix of values where each pixel colour depends on the combination of RGB intensities. Conversion of the colour image to a gray scale image involves mapping of the values in the 3D matrix to a 2D matrix, where each pixel value is in the range [0, 255]. This 2D matrix of values may be scaled to the range [0, 1] by dividing each element of the matrix by 255. The scaled value represents the membership value of each pixel to the fuzzy sets labelled as WHITE and BLACK. The number of elements in each fuzzy set is equal to the number of elements in the said 2D matrix. If a value 0 represents a completely black pixel and 1 a completely white one, then value of each element depicts the degree of membership $\mu_W(p_i)$ of the i^{th} pixel p_i to the fuzzy set WHITE. The degree of membership $\mu_B(p_i)$ of the pixel p_i to the set BLACK can be represented as $\mu_B(p_i) = 1 - \mu_W(p_i)$. Incorporating the fuzzy set theory, Bhattacharyya et al. [2] have proposed the fuzzy hostility index for detecting the point of interest in an image, which is useful for high speed target tracking. As an extension of this concept, a new index is proposed for obtaining the edge map of a video frame as explained in the next sub-section.

2.2. Edge Map using Spatio-Temporal Fuzzy Hostility Index (STFHI)

Fuzzy hostility index [2] indicates the amount of variation in the pixel neighbourhood with respect to itself. The pixel hostility has high value if the surrounding pixels have greater difference of values as compared to the candidate pixel i.e. heterogeneity in its neighbourhood is more. In an n-order neighbourhood the hostility index (ζ) of a pixel is defined as:-

$$\zeta = \frac{3}{2^{n+1}} \sum_{i=1}^{2^{n+1}} \frac{|\mu_p - \mu_{qi}|}{|\mu_p + 1| + |\mu_{qi} + 1|} \qquad (1)$$

where μ_p is the membership value of the candidate pixel and μ_{qi} ; i =1, 2, 3, . . . , 2^{n+1} are the membership values of its fuzzy neighbours in a second-order neighbourhood fuzzy subset. The value of the fuzzy hostility index ζ lies in [0, 1], with $\zeta = 1$ signifying maximum heterogeneity and $\zeta = 0$ indicating total homogeneity in the neighbourhood. The concept of fuzzy hostility index can be effectively extended to accommodate temporal changes in the time sequenced frames of a video.



(a) (b) (c)

Figure 1. (a) Pre-frame (f_{i-1}) (b) Present frame (f_i) (c) Post-frame (f_{i+1})

The STFHI (λ) of a pixel in the i^{th} image frame f_i of a video is a function ω of the fuzzy hostility index of the candidate pixel in f_i (marked red in figure 1(b)) and the corresponding pixels in the previous f_{i-1} and post f_{i+1} frames (marked blue in figures 1(a) and 1(c)), can be expressed as follows:- $\lambda_{f_i} = \omega(\zeta_{f_{i-1}}, \zeta_{f_i}, \zeta_{f_{i+1}})$ (2)

In other words, λ of a pixel is a function of the second order neighbourhood of its corresponding pixels (marked yellow in figures 1(a) and 1(c)) and itself (marked green in figure 1(b)). λ_{f_i} is computed as the average of $\zeta_{f_{i-1}}$, ζ_{f_i} and $\zeta_{f_{i+1}}$ except for the first and last frames of a video where $\zeta_{f_{i-1}}$ and $\zeta_{f_{i+1}}$ are not present respectively. The 2D matrix thus formed by computing the λ of each pixel will represent the edge map of an image with profound edges of all objects present in the original image as depicted in figure 3(b).

2.3. Pixel Intensification Function

Pixel intensity scaling is performed to make the edges more prominent compared to other portions of the image as shown in figure 3(c). The intensity scaling function (ϕ) used in this work is shown in figure 2 and mathematically represented as:-

$$\begin{aligned} \phi &= (\lambda_{ij})^2, \text{ if } \lambda_{ij} < 0.5 \\ &= (\lambda_{ij})^{1/2}, \text{ if } \lambda_{ij} \geq 0.5 \end{aligned}$$

where λ_{ij} is the STFHI of the pixel at i^{th} row and j^{th} column.

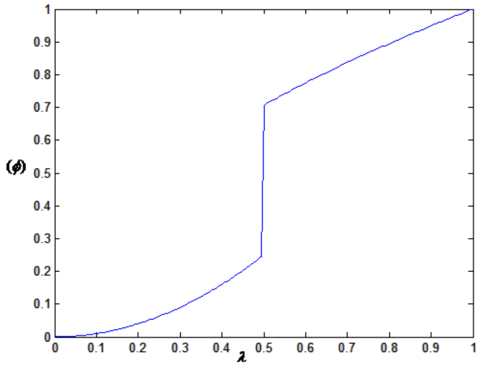
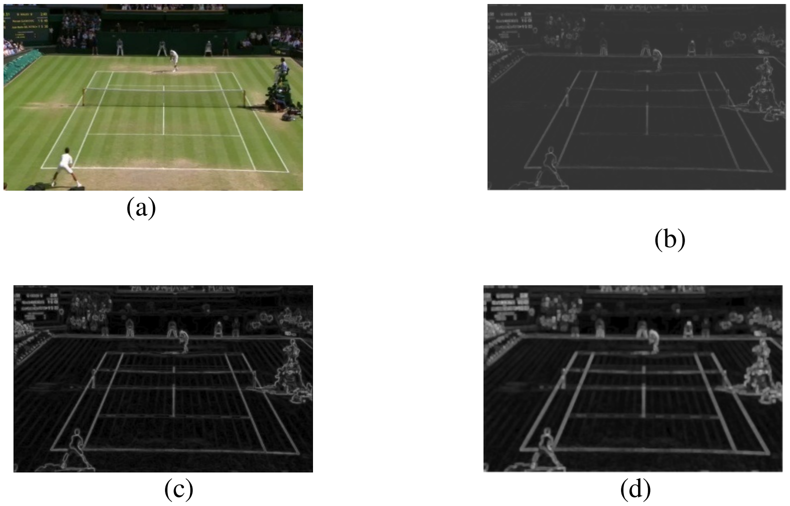


Figure 2. Pixel Intensity Function

2.4. Edge Dilation

Edge dilation is a technique which is used to enlarge the boundaries of the objects in a grayscale image as depicted in figure 3(d). This may be used to compensate for camera and object movement. A 3×3 square structuring element is used to dilate the edges of the grayscale image so generated from the fuzzy hostility map. The value of correlation between the similar images is increased as a result of edge dilation.



(a)

(b)

(c)

(d)

Figure 3. (a) Original image frame (b) Edge map using STFHI (c) Pixel intensified frame (d) Edge dilated frame

2.5. Computing Edge Map Similarity

The similarity between two edge maps can be considered as computing the similarity of two 2D matrices. Therefore, computing the edge map similarity can be achieved by finding the correlation between the matrices of the edge maps. Computing the Pearson's correlation coefficient ($\rho_{X,Y}$) between two matrices X and Y of same dimensions, may be represented as:

$$\rho_{X,Y} = \frac{\sum(x_{ij}-\bar{x})(y_{ij}-\bar{y})}{\sqrt{\{\sum(x_{ij}-\bar{x})^2\}(\sum(y_{ij}-\bar{y})^2\}}} = \frac{\sum x_{ij}y_{ij} - \frac{\sum x_{ij}\sum y_{ij}}{n}}{\sqrt{\{x_{ij}{}^2 - \frac{(\sum x_{ij}{}^2)}{n}\}\{y_{ij}{}^2 - \frac{(\sum y_{ij}{}^2)}{n}\}}} \quad (3)$$

where x_{ij} and y_{ij} are the elements in the i^{th} row and j^{th} column of matrices X and Y respectively, $\bar{x}$ is the mean value of elements of X , $\bar{y}$ is the mean value of elements of Y and n is the total number of elements in the matrix under consideration. The correlation is defined only if both of the standard deviations are finite and both of them are nonzero. The correlation is 1 in the case of an increasing linear relationship, -1 in the case of a decreasing linear relationship, and some value in between in all other cases, indicating the degree of linear dependence between the matrices. It is appropriate to mention here that a high value of correlation value indicates high similarity between the image frames.

2.6. Three Sigma Rule

The standard deviation (σ) of a dataset or probability distribution denotes the variation or deviation from the arithmetic mean (M) or expected value. The three-sigma rule in statistics is used to signify the range in which the values of a normal distribution will lie. According to this rule (refer figure 4), 68.2% values in a normal distribution lie in the range $[M-\sigma, M+\sigma]$, 95.4% values in $[M-2\sigma, M+2\sigma]$ and 99.6% in the range $[M-3\sigma, M+3\sigma]$. Hence, this empirical rule may be reliably used to compute a threshold to detect values which represent abrupt changes. In the proposed method, three-sigma rule has been used to detect the hard cuts at shot boundaries.

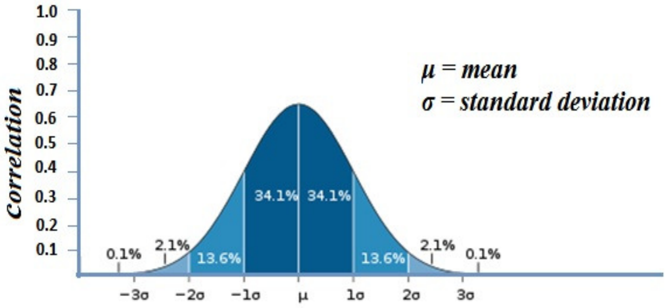


Figure 4. Normal Distribution with three standard deviations from mean

2.7. Real time updating of parameters used for dynamic threshold

The correlation between the edge maps of consecutive image frames provides an indication about the similarity of the video frames. However, for detection of a video segment, the correlation gradient is computed from the stream of consecutive correlation values obtained. As mentioned in the previous sub-section, the threshold is computed from the correlation gradient values using the three sigma rule. The parameters involved in calculating the threshold are mean (M) and standard deviation (σ) of the correlation gradient values. It must be noted that these parameters are to be updated in real time as new video frames are streamed in. The new mean (M_{new}) and (σ_{new}) standard deviation may be obtained as follows:-

$M = \frac{\sum_{i=1}^{N} C_i}{N}$ where, C_i is the correlation gradient and N is the number of correlation gradient values calculated from the frames received. On arrival of a new frame the value of the new mean will be:-

$$M_{new} = \frac{\sum_{i=1}^{N+1} C_i}{N+1} = \frac{\sum_{i=1}^{N+1} C_i + C_{N+1}}{N+1} = \frac{MN + C_{N+1}}{N+1} \Rightarrow M_{new} = (\frac{N}{N+1})M + (\frac{1}{N+1})C_{N+1} \qquad (4)$$

The new value of standard deviation (σ_{new}) may be calculated as follows:-

Since, $\sigma = \sqrt{\frac{\Sigma(C_i - M)^2}{N}} \Rightarrow N\sigma^2 = \sum (C_i - M)^2$

$$\Rightarrow (N+1)\sigma_{new}^2 = \sum_{i=1}^{N+1} (C_i - M_{new})^2 \quad \Rightarrow (N+1)\sigma_{new}^2 = \sum_{i=1}^{N+1} C_i^{\,2} - (N+1)M_{new}^2$$

$$\Rightarrow \sigma_{new}^2 = \frac{1}{N+1}\left[\sum_{i=1}^{N} C_i^{\,2} + C_{N+1}^2\right] - M_{new}^2 \qquad (5)$$

Thus, equations (4) and (5) may be used to update the parameters required for calculating the new threshold. The two equations are significant because the new values of mean and standard deviation can be calculated in real time from the new correlation value computed and earlier values of the two parameters, without having to recalculate the parameters over the whole span.

3. Proposed Method For Real Time Video Segmentation

The proposed detection mechanism for real time video shot segmentation is shown as a flow diagram in figure 5. The steps are described in the following subsections.

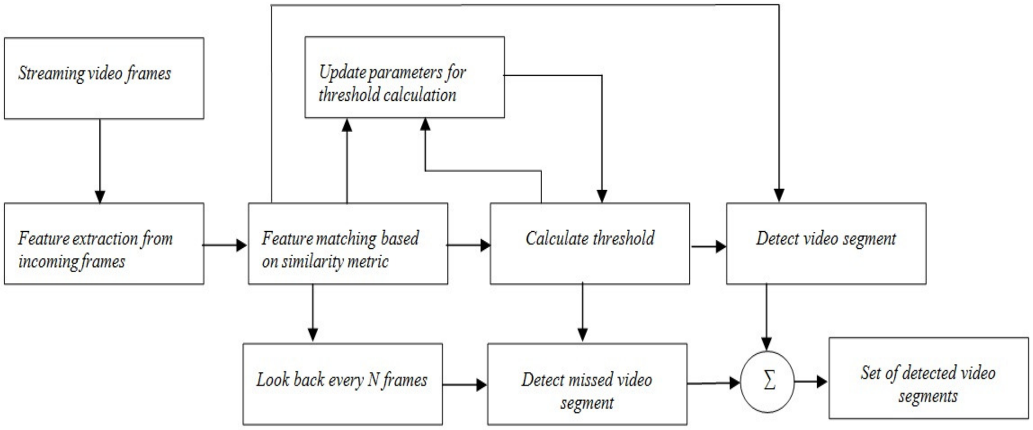

Streaming video frames
Update parameters for threshold calculation
Feature extraction from incoming frames
Feature matching based on similarity metric
Calculate threshold
Detect video segment
Look back every N frames
Detect missed video segment
Σ
Set of detected video segments

Figure 5. Flow diagram of the video segment detection process in real time

3.1 Extraction of time sequenced image frames from a video

The streamed video frames are decomposed into its constituent image frames in a time sequenced manner by a standard codec corresponding to the file type i.e. AVI, MPEG, MP4 etc. The extracted images are in uncompressed format and are stored as bitmaps for further processing.

3.2 Feature extraction from the image frames

The feature extraction process consists of generating a fuzzy hostility map using the STFHI for each image frame, as explained in section 2.2. The fuzzy hostility map indicates the amount of coherence/incoherence in the movement of the objects and is a 2D matrix which is used to generate the edges of the objects of the gray scale image. Thereafter, an intensity scaling function is used to make the edges more prominent as explained in section 2.3. To compensate for the rapid object or camera movement, edge dilation is performed as explained in section 2.4.

3.3 Feature matching based on similarity metric

In this step, the Pearson's correlation between successive fuzzy hostility maps is used as the similarity metric as explained in section 2.5. The correlation values thus computed are stored in a row matrix C_M . The shot boundaries occur at points of abrupt change in the correlation values. In order to detect a shot boundary, the gradient of the correlation values (which is a row vector) is computed. The correlation gradient plot is depicted in figure 6, consists of steep spikes at the points of shot boundary.

3.4 Calculation of threshold and updating of threshold parameters

Segments in the streaming video are detected by using the three-sigma rule as explained in section 2.6. If the correlation gradient exceeds the upper or lower threshold, a video segment is detected. Since threshold is a function of mean and standard deviation, it has to be updated as new frames are streamed in real time. The new threshold is calculated using equations (4) and (5) explained in section 2.7.

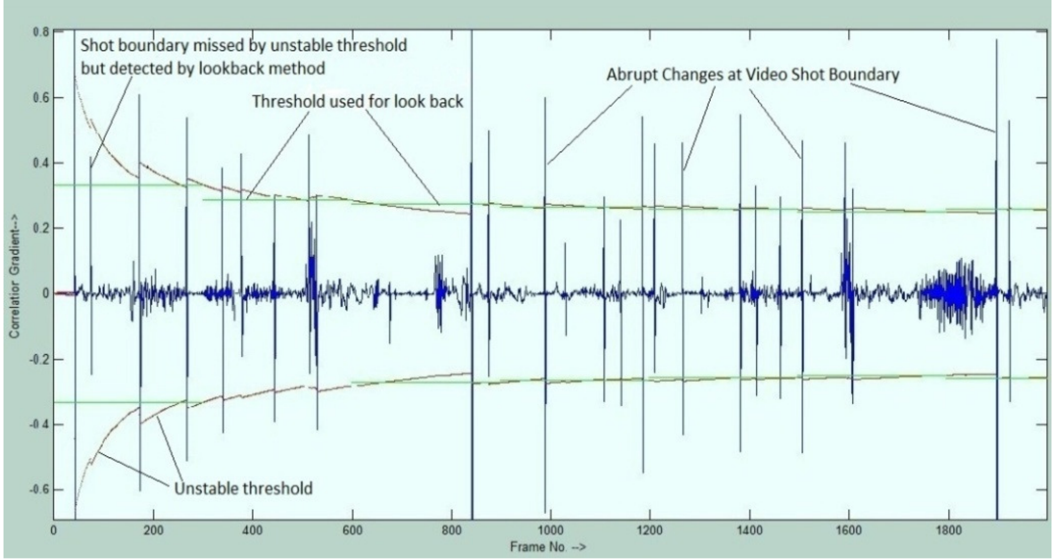


Figure 6. Plot of Correlation gradient values.

3.5 Look back N frames to detect missed shot boundaries

In the proposed video shot detection mechanism, the threshold for detection of shot boundary is set dynamically without any prior knowledge about the type or content of the video. At the initial stage, the threshold fluctuates due to smaller number of data points owing to the lesser number of arrived video frames as shown in figure 6. However, the threshold becomes more stable as more frames arrive in real time. Although the threshold is updated with the arrival of each frame, but the system looks back only after the passage of *N* frames, to check if any shot boundaries have been missed due to the unstable threshold. In this work, *N*=300 has been taken, i.e. look back will occur every 12 seconds in a video with frame rate 25 fps or every 10 seconds for 30 fps video.

4. Experimental Results And Analysis

The proposed method for shot segmentation in real time videos has been tested on a video data set consisting of eleven videos with varied features (Table I and Table II). All the videos in the test set have a resolution of 640×360 at 25 fps and are in MP4 format. The performance of the proposed method is evaluated by taking into consideration two parameters, recall (*R*) and precision (*P*) defined as follows:-

$$Recall = (B_d - B_f) / B_t$$

$$Precision = (B_d - B_f) / B_d$$

where, B_d : Shot boundaries detected by algorithm; B_f : False shot boundaries detected and B_t : Actual shot boundaries present in the video

4.1 The Video Dataset

The proposed method has been tested on a dataset consisting of two subsets. The first subset comprised five videos which were of long duration of average length of more than one hour (Table 1). The four videos V1 to V4 are documentaries taken from different TV channels. The video V5 is a highlights video taken from the Cricket World Cup 2011 final match between India and Sri Lanka.

The second subset composed of short videos (Table 2). The videos V6 and V7 are sports videos from cricket and tennis. The reason for including this in the dataset was because of the rapid object movement and small length shots. In contrast, V8 and V9 are movie songs from Hindi films. V8 has shots taken mostly outdoors in daylight whereas V9 has real life shots are mixed with some computer generated frames. The average shot duration in V9 is the least among all videos of the dataset. The video V10 is based on a violin track by Lindsey Stirling which is characterized by simultaneous movement of the performer as well as camera. The motivation for including this video is the rapid zoom-in and zoom-out sequences. The video V11 is the official song of the 2010 FIFA World Cup called “Waka Waka”. The video comprises of varied background and illumination, intermixed with match sequences taken from FIFA World Cup history.

Table 1. Test Video Dataset-I

Video	**V1**	**V2**	**V3**	**V4**	**V5**
Duration (mm:ss)	51:20	28:40	58:06	59:29	111:19
No. of Frames	74020	43018	87150	89225	166990
No. of Hard Cuts	941	406	807	1271	2807
Average no. of frames in each shot	78.57	105.69	107.85	70.14	59.46

Table 2. Test Video Dataset-II

Video	V6	V7	V8	V9	V10	V11
Duration (mm:ss)	02:33	02:58	02:42	04:10	03:27	03:31
No. of Frames	3833	4468	4057	6265	4965	5053
No. of Hard Cuts	46	43	70	172	77	138
Average no. of frames in each shot	81.55	101.54	57.14	36.21	63.65	36.35

4.2 Experimental Results

The proposed method for video shot segmentation is found to work accurately on video frames streamed in at real time. The results obtained by performing the experiments on the video data set are summarized in Table 3. The shot boundaries obtained are the summation of two phases. Video segments are detected using the threshold and method as explained in sections 3.2, 3.3 and 3.4. However some shot boundaries may be missed due to unstable mean and standard deviation. These missed shot segments are detected by updating the threshold and looking back after every *N* frames have elapsed. The methodology has been discussed in section 3.5. Effectiveness of the proposed method is seen from the high recall and precision values obtained for each of the videos in the test set.

Table 3. Experimental Results of Test Video Dataset

Video	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
Hard Cuts present	941	406	807	1271	2807	46	43	70	172	77	138
Hard Cuts detected	940	406	806	1269	2796	45	43	67	165	75	136
Detected by Lookback	1	0	1	2	5	2	0	2	5	1	3
False detection	0	0	0	0	1	2	0	1	1	0	1
Recall (%)	100	100	100	100	99.75	97.82	100	97.14	98.25	98.70	100
Precision (%)	100	100	100	100	99.96	95.74	100	98.55	99.41	100	99.28

4.3 Comparison with other existing methods

Several existing methods for video segmentation like Mutual Information (MI) [8], Edge Change Ratio (ECR) [13] and Color Histogram Differences (CHD) [14] have been applied for non-real time videos. Shot boundary detection for streaming videos is hard to find in the literature. The problem of automatic computation of a threshold has been addressed in the literature [1, 15] and strength of the proposed method lies in the fact that the threshold is computed and updated automatically without manual intervention, unlike the other existing methods. Hence, this method can be applied for on-the-fly detection of shot boundaries in real time videos. The comparative results of the proposed method with its non-real time counterparts are shown in Table 4.

Table 4**.** Comparison with Existing Methods

Video	**Proposed Method**		**MI**		**CHD**		**ECR**	
	R	**P**	**R**	**P**	**R**	**P**	**R**	**P**
V1	100%	100%	85.97%	92.98%	75.98%	95.96%	90.96%	92.98%
V2	100%	100%	83%	88.91%	78.07%	87.93%	92.11%	96.05%
V3	100%	100%	91.94%	93.55%	83.02%	88.97%	89.96%	85.99%
V4	100%	100%	87.96%	95.98%	76%	91.03%	95.98%	92.99%
V5	99.75%	99.96%	85.99%	93.97%	81.97%	91.98%	87.99%	94.01%
V6	97.82%	95.74%	86.95%	90.90%	73.91%	80.95%	91.30%	95.45%
V7	100%	100%	88.37%	92.68%	81.57%	96.87%	90.69%	88.63%
V8	97.14%	98.55%	91.42%	94.11%	75.71%	92.98%	95.71%	89.33%
V9	98.25%	99.41%	84.88%	94.80%	77.90%	94.36%	92.44%	88.33%
V10	98.70%	100%	81.81%	92.64%	74.02%	95%	93.50%	90%
V11	100%	99.28%	84.05%	95.08%	80.43%	94.87%	94.92%	91.60%

5. Conclusions And Remarks

The proposed method for real time video segmentation was tested on a diverse video test set. It is seen to outperform the existing methods in terms of both the recall and precision. As compared to the state-of-the-art techniques, the proposed method achieves nearly 100% accuracy in terms of both recall and precision. Also, the number of false detections is very less. The problem of automatically setting the threshold, without any human interference, has been addressed and results are very encouraging. The number of false hits is almost negligible as compared to the other existing methods. A major challenge would be to detect dissolve edits in streaming videos.

Acknowledgements

The authors acknowledge the contributions of Surajit Dutta, M.Tech student at RCC Institute of Information Technology, Kolkata, during the experimentation phase of the proposed method.

References

[1] Alan Hanjalic "Shot-Boundary Detection: Unraveled and Resolved," Circuits and Systems for Video Technology, IEEE Transactions, Volume:12, Page(s):90-105, February 2002.

[2] Siddhartha Bhattacharyya, Ujjwal Maulik, and Paramartha Dutta "High-speed target tracking by fuzzy hostility-induced segmentation of optical flow field," Applied Soft Computing ,Science Direct, 2009.

[3] Hattarge A.M., Bandgar P.A., and Patil V.M. "A Survey on Shot Boundary Detection Algorithms and Techniques", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 2, February 2013.

[4] Biswanath Chakraborty, Siddhartha Bhattacharyya, and Susanta Chakraborty "A Comparative Study of Unsupervised Video Shot Boundary Detection Techniques Using Probabilistic Fuzzy Entropy Measures", DOI: 10.4018/978-1-4666-2518-1.ch009 in Handbook of Research on Computational Intelligence for Engineering, Science, and Business, 2013.

[5] John S. Boreczky, and Lawrence A. Rowe "Comparison of video shot boundary detection techniques", Journal of Electronic Imaging, Page(s):122–128, March,1996.

[6] Swati D. Bendale, and Bijal. J. Talati "Analysis of Popular Video Shot Boundary Detection Techniques in Uncompressed Domain," International Journal of Computer Applications (0975 – 8887) ,Volume 60– No.3, IJCA, December, 2012.

[7] Ullas Gargi, Rangachar Kasturi, and Susan H. Strayer "Performance Characterization of Video-Shot-Change Detection Methods," IEEE Transaction on Circuits and Systems for Video Technology, Vol. 10, No. 1, IEEE, February 2000.

[8] Aarti Kumthekar and Mrs.J.K.Patil "Comparative Analysis Of Video Summarization Methods," International Journal of Engineering Sciences & Research Technology, ISSN: 2277-9655,2(1): Page(s): 15-18, IJESRT January, 2013.

[9] R. Lienhart, S. Pfeiffer, and W. Effelsberg "Scene determination based on video and audio features," Proceedings of IEEE International Conference on Multimedia Computing and Systems, Volume:1, Page(s):685 -690 IEEE, 1999.

[10] H. Zhang, A. Kankanhalli, and S.W. Smoliar "Automatic partitioning of full-motion video," Multimedia Systems, Volume: 1, no. 1, Page(s): 10–28, 1993.

[11] A. Nagasaka and Y. Tanaka "Automatic video indexing and full-video search for object appearances," Proceedings of IFIP 2nd Working Conference on Visual Database Systems, Page(s): 113-127, 1992

[12] A. Hampapur, R. C. Jain, and T. Weymouth "Production Model Based Digital Video Segmentation," Multimedia Tools and Applications, Vol.1, No. 1, Page(s): 9-46, March 1995.

[13] R. Zabih, J. Miller, and K. Mai "A Feature-Based Algorithm for Detecting and Classifying Scene Breaks," Proceedings of ACM Multimedia 1995, San Francisco, CA, Page(s):189-200, November, 1995.

[14] Vrutant Hem Thakore "Video Shot Cut Boundary Detection using Histogram," International Journal of Engineering Sciences & Research Technology, ISSN: 2277-9655, 2(4): Page(s): 872-875, IJESRT, April, 2013.

[15] L. Ranathunga, R. Zainuddin, and N. A. Abdullah "Conventional Video Shot Segmentation to Semantic Shot Segmentation," 6th IEEE International Conference on Industrial and Information Systems (ICIIS), Page(s):186-191, August 2011.

AUTHORS

Hrishikesh Bhaumik is currently serving as an Associate Professor and HOD of Information Technology Department at RCC Institute of Information Technology, Kolkata, India. He did B.Sc. from Calcutta University in 1997, AMIE in Electronics and Comm. Engg in 2000 and M.Tech in Information Technology from Bengal Engineering and Science University, Shibpur in 2004. In 2008 he received sponsorship for working on a Business Process Sniffer Tool developed at Infosys, Bhubaneswar. He made significant contributions to the EU-INDIA grid project in 2010 and 2011. His research interests include Content Based Video Retrieval Systems, Text Mining and High Performance Computing.

Siddhartha Bhattacharyya did his Bachelors in Physics, Bachelors in Optics and Optoelectronics and Masters in Optics and Optoelectronics (Gold Medal) from University of Calcutta, India in 1995, 1998 and 2000 respectively. He completed PhD (Engg.) in Computer Science and Engineering from Jadavpur University, India in 2008. He is currently an Associate Professor and Dean (R & D) in the Department of Information Technology of RCC Institute of Information Technology, Kolkata, India. He is a co-author of two books, co-editor of a book and has more than 90 research publications. He was the member of the Young Researchers' Committee of the WSC 2008 Online World Conference on Soft Computing in Industrial Applications. He was the convener of the AICTE-IEEE National Conference on Computing and Communication Systems (CoCoSys-09) in 2009. He is the Assistant Editor of International Journal of Pattern Recognition Research since 2010. He is the Associate Editor of International Journal of BioInfo Soft Computing since 2013. He is the member of the editorial board of International Journal of Engineering, Science and Technology and the member of the editorial advisory board of HETC Journal of Computer Engineering and Applications. He is a senior member of IEEE and a member of ACM, IRSS and IAENG. He is a life member of OSI and ISTE, India.His research interests include soft computing, pattern recognition and quantum computing.

Dr. Susanta Chakraborty received the B. Tech, M.Tech and Ph.D(Tech) in Computer Science in 1983, 1985 and 1999 respectively from the University of Calcutta. He is currently a Professor in the department of Computer Science and Technology at the Bengal Engineering Science and University, Shibpur, West Bengal, India. Prior to this he served at University of Kalyani as a **Dean of Engineering, Technology and Management faculty.** He has published around 31 research papers in reputed International Journals including IEEE Transactions on CAD and refereed international conference proceedings of IEEE Computer Science Press. He was awarded **INSA-JSPS Fellowship** of Indian National Science Academy (INSA) in the session 2003-2004. He has collaborated with leading scientists around the world in areas of Test Generation of Sequential Circuits and Low Power Design, VLSI testing and fault diagnosis, Quantum circuit and Testable Design of Nano-Circuits and Micro Fluidic Bio-chips. He has more than 25 years of research experience. He has served as the Publicity Chair, Publication Chair and Program Committee members of several International Conferences.

OPTICAL CHARACTER RECOGNITION PERFORMANCE ANALYSIS OF SIF AND LDF BASED OCR

Pradeep Kumar Jena[1], Charulata Palai[2], Lopamudra Sahoo[3]
and Anshupa Patel[4]

[1]Department of MCA, National Institute of Science and Technology,
Berhampur, Odisha, India
pradeep1_nist@ yahoo.com
[2]Department of CSE, National Institute of Science and Technology,
Berhampur, Odisha, India
charulatapalai@ gmail.com
[3,4]National Institute of Science and Technology, Berhampur, Odisha, India
06lopamudra@gmail.com
anshupa@gmail.com

ABSTRACT

The Optical Character Recognition (OCR) is becoming popular areas of research under pattern recognition and smart device applications. It requires the intelligence like human brain to recognize the various handwritten characters. Artificial Neural Network (ANN) is used to gather the information required to recognize the characters adaptively. This paper presents a performance analysis of character recognition by two different methods (1) compressed Lower Dimension Feature(LDF) matrix with a perceptron network, (2) Scale Invariant Feature (SIF) matrix with a Back Propagation Neural network (BPN). A GUI based OCR system is developed using Matlab. The results are shown for the English alphabets and numeric. This is observed that the perceptron network converges faster, where as the BPN can handle the complex script recognition when the training set is enriched.

KEYWORDS

Character recognition, perceptron network, back-propagation neural network, scale invariant feature, low dimensional feature.

1. INTRODUCTION

Automatic character recognition is a well-accepted area of research under pattern recognition. In handwritten character recognition, characters are written by different individuals that vary drastically from person to person due to variation in the writing style, its size and orientation of characters[1]. This makes the system difficult to recognize the characters. Artificial Neural Network (ANN) helps to solve the problem of identifying handwritten characters in an automated manner. ANN is an adaptive   computational model which is activated by set of pixels of a specific character as features, processing of the similar and divergence information available in the features used to recognize the character.

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 319–328, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4532

Various methods have been developed for the recognition of handwritten characters, such as the compressed Column-wise Segmentation of Image Matrix (CSIM) [2] using Neural Network. In this method, the image matrix is compressed and segmented column-wise then training and testing using a neural network for different character is performed. The Multi-scale Technique (MST)[2][3] used for high resolution character sets.

In case of feature based method, Scale Invariant features of characters such as height, width, centroid, number of bounded regions and textual features such as histogram information are used for character recognition. The features are feed to a Neural Network [4] for training and testing purpose. Hand written character recognition using Row-wise Segmentation Technique (RST) approach used to find out common features along the rows of same characters written in different hand writing styles and segmenting the matrix into separate rows and finding common rows among different hand writing styles[5]. Block-wise segmentation technique is also used for the character recognition [6] by matching the similarity among blocks of the characters.

The complex character such as Hindi, Oriya and Bangla character recognition [7][8] is more challenging at it produces very alike feature matrices for different characters due to their structural complexity. Hybrid methods are also applied to recognize the hand written characters. One such method is a prototype learning/matching method that can be combined with support vector machines (SVM) in pattern recognition [9].

It is observed that the finding the ideal feature set for particular language and normalizing the feature matrix is not easy, it requires substantial amount of processor time. In this work we have done a comparative study on character recognition using feature matrix with a Back Propagation Neural network (BPN) verses the character recognition using reduced dimensional block matrix(8x8) with a Perceptron Neural network for English hand written characters i.e. the upper case, lower case alphabet as well as digits.

2. METHODOLOGY

The sample handwritten characters are collected form ten different persons using black gel pen i.e. 10 samples of each letter each having different style. These blocks of characters were digitized using a scanner. Then each character is extracted from the scanned image automatically and saved with an appropriate name.

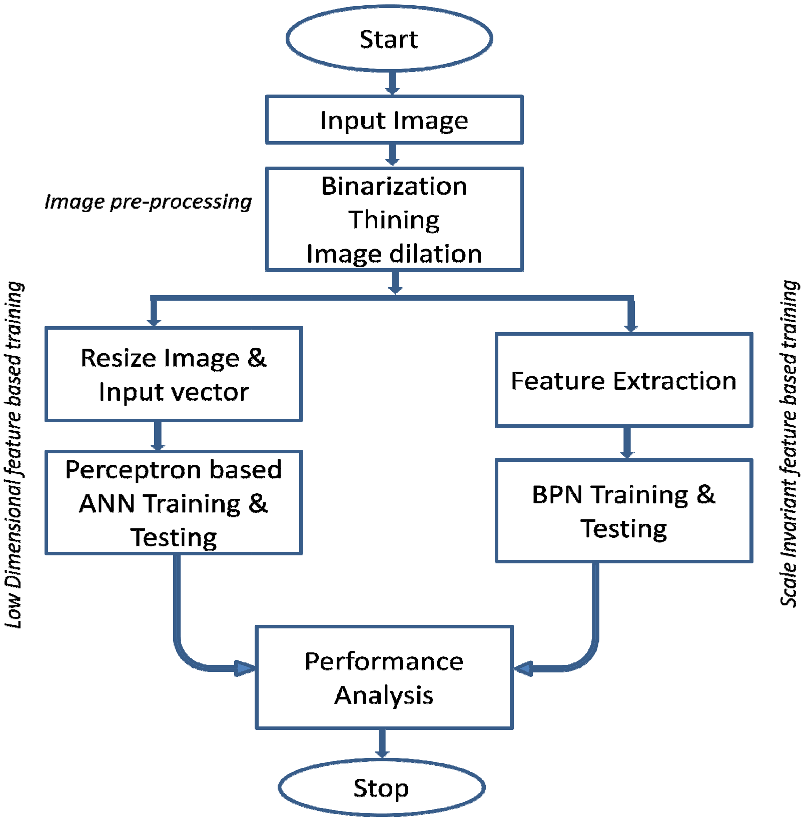


Figure 1: The Proposed Model for Analysis

2.1. Image Pre-processing

The individual character image is pre-processed to produce a skeletal template of the handwritten character. These involve various tasks such as (1) Binarization to reproduce the image with 0 (black) or 1(white), (2) Thinning to remove the thickness artefact of the pen used for writing characters, (3) image dilation to restore the continuity of the image pixels. Figure-2 show the inverted binarized data sheet of the set of handwritten characters.

2.2. Approach-1: Low dimensional feature based recognition

The images are resized into unique standard since the sample character images are different in dimension. They are converted to a reduced dimensional image matrix of size 8x8. It preserves only the highly significant features of the character that are used for the character recognition. The input vector i,e. a [64x1] matrix is prepared from the 8x8 image. It is used for the training and testing of perceptron neural network.

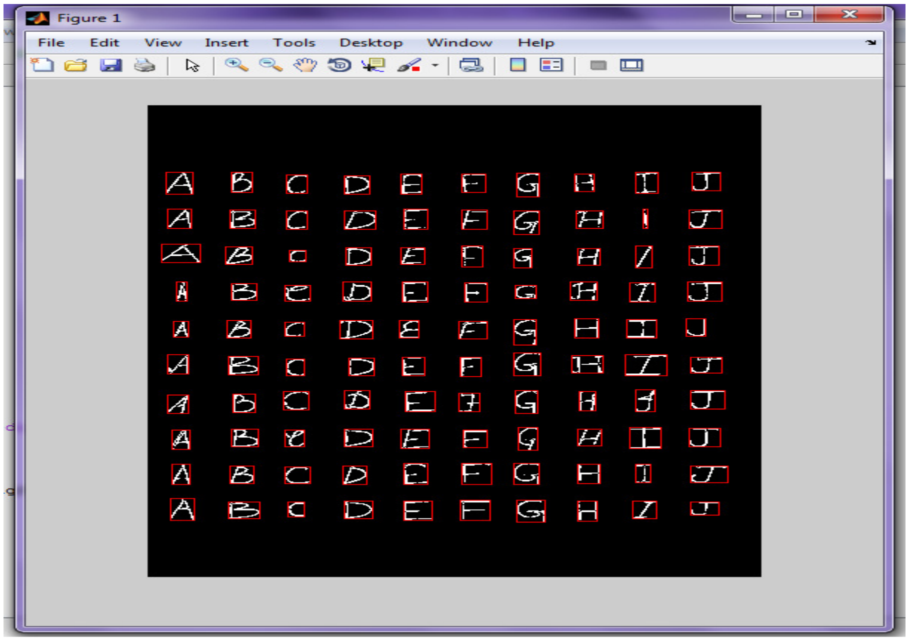


Figure 2: Processed image with bounding box

2.3. Rescaling of image matrix

Case-1: When the original image matrix is a multiple of 8.

i) Input image matrix
ii) Dimension of the original image is divided by 8 i.e. 64 x 32 will be 8 x 4
iii) Original matrix is split into 'n' uniform blocks of new dimension
iv) A uniform block is assigned to '1' if the number of 1's is greater than or equal to the number of 0's. Otherwise, a uniform block is assigned to '0'.

Case-2: When the original image matrix is not a multiple of 8.

i) Input image matrix
ii) Dimension of the original image is first converted to the nearest multiple of 8 by appending dummy (zeros) rows and columns i.e. 60 x 50 will be 64 x 56
iii) Dimension of the revised image is divided by 8
iv) Revised matrix is split into 'n' uniform blocks of new dimension
v) A uniform block is assigned to '1' if the number of 1's is greater than or equal to the number of 0's. Otherwise, a uniform block is assigned to '0'.

In this way images of different dimensions are resized into an 8 x 8 binary matrix. The columns of 8 x 8 matrixes are stored in a single column matrix one after other. Likewise, 10 samples of each 26 characters are considered and transformed it into column of size 64 each. As a result of a training set of 64 x 260 matrix is obtained. Accordingly the target set of 5 x 260 is generated.

2.4. Perceptron based ANN training

An Artificial Neural Network (ANN) is an adaptive computational system, it follows perceptron learning technique. The input layer consists of 64 neurons that represent one character as input, the hidden layer consists of 32 neuron, where as the output consists of 7 neurons that represents pattern of 0 and 1 which maps to an individual character. Each neuron is connected to other neurons by a link associated with weights. The weight contains information about the input, which is updated during each epoch.

2.5. Approach-2: Scale Invariant feature based recognition

Aspect Ratio: Height and width of character is obtained. The ratio of height and width remain approximately same for same person for the different characters.

$$AR = \frac{L}{W}$$

Where, AR=Aspect Ratio
L=Length of Character
W=Width of Character

Occupancy Ratio: This feature is the ratio of number of pixels which belong to the character to the total pixels in the character image. This feature provides information about character density.

Number of Horizontal Lines: It's the number of horizontal lines in a character. It's found out using a 3x3 horizontal template matrix.

Number of Vertical Lines: It's the number of vertical lines in a character. It's found out using a 3x3 vertical template matrix.

Number of Diagonal Lines: It's the number of diagonal-1 lines in a character. It's found out using a 3x3 diagonal-1 and diagonal-2 template matrix.

Number of Bounded Regions: It's the number of bounded areas found within a character image.

Number of End Points: End points are defined as those pixels, which have only one neighbor in its eight way neighborhood. Figure-3 the two end points of the image.

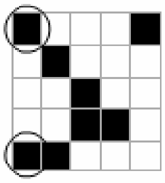


Figure 3: End point in an image

Vertical Center of Gravity: Vertical centre of gravity shows the vertical location of the character image. Vertical centre of gravity of image is calculated as follow

$$cog(v) = \frac{\sum y.Ny}{\sum Ny}$$

Where, Ny: the number of black pixels in each horizontal line Ly with vertical coordinate y,
Horizontal Center of Gravity: Horizontal centre of gravity shows the horizontal location of the character image. Horizontal centre of gravity of image calculated as follow:

$$cog(h) = \frac{\sum x.Nx}{\sum Nx}$$

Where, Nx: the number of black pixels in each vertical line Lx with horizontal coordinate x

2.6. BPN based ANN Training

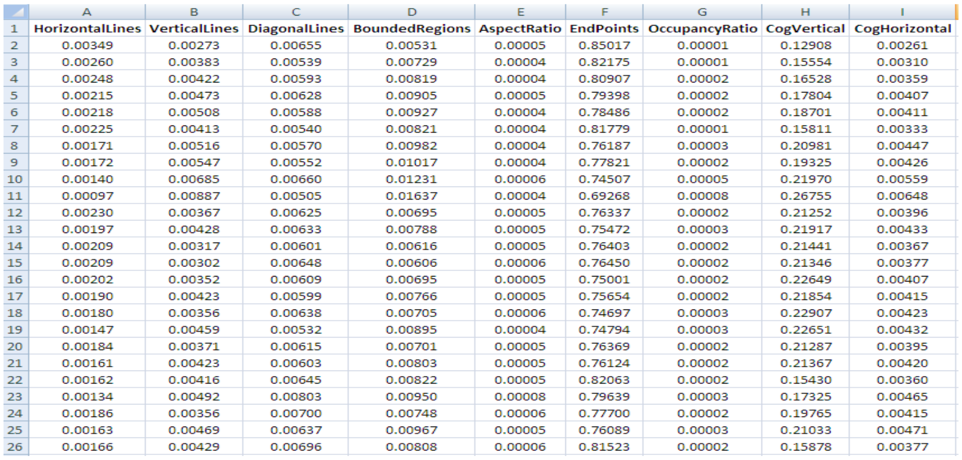
Back-propagation neural network is mostly used for the handwritten character recognition since it support the real values as input to the network. It is a multi-layer feed-forward network which consists of an input layer, hidden layer and output layer. The normalized values of the scale-invariant feature matrix are given as the input for the training, the output is pattern of 0's and 1's which maps to an individual character.

2.7. Testing and Recognition

Separate test sets are prepared for testing purpose. After the training process is completed, the test pattern is fed to the neural network to check the learning ability of the trained net. The output of the simulation is compared with the specified target set. The character is recognized by selective thresholding technique.

3. RESULT ANALYSIS

The OCR system is developed using MATLAB. The experimental results are shown below which is carried out to recognize the "A" and "5" as the inputs. The Figure-4 represents the normalized values of the scale invariant features of the English Alpha-numerics. Figure-5 indicates the performance analysis of BPN network as trained with normalized feature values. The Figure-6 shows the interface to load the test image and to select the training type such as BPN based training. Figure-7 indicates the recognition of "A", The Figure-8 shows the LDF matrix, Figure-9 represents the perceptron based training and testing. The recognition of "5" is shown in Figure-10.



	A	B	C	D	E	F	G	H	I
1	HorizontalLines	VerticalLines	DiagonalLines	BoundedRegions	AspectRatio	EndPoints	OccupancyRatio	CogVertical	CogHorizontal
2	0.00349	0.00273	0.00655	0.00531	0.00005	0.85017	0.00001	0.12908	0.00261
3	0.00260	0.00383	0.00539	0.00729	0.00004	0.82175	0.00001	0.15554	0.00310
4	0.00248	0.00422	0.00593	0.00819	0.00004	0.80907	0.00002	0.16528	0.00359
5	0.00215	0.00473	0.00628	0.00905	0.00005	0.79398	0.00002	0.17804	0.00407
6	0.00218	0.00508	0.00588	0.00927	0.00004	0.78486	0.00002	0.18701	0.00411
7	0.00225	0.00413	0.00540	0.00821	0.00004	0.81779	0.00001	0.15811	0.00333
8	0.00171	0.00516	0.00570	0.00982	0.00004	0.76187	0.00003	0.20981	0.00447
9	0.00172	0.00547	0.00552	0.01017	0.00004	0.77821	0.00002	0.19325	0.00426
10	0.00140	0.00685	0.00660	0.01231	0.00006	0.74507	0.00005	0.21970	0.00559
11	0.00097	0.00887	0.00505	0.01637	0.00004	0.69268	0.00008	0.26755	0.00648
12	0.00230	0.00367	0.00625	0.00695	0.00005	0.76337	0.00002	0.21252	0.00396
13	0.00197	0.00428	0.00633	0.00788	0.00005	0.75472	0.00003	0.21917	0.00433
14	0.00209	0.00317	0.00601	0.00616	0.00005	0.76403	0.00002	0.21441	0.00367
15	0.00209	0.00302	0.00648	0.00606	0.00006	0.76450	0.00002	0.21346	0.00377
16	0.00202	0.00352	0.00609	0.00695	0.00005	0.75001	0.00002	0.22649	0.00407
17	0.00190	0.00423	0.00599	0.00766	0.00005	0.75654	0.00002	0.21854	0.00415
18	0.00180	0.00356	0.00638	0.00705	0.00006	0.74697	0.00003	0.22907	0.00423
19	0.00147	0.00459	0.00532	0.00895	0.00004	0.74794	0.00003	0.22651	0.00432
20	0.00184	0.00371	0.00615	0.00701	0.00005	0.76369	0.00002	0.21287	0.00395
21	0.00161	0.00423	0.00603	0.00803	0.00005	0.76124	0.00002	0.21367	0.00420
22	0.00162	0.00416	0.00645	0.00822	0.00005	0.82063	0.00002	0.15430	0.00360
23	0.00134	0.00492	0.00803	0.00950	0.00008	0.79639	0.00003	0.17325	0.00465
24	0.00186	0.00356	0.00700	0.00748	0.00006	0.77700	0.00002	0.19765	0.00415
25	0.00163	0.00469	0.00637	0.00967	0.00005	0.76089	0.00003	0.21033	0.00471
26	0.00166	0.00429	0.00696	0.00808	0.00006	0.81523	0.00002	0.15878	0.00377

Figure 4: Scale Invariant normalized feature matrix for 'A'-'z' & '0'-'9'

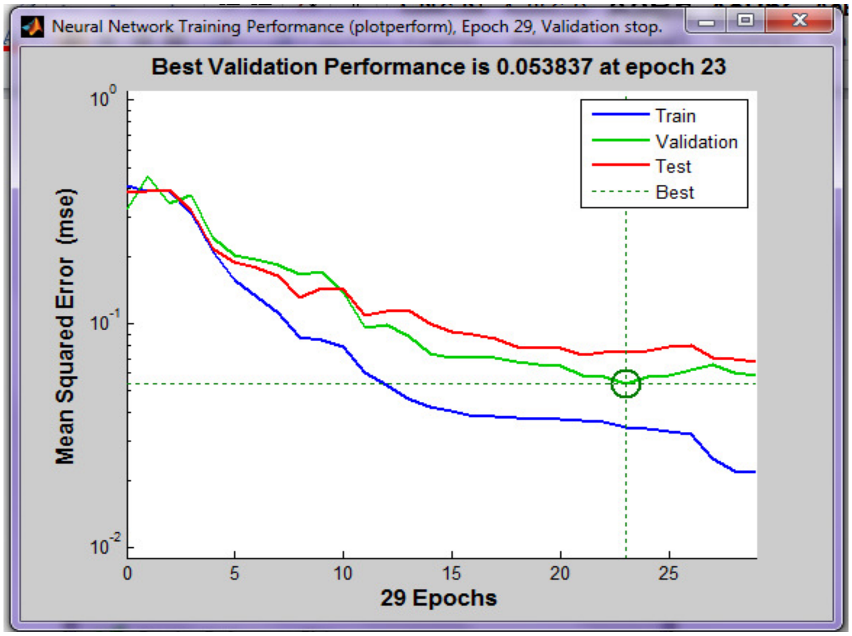

Neural Network Training Performance (plotperform), Epoch 29, Validation stop.

Figure 5: BPN based training scale invariant features

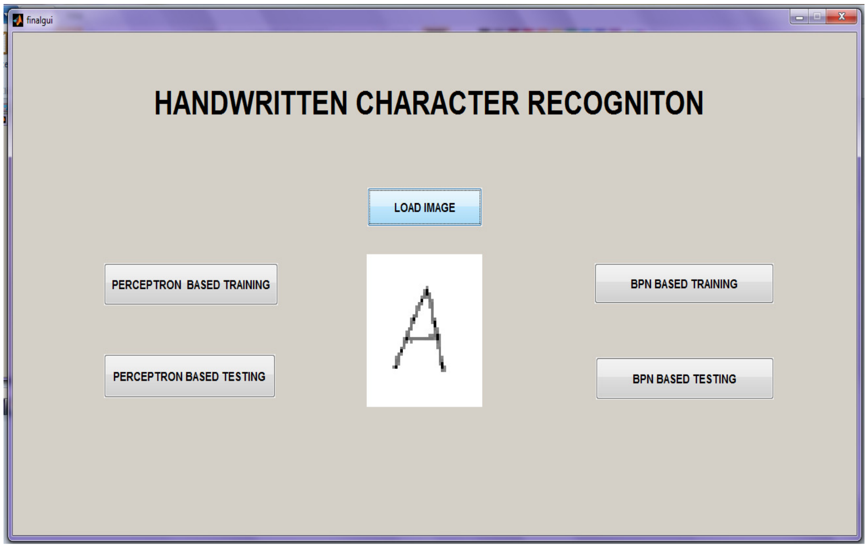
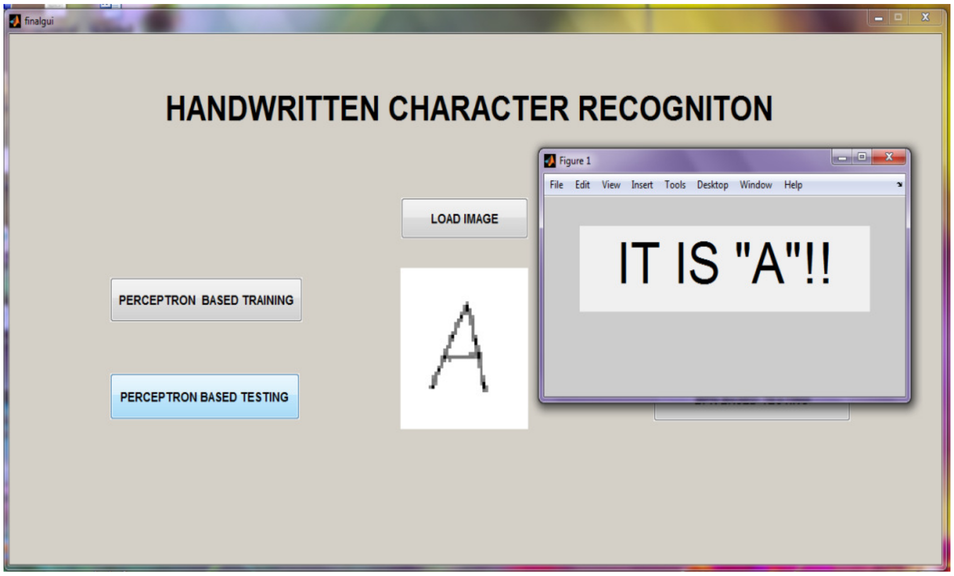
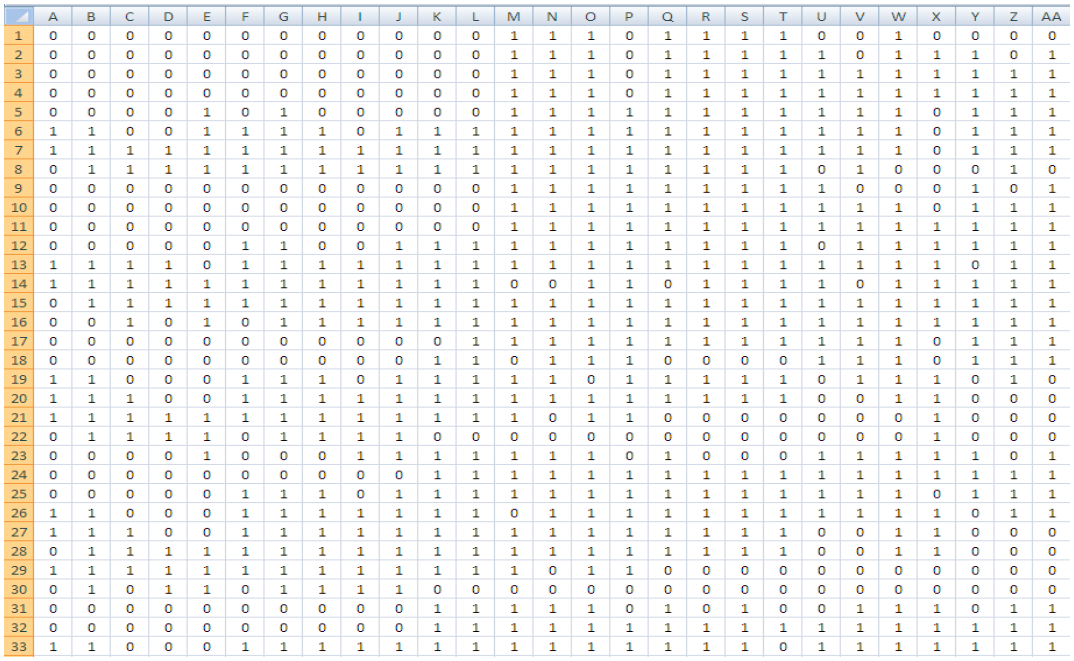


Figure 6: User Interface for Load Image


finalgui
HANDWRITTEN CHARACTER RECOGNITON
Figure 1
File Edit View Insert Tools Desktop Window Help
LOAD IMAGE
PERCEPTRON BASED TRAINING
PERCEPTRON BASED TESTING

IT IS "A"!!

Figure 7: Recognition of character ‘A’ using BPN



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	1	0	0	1	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1
3	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
4	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
5	0	0	0	0	1	0	1	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
6	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
8	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	1	0
9	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	0	0	1	0	1
10	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
11	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
13	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
14	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	1	1	1	0	1	1	1	1	1
15	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16	0	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
17	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
18	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	0	0	0	0	1	1	1	0	1	1	1
19	1	1	0	0	0	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	0
20	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0
21	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0	1	0	0	0
22	0	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
23	0	0	0	0	1	0	0	0	1	1	1	1	1	1	1	0	1	0	0	0	1	1	1	1	1	0	1
24	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
25	0	0	0	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
26	1	1	0	0	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1
27	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0
28	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0
29	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0
30	0	1	0	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	1	0	1	0	0	1	1	1	0	1	1
32	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
33	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1

Figure 8: Reduced dimensional feature matrix[64x1] for ‘A’-‘z’ & ‘0’-‘9’

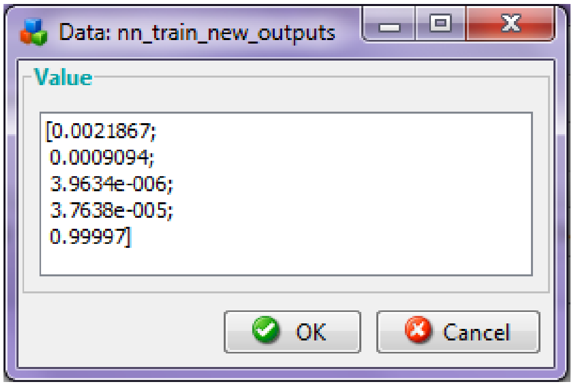
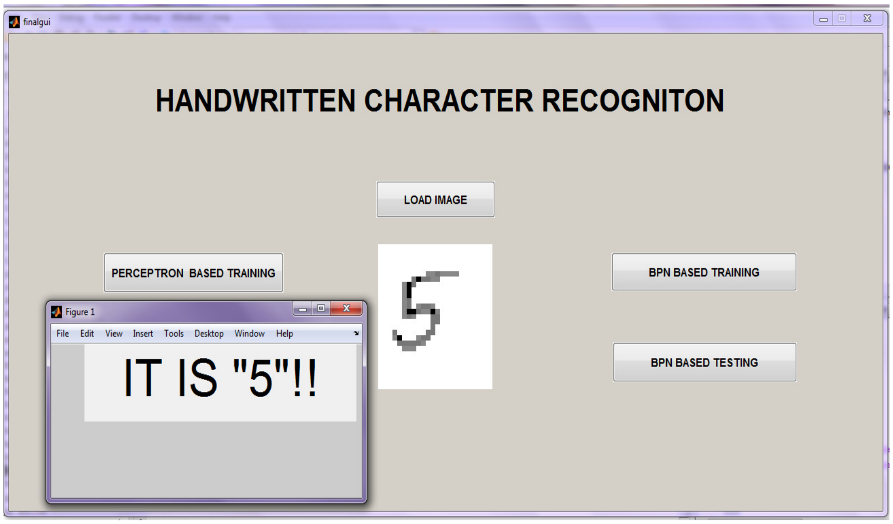
The target matrix of 'A' is [0 0 0 0 1]	The output matrix of 'A' is  Data: nn_train_new_outputs Value [0.0021867; 0.0009094; 3.9634e-006; 3.7638e-005; 0.99997] OK Cancel

Figure 9: Perception based training & testing


finalgui
HANDWRITTEN CHARACTER RECOGNITON
LOAD IMAGE
PERCEPTRON BASED TRAINING
BPN BASED TRAINING
BPN BASED TESTING
Figure 1
File Edit View Insert Tools Desktop Window Help

IT IS "5"!!

Figure 10: Recognition of character '5' using perceptron network

4. CONCLUSION

In this work it has been observed that finding the reduced dimensional feature matrix of an image is easy in comparison with the scale invariant feature matrix. The training performance of the SIF matrix is much faster and reliable. The performance of the network depends upon selection of the features into the SIF matrix. The feature selection is more challenging in case of structurally complex scripts such as Bangla, Hindi and Oriya. It has been observed that using the SIF features, the English alphabets and numeric's matches with an accuracy of 95% on average matching while the LDF match accuracy varies from 78% to 96% for different characters.

The work may be further extended to test the regional language scripts. It can be tested with different reduced dimensional matrix of different dimensions.

REFERENCES

[1] Vijay Patil and Sanjay Shimpi, “Handwritten English Character Recognition Using Neural Network”, Elixir Comp. Sci. & Engg. 41 (2011) 5587-5591, November 2011.

[2] Velappa Ganapathy, and Kok Leong Liew, “Handwritten Character Recognition Using Multi-scale Neural Network Training Technique”, World Academy of Science, Engineering and Technology 39 2008.

[3] Rakesh Kumar Mandal, N R Manna, “Hand Written English Character Recognition Using Column-Wise Segmentation of Image Matrix (CSIM)”, Issue 5, Volume 11, May 2012.

[4] Rakesh Kumar Mandal, N R Manna, “Hand Written English Character Recognition Using Row-Wise Segmentation (RST)”, International Symposium on Devices MEMS, Intelligent Systems & Communication (ISDMISC) 2011.

[5] Apash Roy, N. R. Manna, “Handwritten Character Recognition Using Block wise Segmentation Technique (BST) in Neural Network”.

[6] Soumya Mishra, Debashish Nanda, SanghamitraMohanty, “Oriya character recognition using Neural Networks”, Special Issue of IJCCT Vol. 2 Issue 2, 3, 4; 2010 for International Conference [ICCT-2010], 3rd-5th December 2010.

[7] B. Kumar, N. Kumar, C. Palai,.P. K. Jena, S. Chattopadhyay, “Optical Character Recognition using Ant Miner Algorithm: A Case Study on Oriya Character Recognition”, International Journal of Computer Applications (0975 – 8887) Volume 61– No.3, January 2013.

[8] Fu Chang, Chin-Chin Linand Chun-Jen Chen,Institute of Information Science, Academia Sinica , Taipei, Taiwan Dept. of Electrical Engineering, National Taipei University of Technology, Taipei, Taiwan”Applying A Hybrid Method To Handwritten Character Recognition”.

AUTHORS

Pradeep Kumar Jena Working as an Associate Professor in the Dept. of MCA at National Institute of Science and Technology, Berhampur. His domain of research includes Biometric AuthenticationImage Processing, Pattern Recognition and Data Mining.

Charulata Palai Working as an Assistant Professor in the Dept. of CSE at National Institute of Science and Technology, Berhampur. Her domain of research includes Soft ComputingImage Processing, Pattern Recognition.

Lopamudra Sahoo is a BTech Final Year student of NIST, Berhampur.

Anshupa Patel is a BTech Final Year student of NIST, Berhampur.

ANALYSIS OF NEAR FIELD DISTRIBUTION VARIATION USING AWAS ELECTROMAGNETIC CODE FOR WIMAX

Chhaya Dalela[1], MVSN Prasad[2], Rahul Namawat[3]

[1] Department of Electronics & Communication Engineering, JSSATE, Noida, India
[1]chhayadalela@jssaten.ac.in
[2]National Physical Laboratory, New Delhi, India,
[2]mvprasad@mail.nplindia.ernet.in
[3]JECRC University, Jaipur, Rajasthan, India
[3]rahulnamawat@gmail.com

ABSTRACT

Rapid Fluctuations and variations of signal strength at higher frequency range in Near Field zone, is a common difficulty to achieve higher data rate. As signal varies continuously, it starts decaying by the interference of the atmospheric obstructions and the electric field intensity gradually decreases with the distance. This effect is observed by AWAS Electromagnetic Code which predicts the rapid variations in electric field intensity irrespective of environment, whereas statistical models do not capture the fundamental physics and variations as per Environment. An Adequate and optimum values of these external parameters is essential for controlled and efficient transmission.

KEYWORDS

WiMAX, Propagation Modelling, Near Field Zone, Electric Field Intensity, Electromagnetic Code

1. INTRODUCTION

Worldwide Interoperability for Microwave-Access (WiMAX) has emerged as wireless access technology that is capable of providing fixed and mobile broadband connectivity. Fixed WiMAX is targeted for fixed and nomadic broadband services while mobile WiMAX are designed to provide high mobility services. Operators do drive-tests on a continuous basis, collect signal levels, network quality and performance which are then used to refine empirical propagation models for system-planning and/or existing network optimization. The fast evolution of wireless communications has led to the use of higher frequency bands, smaller cell sizes, and smart antenna systems, making the propagation prediction issues more challenging since wireless communication channels are inherently frequency dispersive, time varying, and space selective. These data rates can be further increased by employing multiple antennas both at the transmitter and receiver.

In India, WiMAX is operating at 2300 MHz frequency band which is tends to provides an access to operate the wireless devices at very higher data rates and ubiquitous access in a large coverage

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 329–336, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4533

area [1, 2]. Propagation Models are developed to estimate various parameters viz field strength, path loss etc. in different environments. Propagation Models are for telecommunication providers to improve their services for better signal coverage and capacity for mobile user satisfaction in the area. Prasad *etal.* [3] reported that the AWAS Electromagnetic Code did not require any building information and was able to compete with other empirical methods. An attempt has been made by the authors to realize the objective by interpreting available experimental data to get a better understanding of the propagation conditions models for different environments to provide guidelines for cell planning of WiMAX transmissions in the Indian urban zones in general. The experimental data utilized in this study corresponds to 2300 MHz WiMAX radio measurements in different environments, carried out in Western India. AWAS Electromagnetic code, which is based on Sommerfield's approach for ground, is used to compute the near field signal strength of propagation link and significant changes with height of transmitting antenna is identified to radiate efficient signal for WiMAX.

In Section II, the details of AWAS Electromagnetic code and Environmental details are provided. In section III, we have analysed AWAS Electromagnetic code with Existing Prediction Methods. Conclusions are presented in Section IV.

2. EXPERIMENTAL DETAILS

2.1. The AWAS Numerical Electromagnetic Code

AWAS Numerical Electromagnetic Code [4] is a computer program which evaluates the current distribution of a conductor by analysing the polynomial coefficients. This program is based on a two potential equations which is solved numerically using method of moments with polynomial approximation for the current distribution. The influence of the ground is taken into account using Sommerfield's approach, with numerical integration algorithm developed for this program. It was utilized to compute the path-loss values for different values of dielectric constant and a conductivity of 2 x 1 0-4 over real ground. This commercially available computer program is capable of analyzing wire antennas operating in transmitting and receiving modes, as well as analyzing wire scatterers. Different values of dielectric constant for dense urban, urban, and suburban regions were incorporated into the computation of the A WAS simulation. Reference [5] gave relative dielectric constants for various types of ground and environments. It gave values of 3 to 5 for city industrial areas. Hence, a value of 4 was used in the A WAS simulation. In the AWAS simulation, the computation of the electric field was carried out by taking a vertical dipole located over an imperfect ground plane, whereas in the measurements, a high transmitting- gain antenna was employed.

2.2. Environmental Details

The experimental sites AAC, AHT, BTW, KTB, GRJ, JVD,and OLK [6] are situated in the dense urban area of Mumbai, India, except AAC, JVD, and OLK, which are located in an urban area [Fig. 1(a) and (b)]. The clutter environments of these sites are shown in different colors in the legend of Fig. 1(a) and (b). Since AAC [Fig. 1(a)] and OLK [Fig. 1(b)] are surrounded by skyscraper buildings, they represent a typical dense urban environment. AHT [Fig. 1(a)] shows the presence of skyscrapers at north, east, and west sides, while the remaining other areas BTW and GRJ are fully surrounded by dense environment. They are surrounded by industrial environments at eastern side of 0.7 km and at eastern side after 0.9 km respectively [6]. The parameters of these base stations are shown in Table I and Table II.

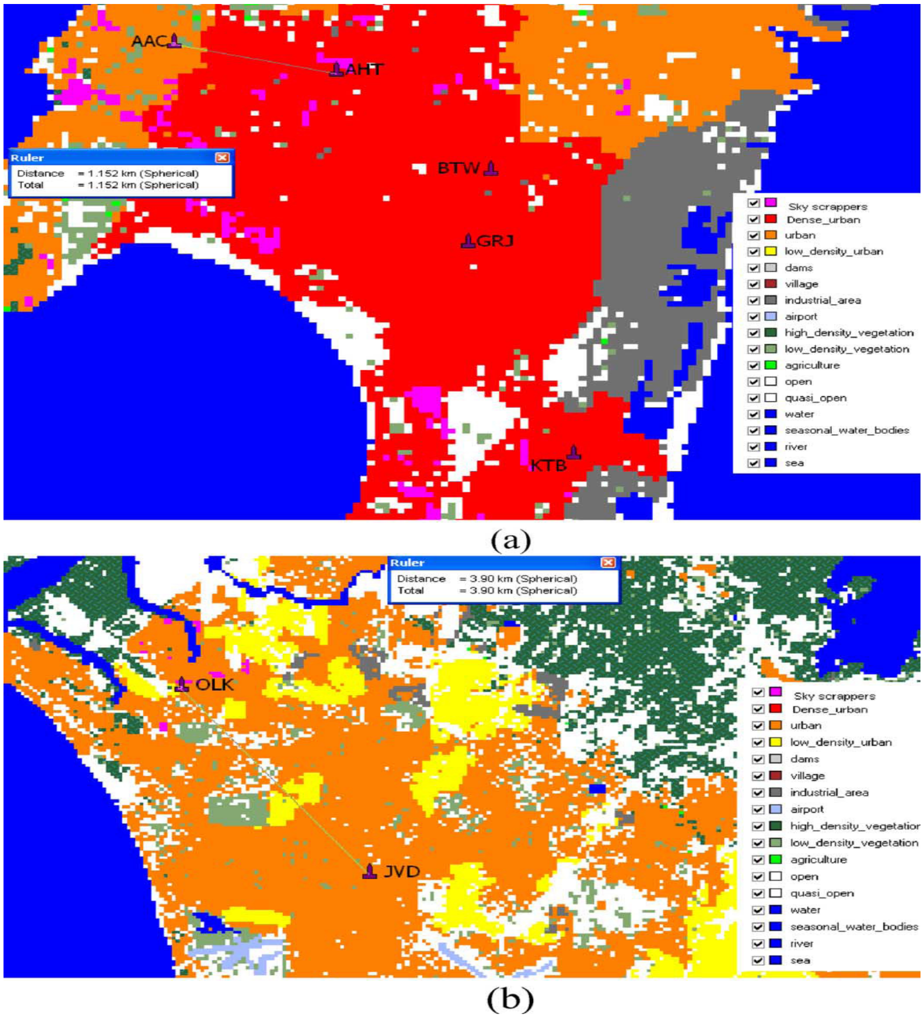


Fig. 1. Clutter environment for experimental sites (a) AAC, AHT, KTB, BTW, and GRJ and (b) OLK and JVD.

TABLE I
Base Station details

Sr. No.	Name of Base Stations	Height of Transmitting Antenna	Near Field Distance (in Km)
1.	Ajay-Amar (AAC)	37m	1.70
2.	Arihant (AHT)	32m	1.47
3.	Bootwala Bldg (BTW)	46m	2.11
4.	Khethan Bhabhan (KTB)	31m	1.42
5.	Giriraj (GRJ)	28m	1.28
6.	Jeevan Dhara (JVD)	27m	1.24
7.	Obelisk (OLK)	30m	1.38

TABLE III
Other details of experimental site

Sr. No.	Other Details	
1.	Height of Receiving Antenna	1.5m
2.	Transmitted Power	43.8 dBm
3.	Average Height of Building	25m
4.	Average Street Width	15m
5.	Average Separation Between Buildings	30m

2.3. Electric Field Distribution Analysis

Electric Field is an important factor in analysing the path loss and its effects in WiMAX. AWAS Electromagnetic code is implemented to calculate the electric field density (V/m) and then further sustained path loss of the Base Stations in this environment can be calculated. The Near Field distribution of base stations is determined for near field distance (in Km) (Table I) for various base stations and are substituted in AWAS Electromagnetic Code in order to estimate the Electric Field Distribution.

Near Field distance is calculated by :

$$D = 4 *h_t* h_r / \lambda \qquad (1)$$

where D is the near field distance and h_t, h_r is the height of transmitting and receiving antennas respectively. Fig. 1(a) and 1(b) shows the rapid fluctuations in near field at distance of 100 m to 200m and at 200m to 400m of GRJ Base Station. A Base Station like GRJ base station is assumed, so the environmental parameters of both the base stations are considered same and treated as a fully dense environment, whose height is considered as 20 meters. Fig 2(a) and 2(b) are near field variations of this base station. By Comparing Fig 1(a), 1(b) and 2(a), 2(b), it was observed that the Field intensity in the Figure 2(a) and 2(b) has less variations as consider to GRJ Base Station.

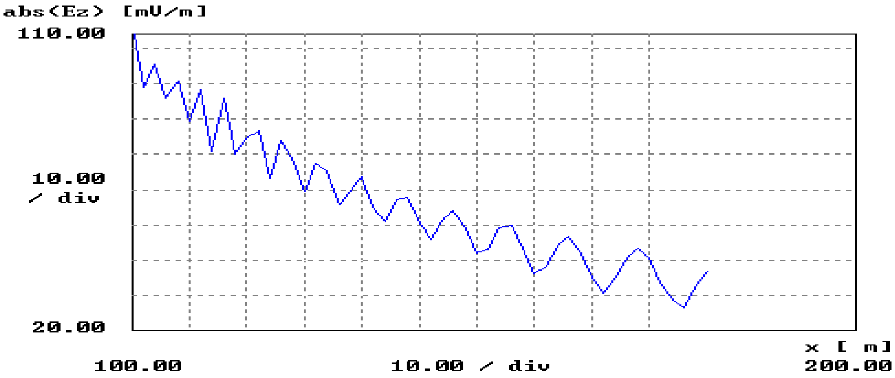


Fig 1(a) : Near Field Distribution of GRJ Base Station from 100m to 200m

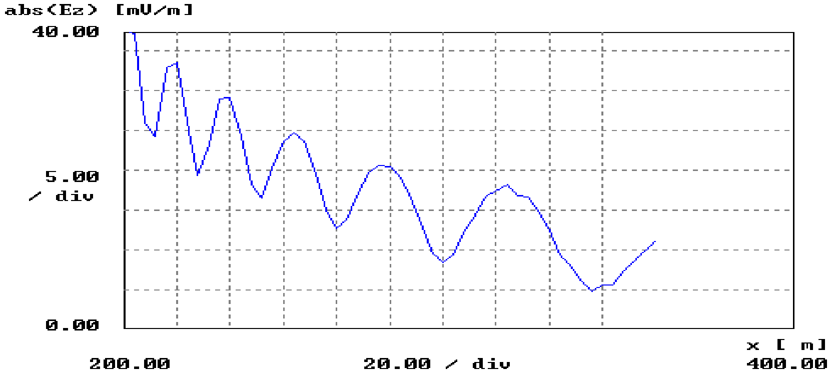


Fig 1(b) : Near Field Variation of GRJ Base Station from 200m to 400m

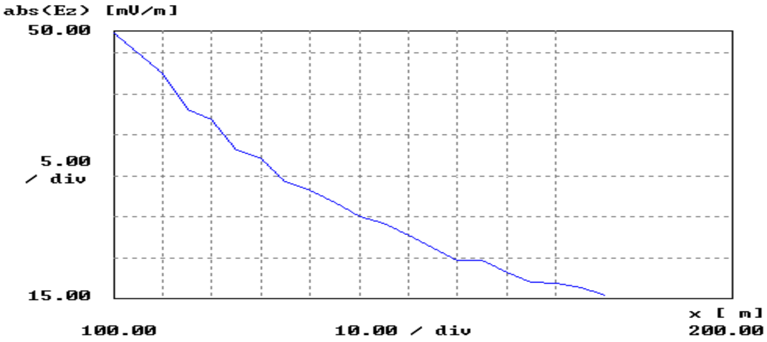


Fig 2(a) : Near Field Distribution of Base Station Antenna of height 20 meters from 100m to 200m

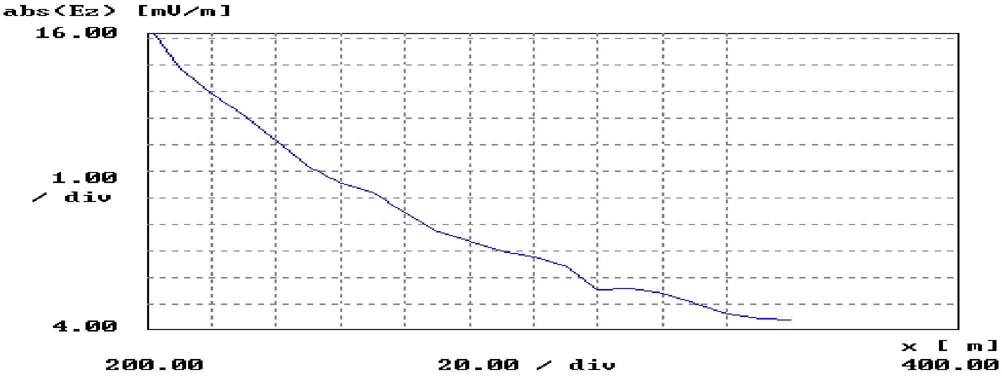


Fig 2(b) : Near Field Distribution of Base Station Antenna of height 20 meters from 200m to 400m

2.4. Analysis of Variations in Electric Field

As observed from figure 1(a) and 1(b), there is a rapid variations in Electric Field Distribution and can be represented by Δ which is defined as :

$$\Delta = (E_{max} - E_{min}) /d \qquad (2)$$

where d is the distance in a single division. Table III shows the variation in electric field which is numerically decreasing with the distance.

TABLE IIIII

RAPID VARIATION IN THE NEAR FIELD OF GRJ BASE STATION AT 2300 MHZ OPERATING FRQUENCY

Sr. No.	Near Field distance (meters)	Electric field Variations(mV/m)	
		H_t = 28m	H_t = 20m
1.	100-110	2	0.8
2.	120-130	1.5	0.6
3.	150-160	1.1	0.3
4.	170-180	1.0	0.13
5.	200-220	0.8	0.1
6.	240-260	0.65	0.08
7.	300-320	0.58	0.03
8.	400-500	0.14	0.015

Basically there are various conventional methods are used to predict the path loss and electric field distribution. But they are unable to predict the variation due to increasing height of the Transmitting Antenna. The reason behind rapid variation in Electric Field in Near Field zone is the atmospheric fluctuation and environmental obstruction. The reduction in variations in Electric field intensity with reduction in height of Transmitting Antenna can be explained as shown in Fig 3. Several repeaters with reduced transmitting antenna height of Transmitting Stations can play a vital role in this situation to remove the rapid fluctuations.

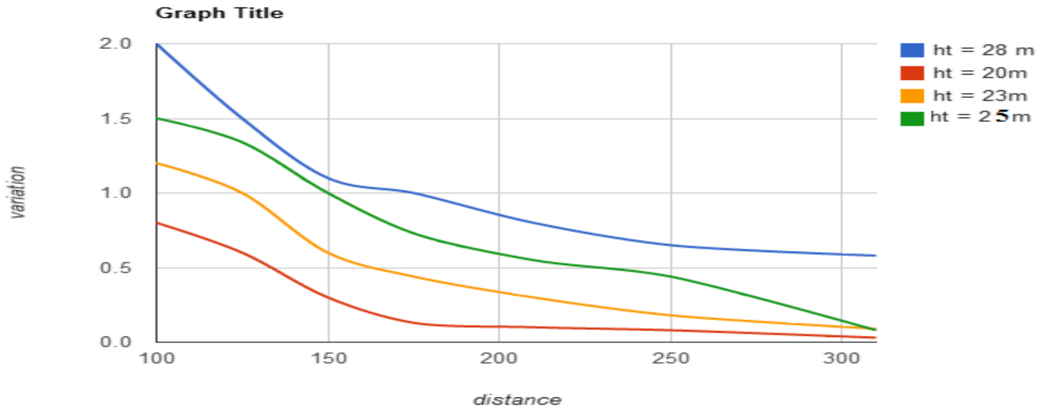


Fig 3 : Variations in Near Field Distribution with change in height of Base Station antenna

3. CONCLUSIONS

The near field distribution for WiMAX at 2300MHz is analyzed for dense urban region of Western India by AWAS Numerical Electromagnetic Code. The advantage of using AWAS electromagnetic code is that it predicts the fundamental atmospheric variations in the field strength as per the environmental parameters of that region. It has been found that in near field region, signal fluctuations are very high, and as the height of the transmitting antenna increases, near field distance will be larger and the signal remain stable in far field. Hence, it is advisable to keep the antenna closed to the ground and employee more repeater stations. Thus, the reduction in transmitting antenna height will ultimately reduces variation in near field distribution and is achieved to produced efficient radiated signal. Also AWAS numerical Electromagnetic Code predicts all the variations of path loss irrespective of environment whether it is urban, suburban or rural whereas statistical models do not capture the fundamental physics and it has separate models for urban, suburban or rural environment.

REFERENCES

[1] "IEEE Standards for local and metropolitan area network – Part 16 : Air Interface For fixed broadband wireless access systems" ,2011.

[2] CEPT ECC Report 172, "Broadband Wireless Systems usage in 2300-2400MHz" ,p. 11, March 2012.

[3] M.V.S.N. Prasad, Saurabh Gupta, M.M. Gupta "Comparison of 1.8GHz Cellular Outdoor measurement with AWAS Electromagnetic Code and conventional Models over urban and suburban regions of Nothern India"Antenna and Propagation Magazine, IEEE, 53 (2011) 76-85

[4] "Antenna and Wired Scatters"

[5] Soil Dielectric Properties (Dielectric Materials and Applications)," NEC list and Web site: http://pe2bz.philpem. me.uklComm/-%20AntennalInfo-905-Misc/soildiel.htm; also from Arthur R. von Hippel (ed.), Dielectric Materials and Applications, Cambridge, MA, MIT Press, 1 954.

[6] Chhaya Dalela, M.V.S.N Prasad, P.K. Dalela and Rajeev Saraf "Analysis of WiMAX Radio Measurements and Comparison With Some Models Over Dense Urban Western India at 2.3 GHz" IEEE Antenna and Wireless Propagation, 10 (2011)

AUTHORS

Chhaya Dalela received the B.Tech. degree in Electronics Engg. from H.B.T.I.,Kanpur, M.Tech. in Digital Communication from .P.T.U.,Lucknow and completed her Ph.D. In channel characterisation and modelling. resently, she is working with JSS Academy of Technical Education, Noida, as Associatet Professor in Electronics Engineering Department. Her areas of research interest are channel measurements and modeling for broadband communications, Cognitive Radio, Telecommunication network planning etc. He has published more than 30 research papers in national and international journals and conference proceedings.

Dr M V S N Prasad is presently working as a scientist in National Physical Laboratory. His research areas are radio channel measurements and modeling for mobile and fixed communications, mobile commnications in railway tunnels, microwave propagation, radiowave propagation related to broadcasting etc. He has developed active links with various user organizations in the area of telecommunications like VSNL, Railways,Dept. of of Telecommunications, three wings of defense and rendered consultancy services in these areas and established collaborations with many universities. He received the URSI young scientist award in 1990, Best paper award from National Space Science Symposium in 1990, Best paper award from Broadcast engineering society(India) in 1998 and 2001. Elected as a member of American Geophysical union under the Lloyd V.Berkner fund. He participated in telecommunication and radio wave propagation workshops at the International centre for theoretical physics, Trieste, Italy. He has published several papers in national and international journals and acted as a reviewer for many journals in this field.

Rahul Namawat received a B.Tech. Degree in Electronics & Communication Engg. From JECRC UDML College of Engg, R.T.U, Kota. He’s currently pursuing in Masters of Technology in Digital Communications from JECRC University, Jaipur.

SEMANTIC TAGGING FOR DOCUMENTS USING 'SHORT TEXT' INFORMATION

Ayush Singhal[1] and Jaideep Srivastava[1]

[1]Department of Computer Science & Engineering,
University of Minnesota, Minnesota, USA
singh196,srivasta@umn.edu

ABSTRACT

Tagging documents with relevant and comprehensive keywords offer invaluable assistance to the readers to quickly overview any document. With the ever increasing volume and variety of the documents published on the internet, the interest in developing newer and successful techniques for annotating (tagging) documents is also increasing. However, an interesting challenge in document tagging occurs when the full content of the document is not readily accessible. In such a scenario, techniques which use "short text", e.g., a document title, a news article headline, to annotate the entire article are particularly useful. In this paper, we propose a novel approach to automatically tag documents with relevant tags or key-phrases using only "short text" information from the documents. We employ crowd-sourced knowledge from Wikipedia, Dbpedia, Freebase, Yago and similar open source knowledge bases to generate semantically relevant tags for the document. Using the intelligence from the open web, we prune out tags that create ambiguity in or "topic drift" from the main topic of our query document. We have used real world dataset from a corpus of research articles to annotate 50 research articles. As a baseline, we used the full text information from the document to generate tags. The proposed and the baseline approach were compared using the author assigned keywords for the documents as the ground truth information. We found that the tags generated using proposed approach are better than using the baseline in terms of overlap with the ground truth tags measured via Jaccard index (0.058 vs. 0.044). In terms of computational efficiency, the proposed approach is at least 3 times faster than the baseline approach. Finally, we qualitatively analyse the quality of the predicted tags for a few samples in the test corpus. The evaluation shows the effectiveness of the proposed approach both in terms of quality of tags generated and the computational time.

KEYWORDS

Semantic annotation, open source knowledge, wisdom of crowds, tagging.

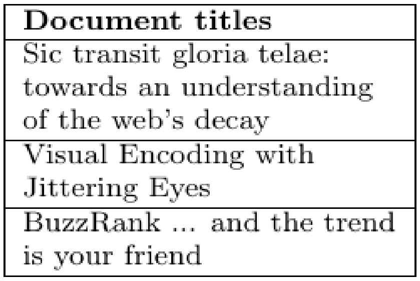
1. INTRODUCTION

Tagging documents with relevant and comprehensive keywords offer an invaluable assistance to the readers to quickly overview any document [20]. With the ever increasing volume and variety of the documents published on the internet [12], the interest in developing newer and successful techniques for tagging documents is also increasing. Tagging documents with minimum words/key-phrases have become important for several practical applications like search engines, indexing of databases of research documents, comparing the similarity of documents, ontology creation and mapping and in several other stages of important applications [4]. Although

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 337–350, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4534

document tagging is a well-studied problem in the field of text mining, but there are several scenarios that have not drawn sufficient attention from the scientific community.

Table 1: A few examples of document titles which do not try to capture the essence of the document's content.



Document titles
Sic transit gloria telae: towards an understanding of the web's decay
Visual Encoding with Jittering Eyes
BuzzRank ... and the trend is your friend

A few of the challenges regarding document tagging, which is not well addressed in the literature are: (1) entire content of the document is not accessible due to privacy or protection issues (2) document heading does not summarize the content of the document (3) reading entire document is time consuming. The first challenge requires techniques to generate tags using only a short description of the document (document heading/title, snippet). The second challenge requires 'intelligence' to figure out the context represented by the heading or title. As an example, consider the examples shown in table 1. This table shows a few examples of 'catchy' titles used in scientific research articles to provide headings of the articles. Only using such title information it would be hard to delve into the subject matter of these articles. The third challenge is particularly relevant in situations when the document itself is quite large and thus, requires tagging using only partial information from the document for quick annotation. The third challenge is particularly relevant in the case of real-time response systems.

To the best of our knowledge, the above mentioned challenges have not been well addressed in the literature. Most of the current literature provide efficient techniques for key- word extraction using the text content from single or multiple documents [17, 8]. Such techniques are not suitable if the text content of the document is very short or unavailable. Another class of problems which is of increasing interest is that of key-phrase abstraction. While these techniques do not extract keywords directly from the text content, the text content is required to build models for keyword abstraction [7]. However, the area of keyword extraction is still in the developing stage. Eventually, the overall goal of these research directions is to automate the annotation of documents with key phrases that are very close to what a human could generate. We further elaborate upon the specific research works and milestones in section 7.

In this work we propose a novel approach to address the above mentioned challenges. We propose a novel approach that takes as input only a 'short text' from the query document and leverages intelligence from the Web2.0 to expand the context of the 'short text'. We have used academic search engines to expand the context of the 'short text'. The expanded context utilizes the intelligence of the web to find relevant documents to overcome the 'catchiness' of the title. The tags are generated using the world knowledge from DBpedia, Freebase, Yago and other similar open source crowd-sourced databases. Moreover, using the crowd- sourced knowledge bases ensures that the tags are up-to-date as well as popular. Finally, we propose an unsupervised algorithm to automatically eliminate the 'noisy' tags. The un- supervised approach uses web-based distances (also famous as 'wisdom of crowd' [10]) to detect outlier tags. The overall framework is fully unsupervised and therefore, suitable for real-time applications for any kind of documents.

In order to demonstrate the effectiveness of the proposed approach in real-world applications, we have used a sample of the dataset from the DBLP digital archive of computer science research articles [15]. We evaluate the performance of the proposed approach using 50 test documents. We also compare the performance of the proposed approach with a baseline approach which uses the full content of the documents in order to generate tags. Surprisingly, we find that the tags generated by the proposed approach, which uses only the title/heading information of the document to predict tags, has a greater overlap (measured via the Jaccard index) with the ground truth tags (0.058 vs. 0.044) in comparison to the baseline. We also find that the proposed approach is computationally at least 3 times faster than the baseline approach. A qualitative analysis of the generated tags for a few sample test documents further reveal the effectiveness of the proposed approach for semantic tagging using only 'short text' information from the document. Although the proposed approach is tested only on the DBLP dataset, the approach, however, is generic enough to be used for various types of documents like news articles, patents and other large content documents.

We have made the following contributions in this work:

- A novel approach for using 'short text' for context expansion using web intelligence.
- A novel approach for tag generation using crowd-sourced knowledge.
- A novel approach to eliminate 'noisy' tags using web- based distance clustering.
- We provide a quantitative and a qualitative validation on a real world dataset.

The rest of the paper is organized in the following manner. In Section 2, we define the problem statement. In section 3, we mention some of the important features of crowd-sourced knowledge bases. The details of the proposed approach are discussed in Section 4. Experimental design and the results are discussed in Section 5 and 6. Section 7 describes the related literature. Finally, the summary of the work and a few directions for future research are presented in Section 8.

2. PROBLEM FORMULATION

The problem of document tagging is formulated in the following manner. Given a document's text content S, the research problem is to identify k keywords/phrases based on the content S of the given document. In this case $k \ll \text{size}(S)$.

In the present work, we are studying a slightly different problem from the one describe above. We consider the title of the document as the only available text content (S'). We call this information about the document as the 'short text' since it is only a short description of the document. Also the $\text{size}(S') \ll \text{size}(S)$. The research problem is to find k keywords/phrases to describe the main topics/themes of the document. The keywords/phrases may not be directly present in the content of the document. Here, k is not known a priori.

3. BACKGROUND OF OPEN SOURCE KNOWLEDGE BASES

3.1 Crowd-sourced knowledge

Wikipedia is currently the most popular free-content, online encyclopedia containing over 4 million English articles since 2001. At present Wikipedia has a base of about 19 million registered users, including over 1400 administrators. Wikipedia is written collaboratively by largely anonymous internet volunteers. There are about 77,000 active contributors working on the articles in Wikipedia. Thus the knowledge presented in the articles over the Wiki are convinced upon by editors of similar interest.

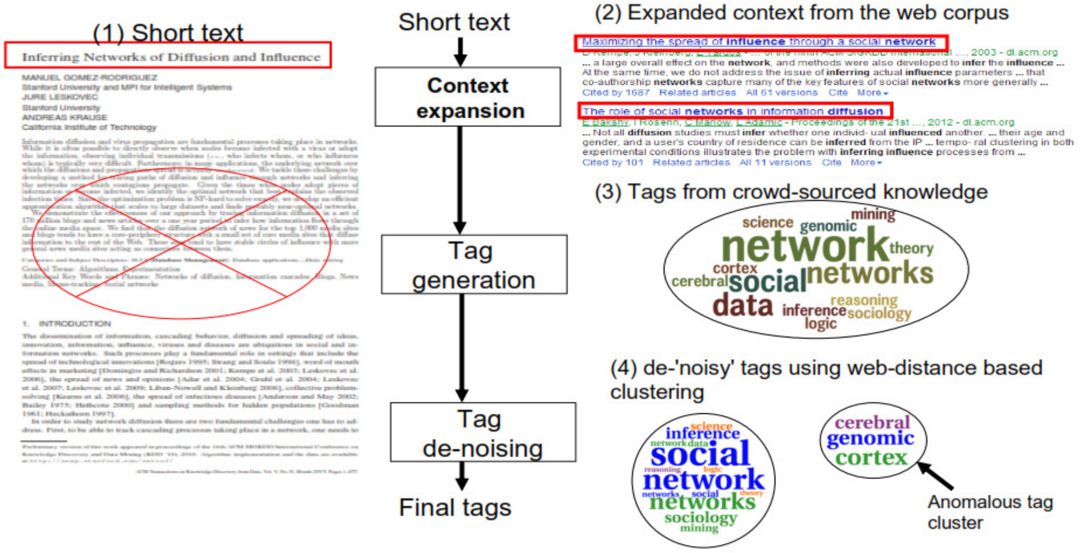

(1) Short text
Inferring Networks of Diffusion and Influence
Short text
Context expansion
Tag generation
Tag de-noising
Final tags
(2) Expanded context from the web corpus
Maximizing the spread of influence through a social network
The role of social networks in information diffusion
(3) Tags from crowd-sourced knowledge
science genomic mining network theory cortex cerebral social networks data reasoning inference sociology logic
(4) de-'noisy' tags using web-distance based clustering
science inference network data social reasoning logic network networks social theory networks sociology mining
cerebral genomic cortex
Anomalous tag cluster

Figure 1. A systematic framework of the proposed approach. An example is illustrated to explain the proposed approach

DBpedia is another crowd-sourced community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries against Wikipedia, and to link the different data sets on the Web to Wikipedia data. The English version of the DBpedia knowledge base currently describes 4.0 million things, out of which 3.22 million are classified in a consistent ontology. For example DBpedia knowledge base allows you to ask quite surprising queries against Wikipedia, for instance “Give me all cities in New Jersey with more than 10,000 inhabitants” or “Give me all Italian musicians from the 18th century”.

Yago is similar to DBpedia. In addition to Wikipedia, Yago combines the clean taxonomy of WordNet. Currently, YAGO2s has knowledge of more than 10 million entities (like persons, organizations, cities, etc.) and contains more than 120 million facts about these entities. Moreover, YAGO is an ontology that is anchored in time and space as it attaches a temporal dimension and a spatial dimension to many of its facts and entities proving a confirmed accuracy of 95%.

Freebase is another online collection of structured data collected from various sources such as Wikipedia, ChefMoz, and MusicBrainz, as well as individually contributed user information. Its database infrastructure uses a graph model to represent the knowledge. This means that instead of using tables and keys to define data structures, its data structure is defined as a set of nodes and a set of links that establish relationships between the nodes. Due to its non-hierarchical data structure, complex relationships can be modeled be- tween individual entities.

3.2 Academic search engines

Academic search engines provide a universal collection of research documents. Search engines such as Google scholar and similar other academic search engines have made the task of finding relevant documents for a topic of interest very fast and efficient. We use the capacity of search engines to find relevant documents for a given query document. We have used the Google search engine for this purpose.

4. PROPOSED APPROACH

In this section, we discuss the framework of the proposed approach. The approach consists of three main components which will be discussed in detail in this section. The overall framework is summarized in the schematic (figure 1). As shown in this figure, the proposed approach for semantic annotation of a document is a three step procedure: (1) Context expansion using academic search engine, (2) candidate tag generation using crowd-sourced knowledge and (3) de-noising tags using web-based distance (a.ka. ’wisdom of crowd’) clustering technique. Given a document as a short text S’, the final results are k semantic tags, where k is not fixed apriori.

4.1 Context expansion

As mentioned earlier, the problem of reading the entire text content of the document or the lack of availability of the full text content restricts the task of tagging based on the document’s text content. Moreover, techniques utilizing the text content of the document generate tags or keywords only from within the document’s text content. While such key- word extraction approaches are necessary, but this might often restrict the keyword usage for the document. In such a scenario, it is helpful to generate keywords that are more popular and widely accepted for reference. To accomplish this goal, we propose a web-based approach to generate an expanded context of a document.

Given the ’short text’ S information of a document, the expanded context is generated by mining intelligence from the web using an academic search engine. As shown in figure 1, the context of the ’short text’ (S’) is expanded using the results obtained by querying the web corpus with an academic search engine. The ’short text’ is used as a query for the search engine. The new context of the ’short text’ include the titles/heading (h) of the top-k results returned by the search engine. It is also possible to use other contents of the results like the snippets, author names, URLs to create an extended context. However, for this work, the approach is kept generic such that it is applicable to all sorts of search engines. The value of k is not fixed and can be a parameter to the approach. In the later section, the results are evaluated by varying the value of k.

The extracted results headings (h) that form the expanded context of the ’short text’ are transformed into a bag of words representation. As a basic step in text mining, the bag of words is pre-processed by applying stop-word removal, non-alphabetic character removal and length-2 word removal techniques. In the rest of the paper, the expanded context of S is referred as C (S’) for the sake of convenience and consistency. The final context created using the search engine is expected to contain a wider variety.

4.2 Tag generation

In this section, we describe the procedure to utilize crowd- sourced knowledge to generate tags from the expanded context C (S’). As described earlier, the crowd-sourced knowledge is available in well-structured format unlike the un- structured web. The structured nature of knowledge from sources such as DBpedia, Freebase, Yago, Cyc provides opportunity to tap in the world knowledge from these sources. The knowledge of these sources is used in the form of concepts and named entity information present in them, since the concepts and named entities consists of generic terms useful for tagging. We have used the AlchemyAPI [1] to access these knowledge bases. A tool such as this provide a one-stroke access to all these knowledge bases at once and returns a union of the results from all the various sources.

Given the expanded context (C (S’)) as the input to the AlchemyAPI, which matches the C (S’) against the indices of these knowledge sources to match C (S’), using the word frequency distribution, with concepts and named entities stored in the knowledge bases. The output for an API query C (S’) is a list of concepts and named entities. Using the open source knowledge bases and the word frequency information from the input, the API returns a list of concepts related to the content. The named entity list returned from a query C (S’) consist of only those named entities of type ’field terminologies’. There are other types of named entities such as ’person's name’, ’job title’, ’institution’ and a few other categories but those are not generic enough to be used as tags. The concepts and named entities for C (S’) together form a tag cloud T.

Figure 1 highlights a tag cloud consisting of tags generated using the above described technique. As shown in the figure, the tags are weighted based on the word distribution in C (S’). This example also shows a few tags like ’cerebral’, ’cortex’, ’genomic’ that appear to be inconsistent with the overall theme of the tag cloud for C (S). The next step describes an algorithm to handle such situations in the tagging process.

4.3 Tag cloud de-noising

As described in the previous step, the tag cloud T for C (S’) may contain some inconsistent or ’noise’ tags in it. In this section, we describe an algorithmic approach to automatically identify and prune ’noisy’ tags in the tag/keyword cloud. This step is therefore termed as tag cloud de-noising.

Given the tag cloud T for C (S’), noisy tags are pruned in the following manner. The tags in T are clustered using a pairwise semantic distance measure. Between any two tags in T, the semantic distance is computed using the unstructured web in the following way. For any two tags t_1 and t_2 in T, *dis* (t_1, t_2) is defined as the normalized Google distance (NGD) [2]:

$$NGD(t_1, t_2) = \frac{max\{logf(t_1), logf(t_2)\} - logf(t_1, t_2)}{logM - min\{logf(t_1), logf(t_2)\}}$$

where M is the total number of web pages indexed by the search engine: f(t_1) and f(t_2) are the number of hits for search terms t_1 and t_2, respectively; and f(t_1,t_2) is the number of web pages on which both t_1 and t_2 occur simultaneously.

Using the NGD metric, a pairwise distance matrix (M) is generated for the tag cloud T. The pairwise matrix M is used to identify clusters in the tag cloud. Finally, the tag cloud is partitioned into two clusters using hierarchical clustering techniques. Here, we assume that there is at least one ’noise’ tag in the tag cloud T. Out of the two clusters identified from the tag cloud T, the one cluster with majority tags is called a normal cluster, whereas the other cluster is called as an outlier cluster (or noisy cluster). In case of no clear majority the tie is broken randomly.

The algorithm is illustrated through an example shown in figure 1 step 4. This step shows that the tags in the tag cloud T generated in step 3 are partitioned into two clusters as described above. The tags in one cluster are semantically closer than the tags in the other clusters, as per the ’wisdom of crowd’ semantics. As shown in this example, the outlier tags ’cerebral’, ’cortex’, ’genomic’ are clustered together while the remaining normal tags cluster together. Since the former is a smaller cluster, it is pruned out from the tag cloud. Lastly, the final tag cloud consisting only the larger cluster of tags is returned as the output.

5. EXPERIMENT ANALYSIS

This section describes the experimental design used to evaluate the performance of the proposed approach. In this section, we discuss the test dataset used for evaluation, the ground truth for evaluation, the baseline and the evaluation metrics used for evaluation of the proposed approach.

5.1 Test dataset description

For the purpose of evaluating our approach we use a test set consisting of 50 research documents from top tier computer science conferences constructed from the DBLP corpus [15]. The 50 papers were selected to capture the variety of documents in computer science research. Several of the documents had catchy titles (examples given in Table 1). The test data are accessible here (https://www.dropbox.com/sh/iqnynrixsh2oouz/8dWnbXhh7B?n=62599451).

5.2 Ground truth

In the absence of any gold standard annotations for the test documents, the ground truth of the documents was collected by the author assigned keywords to these documents. We collected this information by parsing these documents. We assume that the keywords assign by the authors are representative of the annotations for the document. The proposed approach and the baseline were evaluated on this ground truth.

5.3 Baseline approach

We have compared the performance of the proposed approach with a baseline, which uses the full text content of the test documents. In order to evaluate the claim that the ’short text’ information in combination with web intelligence is sufficient to semantically tag a document, it is important to consider a baseline which takes the full text content of the document for tagging. The full text information is generated from the pdf versions of the test documents. The PDF documents were converted to text files using PDF conversion tools. As a basic pre-processing step, stop-words, non- alphabetical characters and special symbols were removed from the text to generate a bag of word representation of the full text.

For the purpose of comparison, the full text context was used to generate tags using the proposed approach and at the final step, de-noising of tags was done using the proposed algorithm. The purpose of this baseline is to see the effectiveness of the ’short-text’ expansion approach for semantic tagging in comparison to full-text approach.

5.4 Evaluation metrics

Given that the topics/keywords for a document are assigned in natural language, evaluating accuracy of any algorithm for tagging is a challenging task. Though, solutions such as expert’s evaluation exist, but for this project expert assistance was a challenge. In the absence of expert evaluation, we evaluated the results of our approach in the following ways. We evaluated the effectiveness of the proposed approach in the following three ways.

5.4.1 Jaccard similarity with baseline

The Jaccard similarity between two sets A and B is defined as the ratio of the size of the intersection of these sets to the size of the union of the sets. It can be mathematically stated as:

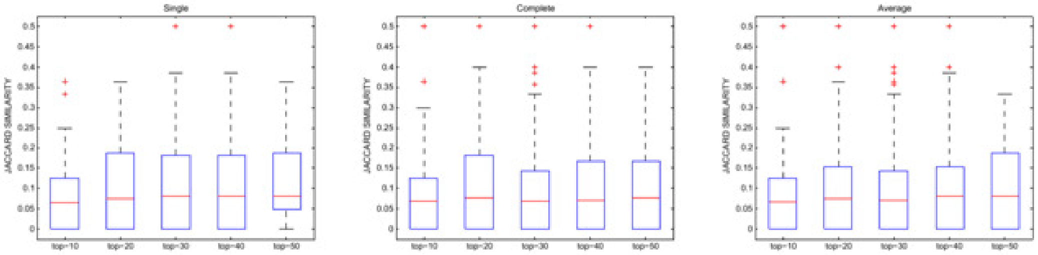


Figure 1. Boxplots showing the distribution of Jaccard Index (Overlap of tags generated from expanded context vs the full content) for 50 documents . The following hierarchical clustering criterion are used : (a) single (b) complete (c) average .

$$J(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|}$$

We used the Jaccard similarity to quantify the similarity between the tags, predicted by the proposed approach versus the baseline approach. This metric represents the magnitude of overlap between the tags generated by the two approaches.

5.4.2 Jaccard index

The Jaccard index is same as the Jaccard similarity except that it is used for a different purpose. Instead of computing the Jaccard index between the results of the proposed approach and the baseline, the Jaccard index for both the approaches is computed with the ground truth tags. This metric gives us the overlap of predicted tags using proposed approach and the baseline with the ground truth tags for each of the test documents. The Jaccard index is averaged over the total number of documents in the test dataset.

5.4.3 Execution time

The final metric for comparing the proposed approach with the baseline is the execution times. Since the main overhead of the approach is in the first step of tag generation due to differences in the sizes of the input context. The execution time is computed as the time taken in seconds to generate tags for the 50 test documents given their input context. For the proposed approach the context is derived using web intelligence whereas for the baseline the context is the full text of the test document. Pre-processing overheads are not taken into account while computing execution timings.

6. RESULTS AND DISCUSSION

This section is sub-categorized into two sections. The first section discusses the quantitative evaluation of the proposed work. In the next section we present a qualitative discussion about the results of the proposed algorithm for some of the documents in the test corpus.

6.1 Quantitative evaluations

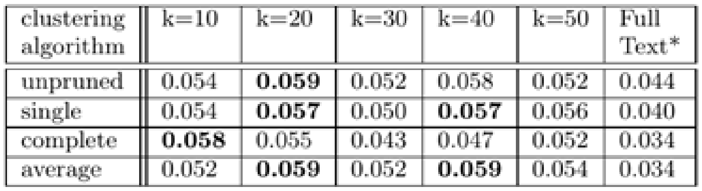
In this section, we describe three experiments conducted to quantitatively evaluate the proposed

approach. The first experiment compares the similarity of the results from the pro- posed and the baseline approaches. The second experiment gives insights about the differences between the proposed approach and the baseline using the ground truth information. The last experiment compares the proposed approach and the baseline based on the execution time performances. These experiments are described as follows.

6.1.1 Experiment 1

Figure 2 shows the distribution of Jaccard similarity for 50 documents for different clustering algorithms. Figure 2 (a), (b) and (c) corresponds to the results of single, complete and average hierarchical clustering based de-noising algorithms. The x-axis of these plots show the variation over k (the top-k headings incorporated in the expanded context). The value of k varies from 10 to 50 in steps of 10. A context made of top-50 web search results are referred as 'top-50' in the plots. The y-axis shows the Jaccard similarity value. The box in the plots are distribution of the Jaccard similarity values for the 50 test documents. The red bar in the box corresponds to the median of the similarity value, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually as '+'. As shown in these figures, the value of Jaccard similarity is not very high.
On an average this value is lesser than 0.10 which signifies a low order of similarity between the tags generated using expanded context versus the tags generated using the full text. However, for the sake of comparison, we find that the Jaccard similarity between the expanded context tagging and full text tagging is higher when the value of k is low. There are very few test documents which have a high Jaccard similarity as shown by the outliers. The maximum similarity is 0.5 for almost all the values of k. We also see that using different clustering algorithms do not make a significant difference in the Jaccard similarity.

Table 2: Table showing Jaccard Index measure for the pro-posed approach (varying k in context expansion) and the full content baseline



clustering algorithm	k=10	k=20	k=30	k=40	k=50	Full Text*
unpruned	0.054	**0.059**	0.052	0.058	0.052	0.044
single	0.054	**0.057**	0.050	**0.057**	0.056	0.040
complete	**0.058**	0.055	0.043	0.047	0.052	0.034
average	0.052	**0.059**	0.052	**0.059**	0.054	0.034

6.1.2 Experiment 2

Based on the Experiment 1, we can say that the tags generated using expanded context and the tags generated using the full text do not overlap significantly. However, this experiment does not conclude about the quality of the tags generated by both the approaches. In order to compare the quality of tags generated by both the approaches, we evaluate the results of the proposed approach and the baseline approach using the ground truth tags for the test documents.

The results of this experiment are shown in table 2. As described earlier, we use the Jaccard index to compare between the qualities of the results. The rows in this table correspond to the results obtained by using different clustering algorithms. The first five columns correspond to the expanded context extracted using k as 10, 20, 30, 40 and 50. The last column contains the results for the baseline referred as Full Text since the context consists of the full text of the document. For the first row (unpruned), tags are not de-noised using any algorithm. The Jaccard index of the baseline (Fulltext) with the ground truth is 0.044 whereas the Jaccard index for all the expanded context (proposed approach) over all values of k is greater than 0.50. The highest Jaccard index is 0.059 at k=20.

When we use the single hierarchical clustering algorithm for de-noising, the Jaccard index is only reduced to 0.040 for Fulltext baseline. The Jaccard index in the expanded context with k=20, 40 is 0.057 which is clearly higher to the baseline results. Similarly for the complete hierarchical clustering based de-noising, the Jaccard index is 0.058 for k=10 whereas it is only 0.034 for the full text baseline. The same scenario is found for average hierarchical clustering based de-noising. The Jaccard index is 0.059 for k=20, 40 while it is only 0.034 for Fulltext baseline.

The above described experiment shows a quantitative approach for comparing the quality of resultant tags from the proposed and the baseline approaches. The results shows above, surprisingly, favor the tags generated by the proposed approach which uses only the title/heading information about the document and web intelligence to annotate the document with relevant tags. The baseline approach uses the full text of the document in order to generate annotations. Although the degree of overlap between the predicted tags with the ground truth tags is low (due to the inherent challenge of natural language), the results are useful to show the difference in the quality of predicted tags by the proposed and baseline approaches. An explanation for the observed results can be attributed to the fact that context derived from the web contains a wide spectrum of terms useful for generating generalized tags for the document. While on the other hand, the full text approach uses only the terms local to the specific document which might not be diverse enough to generate generalized tags.

Since an exact match evaluation (as done above) might not fully account for the quality of tags, we demonstrate the results of a few sample documents from the test dataset in the qualitative evaluation section.

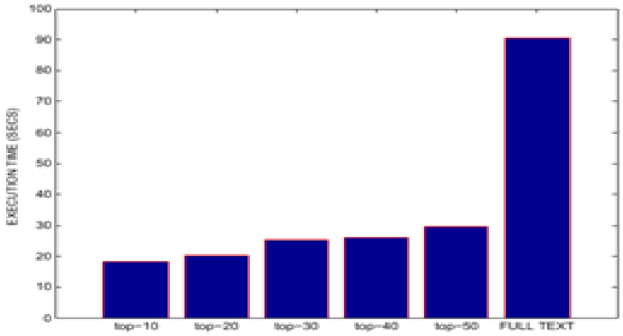


Figure 3: Figure showing execution time comparison for the tag generation step using the expanded context (varying k) vs the full text for 50 documents .

6.1.3 Experiment 3

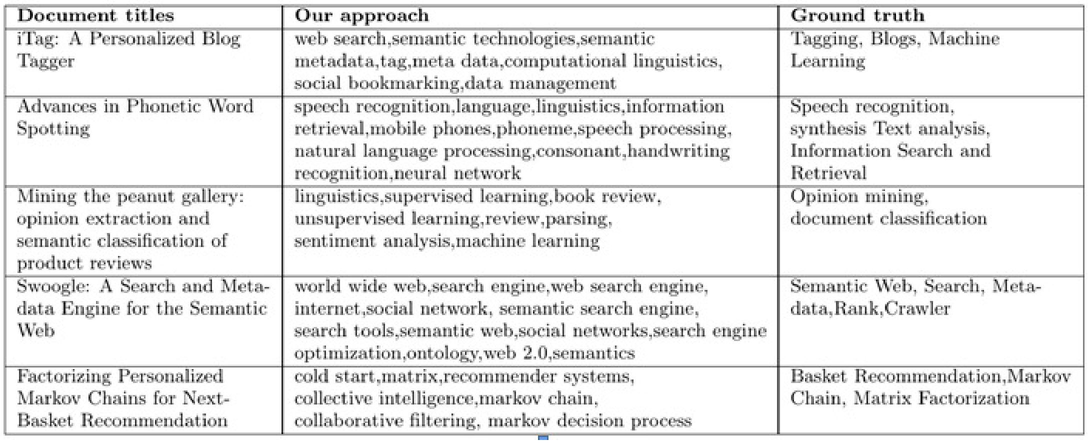
One of the challenges described about using the full text approach for tagging is the issue of time consumption for reading the full text in case the document is large. In Figure 3, we show the results of an experiment conducted to compare the execution time of the tag generation step for the pro- posed and the baseline approaches. The x-axis in the figure shows the expanded context (using different values of k) and the baseline (Full text). The y-axis corresponds to the total execution time (in seconds) for 50 documents. AS shown in the figure, the execution time for the baseline is approximately 90 seconds for 50 documents, whereas the maximum execution time is only 30 seconds for the expanded context where k=50. As shown earlier, that the quality of tags generated using expanded context with k=10 or k=20 is as good as higher values of k. This implies that good quality tags for a document can be generated 4.5 times faster using the proposed approach than using the full text of the document. This shows the effectiveness of the proposed approach to be useful in real time systems.

6.2 Qualitative evaluation

In this section, we discuss the results of the proposed approach by qualitatively analyzing the results of the proposed algorithm. The last section highlighted the performance of the proposed algorithm and quantitatively compared the results with the baseline using the Jaccard index. However, using a quantitative measure like Jaccard fails to account for the subjective accuracy of the tags other than those which do not match the ground truth exactly. Here we analyze the results in a subjective manner.

Table 3 shows the tags, predicted by the proposed approach and the ground truth tags for a few sample documents from the test dataset. The titles shown in this table (in column 1) in general capture the core ingredients of the document. The second column captures the results of the proposed approach and the column 3 captures the ground truth tags.

Table 3: Table showing results for a few of the sample documents. This table shows that several of the topics in the second column (our approach) are very closely related to the keywords in the ground truth (column 3).



Document titles	Our approach	Ground truth
iTag: A Personalized Blog Tagger	web search,semantic technologies,semantic metadata,tag,meta data,computational linguistics, social bookmarking,data management	Tagging, Blogs, Machine Learning
Advances in Phonetic Word Spotting	speech recognition,language,linguistics,information retrieval,mobile phones,phoneme,speech processing, natural language processing,consonant,handwriting recognition,neural network	Speech recognition, synthesis Text analysis, Information Search and Retrieval
Mining the peanut gallery: opinion extraction and semantic classification of product reviews	linguistics,supervised learning,book review, unsupervised learning,review,parsing, sentiment analysis,machine learning	Opinion mining, document classification
Swoogle: A Search and Meta-data Engine for the Semantic Web	world wide web,search engine,web search engine, internet,social network, semantic search engine, search tools,semantic web,social networks,search engine optimization,ontology,web 2.0,semantics	Semantic Web, Search, Meta-data,Rank,Crawler
Factorizing Personalized Markov Chains for Next-Basket Recommendation	cold start,matrix,recommender systems, collective intelligence,markov chain, collaborative filtering, markov decision process	Basket Recommendation,Markov Chain, Matrix Factorization

For the first document in the table ('iTag: A Personalized Bog Tagger'), the keywords (our ground truth) assigned by the used contains terms like 'tagging', 'blogs' and 'Machine learning'. The tags generated by the proposed approach are shown in the middle column. Although there are no exact match between the proposed tags and the ground truth tags yet the relevance of the proposed tags is striking. Tags such as 'semantic meta data', 'social bookmarking', 'tag', 'computational linguistics' are similar others in this list are clearly good tags in this document. Another example is shown in the next row. The ground truth tag 'speech recognition' exactly match the tag in the proposed list. However, most of the other tags in the list of proposed tags are quite relevant. For example, tags such as 'linguistics', 'natural language processing' are closely related to this document. A few tags such as 'mobile phones', 'consonant, 'hand writing recognition' may not be directly related. The third example shown in this table also confirms the effectiveness of the proposed approach. The ground truth consists of only two tags: 'opinion mining' and 'document classification' while the proposed tag list consists several relevant tags though there is no exact match.

The last two examples shown in this table demonstrate the effectiveness of the approach to expand the annotation with meaningful tags. The fourth example is originally tagged with tags like ’semantic web’, ’search’, ’meta-data’, ’rank’ and ’crawler’. But the proposed tag list consists of highly relevant tags like ’ontology’, ’search optimization’ which capture even the technique used in the particular research document. Similarly, for the last example the not overlapping tags are relevant for annotating the research document.

From the above qualitative analysis, we get a better understanding about the quality of tags generated using the proposed approach. Although it is an interesting challenge to quantitatively describe the quality of the proposed tags, this problem is not addressed in the current version of the work.

7. RELATED WORK

As described earlier, the literature under document annotation can be divided into two broad classes. The first class of approaches studies the problem of annotation using extraction techniques [5, 6]. The main objective of such techniques is to identify important words or phrases from within the content of the document to summarize the document. This class of problem is studied in the literature by several names such as “topic identification” [3],”categorization” [19, 13], “topic finding” [14],”cluster labelling” [18, 16, 21, 24] and as well as “keyword extraction” [5, 6].

Researchers working on these problems have used both supervised and unsupervised machine learning algorithms to extract summary words for documents. Witten et al. [23] and Turney [22] are two key works in the area of supervised key phrase extraction. In the area of unsupervised algorithms for key phrase extraction, Mihalcea and Tarau [17] gave a textRank algorithm which exploits the structure of the text within the document to find key-phrases. Hasan and Ng [8] give an overview of the unsupervised techniques used in the literature.

In the class of key phrase abstraction based approaches. There can be two approaches for document annotation or document classification: single document annotation and multiple document annotation. In the single document summarization, several deep natural language analysis methods are applied. These strategies of document summarization use ontology knowledge based summarization [9, 11]. The ontology sources commonly used are WordNet, UMLS. The second approach widely used in single document summarization is feature appraisal based summarization. In this approach, static and dynamic features are constructed from the given document. Features such as sentence location, named entities, semantic similarity are used for finding documents similarity.

In the case of multi-document strategies, the techniques in- corporate diversity in the summary words by using words from other documents. However, these techniques are limited when the relevant set of documents is not available. Gabrilovich et. al [7] proposed an innovative approach for document categorization which uses of Wikipedia knowledge base to overcome the limitation of generating category terms which are not present in the documents. However, this approach uses the entire content of the document and extend the context using Wikipedia.

8. CONCLUSIONS

In summary, there are three main conclusions in this work. Firstly, we showed an automated approach for tag generation using only a short text information from the document and intelligence from the web. Secondly, we quantitatively evaluated and compared the results of the proposed approach against the baseline approach which uses the full text of the document. We used different metrics to compare and contrast the results. We found that the proposed approach performs better than the baseline approach in terms of the Jaccard index with the ground truth tags. We also found that the proposed approach is at least 3 times faster than the baseline approach and thus, useful for real time system. Thirdly, we evaluated the quality of the proposed tags for documents against the ground truth tags in a qualitative fashion. This analysis reveals the qualitative effectiveness of the proposed approach for meaningful tag generation using only ’short text’ information from the document.

There are several areas in this work which we would extend in the near future. One of the areas of improvement in the current work is the de-noising algorithm which uses hierarchical clustering to pruning. However, hierarchical clustering has its limitations and it is worth to explore other algorithms such as density based clustering or some novel anomaly detection algorithm. We would also test the pro- posed approach for other document corpus like news, patents etc. Finally, we also plan to quantitatively validate the accuracy of the results in the case when the results do not exactly match the ground truth.

ACKNOWLEDGEMENTS

This work was supported by the Army Research Laboratory and DARPA and was accomplished under Cooperative Agreement Numbers W911NF-09-2-0053 and W911NF-12-C-0028. We also thank the lab members of DMR lab, University of Minnesota for their valuable suggestions regarding experimental design and comparison.

REFERENCES

[1] AlchemyAPI. Text analysis by alchemyapi, 2013.

[2] R. L. Cilibrasi and P. M. Vitanyi. “The google similarity distance”. IEEE Transactions on Knowledge and Data Engineering, Vol 19(3):370–383, 2007.

[3] C. Clifton, R. Cooley, and J. Rennie. “Topcat: data mining for topic identification in a text corpus”. IEEE Transactions on Knowledge and Data Engineering, 16(8):949–964, 2004.

[4] O. Corcho. “Ontology based document annotation: trends and open research problems”. International Journal of Metadata, Semantics and Ontologies, 1(1):47–57, 2006.

[5] L. Ertöz, M. Steinbach, and V. Kumar. “Finding topics in collections of documents: A shared nearest neighbor approach”. Clustering and Information Retrieval, 11:83–103, 2003.

[6] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. “Domain-specific keyphrase extraction”. 1999.

[7] E. Gabrilovich and S. Markovitch. “Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge”. In AAAI, volume 6, pages 1301–1306, 2006.

[8] K. S. Hasan and V. Ng. “Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art”. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 365–373. Association for Computational Linguistics, 2010.

[9] M. M. Hassan, F. Karray, and M. S. Kamel. “Automatic document topic identification using wikipedia hierarchical ontology”. In 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), 2012, pages 237–242. IEEE, 2012.

[10] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles. “Context-aware citation recommendation”. In Proceedings of the 19th international conference on World wide web, pages 421–430. ACM, 2010.

[11] S. Jain and J. Pareek. “Automatic topic (s) identification from learning material: An ontological approach”. In Second International Conference on Computer Engineering and Applications (ICCEA), 2010, volume 2, pages 358–362. IEEE, 2010.

[12] N. H. Jesse Alpert. “We knew the web was big...”, July 2008.

[13] T. Joachims. “Text categorization with support vector machines: Learning with many relevant features”. Springer Berlin Heidelberg (pp., 137-142) 1998.

[14] D. Lawrie, W. B. Croft, and A. Rosenberg. “Finding topic words for hierarchical summarization”. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pages 349–357. ACM, 2001.

[15] M. Ley and P. Reuther. “Maintaining an online bibliographical database: The problem of data quality”. In EGC, pages 5–10, 2006.

[16] C.-Y. Lin. “Knowledge-based automatic topic identification”. In Proceedings of the 33rd annual meeting on Association for Computational Linguistics, pages 308–310. Association for Computational Linguistics, 1995.

[17] R. Mihalcea and P. Tarau. “Textrank: Bringing order into texts”. In Proceedings of EMNLP, volume 4. Barcelona, Spain, 2004.

[18] M. F. Moura and S. O. Rezende. “Choosing a hierarchical cluster labelling method for a specific domain document collection”. New Trends in Artificial Intelligence, pages 812–823, 2007.

[19] F. Sebastiani. “Machine learning in automated text categorization”. ACM computing surveys (CSUR), 34(1):1–47, 2002.

[20] A. Singhal, R. Kasturi, and J. Srivastava. “Automating document annotation using open source knowledge”. In IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013, volume 1, pages 199–204. IEEE, 2013.

[21] S. Tiun, R. Abdullah, and T. E. Kong. “Automatic topic identification using ontology hierarchy”. In Computational Linguistics and Intelligent Text Processing, pages 444–453. Springer, 2001.

[22] P. D. Turney. “Learning algorithms for keyphrase extraction”. Information Retrieval, 2(4):303–336, 2000.

[23] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning. “Kea: Practical automatic keyphrase extraction”. In Proceedings of the fourth ACM conference on Digital libraries, pages 254–255. ACM, 1999.

[24] O. Zamir and O. Etzioni. “Web document clustering: A feasibility demonstration”. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 46–54. ACM, 1998.

AUTHORS

Ayush Singhal was born in India in the year 1990. He is currently a second year PhD student in the computer science department at the University of Minnesota, USA. He completed his under graduation from the Indian Institute of Technology Roorkee, India in 2011. His major is computer science. His current research interests are data mining, information retrieval, web mining and social network analysis.As an under graduate he has co-authored two conference papers in prestigious national and international IEEE conference. He has also published a journal article in Springer journal (Real time image processing). He has been working as a research assistant in the University of Minnesota for 2 years now. He has also worked in IBM Research labs, New Delhi India as a summer intern in year 2010.

Jaideep Srivastava received the Btech degree in computer science from The Indian Institute of Technology, Kanpur, India, in 1983, the MS and PhD degrees in computer science from the University of California, Berkley, in 1985 and 1988 respectively. He has been on the faculty of the Department of Computer Science and Engineering of the University of Minnesota, Minneapolis, since 1988 and is currently a professor.He served as a research engineer with Uptron Digital Systems in Lucknow, India, in 1983. He as published more than 250 papers in refereed journals and conference proceedings in the areas of databases, parallel processing, artificial intellig3ence, multimedia and social network analysis; and he has delivered a number of invited presentations and participate in panel discussions on these topics. His professional activities have included service on various program committees and he has refereed papers for varied journals and proceedings, for events sponsored by the US National Science Foundation. He is a Fellow of the IEEE, and a Distinguished Fellow of Allina Hospitals’ center for Healthcare Innovation. He has given over 150 invited talks in over 30 countries, including more than a dozen keynote addresses at major conferences.

AUTHOR INDEX