Computer Science & Information Technology 23

Natarajan Meghanathan
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

Second International Conference of Advanced Computer Science & Information Technology (ACSIT 2014)
Zurich, Switzerland, June 14 ~ 15 - 2014

AIRCC

Volume Editors

Natarajan Meghanathan,
Jackson State University, USA
E-mail: nmeghanathan@jsums.edu

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

ISSN : 2231 - 5403
ISBN : 978-1-921987-25-0
DOI : 10.5121/csit.2014.4601 - 10.5121/csit.2014.4612

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

Preface

Second International Conference of Advanced Computer Science & Information Technology (ACSIT 2014) was held in Zurich, Switzerland, during June 14~15, 2014. Second International conference on Signal Image Processing and Multimedia (SIPM 2014), Second International Conference on Foundations of Computer Science & Technology (FCST 2014), Second International Conference of Information Technology, Control and Automation (ITCA 2014), Second International Conference on Software Engineering (SE 2014), Second International Conference of Managing Information Technology (CMIT 2014), Second International Conference on Information Technology in Education (ICITE 2014) were collocated with the ACSIT-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 ACSIT-2014, SIPM-2014, FCST-2014, ITCA-2014, SE-2014, CMIT-2014, ICITE-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, ACSIT-2014, SIPM-2014, FCST-2014, ITCA-2014, SE-2014, CMIT-2014, ICITE-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 ACSIT-2014, SIPM-2014, FCST-2014, ITCA-2014, SE-2014, CMIT-2014, ICITE-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.

Natarajan Meghanathan
Dhinaharan Nagamalai

Organization

Program Committee Members

Aakanksha Pandey	Chouksey Engineering College, India
Aarti Milind Karande	S.P.I.T. University of Mumbai, India
Abd El-Aziz Ahmed	Anna's University, Egypt
Abdolreza Hatamlou	Islamic Azad University, Iran
Abdul Aziz	Slamia University of Bahawalpur, Pakistan
Abdul Raouf Khan	King Faisal Unversity, Saudi Arabia
Abdurrahman Celebi	Bedër University, Albania
Abha Kiran Rajpoot	Sharda University, India
Abhay Saxena	Dev Sanskriti Vishwavidyalaya, India
Abirami	Anna University, India
Adam Taylor	Trinity College Dublin, Ireland
Aditya Goel	Indian Institute of Technology Bombay, India
Aditya Goel	TIT&S Bhiwani, India
Ahmed Arara	University of Tripoli, Ahmed
Ahmet Cinar	FıRat University Faculty of Engineering, Turkey
Ajit Shrivastava	Rajiv Gandhi Technical University, India
Akhil Jabbar M	Aurora's Engineering College, India
Akhil jabbar Meerja	Aurora's Engineering College, India
Ali Azimi	Ferdowsi University of Mashhad, Iran
Alia Ghaddar	Lebanese University, Lebanon
Alireza Masoum	University of Twente, India
Alireza Souri	Islamic Azad University, Iran.
Aman Singh	Lovely Professional University, India
Amandeep Singh	NIT Jalandhar, India
Amanpreet Kaur	ITM university, India
Ambresh	Mangalore University, India
Ameur Mohamed El Amine	Amar Telidji Laghouat, Algeria
Amir Khusru Akhtar	Cambridge Institute of Technology, India
Amita Sharma	The IIS University, India
Amitava Mukherjee	IBM India Private Limited, India
Amr Rekaby	ERSIL Research Lab, Egypt
Anandkumar Mani	Anna University, India
Anil Kumar Dubey	Govt. Engineering College Ajmer, India
Anitha	Karunya University, India
Anju S. Pillai	Amrita Vishwa Vidyapeetham, India
Anna Tcherkassof	Universite Pierre Mendes France
Anujvadhera	Lovely Professional University, India
Archana Chougule	MIT, India
Aref Tahmasb	Shahid Bahonar University, Iran
Arindam Sarkar	University of Kalyani, India
Arpit Bhardwaj	Scholar Indian Institute of Technology, India
Arun Fera	Thiagarajar College of Engineering, India
Aruna Pathak	Govt Engg College Bharatpur Rajasthan, India

Asha.K	Krupanidhi school of Management, India
Ashok Kumar	CMR Engineering College, India
Ashutosh Dubey	TITR Bhopal, India
Ashutosh Gupta	Amity University, India
Ashutosh Kumar Dubey	TITR, India
Avinash Shankaranarayanan	Vels University, India
Balaji S	Visvesvaraya Technological University, India
Balakannan S.P	Kalasalingam University, India
Balraj Singh	Lovely Professional University, India
Balwinder Singh	CDAC, India
Barbaros Preveze	Cankaya University, Turkey
Barun Parichha	Ericsson Research and Development, India
Basant Verma	AICTE, India
Basit Shahzad	King Saud University, Saudi Arabia
Belsare Yeshwantrao A.D	Chavan College of Engineering, India
Benmohammed Pr. M	University of Constantine, Algeria
BenZidane Moh	University of Constantine, Algerie
Bharani Sethupandian	Madurai Kamaraj University, India
Bharathi Mallapur	Basaveshwar Engineering College, India
Bindu. K R	Amrita School of Engineering, India
Binod Kumar Pattanayak	Siksha 'O' Anusandhan University, India
Chairiawaty Hendar	Bandung Islamic University, India
Chandrakant	Bangalore University, India
Chandramohan.D	Pondicherry University, India
Chandrappa D N	SJBIT, India
Chintan patel	Gujarat Technological University, India
Damien Dupre	Univ. Grenoble Alpes, France
Daniel AK	M.M.M University of Technology, India
Daniel D. Dasig	Jose Rizal University, Philippines
Daniel K	M.M.M University of Technology, India
Danish Abbas	Beijing Institute of Technology, China
Debasish Jana	TEOCO Software Pvt Ltd, India
Debdas Ghosh	IIT Kharagpur, India
Debjani Chakraborty	Indian Institute of Technology, India
Deepa T P	Acharya Institute of Technology, India
Deepti Chopra	Banasthali Vidyapith, India
Derdour Makhlouf	University of Tebessa, Algeria
Dharam Pathak	Chameli Devi Group of Institutions Indore, India
Dharani Andhe-Putandoddi	R.V.College of Engineering, India
Dharmender Singh Kushwaha	MNNIT Allahabad, India
Dhinaharan Nagamalai	Wireilla Net Solutions, Australia
Dias N.G.J	University of Kelaniya, Sri Lanka
Dires, Fasil Fenta	University of Gondar, Ethiopia
Dongale T. D	Shivaji University, India
Dongchen Li	Peking University, P.R.China
Doreswamy	Mangalore University, India
Dripto Bakshi	Indian Statistical Institute, India
Durga Mohapatra	NIT-Rourkela, India

Durgesh Samadhiya	Chung Hua University, Taiwan
Edward David Moreno	Federal University of Sergipe, brazil
Ehsan Ali	University of Lahore, Pakistan
Ehsan Heidari	Islamic Azad University, Iran
Ehsan Saradar Torshizi	Urmia University, Iran
El Miloud Ar-Reyouchi	Abdelmalek Essaadi University, Morroco
Elahe Badiee	Payamenoor University, Iran
Elmahdi Abousetta	University of Tripoli, Libya
Eshan Kapur	Lovely Professional University, India
Fadhil A. Ali	Oklahoma State University, USA
Fatemeh Alidusti	Islamic Azad University, Iran
Fatih Korkmaz	Cankırı Karatekin University, Turkey
Ganesan G	Adikavi Nannaya University, India
Gaurav Kumar Tak	Lovely Professional University, India
Geeta Totad	GMR Institute of Technology, India
Geetha Ramani	Anna University, India
Geetha S	Thiagarajar College of Engginering, India
Geetharamani R	Anna University, India
Ghazali Sulong	Universiti Teknologi Malaysia, Malaysia
Gladis Pushpa Rathi V.P	Sudharsan Engineering College, India
Gondi Lakshmeeswari	GITAM University, India
Goraksh Garje	PVGs's College of Engg. & Tech, India
Govindraj B. Chittapur	Basaveswar Engineering College, India
Gulshan Shrivastava	Uttar Pradesh Technical University, India
Hacene Belhadef	University of Constantine, Algeria
Hamdi Yalin Yalic	Hacettepe University, Turkey
Hameem Shanavas	MVJ College of Engineering, India
Hamid Bentarzi	UMBB University, Algeria
Hari Chavan	Terna Engineering College, India
HASSINI Noureddine	University of Oran, Algeria
Hayati Mamur	Cankiri Karatekin University, Turkey
Hemalatha Sekar	Karpagam University, India
Hemant Darbari	Pune University Campus, India
Hemant Pasusangai Kasturiwale	Mumbai University, India
Hicham Behja	University of Hassan II Casablanca, Morocco
Himanshu Pathak	Amity University Lucknow Campus, India
Hossein Jadidoleslamy	MUT University, Iran
Hota H.S	Guru Ghasidas Central University, India
Hsin-Chou Chi	National Dong Hwa University, Taiwan
Hui Wu	University of New South Wales, Australia
Hussein Ismail Al-Bahadili	Petra University, Jordan
Hyung-Woo Lee	Hanshin University, South Korea
Ijaz Ali Shoukat	King Saud University, Saudi Arabia
Inderpal Singh	Punjab Technical University, India
Indira K	E.S.Engineering College, India
Indr Jeet Rajput	Gujarat Technical University, India
Indra Rajput	Gujarat Technical University, India
Indrajit Bhattacharya	Kalyani Government Engineering College, India

Indrajit Mandal	Sastra University, India
Indumathi J	Anna University, India
Isa Maleki	Islamic Azad University, Iran
Jagadeesh K	National Institute of Rourkela, India
Jagdish Bhatt	Tribhuvan University, Nepal
Jaison B	RMK Engineering College, India
Jan Lindström	SkySQL - The MariaDB Company, Finland
Javed Ali	Glocal University, India
Jayachandran D	KSR College of Engineering, India
Jayakumar C	RMK Engineering College, India
Jayalakshmi V	Sudharsan Engineering College, India
Jayant Gambhir	MM Polytechnic, India
Jaydeep Howlader	National Institute of Technology, India
Jin-Cherng Lin	Tatung University, Taiwan
JohnTenvile	Sunyani Polytechnic, Ghana
José Raniery	Federal University of Alagoas, Brazil
Jungpil Shin	University of Aizu, Japan
Jyothi Pillai	Bhilai Institute of Technology, India
Jyoti Gautam	JSS Academy of Technical Education, India
Kalai Vani	Easwari Engineering College, India
Kamal Sutaria	Gujarat Technological University, India
Kamala Krithivasan	Indian Institute of Technology, India
Kamaraju M	Gudlavalleru Engineering College, India
Kanti Prasad	University of Massachusetts Lowell, Lowell
Karthi S	Sathyabama University, India
Karunakaran S	Kongu Engineering College, India
Kashif Ahmed	CMRIT, India
Kavita Choudhary	ITM University, India
Kavitha Rajamani	St. Aloysius College, India
Keneilwe Zuva	University of Botswana, Botswana
Kenneth Mapoka	Botswana College of Agriculture, Botswana.
Keshavamurthy B.N	Manipal institute of Technology, India
Ketan Goswami	Parul Institute of Technology, India
Khaled Merit	University of Mascara, Algeria
KHAZE S.R	Islamic Azad University, Iran
Khikmat Muminov	Umarov Physical Technical Institute, Tajikistan
Kiran.D.C	BITS Pilani, India
Koushik Majumder	West Bengal University of Technology, India
Krishit	Anna university, India
Krishna Prakash K	MIT-Manipal, ,India
Kunwar Singh Vaisla	SMIEEE, India
Kuppusamy	Alagappa University, India
Kurd gift	IMAM University, Saudi Arabia
Lakshmi C Devasena Radhakrishnan	IFHE University, India
Lakshmi P.R.S.M	Vignan University, India
Lakshmi	Amrita School of Engineering, India
Lathief K.A	Jimma university, Ethiopia.
Leelavathi G	Visvesvaraya Technological University, India

Lokesh Sharma	Manipal University, India
M.Kamaraju	Gudlavalleru Engineering College, India
Madhavi Vaidya	University of Mumbai, India
Mahdieh Ghazvini	Shahid Bahonar University of Kerman, Iran
Mahesh K	Alagappa University, India
Mai Shouman	UNSW@ADFA, Australia
Malliga Raguraman	Dhanalakshmi College of Engineering, India
mamy alain Rakotomalala	Universite d'Antananarivo, Madagascar
Mani Joseph P	Modern College of Business and Science, Oman
Manik Sharma	Sewa Devi S.D. College, India
Manishsingh Chaudhary	Cognizant Technology Pvt Solution, India
Manju Khari	Ambedkar Institute of Technology, India
Manjunath T.C	HKBK College of Engineering, India
Manpreet Singh	M. M. University, India
Mansaf Alam	Jamia Millia Islamia, India
Mansouri Ali	Université Claude Bernard Lyon1, Tunisia
Manu Sood	Himachal Pradesh University, India
Maragathavalli P	Pondicherry Engineering College, India
Maryam K	University of Tabriz, Iran
Maryamsoltanalikhalili	Shahed University, Iran
Masoud Ziabari	Mehr Aeen University, Iran
Maya Ingle	School of Computer Science & IT, India
Md. Amir Khusru Akhtar	Cambridge Institute of Technology, India
Meenakshi A.V	Periyar Maniammai University, India
Meenakshi M	Dr. Ambedkar Institute of Technology, India
Melih KIRLIDOG	Marmara University, Turkey
Meyyappan T	Alagappa University, India
Milind M. Mushrif	Y. C. College of Engineering, India
Minesh Thaker	Indus University, India
Mohamed Alajmi	King Saud University, Saudi Arabia
Mohamed Hashem Abd El-Aziz Ahmed	Ain Shams University, Egypt
Mohammad Arif	Integral University, India
Mohammad Asmat Ullah Khan	Effat University, Saudi Arabia
Mohammad H. Alomari	Applied Science University, Jordan
Mohammad Khanbabaei	Islamic Azad University, Iran
Mohammad Zunnun Khan	Integral University, India
Mohammed AbouBakr Elashiri	Cairo University, Egypt
Mohammed Abufouda	University of Kaiserslautern, Germany
Mohammed Al-kahtani	Salman Bin Abdulaziz University, Saudi Arabia
Mohammed Faizan Farooqui	Integral University, India
Mohammed Youssif	Hewlett-Packard, USA
Mohankumar N	Amrita Vishwa Vidyapeetha, India
Mohd Arif	Integral University, India
Mohiy Mohamed Hadhoud	Menoufia university, Egypt
Morteza Saberi	University of Tafresh, Iran
Mostafa Abo-bakr Abdelmajed	Assiut University, Egpt
Muhammad Ilyas	University of Sargodha, Pakistan
Muhammad Imran Khan	Université de Toulouse, France

Muhammad Saeed	Federal Urdu University, Pakistan
MV Ramana Murthy	Osmania University, India
Nagaraj SV	RMK Engineering College, India
Nandhini G.B	Anna University, India
Narasimha Murthy K N	Vemana Institute of Technology, India
Narendra Kumar Rao B	Sree Vidyanikethan Engineering College, India
Narendra V G	Narendra V G, India
Naresh Sharma	SRM University, India
Natarajan Meghanathan	Jackson State University, USA
Nazirah	Universiti Sultan Zainal Abidin, Malaysia
Neda Darvish	Islamic Azad University, Iran
Neda Enami	Payame Noor University, Iran
Neelam Goyal	PEC University of Technology, India
Neethu Mathai	KMP College of Engineering, India
Neha Chaudhary	Uttar Pardesh Technical University, India
Ngo Minh Vuong	Free University of Bolzano, Italy
Nickolas S	National Institute of Technology, India
Nikunj Domadiya	Gujarat Technologial University, India
Nilesh Dhannaseth	Cognizant Technology Solutions, India
Nilesh Prajapati	Gujarat Technological University, India
Niloofar Khanghahi	Islamic Azad University, Iran
Nirmala Devi L	Osmania University, India
Nirmala Devi M	Amrita Vishwa Vidyapeetham, India
Nishant Doshi	NIT Surat, India
Nisheeth Joshi	Banasthali University, India
Nishkarsh Sharma	Graphic Era University, India
Noureddine Bouhmala	Buskerud and Vestfold College, Norway
Ola Younes	Philadelphia University, Jordan
Olufade F. W. Onifade	University of Ibadan, Nigeria
Orhan Dagdeviren	Ege University, Turkey
Ouarda Barkat	University of Constantine, Algeria
Pallawi Bulakh	Modern College, India
Palsonkennedy Rajagopal	Anna University, India
Panchami Vijayan	TocH Institute of Science and Technology, India
Pankajgupta	Indian Institute of Technology Roorkee, India
Parth Shah	Charusat, India
Patil R.A	College of Engineering, India
Peiman Mohammadi	Islamic Azad University, Iran
Phuc V. Nguyen	Polytechnic Institute Saint Louis, France
Pinaki Bhaskar	Jadavpur University, India
Poonam Saini	PEC University of Technology, India
Prabhu P	Alagappa University, India
Pradeesh Hosea	Bharathiar University, India
Pradnya Kulkarni	Federation University, Australia
Praneeth Kumar G	EiQ Networks India Pvt. Ltd, India
Prasad Halgaonkar	MIT College of Engineering, India
Prasad T. V	Visvodaya Technical Academy, India
Pratibha Singh	IET DAVV Indore, India

Praveen Kumar	JNTU, India
Prieya Dharsini	Anna University, India
Purohit G.N	Banasthali University, India
Pushpa Siva Kumar M	University of Madras, India
Pushpendra Kumar Pateriya	Lovely Professional University, India
Pushpendra Pateriya	Lovely Professional University,, India
Raed A. Alsaqour	Universiti Kebangsaan Malaysia, Malaysia
Rafah M. Almuttairi	University of Babylon, Iraq
Rahmath Safeena Abdullah	Taif University, Saudi Arabia
Rahul Gupta	Manipal Institute of Technology, India
Raj Mohan	IFET College of Engineeirng, India
Rajan Vohra	Guru Nanak Dev University, India
Rajat	JECRC University, India
Rajendra A B	Vidyavardhaka College of Engineering, India
Rajeshwari Hegde	BMS College of Engineering, India
Raji Goutham	M.S.Ramaiah Institute of Technology, India
Rajiv Pandey	Amity University, India
Rajkumar R	Vellore Institute of Technology, India
Ram Gopal L	NSN, USA
Ramachandra Rao Kurada	Shri Vishnu Engineering College, India
Ramakrishna K	Sridevi Women's Engineering College, India
Ramesh Babu	VIT University, India
Ramesh Sunkaria	National Institute of Technology, India
Ramkumar Prabhu M	Anna University, India
Ramu	JNTUA College of Engineering, India
RanjanKumar	Cambridge Institute of Technology, India
Ranjeet Vasant Bidwe	Pune Institute of Computer Technology, India
RAO M N	SCET Enggineering College, India
Rapali d	Rajiv Gandhi Technical University, India
Rashi Agrawal	CSJM University, India
Rasmiprava Singh	MATS University, India
Raveendra Babu B	VNR Vignana Jyothi Inst of Engg & Tech, India
Ravendra Singh	MJP Rohilkhand University, India
Ravindranath Kongara	K.L.University, India
Reena Pagare	MIT College of Engineering, India
Renjith Kurup	Rajagiri College of Social Sciences, India
Reza Ebrahimi Atani	University of Guilan, Iran
Reza Ravanmehr	Islamic Azad University, Iran
Ritu Vijay	Bansthali University, India
Rizwan Beg	Integral University, India
Rohit Jha	Tata Consultancy Services, India
Roseline	Government Arts College-Coimbatore, India
Sachidananda Patnaik	Biju Patnaik University of Technology, India
Sachin Chirgaiya	Oriental University, India
Sachin Tripathi	Indian School of Mines, India
Sadeque Reza Khan	National Institute of Technology, India
Saeid Ghazi	Islamic Azad University, Iran
Safvan. A. Vahora	VGEC Chandkheda, India

Sahab AR	Islamic Azad University, Iran
Sai Kumar	CMR Technical Campus, India
Saikumar Tara	CMR Technical Campus, India
Sakthi Ganesh M	VIT University, India
Salini P	Pondicherry Engineering College, India
Samarendra Nath Sur	Sikkim Manipal Institute of Technology, India
Sandeep Chaware	University of Pune, India
Sandeep Sharma	Gautam Buddha University, India
Sandhya Tarar	Gautam Buddha University, India
Sangeetha Bala	Birla Institute of Technology and Science, India
Sangita Chaudhary	A. C. Patil College of Engineering, India
Sangita Zope-Chaudhari	ACPCE, India
Sanjay K Biswash	San Diego State University, USA
Sanjoy Das	Galgotis University, India
Sankara Malliga	Dhanalakshmi College Of Engineering, India
Santosh Kumar	Victoria University, New Zealand
Sapan Naik	UKA Tarsaida University, India
Saradhi Varma G.P	S.R.K.R.Engineering College, India
Saravanan S	SASTRA University, India
Sarita	University of Rajasthan, India
Sasanko Sekhar Gantayat	GMR Institute of Technology, India
Sasikumar Gurumurthy	VIT University, India
Sassi Abdessamed	Mohamed Khider University, Algeria
Savita Hanji	Basaveshwar Engineering College,India
Sayyed Majid Mazinani	Imam Reza University, Iran
Semih Yumusak	KTO Karatay University, Turkey
Seyed Ziaeddin Alborzi	Nanyang Technological University, Singapore
Shahid Siddiqui	Integral University, India
Shailendra Singh	PEC University of Technology, India
Shamim H Ripon	East West University, India
Shamla Mantri	MIT College of Engineering, India
Shankar T	VIT University, India
Shanmugasundaram Hariharan	TRP Engineering College, India
Shanthi N	Nandha Engineering College, India
Shanthi selvaraj	Rathinam Technical Campus, India
Sharma MK	Amrapali Institute, India
Sharvani Mathad	R V College of Engineering, India
Shervan Fekri Ershad	International Shiraz University, Iran
Shilpi Bose	Netaji Subhash Engineering College, India
Shish Ahmad	Integral University, India
Shiv K. Sahu	Technocrats Institute of Technology, India
Shivani Agarwal	Uttar Pradesh Technical University, India
Shivaputra	Ambedkar Institute of Technology, India
Shubhangi Bhatambrekar	University of Pune, ,India
Siddhivinayak Kulkarni	Federation University, Australia
Smita patel	SKNCOE STES, India
Smitha N. Pai	Massachusetts Institute of Technology, India
Sokyna Mohammad Al-Qatawneh	Al-Zaytoonah University of Jordan, Jordan

Sonika Rathi	College of Engineering Pune, India
Soumen Kanrar	Vehere Interactive (P) Ltd, India
Srinath N.K	R.V. College of Engineering, India
Srinivas Y	GITAM University, India
Srinivasa Rao K	Anurag Group of Institutions, India
Stanka Hadzhikoleva	University of Plovdiv, Bulgaria
Steven Y. Liang	Georgia Institute of Technology, USA
Sudhir G. Akojwar	Rajiv Gandhi College of Engineering, India
Sugantyhi B	Mahalakshmi Engineering College, India
Sujatha B R	Malnad College of Engineering, India
Sultan Alshehri	University of Regina, Canada
Suman Deb	National Institute of Technology Agartala, India
Sundarapandian V	Vel Tech University, India
Suprativ Saha	Adamas Institute of Technology, India
Supriya M	Amrita School of Engineering, India
Suresh Kumar S	Georgia Software Technology Pvt Ltd, India
Sushila Madan	Lady Shri Ram College, India
Swaminathan Bhuvaneswari	Easwari Engineering College, India
Syed Imtiyaz Hassan	Jamia Hamdard (Hamdard University), India
Taruna S	Banasthali University, India
Tawfik T. El-Midany	Mansoura University, Egypt
Thamari S.M	Alagappa Government Arts College, India
Thamo dharan	Thanthai Hans Roever college, India
Thasleema T. M	Central University of Kerala, Kerala
Tinatin Mshvidobadze	Gori University, Georgia
Upendra Kumar M	Mahatma Gandhi Institute of Technology, India
Usha. J	R V College of Engineering, India
Vadivelou G	Mahatma Gandhi Arts & Science College, India
Varaprasad Rao M	Matrusri Institute of PG Studies, India
Vastrad C.M	Mangalore University, India
Venugopal	Visvesaraya Technological University, India
Vibhuti Sikri	Bahra University, India
Vijay Mankar	Government Polytechnic, India
Vijay.V	Infosys Ltd , India
Vijayapal Reddy	Gokaraju Rangaraju Engineering college, India
Vijender Kr Solanki	Anna University, India
Virginia Araujo	University of Vigo, Spain
Vishal Shrivastava	Rajasthan Technical University, India
Vishal Zinjuvadia	StratExcel Technologies, India
Vishnu G. Murthy	Anurag Group of Institutions, India
Vishnu Vardhan B	JNTUH College of Engineering, India
Vivek Arya	JSS Academy of Technical Education, Noida
Wahiba Ben Abdessalem	High Institute of Management of Tunis, Tunisia
Wenisch	Anna University, India
Wenwu Wang	University of Surrey, United Kingdom
Yamunadevi N	Institute of Technology, India
Yogesh Khandagre	Trinity Institute of Technology & Reseach, India
Yuhanis Binti Yusof	Universiti Utara Malaysia, Malaysia

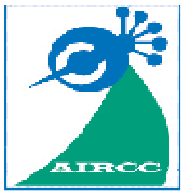
Technically Sponsored by

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

Second International Conference of Advanced Computer Science & Information Technology (ACSIT 2014)

Second International Conference on Foundations of Computer Science & Technology (FCST 2014)

Second International Conference of Information Technology, Control and Automation (ITCA 2014)

Second International Conference on Software Engineering (SE 2014)

Second International Conference on Information Technology in Education (ICITE 2014)

Second International conference on Signal Image Processing and Multimedia (SIPM 2014)

Second International Conference of Managing Information Technology (CMIT 2014)

DESIGNING A DISEASE OUTBREAK NOTIFICATION SYSTEM IN SAUDI ARABIA

Farag Azzedin, Salahadin Mohammed, Jaweed Yazdani and Mustafa Ghaleb

King Fahd University of Petroleum and Minerals
Dhahran 31261, Saudi Arabia
Email: {fazzedin,adam,jaweed,g200905270}@kfupm.edu.sa

ABSTRACT

This paper describes the design and development of a Disease Outbreak Notification System (DONS) in Saudi Arabia. The main function of DONS is to warn for potential outbreaks. A prototype of the DONS was implemented in a hybrid cloud environment as an online/real-time disease outbreak notification system. The system notifies experts of potential disease outbreaks of both pre-listed diseases and totally unknown diseases. The system only accepts cases from pre-registered sources. It is designed to also share information about disease outbreaks with international systems. As soon as the system detects a potential disease outbreak it notifies stakeholders and experts. The system takes feedback from experts to improve the disease detection capabilities and to adapt to new situations.

KEYWORDS

Disease Outbreak Notification System, Saudi Arabia, Prototype, Outbreak

1. INTRODUCTION

Outbreaks such as MERS, SARS [1, 2], the threat of bio-terrorism (e.g. anthrax) [3] and Mad Cow diseases (BSE) [4] as well as the recent different strains of "bird flu" or Influenza (i.e., H1N1 and H5N1) [5] are the most intriguing and complex phenomena that confront scientists in the field of microbiology, virology and epidemiology [6, 7, 8]. The ability of these viruses to mutate and evolve is one of the acute mysteries that puzzle health officials, who are trying to find out the root cause of worldwide pandemics since the late 1880s. Pandemics occur when small changes in the virus over a long period of time eventually “shift” the virus into a whole new subtype, leaving the human population with no time to develop a new immunity.

Unlike bacteria, viruses are sub-microscopic and do not have a cellular structure [9]. Their essential component is genetic material—either DNA (Deoxyribonucleic Acid) or RNA (Ribonucleic Acid)—that allows them to take control of a host cell. Viruses reproduce by invading a host cell and directing it to produce more viruses that eventually burst out of the cell, killing it in the process.

Therefore, a tremendous and an unforeseen threat could mark the start of a global outbreak given the above mentioned scenarios, namely (a) the ability of the virus to shift into a whole new subtype, (b) the time shortage of the human population to develop a new immunity, (c) the

Natarajan Meghanathan et al. (Eds) : ACSIT, FCST, ITCA, SE, ICITE, SIPM, CMIT - 2014
pp. 01–14, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4601

limitation in terms of the effect of immunization, and (d) the viruses' ability to change to a form that is highly infectious for humans and spreads easily from person to person. Furthermore, as outlined by World Health Organization (WHO) and the World Organization for Animal Health (OIE) [10, 11], the international standards, guidelines and recommendations in an event of an outbreak state that member countries are obliged to notify within 24 hours epidemiological information with regards to occurrence/reoccurrence of listed notifiable diseases, the occurrence of a new strain of a listed disease, a significant change in the epidemiology of a listed disease, or the detection of an emerging disease.

The disease outbreak reports within specific time limits are required to be sent on the presence or the evolution of the listed diseases and their strains. It is apparent that the increasing threat of disease outbreak highlights the need to provide timely and accurate information to public health professionals across many jurisdictional and organizational boundaries. Also, the increasing frequency of biological crises, both accidental and intentional, further illustrates that Disease Outbreak Notification System (DONS) needs to be in place to meet the challenges facing today's society. Such DONS should prevent, prepare, and respond to an outbreak having the potential to affect humans and/or animals. The surveillance and management roles and responsibilities should be identified for a unified approach that considers humans, domestic animals, and wildlife.

The importance of such a system to Kingdom of Saudi Arabia (KSA) is tremendous. It is well known that the KSA is a vital hub for two major events: (a) around 2.5 million pilgrims from more than 160 countries take part in the Hajj in the holy city of Makkah every year during a very short time spanning only four weeks, and (b) around six million Muslims perform Umrah every year. These two events make KSA a fertile place for outbreaks as people fly into KSA from overseas every year. As such, a contagious disease outbreak overseas such as MERS, H1N1 or H5N1, whether natural or due to bioterrorism can spread long before an epidemic is recognized. This will not only jeopardize people's health but also impact worldwide because pilgrims and those performing religious duties will eventually return to their countries and can spread the disease worldwide.

This paper, as stated previously, focuses on the design and prototype implementation of a DONS for KSA. The rest of the paper is organized as follows. Section 2 discusses the existing methodology of disease outbreak monitoring in Saudi Arabia. Section 3 describes the proposed architecture of the system. Section 4 provides implementation detail and results from the prototype system implemented. Finally, Section 5 outlines the conclusions reached from our work.

2. THE CURRENT SYSTEM IN KINGDOM OF SAUDI ARABIA (KSA)

The Saudi Ministry of Health (MoH), by way of its objectives, policies and projects seeks to deliver the best-quality integrated and comprehensive healthcare services throughout the Kingdom. According to the MoH website (http://www.moh.gov.sa/), the MoH is committed to the provision of healthcare at all levels, promotion of general health and prevention of diseases. MoH is also accountable for performance monitoring in health institutions, along with research activity and academic training in the field of health investment. In this section, we analyze and present the processes and data flow of the current health systems for infectious diseases.

The process of analysis of the current system started in this project two years back in March 2012. The MoH is presently in the process of implementing a cloud-based health system with IBM which was initiated in May 2013 [55]. The MoH plans to fully integrate to the new cloud-based system for reporting diseases in three to five years. This cloud-based implementation at

MoH aligns with our vision as proposed in our work and outlined in the design and architecture of KSA DONS in this paper.

2.1 PROCESS AND DATA FLOW OF THE CURRENT SYSTEM

The current system is a manual system that consists of four main entities, namely, the Local Healthcare Unit (LHU), the Primary Healthcare Center (PHC), the Directorate of Health Affairs (DHA), and the Ministry of Health (MoH) [12]. A LHU can be a hospital belonging to MoH or other government sectors, private hospitals, private dispensaries, and private clinics. The main tasks of a LHU include: {a} discover new cases and notify PHC about its discovery; {b} isolate the cases of infectious diseases; {c}confirm the diagnosis; {d} conduct treatment; {e} conduct epidemiological survey; and {f}record disease status [12].

The main tasks of a PHC include: {a} epidemiological surveillance of infectious diseases in the affiliated areas of health care centers and work to limit and contain epidemics; {b} review forms survey epidemiological and procedures; {c} follow-up cases that are subject to treatment and provide feedback; {d} take the necessary preventive measures;{e} data collection, conducting statistics, and reporting the cases and mortality rates of infectious disease; and {f} notify the regional directorate of health affairs [12].

The DHA functions at the regional level and its main tasks include: {a} compile the forms of communicable diseases received from PHCs; {b} monitor, analyze and extrapolate epidemiological data received from LHUs and PHCs to identify trends in disease and detect epidemics; {c} take preventive measures; {d} identify people and places most vulnerable to disease; and {e} report to MoH and send feedback to LHUs and PHCs [12]. The main tasks of the MoH include: {a} development and implementation follow-up of policies and plans for the prevention and control of communicable diseases; {b} monitor and control epidemic infectious diseases in the Kingdom; {c} receive regular and monthly reports of infectious diseases with analysis and extrapolation of data; {d} identify and provide vaccinations to the areas most vulnerable to infectious diseases; {e} monitor epidemiological diseases regionally and globally and develop the necessary policies to prevent its arrival and spread in the Kingdom; and {f} report the diseases that are subject to international health regulations to the World Health Organization (WHO) [12].

The flow of reporting among these entities is shown in Figure 1.

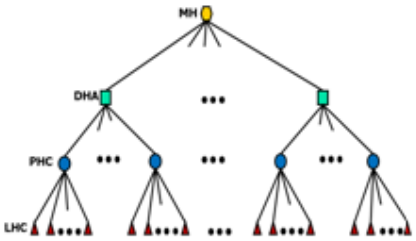


Figure 1 : Information flow among entities of the current KSA System

When a new case is discovered, the Infection Control department collects the data from the LHU where the disease occurred. Then, the data is accumulated at the district level by the PHC followed by the DHA at the province level. Finally, the accumulated data is analyzed and monitored by the preventative sector for communicable diseases at the MoH. This notification process is shown in Figure 2.

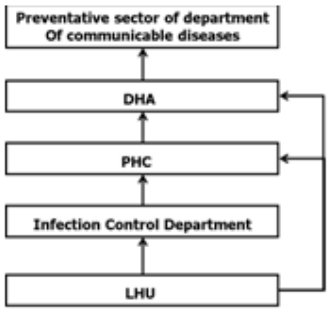


Figure 2 : The notification channel between the 4 entities of the current system

The current system presently handles 35 notifiable diseases [12]. These diseases are classified into 2 classes. The first class consists of quarantinable diseases, such as, Avian influenza, H1N1 (Swine Flu), SARS. Cases of this class of diseases are reported immediately by the LHU or PHU to the corresponding DHA by fax, or telephone. Table 1 lists the diseases of this class.

Table 1 : Quarantinable diseases which are reported immediately

Cholera	Diphtheria	Acute F. Paralysis	Dengue Fever
Yellow fever	Measles	guillain-barré syndrome	Rift V Fever
Plague	Tetanus	myelitis Transverse	Hemmorhagic Fevers Other
Mumps Neonatorum	German measles	Enceph/Mening	

The second class consists of communicable diseases, such as Chicken Pox, Echinococus, Hemolytic Uremic Syndrome, as listed in Table 2. Each case of this class of diseases is reported weekly, by the LHU or the PHU where the disease occurred, to the corresponding DHA. The DHA then reports these cases monthly to the MoH. Each disease is reported with its degree of specificity which is either a suspect (low specificity), or probable, or confirmed (high specificity) [12].

Table 2 : Notifiable diseases which are reported monthly

Tetanus	Hepatitis A	Salmonellosis	Hemolytic Uremic Syndrome
whooping cough	Hepatitis B	Malta Fever	Pneumococcal Meningitis
German Measles in new born	Hepatitis C	Shigellosis	Hemophylus Influenza Meningitis
Typhoid & Para typhoid	Hepatitis Other	Echinococus	Meningitis Other
Amebic dysentery	chicken Pox	Rabies	

The process flow of both classes of diseases is shown in Figure 3. As indicated in the figure, the process flow spans across various entities including the Saudi National Health Authority (NHA) and WHO. The two process flow differ in the frequency of reporting which is immediate in quarantinable while weekly/monthly in communicable diseases.

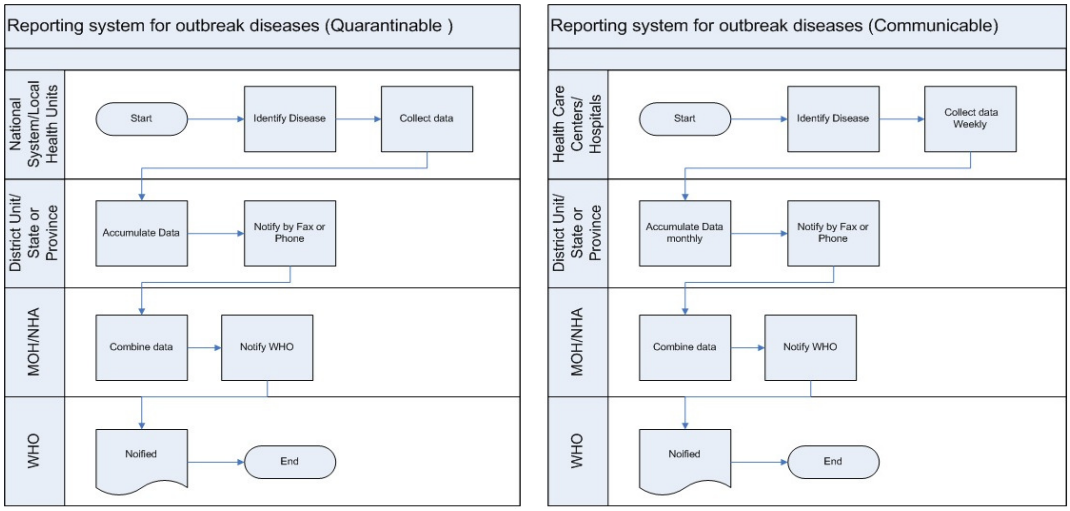


Figure 3 : Process flow for both quarantinable and communicable diseases

2.2 SAUDI HEALTH ELECTRONIC SURVEILLANCE NETWORK (HESN)

The Saudi HESN is a comprehensive, integrated public health information system that helps public health professional work together to efficiently manage individual cases, outbreaks, immunizations, and vaccine inventories. The new cloud-based system will provide public health professionals with tools to better protect the health of the Kingdom's citizens with a secure, easy-to-use application to collect, share and analyze health information critical in managing public health outbreaks such as SARS, influenza and any other communicable diseases [13]. IBM and Saudi Ministry of Health have announced the successful implementation for the first stage in Jeddah, Makkah, Taif and Qunfudah on May 2013 [14]. The program will be in place throughout the entire Kingdom within three to five years.

The goal of the system is to improve public health, as it cuts down the need for paperwork, consolidation of operations and health forms and reports in the entire Kingdom in electronic form, which in turn will increase the accuracy of the data and reduce the difference in monitoring between areas and facilities of the Ministry of Health, and increases the personal and professional skills of the staff.

The HESN system has unique characteristics such as protects public health through the prevention, detection and management of communicable disease occurrences, enables collaboration (interoperability), and follows industry standards (HL7 messaging, SNOMED).
The HESN system is flexible to suit the requirements of health care [15], which vary according to geographical area. The system provides predefined forms and installed on the system and ready to be filled by the users. HESN allows the users to configure a new report form to enter new information about health conditions that are not monitored by the system, and to collect data commensurate with the needs of the region or the program or the situation. The HESN system allows the owners of the powers of the user access to the reports, notification forms and extracts data (Business Object Universes).

The HESN system will not guarantee all units in the public health system to monitor and manage diseases. The major components in the first stage are [15] include communicable disease case management, outbreak management, immunization management, family health materials, vaccine inventory management, notifications and work management.

3. THE PROPOSED KSA SYSTEM

In this section, we present our effort as the proposed Disease Outbreak Notification System (DONS) for the Kingdom of Saudi Arabia (KSA), which we will refer to as KSA DONS. This system is designed to detect potential disease outbreaks within KSA and notifies preregistered experts about the outbreak. In order to come up with a sound design for KSA DONS, our effort was concentrated on the functionalities and features of a typical and generic DONS. Using the taxonomy and critical features generated in that effort, we mapped those features and functionalities to the KSA DONS design.

3.1 CRITICAL FEATURES IN KSA DONS

Based on the taxonomy of various DONS implementation world-wide that we studied, we have identified the critical features that our KSA DONS should have. These features are shown in Tables 3 and 4.

Table 3 : Core features of the proposed KSA DONS

	Collection & Analysis						**Core Functions**							
	Collection				Analysis		Case-Level		Group-Level			Notifications		
System Name	Stakeholders	Data Sources	Collection Strategy	Data Formats	Computation Detection Algorithms	Statistical Analysis	Detection	Confirm & Register	Signal Extraction	Interpretation	Reporting	Communication Procedures	Target Entities	Output/Message Formats
Existing KSA DONS	✓	✓	✓	✓								✓	✓	
Proposed KSA DONS	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 4 : Additional features of the existing KSA DONS

	System Features														**Support Functions & System Interfaces**							
	Attributes														Support Functions					Networking& Partnerships		
System Name	Completeness	Timeliness	Usefulness	Real-time	Coverage	Portability	Sensitivity	Availability	Accessibility	Usability	Interface Design	Mobility	Flexibility	Specificity	Standards	Training	Communication	Evaluation & Monitoring	System Administration	Experts & Institutions	Coordination	Feedback
Existing KSA DONS			✓		✓					✓										✓		
Proposed KSA DONS	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓

4. KSA DONS PROTOTYPE IMPLEMENTATION

A prototype was implemented as a proof-of-concept (PoC). The prototype was a hybrid implementation using a multi-tier architecture spread across physical hosts and the private research cloud infrastructure at KFUPM. The prototype implementation was thoroughly tested for functional and technical performance with a considerably large dataset of diseases and cases. Integration testing between the various modules of KSA DONS was done as well. The prototype KSA DONS is deployed on the KFUPM Cloud Infrastructure service called KLOUD. KFUPM cloud service offers servers and storage as per required specifications from a wide range of available infrastructure templates.

The detection algorithms used in the KSA DONS were selected and adapted after a thorough study of all algorithms from selected DONS systems. We have used five efficient algorithms that are used in CASE system [16]. These five algorithms can detect isolated cases of known diseases and their potential outbreaks. Our preference for these algorithms was based on the fact that the coverage of the list of known diseases included in KSA DONS is handled by these detection algorithms. We have also implemented an outbreak detection algorithm for unknown diseases using data mining techniques. Our implementation uses expert epidemiologists for consultation purposes to confirm outbreaks.

We have adopted a three-tier system architecture which supports features such as scalability, availability, manageability, and resource utilization. Three-tier architecture - consisting of the presentation tier, application or business logic tier and data tier - is an architectural deployment style that describes the separation of functionality into layers with each segment being a tier that can be located on a physically separate entity. They evolved through the component-oriented approach, generally using platform specific methods for communication instead of a message-based approach. This architecture has different usages with different applications. It can be used in web applications and distributed applications. The strength of this approach in particular is when using this architecture for geographically distributed systems. Since our platform is cloud-based and geographically spread across the Kingdom, we were motivated to use this three tier architecture for our prototype implementation.

In our implementation, the methodology used is as follows:

- The presentation tier is through the network using wired and wireless devices.
- The application or business logic tier is entirely hosted on virtual hardware on the cloud platform
- The data tier is hosted on physical servers

The application development tools in implementation of the KSA DONS are JCreator, PHP, and Java. The database development tools used in the implementation of the KSA DONS are MySQL, Oracle SQLPLUS, and Oracle SQL Developer.

The presentation tier is the topmost level of the application. The presentation layer provides the application's user interface. Typically, this involves the use of Graphical User Interface for smart client interaction, and Web based technologies for browser-based interaction. As shown in the Figure 4, the terminals for primary health centres and experts use the DONS browser-based application for data entry and decision making.

The application tier controls an application's functionality by performing detailed processing. The application or the business logic tier is where mission-critical business problems are solved. The components that make up this layer can exist on a server machine, to assist in resource sharing.

These components can be used to enforce business rules, such as business algorithms and data rules, which are designed to keep the data structures consistent within either specific or multiple databases.

Because these middle-tier components are not tied to a specific client, they can be used by all applications and can be moved to different locations, as response time and other rules require. In Figure 4, the web servers and application server constitutes the logic tier. A cluster of web servers and applications servers can be used for load balancing and failover among the cluster nodes.

The data tier consisting of database servers is the actual DBMS access layer. It can be accessed through the business services layer and on occasion by the user services layer. Here information is stored and retrieved. This tier keeps data neutral and independent from application servers or business logic. Giving data its own tier also improves scalability and performance. In Fig. 2, the database servers and the shared storage constitute this tier.

The flow of data in the three tiered architecture is described next. In the presentation layer, the users access the DONS applications over the network through the web browser. The request is securely sent over the network to a firewall. Then the trusted requests from the firewall are forwarded to load balancers. The firewall ensures that trust relationships between the presentation and application tiers are complied with. The trusted requests are sent to a DNS server for name resolution and to load balancers which are capable of distributing the load across the web servers and manage the network traffic.

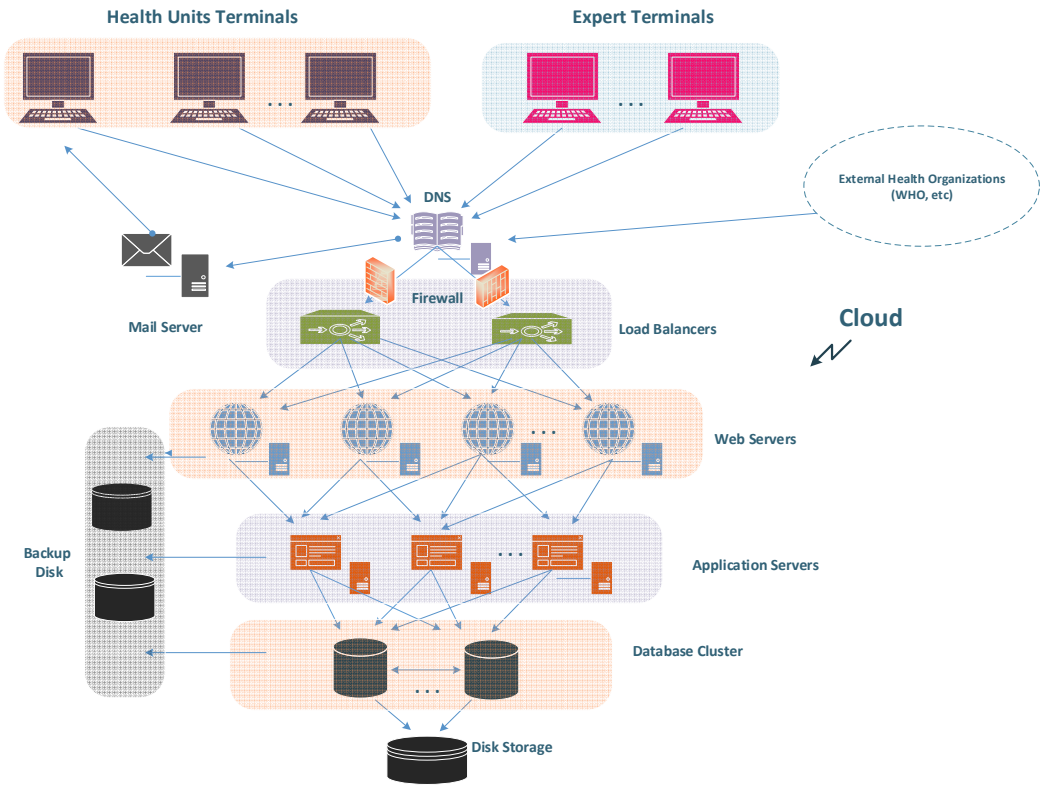


Figure 4 : KSA DONS Prototype System Architecture

In the logic tier of the KSA DONS, web and application servers are deployed to handle all user requests. While the user requests (http or https) are served by the Apache web servers, the

application servers handle all the business logic processing and data processing. The data tier provides all the data needed for the logic tier through SQL queries using database specific protocol over TCP/IP. The database tables are maintained by insertion, updating and deleting of data. To avoid loss of data due to data corruption or system failure, the data backup of all the critical servers and databases is performed periodically in a separate storage location. The data retention policy is applied for timely recovery of data in case of any disaster.

4.1 PROTOTYPE APPLICATION TIER

The prototype implementation of the generic KSA DONS architecture, described in the previous section, utilizes the KFUPM research cloud KLOUD and associated hardware and software tools. In this section we describe all the entities, software and tools used in this tier. A prototype is shown in the Figure 5 and contains the following entities: Health Unit 1, Health Unit 2, Eastern Province Health Department and MoH (Ministry of Heath), Database System, Experts and Users. Health Unit 1 and Health Unit 2 are physical units with Windows 2008 Server operating system (DELL OptiPlex 9010 server). Eastern Province Health Department client is a KLOUD virtual machine with Red Hat Linux 6.4 as its operating system. Ministry of Health (MOH) is also a KLOUD virtual machine with Windows Server 2008 server on it. The database server is again a physical server with Red Hat Linux 6.3 operating system. The Oracle client software was configured on all the servers and clients mentioned in the above figure in order to communicate with each other and to connect to the database server. The web services offered to the clients are configured on an Apache web server on MoH KLOUD virtual server for accessing the application tier of KSA DONS.

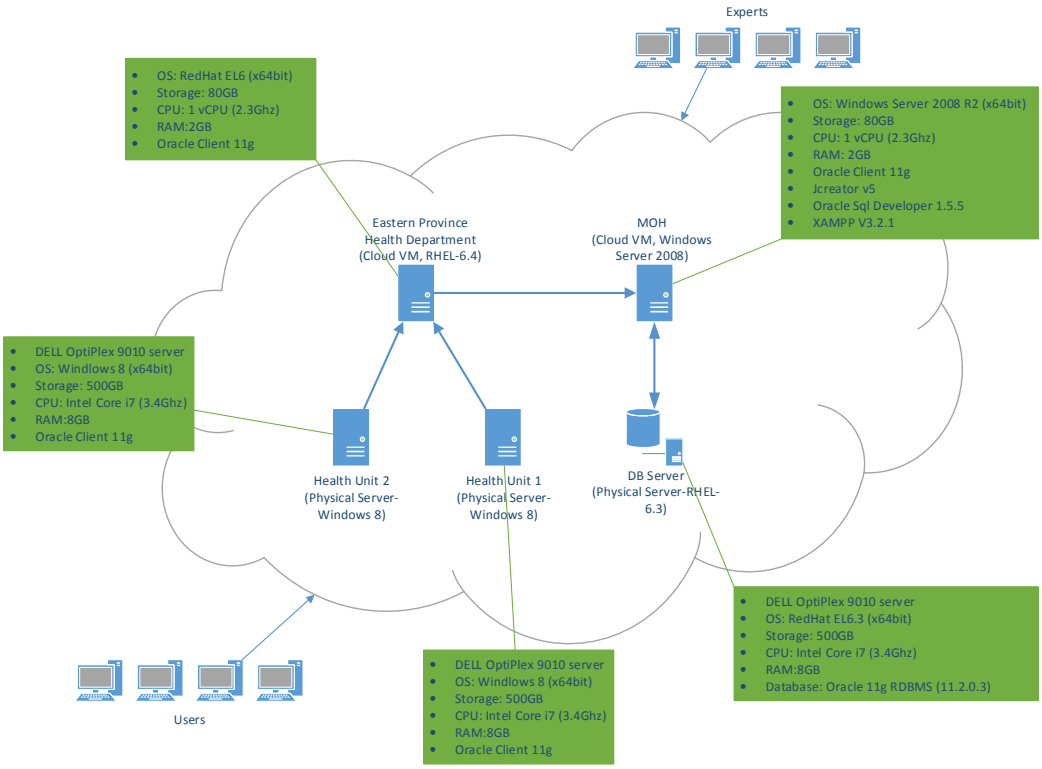


Figure 5 : Prototype System Architecture for the Application Tier

In many disease notification systems, the users had to be manually notified of any new outbreak. In the KSA DONS prototype system, we have developed an application to automate this notification procedure.

Figure 6 shows the data flow of notification delivery system in KSA DONS.

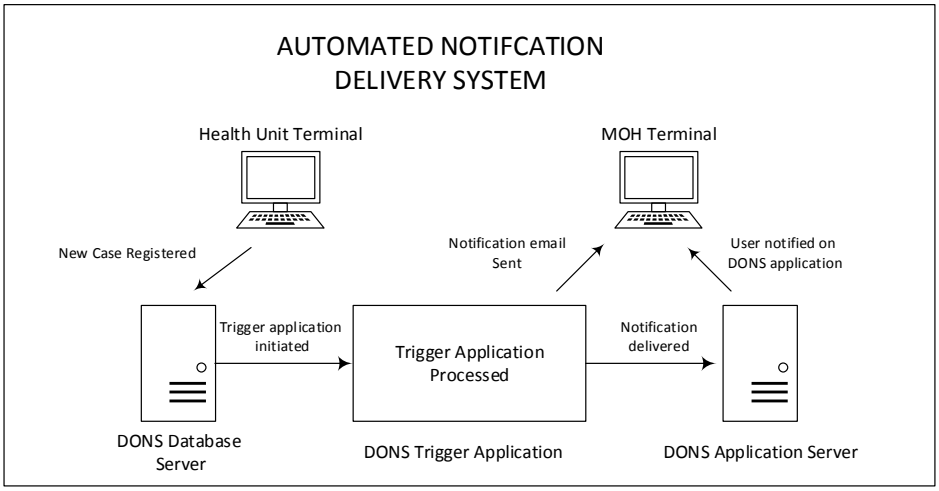


Figure 6 : Data flow in the Notification Delivery System in KSA DONS

Whenever a new disease is registered by the Health Units in the DONS database, a trigger defined in the database runs the trigger application. This application is responsible for delivering the disease notification to the DONS application server over the network, which in turn notifies the appropriate/responsible user in MOH. The trigger application also sends the notification to the email box of the users. The technical details of this implementation of the data flow of notification delivery system in KSA DONS are shown in
Figure 7 .

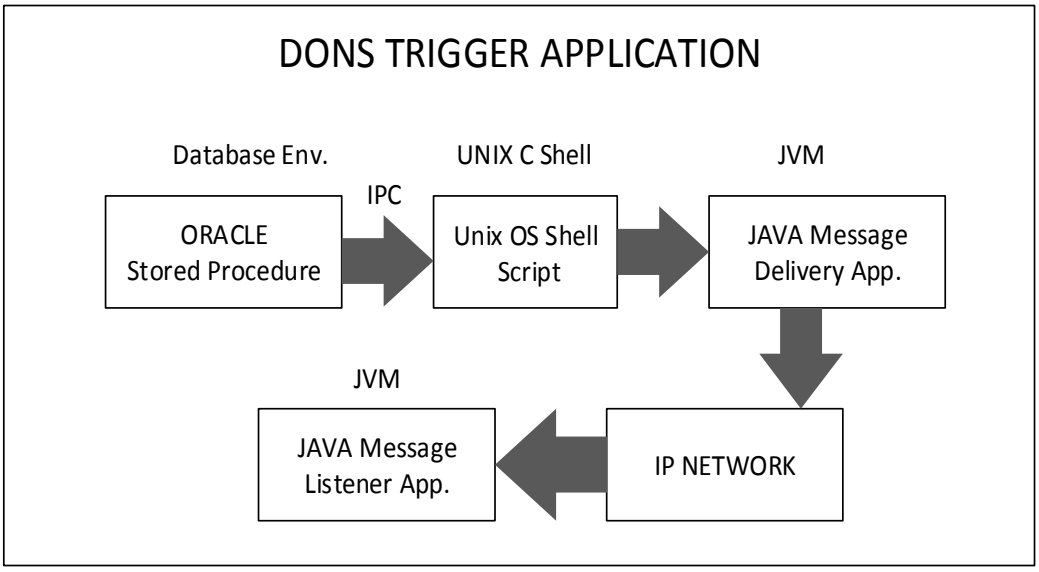


Figure 7 : Technical Implementation of the Notification Delivery System in KSA DONS

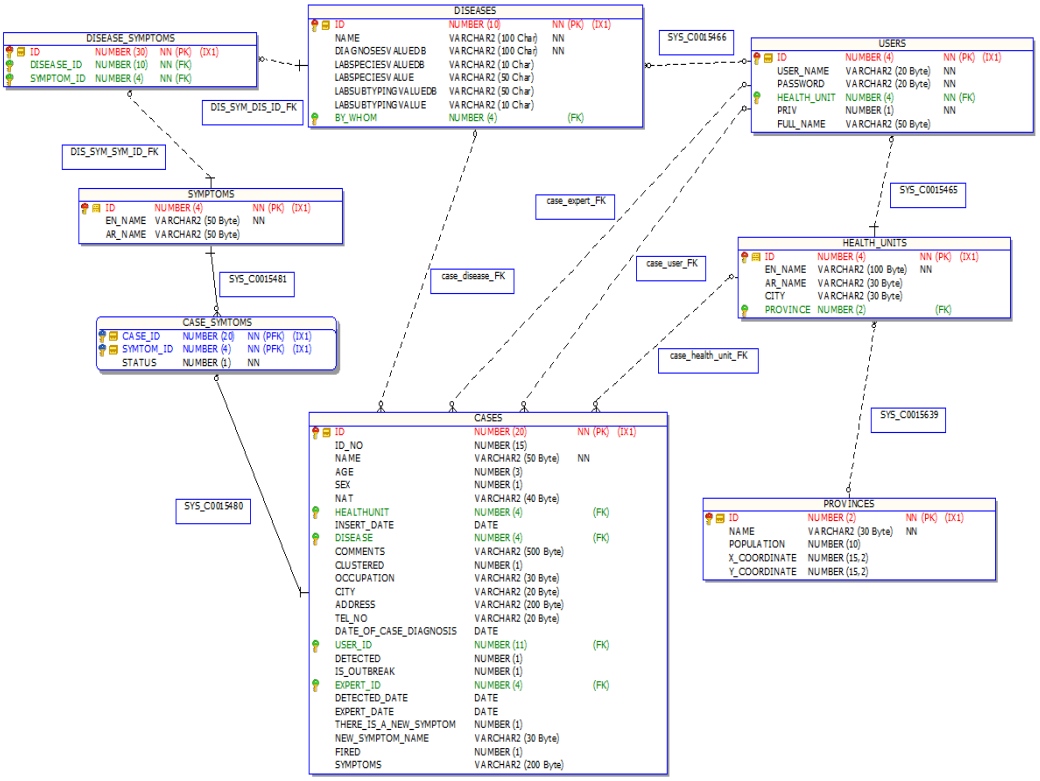
The entire process implemented in the trigger application shown in
Figure 7 is described here. A new disease case is registered in the database by inserting a row of data in a CASE table. A trigger is defined in the DONS database which executes a stored

procedure whenever a row of data is inserted in the CASE table. The stored procedure from the database environment calls a shell script present in the operating system (OS) environment through an Inter Process Communication (IPC).

This shell script is a batch program which can perform multiple tasks sequentially. The shell script runs a Java-based message delivery application within a JVM and also sends the notification e-mail to the end users. The message delivery application sends the notification message over the network to DONS application residing on the application server. At the application server, a message listener application continuously monitors an application port for notification message reception.

4.2 PROTOTYPE DATA TIER

The data tier implemented in the KSA DONS prototype implementation went through two phases of development. Initially, the open database development platform MySQL was used in the prototype development. At a later stage, the entire database with all the associated objects such as the database schema, stored procedures, triggers etc., were migrated to the Oracle database platform. The database server is configured with Oracle Version 11gR2 relational DBMS and was created database using DBCA utility. The data from MySQL database was migrated to the Oracle database using SQL Developer utility offered by Oracle. The complete KSA DONS database schema on the Oracle platform is shown in Figure 8.



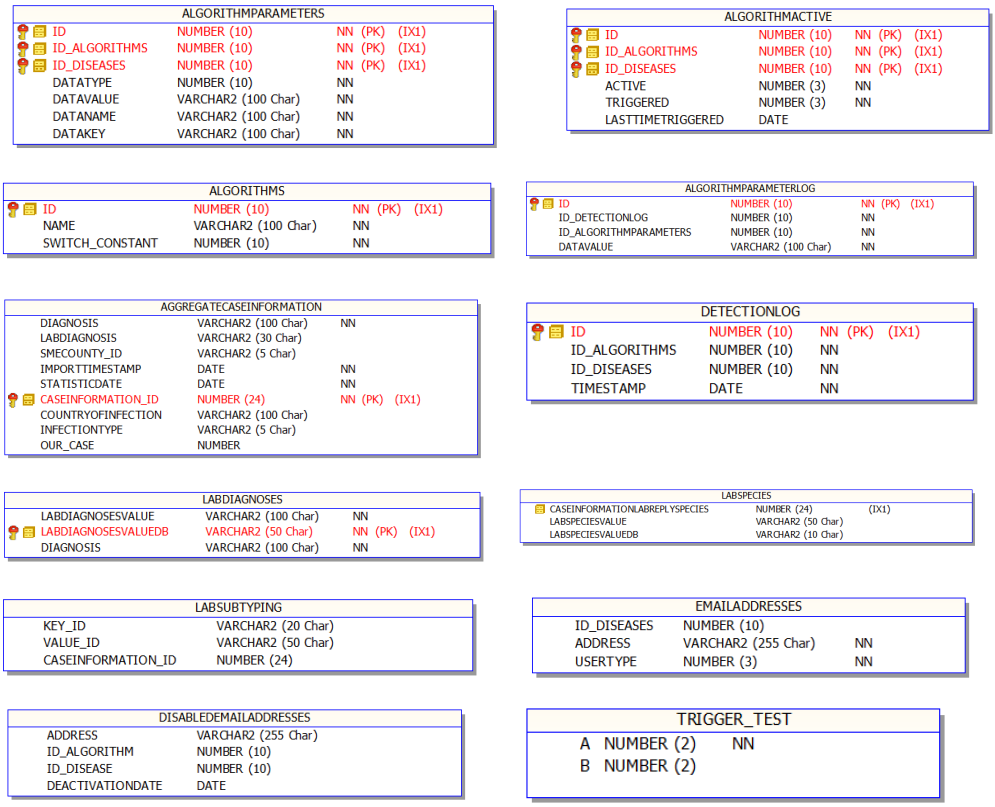


Figure 8 : Complete KSA DONS Database schema on the Oracle platform

5. CONCLUSIONS

This paper describes the design and development of Disease Outbreak Notification System (DONS) in Saudi Arabia. The main function of DONS is to warn for potential outbreaks. The KSA DONS is an online/real-time disease outbreak notification system built for Saudi Arabia. The system notifies experts of potential disease outbreaks of both pre-listed diseases and totally unknown diseases. The system only accepts cases from pre-registered sources. It also shares information about disease outbreaks with international systems. As soon as the system detects a potential disease outbreak it notifies stakeholders and experts. The system takes feedback from experts to improve its disease detection capabilities and to adapt to new situations. A prototype implementation in a hybrid cloud environment was completed to validate the design of the system.

ACKNOWLEDGMENT

The authors would like to acknowledge the support provided by the Deanship of Scientific Research at King Fahd University of Petroleum & Minerals (KFUPM). This project is funded by King Abdulaziz City for Science and Technology (KACST) under the National Science, Technology, and Innovation Plan (project number 11-INF1657-04).

REFERENCES

[1] C. Leung, M. Ho, and A. Kiss, "Homelessness and the response to emerging infectious disease outbreaks: lessons from SARS," J. Urban Heal., vol. 85, no. 3, pp. 402–410, 2008.

[2] M. Muller and S. Richardson, "Early diagnosis of SARS: lessons from the Toronto SARS outbreak," Eur. J. Clin. Microbiol. Infect. Dis., vol. 25, no. 4, pp. 230–237, 2006.

[3] M. Donohoe, "Internists, epidemics, outbreaks, and bioterrorist attacks," J. Gen. Intern. Med., vol. 22, no. 9, p. 1380, 2007.

[4] W. Slenczka, "Mad cow disease.," Emerg. Infect. Dis., vol. 7, no. 3 Suppl, p. 605, 2001.

[5] N. Cohen, J. Morita, D. Plate, and R. Jones, "Control of an outbreak due to an adamantane-resistant strain of influenza A (H3N2) in a chronic care facility," Infection, 2008.

[6] Guidelines for Outbreak Prevention , Control and Management in Acute Care and Facility Living Sites. Alberta Health Services, 2013, pp. 1–54.

[7] G. Cruz-Pacheco, L. Esteva, and C. Vargas, "Seasonality and outbreaks in West Nile virus infection.," Bull. Math. Biol., vol. 71, no. 6, pp. 1378–93, Aug. 2009.

[8] D. Tam, S. Lee, and S. Lee, "Impact of SARS on avian influenza preparedness in healthcare workers," Infection, vol. 3, no. 5, pp. 320–325, 2007.

[9] R. M. Roop II, J. M. Gaines, E. S. Anderson, C. C. Caswell, and D. W. Martin, "Survival of the fittest: how Brucella strains adapt to their intracellular niche in the host," Med. Microbiol. Immunol., vol. 198, no. 4, pp. 221–238, 2009.

[10] "WHO | Influenza at the Human-Animal Interface (HAI)." [Online]. Available: http://www.who.int/influenza/human_animal_interface/en/. [Accessed: 11-Feb-2014].

[11] "Prevention and control of animal diseases worldwide. Economic analysis- Prevention versus outbreak costs. Final Report. Part 1, September 2007." [Online]. Available: http://www.oie.int/fileadmin/Home/eng/Support_to_OIE_Members/docs/pdf/OIE_-_Cost-Benefit_Analysis__Part_I_.pdf.

[12] "Guidelines for epidemiological surveillance and preventive measures for infectious diseases." [Online]. Available: http://qh.gov.sa/uploads/files/.

[13] "Health Electronic Surveillance Network." [Online]. Available: http://www.moh.gov.sa/Hesn/Pages/default.aspx.

[14] "IBM, Saudi Health Ministry rolls out cloud-based system.," May-2013. [Online]. Available: http://www.arabnews.com/news/450055.

[15] "HESN FAQ LEAFLET.," Mar-2013. [Online]. Available: http://www.moh.gov.sa/Hesn/Versions/Documents/HESN\ FAQ\ LEAFLET\ -\ ARABIC.pdf.

[16] B. Cakici, K. Hebing, M. Grünewald, P. Saretok, and A. Hulth, "CASE: a framework for computer supported outbreak detection," BMC Med. Inform. Decis. Mak., vol. 10, no. 1, p. 14, 2010.

AUTHORS

Farag Azzedin is an associate professor at the Department of Information and Computer Science, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia. He received his BSc degree in Computer Science from the University of Victoria, Canada. He received an MSc degree as well as a PhD degree in Computer Science from the Computer Science Department at the University of Manitoba, Canada.

Salahadin Mohammed is an assistant professor at the Department of Information and Computer Science, King Fahd University of Petroleum and Minerals (KFUPM). He received his BS and MS degrees in Computer Science from KFUPM. He received a PhD degree in Computer Science from the Computer Science Department at the Monash University, Melbourne, Australia.

Jaweed Yazdani is a faculty member at the Department of Information and Computer Science and the manager of Administrative Information Systems (ADIS) at King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia. He received his MS degree and BS degrees in Computer Science from KFUPM.

Mustafa Ghaleb earned his BS in computer science at King Khalid University (KKU), Abha, KSA in 2007. At the moment, he is pursuing his MS degree in Information & Computer Sciences at King Fahd University of Petroleum & Minerals (KFUPM).

APPLICATION OF ENHANCED CLUSTERING TECHNIQUE USING SIMILARITY MEASURE FOR MARKET SEGMENTATION

M M Kodabagi[1], Savita S Hanji[2], Sanjay V Hanji[3]

[1,2]Department of Computer Science and Engineering,
Basaveshwar Engineering College, Bagalkot.
malik123_mk@rediffmail.com
savitawali@gmail.com
[3]Department of Management Studies, Basaveshwar Engineering College,
Bagalkot.
sanjayhanji_94@rediffmail.com

ABSTRACT

Segmentation is one of the very important strategic tools used by the marketer. Segmentation strategy is based on the concept that no firm can satisfy all needs of one customer or one need of all the customers. The customers are too numerous and diverse in their buying requirements, hence the marketers or companies cannot cater to the requirements of all customers that too in a broad market such as two-wheelers. Cluster analysis is a class of techniques used to identify the group of customers with similar behaviors given a large database of customer data containing their properties and past buying records. Clustering is one of the unsupervised learning method in which a set of data points are separated into uniform groups. The k-means is one of the most widely used clustering techniques used for various applications. The main drawback of original k-means clustering algorithm is dead centers. Dead centers are centers that have no associated data points. The original k-means clustering algorithm with Euclidian distance treats all features equally and does not accurately reflect the similarity among data points. In this paper, an attempt has been made to apply enhanced clustering algorithm which uses similarity measure for clustering (segmentation) of two-wheeler market data. The enhanced clustering algorithm works in two phases; Seed Point Selection and Clustering. The method adapts new strategy to cluster data points more efficiently and accurately, and also avoids dead centers. The enhanced clustering algorithm is found to be efficient in meaningful segmentation of two-wheeler market data. The results of market segmentation are discussed.

KEYWORDS

Enhanced Clustering, Market Segmentation, Two-wheelers, Similarity Measures, Seed Point Selection.

1. INTRODUCTION

In the last few years, the Indian two-wheeler industry has seen spectacular growth. The country stands next to China and Japan in terms of production and sales respectively. Majority of Indians, especially the youngsters prefer motorbikes rather than cars. Capturing a large share in the two-wheeler industry, bikes and scooters cover a major segment. Bikes are considered to be the favorite among the youth generation, as they help in easy commutation. Large varieties of two

Natarajan Meghanathan et al. (Eds) : ACSIT, FCST, ITCA, SE, ICITE, SIPM, CMIT - 2014
pp. 15–27, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4602

wheelers are available in the market, known for their latest technology and enhanced mileage. Indian bikes, scooters and mopeds represent style and class for both men and women in India [1].

The domestic two-wheeler industry recorded sales volumes of 13.8 million units in 2012-13, a growth of 2.9% over the previous year. India's per capita real GDP growth at 8.6% (CAGR) over the six year period 2005-2011 had contributed substantially towards raising the standard of living of households, which in turn had been one of the key drivers of growth for the country's automobile industry [2].

The general downtrend in the economy appears to have no impact on the two-wheeler industry with many manufacturers reporting robust sales numbers for December 2013 [3].

The growing middle class population, prosperous rural India and the rarity of reliable public transport system is leading to a large number of two wheelers added to the roads every day. Indian roads in most cities, villages and towns are narrow. Two-wheelers allow people to navigate such roads easily. Fuel-efficiency is a huge advantage. With the cost of petrol increasing steadily, two-wheeler makes the daily travel both affordable and convenient. Easy availability of auto finances at attractive schemes has made a two-wheeler a must in most urban and rural homes [4].

A company cannot service all customers in a broad market such as two-wheelers. The customer are too numerous and diverse in their buying requirements. A company needs to identify the market segments which it can serve effectively. Many companies are embracing target marketing. Here sellers distinguish the major market segments, target one or more of these segments, and develop products and marketing programs tailored to each. Instead of scattering their marketing effort, they focus on the buyers they have the greatest chance of satisfying. A market segment consists of a group of customers who share a similar set of needs and wants. The marketer's task is to identify the segments and decide which one(s) to target. The company can create more fine-tuned product or service offering and price it appropriately for the target segment. The company can more easily select the best distribution and communications channels, and it will also have a clearer picture of its competitors, which are the companies going after the same segment [5].

Cluster analysis is a class of techniques used to classify objects into relatively homogeneous groups called *Clusters* [6]. Clustering is the process of organizing objects into groups whose members are similar in some way. Clustering can be applied in many fields: Marketing (finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records), Biology (classification of plants and animals given their features), and Libraries (book ordering), Insurance (identifying groups of motor insurance policy holders with a high average claim cost, identifying frauds), city planning (identifying groups of houses according to their house type, value and geographical locations), Earthquake studies (clustering observed earthquake epicenters to identify dangerous zones), WWW (Classification; Clustering weblog data to discover groups of similar access patterns).

The notion of what constitutes a good cluster depends on the application and there are many methods for finding clusters subject to various criteria, both ad hoc and systematic. These include approaches based on splitting and merging such as ISODATA, randomized approaches such as CLARA, CLARANS, and methods based on neural nets, and methods designed to scale large databases, including DBSCAN, BIRCH, and ScaleKM.

Among clustering formulations the most widely used and studied is *k-means* clustering. Given a set of n data points in real d-dimensional space, and an integer k, the problem is to determine a set of k points in d-dimensional space, called centers, so as to minimize the mean squared distance

from each data point to its nearest center. This measure is often called the squared-error distortion and this type of clustering falls into the general category of variance based clustering.

As with any technology there are challenges while clustering of data. The clustering algorithm works on the assumption that the initial centers are provided. The search for the final clusters or centers starts from these initial centers. Without the proper initialization the algorithm may generate a set of poor final centers and this problem can become serious if the data are clustered using an on-line *k-means* clustering algorithm. In general, there are three basic problems that normally arise during clustering namely dead centers, local minima and centers redundancy [7].
Dead centers are centers that have no members or associated data. These centers are normally located between two active centers or outside the data range. The problem may arise due to bad initial centers, possibly because the centers have been initialised too far away from data. Therefore, it is a good idea to select the initial centers randomly from the data or to set them to some random values within the data range. However, this does not guarantee that all the centers are equally active. Some centers may have too many members and be frequently updated during the clustering process whereas some other centers may have only a few members and are hardly ever updated.

Hence in the current work, a new clustering algorithm [8] that uses similarity measure for clustering of data is used to cluster two-wheeler market data. The enhanced clustering algorithm works in two phases; Seed Point Selection and Clustering. The method adapts new strategy to cluster data points more efficiently and accurately, and also avoids dead centers. The results of clustered data is analysed and found to be more accurate.

2. LITERATURE REVIEW

Some of the related works on clustering and its applications are summarised in the following;
The distinguishing feature to demonstrate a practical procedure for conducting hybrid market segmentation is presented in [9]. In this hybrid segmentation is used as a combination of demographic, psychological, psychographic, socio-cultural, and benefit sought from the product segmentation types.

A system to analyze the performance of students using k-means clustering algorithm coupled with deterministic model is proposed in [10]. The result of analysis will assist the academic planners in evaluating the performance of students during specific semester and steps that need to be taken to improve students' performance from next batch onwards.

A hybrid procedure based on Decision Tree of Data mining method and Data Clustering that enables academicians to predict student's GPA is presented in [11]. Based on predicted students GPA instructor can take necessary step to improve student academic performance.

A hybrid procedure based on Neural Networks (NN) and Data Clustering that enables academicians to predict students' GPA is proposed in [12]. This procedure predicts students' GPA according to their foreign language performance at a first stage then classifies the student in a well-defined cluster for further advising and follows up by forming a new system entry. This procedure has mainly a twofold objective. It allows meticulous advising during registration and thus, helps maintain high retention rate, acceptable GPA and grant management. Additionally, it endows instructors an anticipated estimation of their students' capabilities during team forming and in-class participation. The results demonstrated a high level of accuracy and efficiency in identifying slow, moderate and fast learners and in endowing advisors as well as instructors an efficient tool in tackling this specific aspect of the learners' academic standards and path.

A system for analyzing students' results based on cluster analysis and uses standard statistical algorithms to arrange their scores data according to the level of their performance is described in [13]. The system also implemented k-mean clustering algorithm for analyzing students' result data. The model was combined with the deterministic model to analyze the students' results of a private Institution in Nigeria which is a good benchmark to monitor the progression of academic performance of students in higher Institution for the purpose of making an effective decision by the academic planners.

Data clustering technique named k-means clustering is applied to analyze student's learning behaviour is presented in [14]. The student's evaluation factor like class quizzes, mid and final exam assignment are studied. It is recommended that all these correlated information should be conveyed to the class advisor before the conduction of final exam. This study will help the teachers to reduce the drop out ratio to a significant level and improve the performance of students.

A heuristic method to find better initial centroids as well as more accurate clusters with less computational time is proposed in [15]. Experimental results show that the proposed algorithm generates clusters with better accuracy thus improve the performance of *k-means* clustering algorithm.

A new method to compute initial cluster centers for *k-means* clustering is presented in [16]. The method is based on an efficient technique for estimating the modes of a distribution. The new method is applied to the *k-means* algorithm. The experimental results show better performance of the proposed method.

K-means clustering algorithm based on coefficient of variation (CV-*k-means*) is proposed in [17]. The CV-*k-means* clustering algorithm uses variation coefficient weight vector to decrease the affects of irrelevant features. The experimental results show that the proposed algorithm can generate better clustering results than *k-means* algorithm do.

The empirical study to provide up-to-date assessment of cluster analysis application in marketing research and to examine the extent to which some of the ubiquitous problems associated with its usage have been addressed by marketing researchers is presented in [18]. Therefore, more than 200 journal articles published since 2000 in which cluster analysis was empirically used in a marketing research setting were analyzed.

A study in revealing typical patterns of data driven segmentation is presented in [19]. It also provides a critical analysis of emerged standards and suggests improvements.

A generalize discriminative clustering to structure and complex output variables that can be represented as graphical models is presented in [20].

It is observed from the literature survey that existing clustering algorithms based on *k-means* use kd-tree data structure. The most obvious source of inefficiency in these algorithms is that they pass no information from one stage to the next. Presumably, in the later stages, as the centers are converging to their final positions, one would expect that the vast majority of the data points have the same closest center from one stage to the next. These algorithms perform badly with increases in dimensionality. This is because the most important data structure used in the algorithm, the kd-tree, does not scale well with increases in dimension.

A good algorithm should exploit coherence to improve the running time and has to repeat number of times producing different results in different independent runs. The method should be scalable

and can be coupled with a scalable clustering algorithm to address the large-scale clustering problems as in data mining.

The standard *k-means* clustering algorithm [21] generates null clusters and also it treats all features equally and does not reflect the similarity among data. It does not maintain specified gap between cluster centroids. Hence, the current work a new enhanced *k-means* clustering algorithm [19] is taken for segmenting two-wheeler market data.

3. METHODOLOGY FOR TWO-WHEELER MARKET DATA COLLECTION

The demographic and customer need based variables were taken for the study. These variables were used to segment the two-wheeler market on the bases of customer demographics and their needs. The sample size involved was 200 respondents. Simple random sampling technique was employed for selecting the samples. The data collection tool was a well structured questionnaire which followed after several pre-tests. The questionnaire consisted mix of interval scales and nominal scales. The interval scale was used to collect the customer benefit sought (need based) segmentation data such as style, power, mileage, price, features, and low maintenance. A 5 point Likert-type Scale was used to set the interval data. The nominal scale was used to describe the demographic profiles of respondents such as age, gender, qualification, occupation, marital status, and monthly income. Data was collected from northern districts of Karnataka, India. The two wheeler market segmentation was done using newly defined similarity measure clustering technique which is discussed below.

4. CLUSTERING TECHNIQUE USING SIMILARITY MEASURE

The enhanced clustering method uses newly defined similarity measure for the purpose of clustering data points in *d*-dimensional space. The methodology basically has two subsystems namely Seed Point Selection and Clustering. The block diagram of proposed model is given in figure 1. The subsystems of method are explained in detail in the following section.

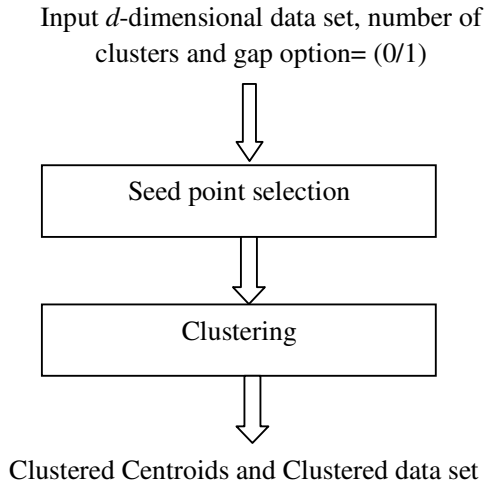


Figure 1. Block Diagram of Enhanced Clustering Method

4. 1. Seed Point Selection

In this phase, seed points i.e. initial cluster centroids are selected based on the value *k* (number of clusters specified by user). The minimum distance between cluster centroids can be specified by user which is optional. If user has specified the minimum distance between cluster centroids, the first *k* data points from the given data set that satisfy following two conditions are taken as seed points:

1. The data points should be unique.
2. The distance between data points should be equal to or greater then minimum distance specified by user.

If user has not specified the minimum distance between cluster centroids, then the first *k* different data points from the given data set are taken as seed points.

The distance between two data points $d_1 (x_1, x_2 \ldots x_n)$ and $d_2 (y_1, y_2 \ldots, y_n)$ in *d*-dimensional space is evaluated using Euclidian Distance as described in equation (1):

$$Euclidian\ distance = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (1)$$

where x_i and y_i are attributes of data points d_1 and d_2 in i^{th} dimension respectively. *n* is number of data points.

Procedure for Seed Point Selection is given in Algorithm1 below:

Algorithm1: Seed Point Selection Algorithm

Input: Data set, *k* value and minimum distance between cluster centroids *dc* (Optional).

Output: Array containing Seed Points.

Step 1: Initialize *i*=1;
Store i^{th} data point from given data set in array containing seed points.

Step 2: Repeat till *k* seed points are selected
Increment *i*;
Select i^{th} data point d_i from given data set;
Compare d_i with all seed points stored in array containing seed points.
if d_i is equal to any of the seed point in array containing seed points
Repeat Step 2;
else if *dc* is specified
for each seed point *s* in array containing seed points
Compute distance *g* between d_i and *s*
if *g* is equal to or greater then *dc*
Store the data point d_i as seed point in array containing seed points.

else

Repeat Step 2

end

end

else

Store the data point d_i as seed point in array containing seed points.

End

STOP

4.2. Clustering

In this phase, clustering of data points is carried out using newly defined similarity measure. Each data point is taken from given data set and a similarity measures is calculated between taken data point and all seed points. This process is repeated for all n data points and the result is stored in $k \times n$ sized matrix which is called Similarity Measure Matrix. Once similarity measure matrix is built, based on the value of similarity measure, each data point is assigned the group. In similarity measure matrix, for each data point similarity measures are calculated with respect to all k cluster centroids. Higher the similarity measure value more closely is the data point associated to the cluster centroid. The data point belongs to that centroid which has maximum similarity measure. Given a set of n data points ($d_1, d_2 \ldots d_n$), where each data point is a d-dimensional real vector, *k-means* clustering aims to partition the n data points into k sets ($k \leq n$). The similarity measure between cluster centroid c ($x_1, x_2 \ldots x_m$) and the data set s ($y_1, y_2 \ldots y_m$) is calculated using formula depicted in equation (2).

$$Similarity\ measure = 1 - \sum_{i=1}^{m} \frac{(x_i - y_i)}{(x_i + y_i)} \qquad (2)$$

Once clustered data matrix is built, the new centroids are evaluated for clustered data. The centroid of m data points i.e. $d_1, d_2, d_3 \ldots\ldots d_m$ each with n dimensions is given using following formula:

$d_{11}, d_{12}, d_{13} \ldots\ldots d_{1n}$ are feature of d_1.

$d_{21}, d_{22}, d_{23} \ldots\ldots d_{2m}$ are feature of d_2 and soon for n data points.

$$Centroid = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} \qquad (3)$$

This process is repeated till convergence is reached. Clustering will reach *convergence point* when clustered data matrix of successive loops are same i.e. data points does not change their cluster in further clustered data matrix evaluation. Suppose G_n is the clustered data at n^{th} iteration and G_{n+1} is the clustered data matrix at n+1 iteration. The convergence data points will be reached when:

$$G_n = G_{n+1} \qquad (4)$$

Algorithm

Procedure for Clustering is given in Algorithm2 below:

Algorithm2: Clustering Algorithm

Description: The proposed clustering algorithm.

Input: The proposed clustering algorithm takes three inputs.

1. Set of *n* data points in *d*-dimensional space.
2. Number of clusters (*k*).
3. Gap option set to either 0 or 1. 1 is set to cluster data points maintaining minimum distance between from clustered centroids and 0 to cluster data points without maintaining minimum distance between clustered centroids.

Output: Centroids of final clusters and clustered data points.

Step 1: Initialize clustered data matrix to zero matrix. If third parameter is set to 1, accept minimum distance between cluster centroids from user.

Step 2: Select Seed Points based on the third parameter. If third parameter is set to 1, then select first *k* data points from given data set that satisfy two conditions described in Section 2.1.1 else directly select first *k* unique data points from given data set.

Step 3: For each data point, calculate the Similarity Measures to all the centroids.

Step 4: Assign each data point to the centroid that has maximum similarity measure (i.e Building clustered data matrix).

Step 5: Check the Convergence. If convergence is reached, go to Step 8 else go to Step 6.

Step 6: Calculate new mean for newly clustered data points to be centroids of clusters. If third parameter is set to 1 then check the newly calculated centroids maintain minimum distance between cluster centroids. If cluster centroids fails to maintain minimum then go to Step 8 else go to Step 7

Step 7: Repeat Step 3 to Step 5.

Step 8: STOP.

5. EXPERIMENTAL RESULTS AND ANALYSIS

The enhanced clustering algorithm was tested for two-wheeler market. It is observed that the enhanced clustering mechanism is efficient when compared to original *k-means* clustering algorithm. Experimental results showed that the enhanced clustering algorithm solved the problem dead centers that were found in original *k*-means clustering algorithm as shown in figure 2 given below;

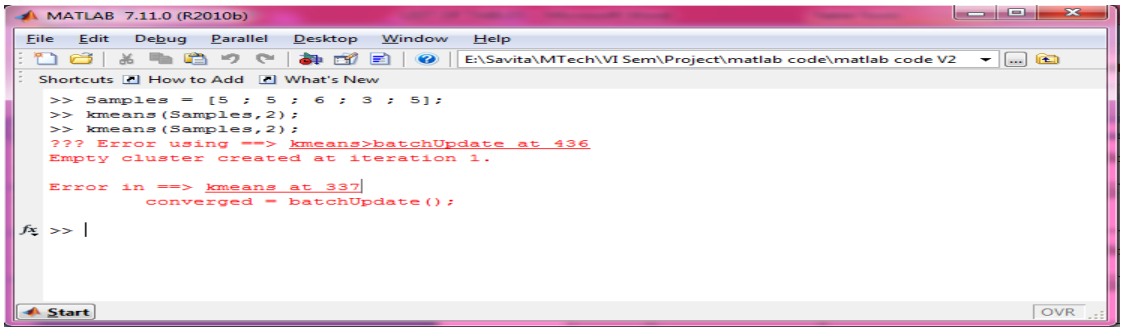


Figure 2: Error in Original k-means clustering algorithm

Table 1. Cluster solutions for the two-wheeler data

Final Cluster Centers					
S.No	**Variables**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
1	Age	1.5	1.2	2.2	1.8
2	Gender	1.2	1.1	1.1	1.2
3	Qualification	1.7	1.4	2.0	1.9
4	Occupation	1.9	2.7	1.5	2.2
5	Marital Status	1.7	1.9	1.2	1.5
6	Monthly Income	1.2	1.1	1.7	1.3
7	Style	4.2	4.5	3.5	2.2
8	Power	4.3	4.2	4.1	1.9
9	Mileage	3.3	3.0	3.9	3.9
10	Price	2.6	3.4	3.6	4.2
11	Features	2.7	4.1	4.0	3.9
12	Maintenance	2.8	4.2	4.4	4.3

The table 1 shows four cluster solutions for the two-wheeler data. The demographic and customer need based variables along with their final cluster centers for each of the clusters (or segments) is shown. After analysing the results, more detailed interpretation of the final cluster centers for each of the clusters is displayed in table 2.

Table 2. Interpretation of Final Cluster Centers

Final Cluster Centers					
S.No	**Variables**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
1	Age	20 to 30 years	20 to 25 years	25 to 35 years	25 to 35 years
2	Gender	Majorly Male	Majorly Male	Majorly Male	Majorly Male
3	Qualification	Many are Graduates & Some are Undergraduates	Both Graduates & Undergraduates	Graduates	Graduates
4	Occupation	Job-Holders	Both Job-Holders & Students	Both Job-Holders & Businessmen	Job-Holders

5	Marital Status	Many Unmarried and Some Married	Unmarried	Many are Married & Some are Unmarried	Both Married & Unmarried
6	Monthly Income	0 to Rs.15000/-	0 to Rs.15000/-	Most have Rs.15000 to Rs.30000 & few have 0 to Rs.15000.	Most have 0 to Rs.15000 & few have Rs.15000 to Rs.30000
7	Need for Style	Agree	Somewhat Strongly Agree	Somewhat Agree	Disagree
8	Need for Power	Agree	Agree	Agree	Disagree
9	Need for Mileage	Somewhat Agree	Neither Agree Nor Disagree	Agree	Agree
10	Need for Low Price	Somewhat Agree	Somewhat Agree	Somewhat Agree	Agree
11	Need for More Features	Somewhat Agree	Agree	Agree	Agree
12	Need for Low Maintenance cost	Somewhat Agree	Agree	Somewhat Strongly Agree	Somewhat Strongly Agree

Cluster 1: The first cluster or the segment consists of people who are between age limits of 20 to 30 years and majority of them are male. Most of the people in this segment are graduates along with few are undergraduates and the segment is solely dominated by job holders. Here many people are unmarried with few are married ones, and their monthly income ranges to a maximum of Rs.15,000. The first thing that these people look for in a two-wheeler is its style and power followed by price, features and maintenance cost. Most of the people in this segment are unmarried young graduates and have a job so they earn their money but with no commitments. Hence these people prefer two-wheelers with style and power, and give little less importance to price of the two-wheelers, their features and maintenance cost.

Cluster 2: The second cluster or the segment comprises of people who are between 20 to 25 years of age; majority are male; same number of graduates and undergraduates are found in this cluster; and it is dominated by both jobholders and students. Mainly unmarried youths constitute this segment with monthly income up to Rs.15000. These people give very high importance to style followed by power, features and maintenance cost. They are not much bothered about the mileage part of the two-wheeler. Since most of these people are unmarried youths who are either undergraduate students with some pocket money and/or fresh graduated jobholders with some basic start-up salaries, these people assign high priorities to style followed by power, features, and low maintenance but don't care much about mileage as they don't seem to have any major commitments.

Cluster 3: The third cluster or the segment includes people who are between 25 to 35 years of age; majority are male; graduates; with both jobholders and businessmen; many married and some unmarried people. Most of these people have their monthly income range between Rs.15000 to Rs.30000 and some have income up to Rs.15000. These people pay more importance to low maintenance cost of the two-wheeler followed by power, mileage, and features. They are not much bothered about initial purchase price of the two-wheeler. This trend in their preferences

is quite obvious because these people are little more elderly than previous segments, graduates with all earning class and most of them are married along with some unmarried people. This segment has more number of people whose earning is up to Rs.30000 per month. Therefore their commitments makes them to be conscious about maintenance cost and mileage of the two-wheeler, their age drives them to look for power and features in the two-wheelers and their income substantiates their buying power, so price of the two-wheeler is not contemplated.

Cluster 4: The fourth cluster or the segment holds people who are between 25 to 35 years of age, majority are males; graduates who are doing jobs, and there are both married and unmarried people. Most of them have their monthly income ranging from 0 to Rs.15000 along with some people who have their income between Rs.15000 to Rs.30000. These people give more weight-age to maintenance cost of the two-wheeler followed by mileage, price and features of the two-wheeler. They completely ignore styling and power of the two-wheeler. These people are graduates who are both married and unmarried people and seem to be more matured when its is compared to other segments. Many don't enjoy huge salaries so they are more concerned about the maintenance cost, mileage, price of the two-wheeler. They also lookout for good features in the two-wheelers but not much bothered about style and power in the two-wheelers as long as they are economical.

Table 3 shows the number of data points which actually belong to each of the four clusters. Out of total 200 data points (respondents), 23 belong to Cluster 1, 66 belong to Cluster 2, 59 belong to Cluster 3 and 52 belong to Cluster 4.

Table 3. Clustered Data Set

Number of data points in each cluster	
Cluster	**Number of data points**
1	23
2	66
3	59
4	52

6. CONCLUSION AND FUTURE SCOPE

In this paper, the two-wheeler market segmentation is carried using enhanced clustering algorithm that uses similarity measure for clustering of data. The enhanced clustering algorithm works in two phases; Seed Point Selection and Clustering. The method adapts new strategy to cluster data points more efficiently and accurately, and also avoids dead centers. The enhanced clustering algorithm is found to be efficient in meaningful segmentation of two-wheeler market data. The two-wheeler market was divided into four segments. The results of market segmentation are discussed. Each of the segments reveal some meaningful information which will be helpful for marketing departments of two-wheeler companies to decide which segments to target for their two-wheeler products and which segments to ignore. In that way, the companies will be in a better position to design proper positioning strategies for their products in the selected segments. These segments will also help the marketers to device their marketing mix strategies i.e., their product, price, place and promotion decisions. Future scope exists to incorporate such new clustering mechanisms to improve efficiency of clustering in various fields.

REFERENCES

[1] Automotive Industry in India, *Auto Guide*, (1996-2014). Viewed on January 21, 2014. http://auto.indiamart.com/two-wheelers.

[2] Indian Two Wheeler Industry, *ICRA Research Services*, (2013). Viewed on December 21, 2013. http://icra.in/Files/ticker/SH-2013-Q2-1-ICRA-Two-Wheeler.pdf.

[3] Two Wheeler Makers Report Robust Sales in December, *The Hindu*, (2014). View on January 10, 2014.http://www.thehindu.com/business/Industry/twowheeler-makers-report-robust-sales-in-december/article5530929.ece.

[4] Bagirathi Iyer, (June 2011), A Study on Reasons for Low Acceptability of Battery Operated 2 Wheelers (E-Bykes) in Nagpur Region with Special Reference to Electrotherm India Limited's Yobykes, *International Journal of Research in Finance & Marketing*, Volume1, Issue2.

[5] Philip Kotler, (2003), *Marketing Management*, 11th Edition, Pearson Eduction, Delhi, India, pp: 278-299.

[6] Naresh K. Malhotra and Satyabhushan Dash, (2007), *Marketing Research- An Applied Orientation,* Fifth Edition, Pearson Education, New Delhi, India, pp. 636-661.

[7] M.N. Tuma, S. Scholz, and R. Decker, (2009), The Application of Cluster Analysis in Marketing Research: A Literature Analysis, *Business Quest Journal*, vol. 14.

[8] M M Kodabagi, Savita S Hanji, Ravatappa A.B, (2013) "This is my paper", A Novel Clustering Technique Using Similarity Measure, *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, Special Issue IDEAS-2013, ISSN: 2278-621X.

[9] H. N. Ramesh, Vinod S., Sanjay Hanji, (Jan-June 2012) "This is my paper", Hybrid Segmentation At Two Wheeler Market In India, *IEMS Jounal of Management Research*, ISSN : 2249-569X, Volume1, Issue-1.

[10] Rakesh Kumar Arora, Dr. Dharmendra Badal, (April - June 2013), Evaluating Student's Performance Using k-Means Clustering, *International Journal of Computer Science And Technology(IJCST),* ISSN(ONLINE) : 0976-8491 ISSN(PRINT) : 2229-4333 Vol. 4, Issue-2.

[11] Hedayetul Islam Shovon, Mahfuza Haque, (2012), An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree, *International Journal of Advanced Computer Science and Applications(IJACSA)*, Vol. 3, No. 8.

[12] Chady El Moucary, Marie Khair, Walid Zakhem, (2011), Improving Student's Performance Using Data Clustering and Neural Networks in Foreign-Language Based Higher Education, *The Research Bulletin of Jordan ACM*, Vol. I (III).

[13] Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C, (2010), Application of k-Means Clustering algorithm for prediction of Students' Academic Performance, International *Journal of Computer Science and Information Security (IJCSIS)*, Vol. 7.

[14] Md. Hedayetul Islam Shovon, Mahfuza Haque, (July 2012), Prediction of Student Academic Performance by an Application of K-Means Clustering Algorithm, International *Journal of Advanced Research in Computer Science and Software Engineering*, ISSN: 2277 128X Vol 2, Issue 7.

[15] Mahmud M.S, Rahman M.M, Akhtar M.N, (2012), Improvement of K-means clustering algorithm with better initial centroids based on weighted average, *Electrical & Computer Engineering (ICECE), 7th International Conference*, Pages:647-650.

[16] Xiaoping Qin, Shijue Zheng, (2009), A New Method for Initialising Clustering Algorithm, *Knowledge Acquisition and Modelin,. KAM '09. Second International Symposium on Nov 30 2009* Volume: 2, Pages:41-44.

[17] Shuhua Ren, Alin Fan, (2011), K-means clustering algorithm based on coefficient of variation, *Image and Signal Processing (CISP), 4th International Congress*, Volume:4, Pages :2076-2079.

[18] M. Y. Mashor, (2000), Hybrid training algorithm for RBF network, *International Journal of The Computer, The Internet and Management*, vol. 8, no. 2, pp. 50-65.

[19] Sara Dolnicar, (2003), Using Cluster analysis for market segment-typical misconceptions, established methodological weakness and some recommendations for improvements. *Australasian Journal of Market Research*.

[20] Peter Haider, Luca Chiarandini, Ulf Brefeld, (2012), *Discriminative for Market Segmentation, KDD'12*, August 12-16, Beijing, China.

[21] Kardi Teknomo, *K-means clustering*, viewed website on June 2013. http://www.planetsourcecode.com/vb/scripts/ShowCode.asp?txtCodeId=26983&lngWId=1.

AUTHORS

M M Kodabagi

Prof M M Kodabagi is currently working as Assistant Professor in Department of Computer Science & Engineering at Basaveshwar Engineering College, Bagalkot. He has more than 10 years of teaching experience in reputed institutes. He has also worked as Scientist 'B' for Electronics Radar Development Establishment, Defence Research & Development Organisation, Bangalore. He has conducted many workshops and also delivered invited lecturers at various engineering colleges. He has published many papers in international journals.

Savita S Hanji

Prof. Savita S. Hanji is working as Assistant Professor in Department of Computer Science & Engineering, Basaveshwar Engineering College, Bagalkot, Karnataka, India. Her areas of interest include Database Management System, Web Developments, Digital Image Processing, Data Mining and Clustering. She has conducted Oracle Certification course and certified many students. She is also certified in Cambridge University for Innovative Teaching Methodologies.

Sanjay V. Hanji

Prof. Sanjay Hanji is working as Associate Professor in Department of Management Studies, Basaveshwar Engineering College, Bagalkot, Karnataka, India. He is a marketing faculty and his subject areas of interest includes marketing research, business statistics, management information systems, consumer behavior, strategic brand management, integrated marketing communications, service marketing, sales and distribution management, and retailing management. He is pursuing his PhD in Business Administration under Visvesvaraya Technological University, Belgaum, Karnataka, India. He has published many papers in national and international journals.

INTENTIONAL BLANK

PARTIAL ORDERS EMBEDDING IS NP-COMPLETE

Dariusz Kalociński

Department of Logic, Institute of Philosophy, University of Warsaw
ul. Krakowskie Przedmieście 3, 00-047 Warsaw, Poland
dariusz.kalocinski@gmail.com

ABSTRACT

Following Barwise, we consider examples of natural language sentences that seem to express that there is an embedding of one partial order into the other. We prove NP-completeness of two versions of partial orders embedding problem. We show that the task of computing the truth value of such sentences in finite models is NP-complete.

KEYWORDS

NP-completeness, partial order, embedding, natural language, computational semantics

1. INTRODUCTION

How to recognize the truth value of natural language sentences? This question may be of interest not only for philosophers, but for computer scientists and engineers as well. We are still far away from realizing the dream of artificial intelligence capable of seamless communication with human beings. Natural language processing is a big challenge. Key questions are (a) how a machine could interpret natural language sentences and (b) compute their truth values? By an interpretation of a sentence we mean assigning a logical form to it. In this paper, we ignore (a) and simply propose some reasonable interpretations for certain natural language sentences. Having a logical form of a sentence, we may approach (b) at least in two ways. One approach is to compute the truth value of a sentence by investigating its inferential meaning, namely its consequences and their logical relations to other, already evaluated, sentences [1]. Another approach is to compute the truth value of a sentence directly in a model. This is the approach we use in this paper. Some interesting results have already been obtained by various authors, see for example [1], [2], [3], [4]. Existing work on the subject clearly indicates that recognizing the truth value of some natural language constructions in finite models is intractable. Our current work supports this view. We consider some interpretations of natural language sentences and show that the problem of recognizing their truth value in finite models is NP-complete. Along the way, we show NP-completeness of two problems concerning embedding of partial orders.

2. VARIATIONS OF "THE…THE…" CONSTRUCTION

In [5] Barwise considers a version of the following natural language sentence:

The richer the country the more powerful are some of its officials. (1)

Natarajan Meghanathan et al. (Eds) : ACSIT, FCST, ITCA, SE, ICITE, SIPM, CMIT - 2014
pp. 29–35, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4603

He observes that (1) seems to express that there is an embedding of one order into the other. Let $\mathrm{A} = (A, >_A)$ denote the set of countries A ordered by richness $>_A$. Let $\mathrm{B} = (B, >_B)$ denote the set of officials B ordered by power $>_B$. According to Barwise, an embedding of A into B is a function $f : A \to B$ such that $\forall x, y \in A(x >_A y \Rightarrow f(x) >_B f(y))$. We use different terminology. Here, functions having the above property are called homomorphisms; embedding is an injective homomorphism that preserves order in both directions: $\forall x, y \in A(x >_A y \Leftrightarrow f(x) >_B f(y))$. Hence – in our terminology – the statement (1) seems to express the fact that there is a homomorphism from A into B.

Now, consider a slightly more complicated example:

The richer the country the more powerful are some of its officials and the more powerful are these officials the richer are countries they represent. (2)

The first conjunct of (2) is the same as (1). Thus, the logical form for (2) starts by saying that there is a function $f : A \to B$ such that $\forall x, y \in A(x >_A y \Rightarrow f(x) >_B f(y))$. If we agree that "these officials" in (2) denotes officials referred to by "some of its officials", namely to elements of the image of f, then it seems that the second conjunct of (2) adds the following condition: $\forall x, y \in A(x >_A y \Leftarrow f(x) >_B f(y))$. Hence, the logical form for (2) reads as follows: there is a function $f : A \to B$ such that $\forall x, y \in A(x >_A y \Leftrightarrow f(x) >_B f(y))$. Observe that f satisfying this condition need not be injective, if we allow the same man to be an official of two countries. However, we may force f to be injective. Consider the following example:

The smarter the student the better are some of her individual presentations and the better are these presentations the smarter are students who performed them. (3)

Here, the syntactical form is exactly the same as in (2). The only difference is that we consider different types of objects and relations. Let A stand for "$\blacksquare$ is a student", P for "$\blacksquare$ is an individual presentation of $\blacksquare$", $>$ for "$\blacksquare$ is smarter than $\blacksquare$" and ϕ for "$\blacksquare$ is better than $\blacksquare$". Observe that since any two students have different individual presentations, any function mapping students to their individual presentations must be injective. (3) could have the following logical form:

$$\exists f \forall x, y[(A(x) \wedge A(y) \wedge x > y) \Leftrightarrow (P(f(x), x) \wedge P(f(y), y) \wedge f(x)\ \phi\ f(y))]. \quad (4)$$

We shall get back to (4) and show that recognizing the truth value of (4) in finite models, where P satisfies certain conditions, is NP-complete.

3. TECHNICAL PART

Definition 1. *Let* $\mathrm{A} = (A, >_A)$ *and* $\mathrm{B} = (B, >_B)$ *be strict partial orders. We say* A *embeds in* B *if there is an injection* $f : A \to B$ *such that*

$$\forall x, y \in A(x >_A y \Leftrightarrow f(x) >_B f(y)). \quad (5)$$

We assume basic knowledge from propositional logic. The reader need to know what are propositional formulae, valuations and satisfiability. To get familiar with the subject, see any introductory book about logic, for example [6].

Now we introduce somewhat specific notions from logic, needed in our complexity analysis. A literal is a sentential variable or a negation of a sentential variable. A clause is an alternative of

literals. A propositional formula is in CNF (conjunctive normal form), if it is a conjunction of clauses. A formula is in 3CNF, if it is a conjunction of clauses each of which consists of exactly three literals.

Definition 2. STISFIABILITY OF 3CNF FORMULAE (3SAT)
Input: a formula φ *in 3CNF.*
Question: is there a valuation satisfying φ *?*

3SAT is an NP-complete problem. For an exhaustive survey of NP-completeness we refer the reader to [7].

The problem of our interest is strict partial orders embedding. We denote it shortly by SPOE.

Definition 3. STRICT PARTIAL ORDERS EMBEDDING (SPOE)

Input: strict partial orders A *and* B.
Question: is there an embedding of A *into* B *?*
Theorem 1. SPOE *is* NP-*complete.*

Proof. First, we prove that SPOE is in NP (this is an easy part). Observe that given strict partially ordered sets A and B and a function $f : A \to B$, checking whether f is an embedding of A into B is in PTIME. Now consider a non-deterministic algorithm with input consisting of strict partially ordered sets A and B. Guess a function $f : A \to B$. Finally, if f is an embedding of A into B, then accept, otherwise – reject. This clearly shows that SPOE is in NP.

We proceed to demonstration that 3SAT is polynomially reducible to SPOE. We describe an algorithm that takes an arbitrary formula φ in 3CNF as input and returns an ordered pair of strict partial orders A_φ and B_φ satisfying the following condition:

$$\forall \varphi \in \mathrm{3CNF}\ (\varphi \in \mathrm{3SAT} \Leftrightarrow \mathrm{A}_\varphi \textit{ embeds in } \mathrm{B}_\varphi). \tag{6}$$

Let φ be an arbitrary formula in 3CNF. φ has the following form:
[
$$\varphi = (a_1 \vee b_1 \vee c_1) \wedge (a_2 \vee b_2 \vee c_2) \wedge \mathrm{K} \wedge (a_k \vee b_k \vee c_k), \tag{7}$$

where k is a natural number of clauses in φ and a_i, b_i, c_i are literals for $i = 1,2,\mathrm{K}\ ,k$.

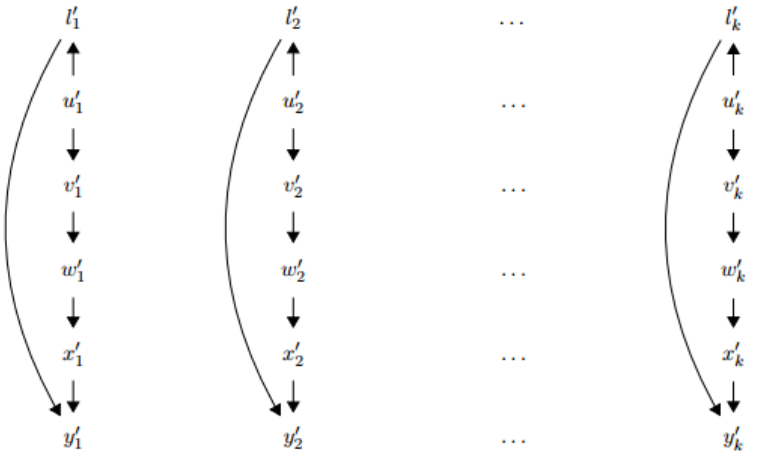


Figure 1. Construction of A_φ.

Construction of $\mathrm{A}_\varphi = (A, <_A)$. Let $A = \{l_1', l_2', \mathrm{K}, l_k'\} \cup \bigcup_{i=1}^{k} \{u_i', v_i', w_i', x_i', y_i'\}$. The strict partial ordering $<_A$ is the transitive closure of the relation presented in Figure 1. We adopt the convention that for any vertices s, t, the relation $s < t$ is graphically represented by an arrow from s to t. Observe that $\mathrm{A}_\varphi = (A, <_A)$ consists of k separate sub-orders. The number k is purposely the same as the number of clauses in φ.

Construction of $B_\varphi = (B, <_B)$. Let $L = \bigcup_{i=1}^{k} \{a_i, b_i, c_i\}$ be the $3k$-element set of all occurrences of literals in φ (different occurrences of the same literal are treated as different). Let $\Gamma = \bigcup_{i=1}^{k} \{u_i, v_i, w_i, x_i, y_i\}$. We put $B = L \cup \Gamma$. We proceed to the construction of $<_B$. At the beginning $<_B$ is empty. Add to $<_B$ all pairs $(s,t) \in B^2$ such that there is a solid arrow from s to t, as indicated in Figure 2. Furthermore, for every l, $l' \in L$, $l <_B l'$ if and only if: l and l' does not occur in the same clause of φ and $\neg l = l'$. This part of the construction is represented by dashed arrows. In this way, we obtain a relation presented in Figure 2. The desired order $<_B$ is the transitive closure of this relation.

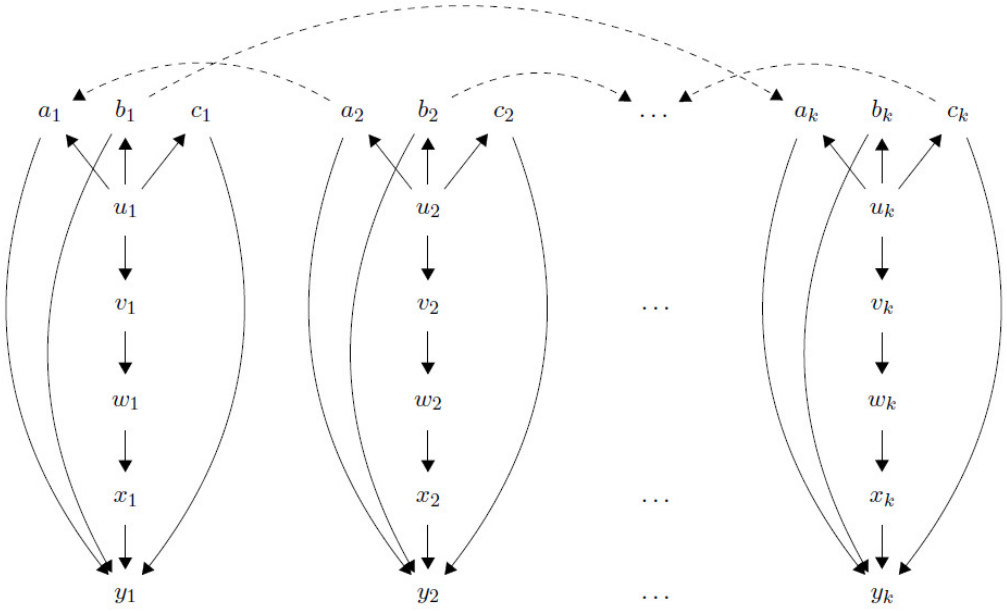


Figure 2. Construction of B_{φ}.

Proof of the equation (6). Let φ be an arbitrary formula in 3CNF with *k* clauses, *k>0.*

($\Rightarrow$) Suppose $\varphi \in$ 3SAT. Let t be a valuation satisfying φ. For each $i = 1,2,\mathrm{K}\,,k$ choose l_i, an occurrence of literal in the *i*-th clause, such that the value of l_i under the valuation t is 1. Let $f : A \to B$ be defined as follows: $f(l_i^{'}) = l_i$, $f(u_i^{'}) = u_i$, $f(v_i^{'}) = v_i$, $f(w_i^{'}) = w_i$, $f(x_i^{'}) = x_i$, $f(y_i^{'}) = y_i$, for $i = 1,2,\mathrm{K}\,,k$. We claim that f is an embedding of A_{φ} into B_{φ}. f is clearly injective. We want to show that $\forall x, y \in A(x >_A y \Leftrightarrow f(x) >_B f(y))$. This is equivalent to the condition that A_{φ} is isomorphic to $(f[A], <_B \big| f[A])$, where $f[A]$ denotes the image of A under the function f and $<_B \big| f[A]$ denotes the restriction of $<_B$ to the set $f[A]$. Observe that for every $1 \le i < j \le k$, neither $\neg l_i = l_j$ nor $\neg l_j = l_i$. For it were the case, then the value of l_i, (l_j) under the valuation t would be 0, which is impossible. Hence, by construction of B_{φ}, no pair of vertices $l_1, l_2, \mathrm{K}\,, l_k$ is joined in B_{φ} by an edge. So A_{φ} and $(f[A], <_B \big| f[A])$ are isomorphic.

($\Leftarrow$) Assume that A_{φ} embeds in B_{φ}. We prove that for each embedding f from A_{φ} to B_{φ} the following conditions hold:

1. For each $i = 1,2,\mathrm{K}\,,k$ there is $j \in \{1,2,\mathrm{K}\,,k\}$ such that $f(l_i^{'}) = l_j$, $f(u_i^{'}) = u_j$, $f(v_i^{'}) = v_j$, $f(w_i^{'}) = w_j$, $f(x_i^{'}) = x_j$, $f(y_i^{'}) = y_j$, where $l_j \in \{a_j, b_j, c_j\}$.

2. For all $i, j \in \{1,2,\mathrm{K}\,,k\}$, if $i \neq j$ then $f(u_i^{'}) \neq f(u_j^{'})$.

Let f be an arbitrary embedding from A_φ to B_φ. To prove the condition 1, note that the only paths of length four in B_φ are $u_j v_j w_j x_j y_j$, for $j = 1,2,\mathrm{K},k$. Hence, for an arbitrary $i \in \{1,2,\mathrm{K},k\}$, it must be that $f(u_i^{'}) = u_j$, $f(v_i^{'}) = v_j$, $f(w_i^{'}) = w_j$, $f(x_i^{'}) = x_j$, $f(y_i^{'}) = y_j$, for some $j \in \{1,2,\mathrm{K},k\}$. Choose such a j. It remains to show that $f(l_i^{'}) \in \{a_j, b_j, c_j\}$. Suppose it is not the case. But then, by construction of B_φ, $f(l_i^{'}) <_B y_j$ does not hold. On the other hand, $l_i^{'} <_A y_i^{'}$ and since f is an embedding, we have $f(l_i^{'}) <_B f(y_i^{'}) = y_j$ which is a contradiction.

To prove the condition 2, let $i, j \in \{1,2,\mathrm{K},k\}$ and assume $i \neq j$. For the sake contradiction, assume $f(u_i^{'}) = f(u_j^{'})$. Since $u_i^{'} <_A v_j^{'}$ does not hold, $f(u_i^{'}) <_B f(v_j^{'})$ does not hold either. However , by condition 1, $f(u_i^{'}) = f(u_j^{'}) <_B f(v_j^{'})$ which is a contradiction.

All in all, every embedding f from A_φ to B_φ chooses a set of k literals $L_f \in \{f(l_1^{'}), f(l_2^{'}),\mathrm{K}, f(l_k^{'})\}$, each literal from a different clause of φ. We claim that L_f is consistent. Observe that, for every $i, j \in \{1,2,\mathrm{K},k\}$, $l_i^{'}$ and $l_j^{'}$ are not connected by an edge and consequently $f(l_i^{'})$ and $f(l_j^{'})$ are not connected by an edge either. This means, by construction of B_φ, that no two elements of L_f are negations of each other. Hence, they all can be made true by an appropriate valuation. This shows that $\varphi \in$ 3SAT.

Complexity. It remains to show that our construction of A_φ and B_φ from an arbitrary 3CNF formula φ is computable in polynomial time in the number of clauses in φ. The construction of the relation from Figure 1 is polynomial in the number of clauses in φ (for each clause we add six vertices and six edges). The construction of the relation from Figure 2 consists of two steps. Initially, for each clause we add eight vertices and ten edges corresponding to solid arrows. Next, we make appropriate interconnections between vertices corresponding to contradictory pairs of occurrences of literals. This can be done by searching through all pairs of occurrences of literals. Hence, the construction of the relation from Figure 2 is polynomial in the number of clauses of φ. Maybe less obvious part is the operation of transitive closure performed as a last step of the construction of A_φ and B_φ. However, given a directed graph (relations presented in Figure 1 and Figure 2 are directed graphs) one can generate its transitive closure using a Floyd-Warshall algorithm [8] which is known to work in polynomial time with respect to the number of nodes.

Definition 4. STRICT PARTIAL ORDERS EMBEDDING IN PARTITION (SPOEP)
Input: strict partial orders A *and* B*, a partition* $\{B_a\}_{a \in A}$ *such that* $\bigcup_{a \in A} B_a \subseteq B$.
Question: is there an embedding f *of* A *into* B *such that* $f(a) \in B_a$*, for every* $a \in A$*?*

Theorem 2. SPOEP *is* NP-*complete*.

Idea of proof. We give an idea of a polynomial reduction from SPOE to SPOEP. Let (A,B) be an arbitrary instance of SPOE. We construct an instance of SPOEP $(\mathrm{A}^{'},\mathrm{B}^{'},\{B_a\}_{a \in A'})$. Set $\mathrm{A}^{'}$ to be A. Let $\{B_a\}_{a \in A'}$ be the set of $|A^{'}|$ disjoint copies of B. Let $B^{'}$ equal $\bigcup_{a \in A} B_a$. Define $<_{B'}$

as follows: for every $x', y' \in B'$, $x' <_{B'} y'$ if and only if there exist $x, y \in A'$ such that x' is a copy of x, y' is a copy of y and $x <_B y$.

Theorem 3. *The problem of recognizing the truth value of* (4) *in finite models of the form* $\mathbf{M} = (U, A, B, P, >, \phi)$*, where* A, B *are unary relations,* $>, \phi$ *are strict partial orders on* A, B *respectively and* P *is a binary relation such that* $\{P_a\}_{a \in A}$ *is a disjoint family of non-empty sets, where* $P_a = \{b \in U : P(b, a)\}$*, is* NP*-complete.*

Proof. We consider only models of the form $(U, A, B, P, >, \phi)$ and satisfying the conditions stated in the theorem. We give an idea of a polynomial reduction of SPOEP to the class of finite models $\mathbf{M}$ that satisfy (4). The algorithm is straightforward: given an instance of SPOEP $\mathbf{X} = ((A', >_{A'}), (B', >_{B'}), \{B_a\}_{a \in A})$, construct a model $\mathbf{M}_{\mathbf{X}} = (U, A, B, P, >, \phi)$ in the following way: let $U = A' \cup B'$, $A = A'$, $B = B'$, $P = \bigcup_{a \in A'} \{B_a \times \{a\}\}$, $>=>_{A'}$, $\phi=\phi_{B'}$.

4. CONCLUSIONS

Assuming Edmond's Thesis and that $\mathbf{P} \neq \mathbf{NP}$, Theorem 3 indicates that the idea of artifitial intelligence capable of effective computation of truth values of some relatively simple natural language sentences in finite models is not realizable.

A somewhat specific questions about computational complexity arise when we take other types of orders into consideration, such as quasi-orders or non-strict orders. Moreover, other kinds of similarity relations between orders may be taken into account, such as homomorphisms and injective homomorphisms. These questions as well as their implications for computational semantics are being considered by the author in collaboration with M.T. Godziszewski and are going to be included in our future work.

REFERENCES

[1] M. Mostowski and D. Wojtyniak, "Computational complexity of the semantics of some natural language constructions," Annals of Pure and Applied Logic, vol. 127, no. 1-3, pp. 219–227, 2004.

[2] J. Szymanik, Quantifiers in Time and Space: Computational Complexity of Generalized Quantifiers in Natural Language. PhD thesis, University of Amsterdam, 2009.

[3] J. Szymanik, "The computational complexity of quantified reciprocals.," in TbiLLC (P. Bosch, D. Gabelaia, and J. Lang, eds.), vol. 5422 of Lecture Notes in Computer Science, pp. 139–152, Springer, 2007.

[4] M. Sevenster, Branches of imperfect information: logic, games and computation. PhD thesis, University of Amsterdam, 2006.

[5] K. J. Barwise, "On branching quantifiers in English," Journal of Philosophical Logic, vol. 8, no. 1, pp. 47–80, 1978.

[6] E. Mendelson, Introduction to Mathematical Logic. 1987.

[7] M. R. Garey and D. S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness (Series of Books in the Mathematical Sciences). W. H. Freeman, first edition ed., 1979.

[8] T. Cormen, C. Stein, R. Rivest, and C. Leiserson, Introduction to Algorithms. McGraw-Hill Higher Education, 2nd ed., 2001.

INTENTIONAL BLANK

AUTO CLAIM FRAUD DETECTION USING MULTI CLASSIFIER SYSTEM

Luis Alexandre Rodrigues and Nizam Omar

Department of Electrical Engineering,
Mackenzie Presbiterian University, Brazil, São Paulo
71251911@mackenzie.br,nizam.omar@mackenzie.br

ABSTRACT

Through a cost matrix and a combination of classifiers, this work identifies the most economical model to perform the detection of suspected cases of fraud in a dataset of automobile claims. The experiments performed by this work show that working more deeply in sampled data in the training phase and test phase of each classifier is possible obtain a more economic model than other model presented in the literature.

KEYWORDS

Fraud Detection, Multi Classifier, Data Mining.

1. INTRODUCTION

The detection of suspected cases of fraud aims to find anomaly patterns in a given population, could be performed in manually or automatically [1]. This task has been applied in various fields like insurance [2], finance [3] and telecommunications [4], etc.

The algorithms used in Data Mining to classification tasks are usually based on heuristics, and thus there is an optimal classifier to perform classification tasks in large datasets [5].

Using a set of 100 samples of training data, this work performs the training and testing of classifiers, whose are applied in an automobile claims dataset that has suspected cases of fraud. After this classifier are combined in a parallel topology that use a combination of results by vote techniques to perform a final classification of each objects.

The classifiers created by this work are evaluated economically. [6] presents a cost matrix that to identified the savings generated by models used in detection of suspected cases of fraud. This cost matrix will used by this work to create a set of classifiers containing the most saving models of detection fraud.

The section 2 from this work presents some researches created to detection suspected cases of fraud. The section 3 presents a methodology used to create the most saving model to detect suspected cases of fraud in an automobile claims dataset. The section 4 presents the results obtained when the classifiers are applied individually and when applied by set of classifiers in the testing dataset to detect suspected cases of fraud.

Natarajan Meghanathan et al. (Eds) : ACSIT, FCST, ITCA, SE, ICITE, SIPM, CMIT - 2014
pp. 37–44, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4604

2. RELATED WORKED

The most common technique to fraud detection by Data Mining is find patterns that shows a behavior uncommon inside of dataset [7]. The Data Mining works with different data exploration models and solutions to specifics fraud cases were proposed [7]:

- Insurance: [6] used individual classifiers and multi classifier system to detect fraud in an automobile claim dataset. The individual classifiers are Decision Tree by C4.5, Naïve Bayes and Artificial Neural Network. The multi classifier system is a combination of Decision Tree, Naïve Bayes and Artificial Neural Network by Stacking-bagging algorithm. The results showed that multi classifier system was the most saving model to detect suspected cases of fraud.
- Credit Card: [8] presents three techniques used in credit card fraud detection, Artificial Neural Network, Logistic Regression and Decision Tree. According [1] the most techniques used in credit card fraud detection are Outliers Detection and Artificial Neural Network.
- Telecommunications: The works to fraud detection in telecommunication field focus on trying to identify the use of services without authorization by Artificial Neural Network, Outliers Visualization and patterns recognition [1].
- Online Auction: [9] presents a model to fraud detection to online auctions. The model used decision tree created by C4.5 algorithm to classifiers suspicious transactions according to the time that they occur. The criteria used to create the decision tree's rules are the average of positives feedbacks that vendors have and the price average of their products.
- Health Insurance: [7] presents fraud cases performed in medical clinics, which impair financially the insurance companies. The cases are detected by a model based on outlier detection by Support Vector Machine.

3. METHODS TO FRAUD DETECTION

This section will present a methodology that this work is using to find the most saving model to detect suspected cases of fraud. Will be presented the classifiers used in fraud detection, the topology and a combination function used to perform a final prediction each objects.

3.1. Classifiers

The classifiers aims to identify the categories set that a object of given dataset belongs [10]. This work selected three algorithms used in related works to perform the classification task in automobile claims dataset:

- *Decision Tree C4.5:* rule induction is one the most used methods used in fraud detection, because is easy to analyze the decisions created by the algorithm [11]. The algorithm C4.5 is used to induction decision tree. The decisions created by this algorithm are performed by the evaluation of dataset's features [12].
- *Naive Bayes:* naive bayes is a static classifier based on Bayes Theorem that mix previous knowledge a class by evidence selected in dataset [13]. The algorithm has a good performance history compared to other algorithms applied in fraud detection of automobile claims [14]. According [6] the algorithm is very efficient in large datasets and very efficient to create classifiers.
- *Support Vector Machines (SVM):* SVM is a binary classifier has been successfully applied in tasks to pattern recognition [15]. The algorithm maximizes the decision limit between two classes using a kernel function [10]. According [16] SVM is used in fraud detection tasks because works very well in datasets with imbalanced class.

The algorithms presented by this work are instable, because can change their forms when the environment and conditions in which they applied change. This feature is important when the combination of classifiers is performed, because each change in training dataset, different classifications will be performed in each new classification model created in the training phase [17].

3.2. Combination of Classifiers

The combination of classifier aims to perform classification tasks by combination of results between different classifiers to predict the final classification of the each object in the dataset [18].

This work combined the result of each algorithm previously presented to detect suspected cases of fraud by parallel topology. According [5] the most systems that used combination of classifiers used a parallel topology, which executes parallel all classifiers and combining their results using a decision function. The Figure 1 presents a structure proposed by this work using a parallel topology.

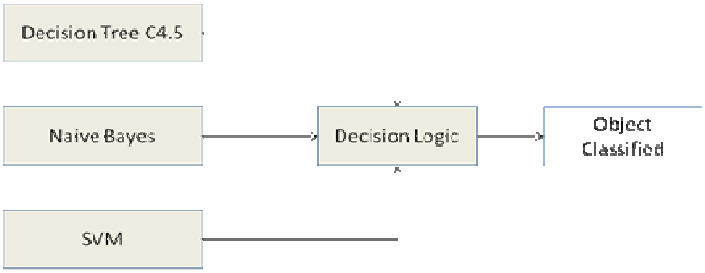


Figure 1. Structure of combination of classifiers using a parallel topology

Accord Figure 1 each object from dataset is applied trough all classifiers and each classifier will present a different classification to object, it can be fraud or legal. The decision function will be responsible for obtaining the classification object provide by each classifier and to perform the final classification to each object.

The decision function used to perform the final classification was the vote technique *AVGVote* [18]. As shown in equation 1, each classifier C_{ji} inside in set of classifiers R, the *AVGVote* function computes one vote in the object x classified as i.

$$AVGVote(x) = argmax_{i=1}^{c}\left(\frac{1}{R}\sum_{i=1}^{R} C_{ji}(x)\right) \quad (1)$$

3.3. Automobile Claim Dataset

The experiments presents in this work used an automobile claims dataset with suspected cases of fraud. Each object from dataset is classified as fraud or legal. This dataset was used in [6] to identify the most saving model to detect suspected cases of fraud.

The dataset has suspected cases of frauds between 1994 and 1996 and has 15.421 objects, and each object has 6 numeric attributes and 25 categorical attributes. The preprocessing data was performed following the orientations proposed by [6].

The dataset was divided in two partitions to training and to testing the classifiers. The training partition has automobile claims between 1994 and 1995 years, and the testing partition has automobile claims of 1996 year.

There are imbalanced classes inside of dataset. This feature indicates that the classes are not distributed in the same quantity inside of dataset [10]. If dataset presents this feature, the generalization in each classifier can be adversely affected, thus classification tasks can be little precise in its test phase.

According [10], one way to solve the imbalanced classes' problem between the classes inside of dataset is the subsample generation from dataset. Thus this work created various subsamples and applied in the training and testing phase of each classifier.

3.4. Creating subsample data to training phase

According [10] the performance of a classifier depends of training data used in training phase. Thus with the goal of finding the best subsample to train algorithms, 100 subsamples were created. Between the first subsample and subsample number 71, there was a variation in the quantity of objects, and the balanced class was between 50% fraud objects and 50% legal objects. The variation of size of first subsample until the subsample 71 was on the order of 20 to randomly selected objects and no repetition of objects. The subsample number 1 was composed of 20 objects and the size of each new subsample was the sum of size of the previous subsample plus 20 objects. Thus the subsample number 71 was composed by 1420 objects.

Because the training dataset has imbalanced classes, the variation of quantity of objects between subsample number 72 until subsample number 100 was performed on the order 10 to 10 objects only to objects that belongs to majority class, and there was not repetition of objects. The subsample number 72 was composed by the sum of size of subsample plus the random selection of 10 objects of majority class. This variation happened until subsample number 100.

3.5. Cost Model

[6] used a cost matrix to identify the most saving model to perform detection suspected cases of fraud in the dataset. Based on the year 1996, the average cost per claim was about USD$ 2,540.00 and the average cost per investigation of suspected cases of fraud was about USD$ 203.00.

Using a confusion matrix [6] defined variables to identify the costs in each experiment performed in his work. According Table 1 the quantity of True Positives (Hits) and the quantity of False Negatives (False Alarm) were used to calculate the cost of suspected fraud claim. The quantity of items classifieds like True Negatives (Normal) and the quantity of items classified like False Negatives (Misses) were used to calculate the cost of each claim.

Table 1. Model Cost to Fraud Detection.

Variable	**Cost**
Hits	Quantity of Hits * Average cost per Investigation.
False Alarms	Quantity of False Alarms * (Average cost per Investigation + Average cost per claim)
Misses	Number of Misses * Average cost per claim
Normal	Quantity of Normal * Average cost per claim

The False Alarm items are the most expensive model, because this variable is defined by the cost per investigation and by the cost per claim. The saving total of each model created was defined in the Model Cost Savings variable by [6], as shown in equation 2.

$$Model\ Cost\ Savings = No\ Action - [Misses\ Cost + False\ Alarms\ Cost + Normals\ Cost + Hits\ Cost] \quad (2)$$

The variable No Action is considering that all claims are Normal. Thus this variable is defined by quantity claims in dataset multiplied by cost per claim. This work used the variable Model Cost Savings to identify the best model created by each algorithm in each testing phase. This variable was used too to compare the cost of combination of classifiers related the classifiers applied manually and related the results presented in [6].

4. EXPERIMENTS

This work performed four experiments to detect suspected cases of fraud in the automobile claim dataset. Three experiments that are divided by algorithm used the set of subsample to find the most saving model, and the last one is related the combination of classifiers created by each algorithm.

The Figure 2 show the process used to create the classification model proposed by this work. In the first moment is performed a preprocessing data according [6]. The preprocessing data was necessary to eliminate missing values and create new attributes that can grow the performance of each classifier created in the training and testing phase.

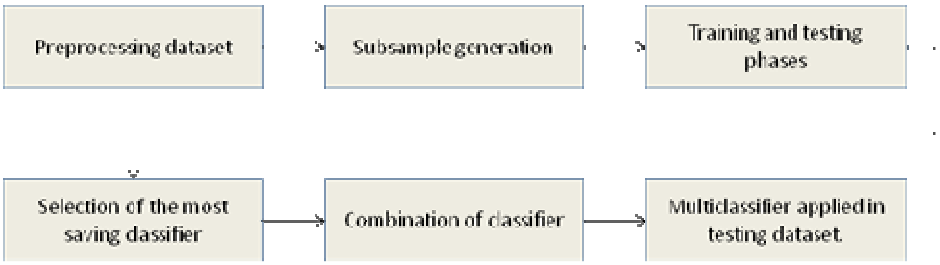


Figure 2. Process to perform fraud detection

When the classifiers were applied in testing dataset a confusion matrix was extracted to calculate all cost variables. It was possible calculate the value of Model Cost Saving variable, as shown in equation 2, and get the most saving model created by each classifier. Each model selected as the most saving was compare with the most saving model shown in [6]. This model is composed by combination of classifiers created by C4.5, Naïve Bayes and Artificial Neural Network, and the max saving cost by this combination was about USD$ 167,000.00.

According Figure 3 the C4.5 algorithm showed the most saving model to detect suspected cases of fraud when it was applied in the subsample number 12. The subsample number 12 consists of 240 objects, with 50% of objects classified as fraud and 50% of objects classified as legal. Reaching a saving of USD$ 177,592.00, the model is the most saving when compared with other classifiers and the most saving when compared the model proposed in [6].

The SVM algorithm created the most saving fraud detection model when it was applied in the subsample number 80. The subsample consists of 1510 objects, with 40% of objects classified as

fraud and 53% of objects classified as legal. The most saving model created by SVM has a saving about 158,732.00, but according Figure 4, the model is not more economic than model proposed in [6].

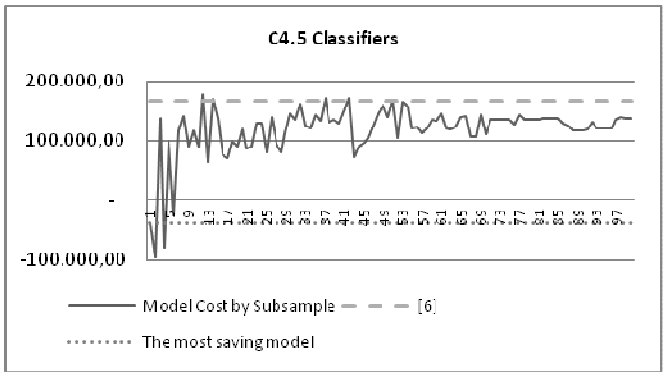


Figure 3. Performance of C4.5 classifier in each subsample

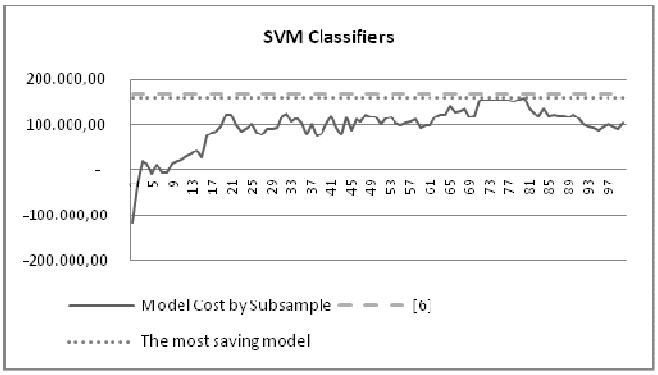


Figure 4. Performance of SVM classifier in each subsample

The most saving model using Naïve Bayes algorithm was created by subsample number 20. This subsample consists of 400 objects, with 50% of objects classifieds as fraud and 50% of objects classifieds as legal. This model has the worst saving compared with other classifiers proposed by this work and the worst performance compared with the most saving model proposed in [6]. According Figure 5, the model presented a saving of USD$ 117,486.00.

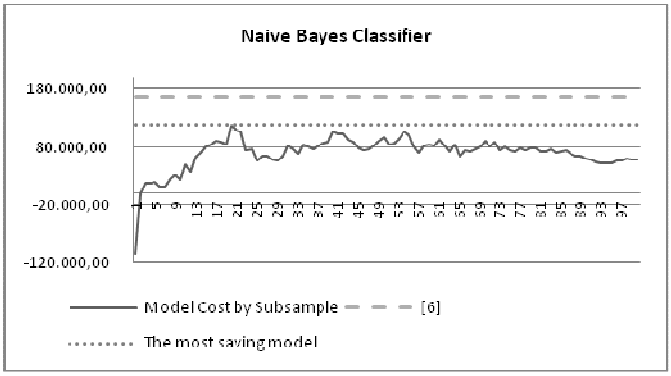


Figure 5. Performance of Naïve Bayes classifier in each subsample

The three models created when applied in testing dataset showed different classification of objects. As shown in Figure 6 there is a diversity of quantity positive class classified by each the most saving model. According [5] the diversity of results is important to make the combination of classifiers and created a final prediction of each object in dataset.

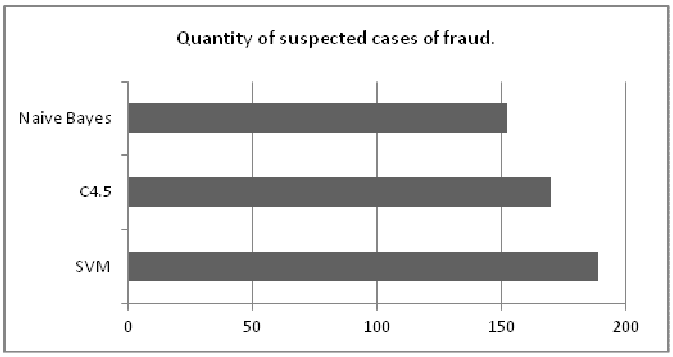


Figure 6. Quantity of suspected cases of fraud identified by each the most saving model

Using this diversity presented by the most saving models, they were combining and applied in testing dataset. The combination of classifiers proposed by this work, which uses the parallel topology and *AVGVote* decision function, presented the most saving model compared to all models applied individually and compared to model proposed in [6]. According Figure 7 the combination of classifiers reaching a saving about USD$ 183,089.00.

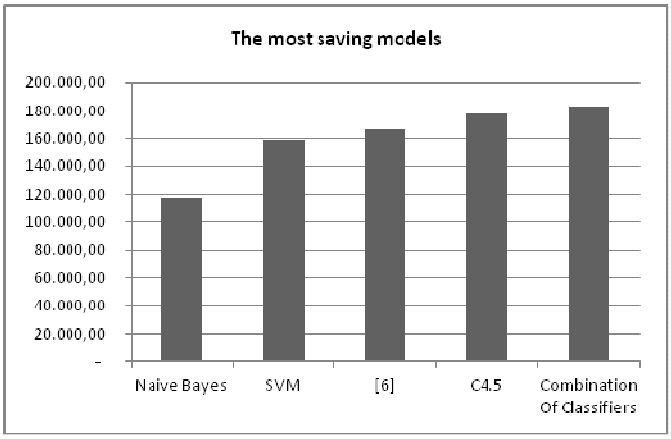


Figure 7. Ranking of the most saving model

5. CONCLUSION

This work combined the classifiers create by C4.5, SVM and Naïve Bayes algorithm to find the most saving model to detect suspected cases of fraud. Working more deep with subsample of training automobile claim dataset, the most saving models were selected, combined and applied in testing dataset. The combination of classifiers was performed by parallel topology and each object was classified by *AVGVote* function decision.

These experiments showed that a good subsample can be efficient to build classifiers and to build a cheaper model to identify suspected cases of fraud. The combination these classifiers presented better performance than the most saving model proposed in [6], which used combination of classifiers.

REFERENCES

[1] Y. Kou, C.-t. Lu, S. Sinvongwattana & Y.-P. Huang, (2004) "Survey of Fraud Detection Techniques," *International Conference on Networking, Sensing & Control*.

[2] K. D. Aral, H. A. Güvenir, I. Sabuncuoğlu & A. R. Akar, (2012) "A prescription fraud detection model," *Computer methods and programs in biomedicine*, Vol. 106, pp37-46.

[3] E. Ngai, Y. Hu, Y. Wong, Y. Chen & X. Sun, (2011) "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems*, Vol. 50, pp559-569.

[4] C. S. Hilas & P. A. Mastorocostas, (2008) "An application of supervised and unsupervised learning approaches to telecommunications fraud detection," *Knowledge-Based Systems*, Vol. 21, pp721-726.

[5] M. Woźniak, M. Grañab & E. Corchadoc, (2014) "A survey of multiple classifier systems as hybrid systems," *Information Fusion*, Vol. 16, pp3-17.

[6] C. Phua, D. Alahakoon & V. Lee, (2004) "Minority Report in Fraud Detection : Classification of Skewed Data," *ACM SIGKDD EXPLORATIONS*, pp50-59.

[7] M. Kirlidog & C. Asuk, (2012) "A Fraud Detection Approach with Data Mining in Health Insurance," *Procedia - Social and Behavioral Sciences*, pp989-994.

[8] A. Shen, R. Tong & Y. Deng, (2007) "Application of Classification Models on Credit Card Fraud Detection," *Service Systems and Service Management*, pp2-5.

[9] W.-H. Chang & J.-S. Chang, (2012) "An effective early fraud detection method for online auctions," *Electronic Commerce Research and Applications*, pp346-360.

[10] P.-N. Tan, M. Steinbach & V. Kumar, (2005) *Introduction to Data Mining*, Addison-Wesley.

[11] Y. Sahin, S. Bulkan & E. Duman, (2013) "A cost-sensitive decision tree approach for fraud detection," *Expert Systems with Applications*, Vol. 40, n. 15, pp5916–5923.

[12] J. R. Quinlan, (1993) *C4.5: programs for machine learning*, San Francisco: Morgan Kaufmann Publishers Inc.

[13] G. H. John & P. Langley, (1995) "Estimating continuous distributions in Bayesian classifiers," *UAI'95 Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp338-345.

[14] S. Viaene, R. A. Derrig, B. Baesens & G. Dedene, (2002) "A comparison of State-of-the-Art Classification Techiniques for Expert Automobile Insurance Claim Fraud Detection," *The Journal of Risk and Insurance*, vol. 69, n. 3, pp373-421.

[15] H.-C. Kim, S. Pang, H.-M. Je, D. Kim & S. Y. Bang, (2003) "Constructing support vector machine ensemble," *Pattern Recognition*, Vol. 36, n. 12, pp2757–2767.

[16] S. Bhattacharyyaa, S. Jhab, K. Tharakunnelc & J. C. Westlandd, (2011) "Data mining for credit card fraud: A comparative study," *Decision Support Systems*, Vol. 50, n. 3, pp602–613.

[17] T. G. Dietterich, (2000) "Ensemble Methods in Machine Learning," *Multiple Classifier Systems*, Springer Berlin Heidelberg, pp1-15.

[18] R. Ranawana & V. Palade, (2006) "Multi-Classifier Systems: Review and a roadmap for developers," *International Journal of Hybrid Intelligent Systems*, Vol. 3, n. 1, pp35-61.

[19] S. Thiruvadi & S. C. Patel, (2011) "Survey of data-mining used in Fraud detection and Prevention," *Information Technology Journal*, pp710-716.

AUTHORS

Luis Alexandre Rodrigues

Degree in Information Systems and MSc student in Electrical Engineering at Mackenzie University. Currently he is working like software architect at Insurance Company and is interested in Data Mining techniques to detect suspected cases of fraud in large datasets.

Nizam Omar

Degree in Mechanical Engineering from the Technological Institute of Aeronautics (ITA), MSc in Applied Mathematics from the ITA and Ph.D. in Applied Informatics by Pontifical Catholic University (PUC). He is currently professor at Mackenzie Presbiterian University. Has experience in Computer Science, Artificial Intelligence, Automata and Formal Languages.

ON INTERVAL ESTIMATING REGRESSION

Marcin Michalak

Institute of Informatics, Silesian University of Technology, Gliwice, Poland
Marcin.Michalak@polsl.pl

ABSTRACT

This paper presents a new look on the well-known nonparametric regression estimator – the Nadaraya-Watson kernel estimator. Though it was invented 50 years ago it still being applied in many fields. After these yearsfoundations of uncertainty theory – interval analysis – are joined with this estimator. The paper presents the background of Nadaraya-Watson kernel estimator together with the basis of interval analysis and shows the interval Nadaraya-Watson kernel estimator.

KEYWORDS

Interval Analysis, Nonparametric Regression, Nadaraya-Watson Kernel Estimator.

1. INTRODUCTION

The main difference between the mathematical and physical interpretation of a number is that from the mathematical point of view the number is a well-defined point in some space while in physics number (a value) cannot be measured without nonzero level of the uncertainty: in the macro world it is the limitation of our eye precision and the rule precision, for example during measuring the apple diameter, and in the micro world it was well described by Werner Heisenberg and his uncertainty principle.

For a long time scientists have been trying to describe the uncertainty in the mathematical way and applying it in the data processing. As the most famous approaches fuzzy sets [1] (with fuzzy numbers [2]) rough sets[3] or interval analysis [4][5] should be mentioned.

The main motivation of the research presented in this paper is the 50^{th} anniversary of very simple nonparametric regression function estimator – the Nadaraya-Watson kernel estimator. It was invented independently by Nadaraya [6] and Watson [7] in 1964. As the aim of the research the application of interval arithmetic into this method of regression analysis was stated. The title of the paper connects directly to the Nadaraya paper.

The paper is organised as follows: it starts from the short reminder of the Nadaraya-Watson kernel estimator and the brief overview of the interval analysis. Afterwards, the interval approach to kernel regression is presented and followed by the results of experiments on the synthetic data. The paper ends with some final conclusions.

Natarajan Meghanathan et al. (Eds) : ACSIT, FCST, ITCA, SE, ICITE, SIPM, CMIT - 2014
pp. 45–53, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4605

2. KERNEL REGRESSION

Some of the commonly known examples of the nonparametric regression estimators are kernel estimators. This group of methods is developed from the solution of nonparametric estimation of the density function. The Nadaraya-Watson kernel estimator is the simplest kernel regression estimator [6][7]. For the one dimensional case it is given by the equation:

$$\tilde{f}(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}$$

where pairs (x_i,y_i) are known, K is a kernel function and h is so called smoothing parameter. This estimator can be explained as the kind of moving average: the kernel function K is responsible for the shape of weights of averaged values and the smoothing parameter defines the range of input values.

In the Table 1. there are presented most popular kernel functions (I is an indicator function). As we can see only one of them, the Gaussian, has the infinite domain what means that it takes into consideration (estimation of the value at x) all given points, even very distant. Other kernels narrow the neighbourhood of x to the value of the smoothing parameter h.In the onedimensional case only pairs from the training set, which first entries belong to the interval $[x-h, x+h]$ are averaged.

Table 1. Most popular kernel functions

Kernel function	Equation
Uniform	$K(x) = 0.5\, I(-1 < x < 1)$
Triangular	$K(x) = (1-\lvert x\rvert)\, I(-1 < x < 1)$
Epanechnikov	$K(x) = 0.75(1-x^2) I(-1 < x < 1)$
Biweight	$K(x) = 0.9375(1-x^2)^2 I(-1 < x < 1)$
Gaussian	$K(x) = (2\pi)^{-0.5} \exp(-x^2/2)$

In practice the selection of kernel function generally influences less than the selection of the smoothing parameter. The less complicated method of estimating its value is approximation of the Mean Integrated Square root Error (MISE). Its final results – optimal values of h – can be evaluated as follows:

$$h_0 = 1.06\, \tilde{\sigma} n^{-0.2}$$

where $\tilde{\sigma}$is the standard deviation of arguments (x), or as follows:

$$h_0 = 1.06 \min\{0.74 \cdot IR, \tilde{\sigma}\} n^{-0.2}$$

where IR is an interquartile range of x. Details of these calculations can be found in [8].More advanced methods of estimating h can be also found in [9][10][11][12][13].

3. INTERVAL COMPUTATIONS

Interval arithmetic is the branch of mathematics where the number is represented as the interval, due to the uncertainty of the measurement that brought the number. As the first use of this kind of number representation the Archimedes approximation of the π can be recalled: Archimedes stated that $223/71 < \pi < 22/7$.

If two interval numbers are considered then it is interesting how the sum, product or other arithmetical operation should be defined, to give the interpretable result. Next subsection brings definitions of most basic interval operations and the following one shows the problem of interval computations.

3.1. Definition of the Interval Arithmetic Operation

If the non-exactness of the number is represented as its lower and upper bound it is necessary to provide new methods of performing calculations on interval numbers. For two interval numbers X and Y their sum must take into consideration all possible values from two intervals as follows:

$$X + Y = \{x + y : x \in X, y \in Y\}$$

This means that sum of two intervals is the set of all possible results of sum of numbers, coming from each particular interval. In the similar way the following simple arithmetic operations may be defined, the difference:

$$X - Y = \{x - y : x \in X, y \in Y\}$$

the product:

$$X \cdot Y = \{x \cdot y : x \in X, y \in Y\}$$

and the quotient:

$$X/Y = \{x/y : x \in X, y \in Y\}$$

The last operation requires to assure that $0 \notin Y$.

All operations become less complicated to perform when we just consider their bounds. Assuming that the interval X is the range $[\underline{X}, \overline{X}](\underline{X} \leq \overline{X})$ and interval Y is the range $[\underline{Y}, \overline{Y}](\underline{Y} \leq \overline{Y})$ we can write simply that:

$$X + Y = [\underline{X} + \underline{Y}, \overline{X} + \overline{Y}]$$

The similar way of defining the subtraction leads to the following formula:

$$X - Y = [\underline{X} - \overline{Y}, \overline{X} - \underline{Y}]$$

which can be derived from the dependence:

$$X - Y = X + (-Y)$$

Situation becomes a little more complicated when the product of two intervals is taken into consideration. Due to conditions of signs of lower and upper bounds of intervals the bounds of the result of the operation take values as it is presented in the Table 2.

Table 2. Definition of product of two intervals.

Case	$\underline{X \cdot Y}$	$\overline{X \cdot Y}$
$0 \leq \underline{X}$ and $0 \leq \underline{Y}$	$\underline{X} \cdot \underline{Y}$	$\overline{X} \cdot \overline{Y}$
$\underline{X} < 0 < \overline{X}$ and $0 \leq \underline{Y}$	$\underline{X} \cdot \overline{Y}$	$\overline{X} \cdot \overline{Y}$
$\overline{X} \leq 0$ and $0 \leq \underline{Y}$	$\underline{X} \cdot \overline{Y}$	$\overline{X} \cdot \underline{Y}$
$0 \leq \underline{X}$ and $\underline{Y} < 0 < \overline{Y}$	$\overline{X} \cdot \underline{Y}$	$\overline{X} \cdot \overline{Y}$
$\overline{X} \leq 0$ and $\underline{Y} < 0 < \overline{Y}$	$\underline{X} \cdot \overline{Y}$	$\underline{X} \cdot \underline{Y}$
$0 \leq \underline{X}$ and $\overline{Y} \leq 0$	$\overline{X} \cdot \underline{Y}$	$\underline{X} \cdot \overline{Y}$

$\underline{X} < 0 < \overline{X}$ and $\overline{Y} \leq 0$	$\overline{X} \cdot \underline{Y}$	$\underline{X} \cdot \underline{Y}$
$\underline{X} \leq 0$ and $\overline{Y} \leq 0$	$\overline{X} \cdot \overline{Y}$	$\underline{X} \cdot \underline{Y}$
$\underline{X} < 0 < \overline{X}$ and $\underline{Y} < 0 < \overline{Y}$	$\min\{\underline{X} \cdot \overline{Y}, \overline{X} \cdot \underline{Y}\}$	$\max\{\underline{X} \cdot \underline{Y}, \overline{X} \cdot \overline{Y}\}$

Definition of division can be obtained from the product of inversion of the second argument, assuming again $0 \notin Y$:

$$1/Y = [1/\overline{Y}, 1/\underline{Y}]$$

and

$$X/Y = X \cdot (1/Y)$$

3.2. Problems of the Interval Arithmetic Operations

One of the problems of interval computations is that we cannot assume two expressions in the real arithmetic to be equivalent in the sense of interval analysis. This will be shown on the following example. Let us consider the following formula in the ordinary (real) arithmetic:

$$1 + \frac{b}{a} = \frac{a+b}{a}$$

Both sides of this formula are equivalent as long as the assumption of $a \neq 0$ is fulfilled. Now let us set a =[3, 5] and b = [7, 10] and calculate both sides with the interval arithmetic:

$$L = 1 + \frac{[7,10]}{[3,5]} = [1,1] + [7,10] \cdot \frac{1}{\left[\frac{1}{5}, \frac{1}{3}\right]} = [1,1] + \left[\frac{7}{5}, \frac{10}{3}\right] = \left[\frac{12}{5}, \frac{13}{3}\right] = [2.4, 4.(3)]$$

$$R = \frac{[3,5] + [7,10]}{[3,5]} = \frac{[10,15]}{[3,5]} = [10,15] \cdot \frac{1}{\left[\frac{1}{5}, \frac{1}{3}\right]} = \left[\frac{10}{5}, \frac{15}{3}\right] = [2,5]$$

$$L \neq R$$

Although both expressions are equivalent in the real arithmetic it occurs that they are not in the interval sense. This difference is caused in general by the phenomenon of *interval dependency*. When we have the interval a in the nominator and the denominator of the fraction has the same value from the same interval, in calculation they are treated as independent. It becomes more apparent when we compare the result of squaring interval number.

From the origin of the idea of interval computation we have the following definition of the square function:

$$f(X) = \{x^2 : x \in X\}$$

This can be expanded as:

$$f(X) = \begin{cases} \left[\underline{X}^2, \overline{X}^2\right] & 0 \leq \underline{X} \\ \left[\underline{X}^2, \overline{X}^2\right] & \overline{X} \leq 0 \\ \left[0, \max\left\{\underline{X}^2, \overline{X}^2\right\}\right] & \underline{X} < 0 < \overline{X} \end{cases}$$

If we are interested in calculating $[-2, 2]^2$ we obtain the interval [0, 4] but if we expand $[-2, 2]^2$ as [-2, 2]· [-2, 2] it will give an interval [-4, 4]. In the first approach an interval is calculated as the

set of all possible squares of values from the given SINGLE input. In the second one – all possible values of product of TWO values from the interval are determined.

4. KERNEL INTERVAL REGRESSION

The first approach of interval kernel regression is to apply the formula from the section 2. of the paper without any transformation. From the other side we know, that formulas that are equal in the ordinary real calculation may give different results when applied for interval numbers. So it is worth to transform the original Nadaraya-Watson kernel regression equation and compare its interval results on the same interval data.

Let us consider the equivalent formula of Nadaraya-Watson kernel estimator, equivalent in the domain of real numbers.

$$\tilde{f}(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)} = \sum_{i=1}^{n} y_i \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)}$$

5. EXPERIMENTS

Experiments were performed on several artificial datasets, where the estimated function was given but there was a random noise, with the zero mean values, added to each function value.The standard deviation of the noise and specification of all datasets are presented in the Table 3. The first two sets come from [14] and the rest from [15].

Table 3. Specification of used datasets.

Dataset	$\boldsymbol{x}$	$\boldsymbol{y}$	$\boldsymbol{\sigma}$
1	$[-\pi;\ \pi]$	$y(x) = e^{\frac{(x-1)^2}{4}}$	0.15
2	$[0; 1]$	$y(x) = 0.3[\sin(5x - 3)]$	$\sqrt{1.5}$
3	$[-2; 2]$	$y(x) = x + 2e^{-16x^2}$	0.4
4	$[-2; 2]$	$y(x) = \sin 2x\ +\ 2e^{-16x^2}$	0.3
5	$[-2; 2]$	$y(x) =\ 0.3e^{-4(x+1)^2} + 0.7e^{-16(x-1)^2}$	0.1
6	$[-2; 2]$	$y(x) = 0.4x + 1$	0.15

Each dataset contained 101 observations, distributed uniformly in the domain. This 101 pairs of observations were recalculated into 201 pairs, which contained original 101 as pairs of interval numbers with their lower and upper bounds equal and 100 new pairs with typically interval numbers whose lower and upper bounds were defined as follows:

Table 4. Transformation of real pairs into interval.

Domain	**Pairs**
real	$\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_{101}, y_{101})\}$
interval	$\{([x_1, x_2], [y_1, y_2]), ([x_2, x_3], [y_2, y_3]), \dots, ([x_{100}, x_{101}], [y_{100}, y_{101}])\}$

Two versions of Nadaraya-Watson kernel estimator were used: the simple (NW1) and the modified, from the section 4 (NW2). As the *h* estimator the equation basing on interquartile range was used.

For the purpose of estimators evaluations the prediction of values in original (non-interval) 101 points was taken into consideration. As the regression error Root Mean Squared Error was used, which formula is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - y_i)^2}$$

6. RESULTS

Twelve experiments were performed (six datasets on two versions of Nadaraya-Watson interval kernel estimator). On the Figure 1. their results are presented. Points are the noised datasets, black points marked with × are the result of standard Nadaraya-Watson estimator and red rectangles present interval output of the kernel regression. Also the original dependence is slightly visible as the black line. The interval output of the interval regressor is shown as the vertical bar.

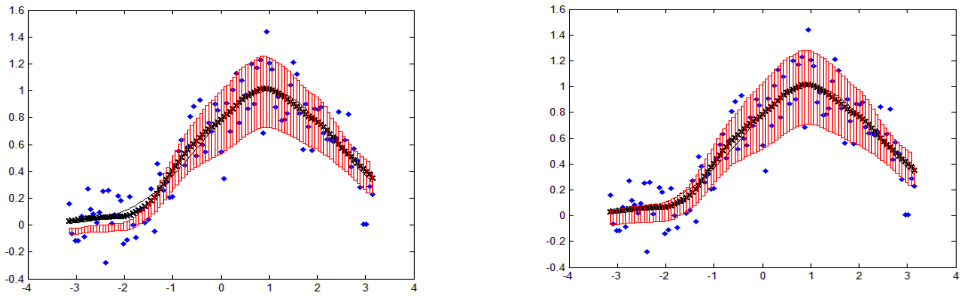


Figure 1. NW1 (left) and NW2 (right) results for dataset #1.

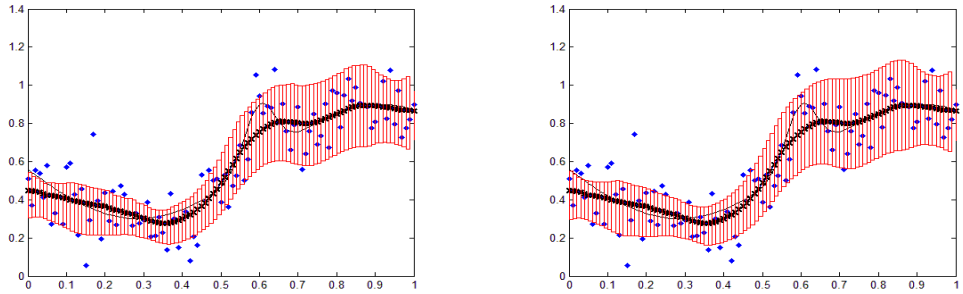


Figure 2. NW1 (left) and NW2 (right) results for dataset #2.

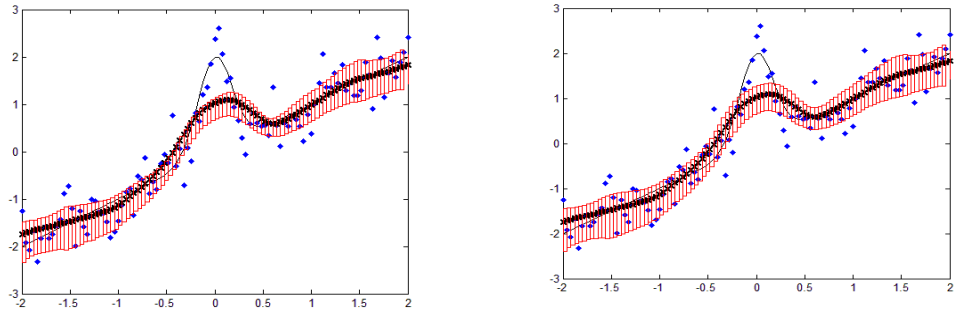


Figure 3. NW1 (left) and NW2 (right) results for dataset #3.

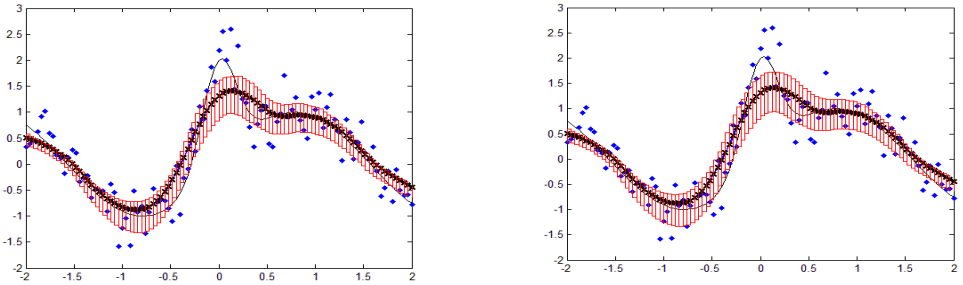


Figure 4. NW1 (left) and NW2 (right) results for dataset #4.

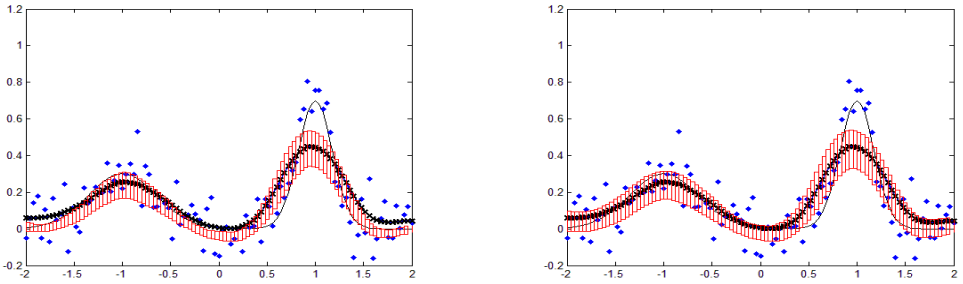


Figure 5. NW1 (left) and NW2 (right) results for dataset #5.

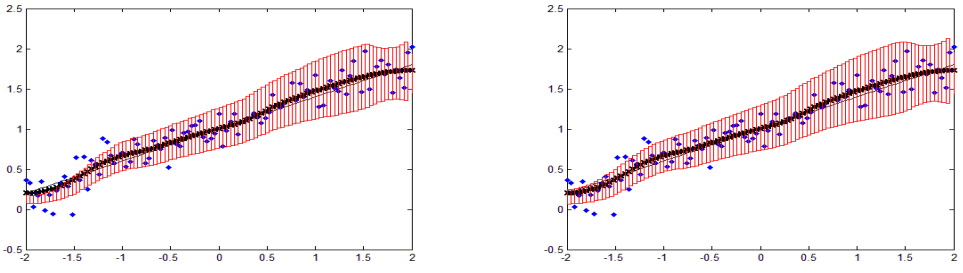


Figure 6. NW1 (left) and NW2 (right) results for dataset #1.

The qualitative evaluation of three models of kernel regression is presented in the Table 5. For standard Nadraya-Watson regression (column marked as NW) a normal (real) RMSE error is presented, while for other two models their interval error is presented.

Table 5. Comparison of three regression models.

Dataset	NW	NW1	NW2
1	0.17134	[0.13636; 0.21691]	[0.13075; 0.22187]
2	0.12390	[0.089199; 0.16802]	[0.088208; 0.16954]
3	0.45991	[0.39968; 0.54332]	[0.38513; 0.55361]
4	0.41109	[0.35302; 0.4886]	[0.34866; 0.49308]
5	0.12824	[0.11465; 0.14434]	[0.10982; 0.15063]
6	0.14425	[0.095705; 0.20777]	[0.090103; 0.21337]

On Figures 1 to 6 we can see that in most of cases (excluding NW1 model for datasets 1 and 5) the interval version of Nadaraya-Watson kernel estimator covers the results of its original version. It means that the value from the original estimator belongs to the interval returned by the interval estimator.

For both of interval estimators the final regression error also contains the value of the error of the non-interval model. It can be explained very simply with the Figures – as interval outputs are "wider" than the real outputs of the estimator it causes that one of the bounds is closer to the original (input) value and the other is further.

Another interesting remark is that the error of the NW2 model is wider than NW1 and is its superset:

$$RMSE(NW1) \subset RMSE(NW2)$$

7. CONCLUSIONS

This paper presents the new approach on the 50 years old Nadaraya-Watson kernel estimator. The novelty is the combination of the kernel estimator and the interval arithmetic. Due to the phenomenon of interval dependency two versions of this kernel estimator in the interval approach were taken into consideration. Application of any of two modifications gives the opportunity to evaluate the level of the uncertainty of the value estimated with the non-interval analysis.

ACKNOWLEDGEMENTS

This work was supported by the European Union from the European Social Fund (grant agreement number: UDA-POKL.04.01.01-106/09).

REFERENCES

[1] Zadeh, Lofti (1965) "Fuzzy sets",*Information and Control*,Vol. 8, No. 3, pp. 338–353.
[2] Dubois, Didier & Prade, Henri (1978) "Operations on Fuzzy Numbers", *International Journal of Systems Science*, Vol. 9, No. 6, pp. 613-626.
[3] Pawlak, Zdzisław (1991) *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer.
[4] Warmus, Mieczysław (1956) "Calculus of Approximations", *Bulletin de l'Academie Polonaise Des Sciences*, Vol. 4, No. 5, pp. 253-257.

[5] Moore, Ramon (1966) *Interval Analysis*, Prentice – Hall, New Jersey
[6] Nadaraya, Elizbar (1964)"On estimating regression",*Theory of Probability and Its Applications*, Vol. 9, pp.141-142.
[7] Watson, Geoffrey (1964)"Smooth regression analysis",*Sankhya - The Indian Journal of Statistics*, Vol. 26, pp. 359-372.
[8] Silverman Bernard (1986) *Density estimation for statistics and data analysis*. Chapman & Hall
[9] Fan, Jianqing& Gijbels Irene (1992) "Variable bandwidth and local linearregression smoothers", *The Annals of Statistics*, Vol. 20, No. 4, pp.2008–2036.
[10] Gasser, Theo& Kneip, Alois & Kohler, Walther (1991) "A flexible and fast methodfor automatic smoothing",*Journal of American Statistical Association*,Vol. 86, No. 415, pp. 643–652.
[11] Terrell, George (1990) "The maximal smoothing principle in densityestimation",*Journal of American Statistical Association*,Vol. 85, No. 410, pp.470–477.
[12] Terrell, George& Scott, David (1992) "Variable kernel density estimation", *The Annals of Statistics*, Vol. 20, No. 3, pp. 1236–1265.
[13] Turlach, Berwin (1993) "Bandwidth selection in kernel density estimation:a review", C.O.R.E. and Institut de Statistique, UniversiteCatholique de Louvain.
[14] Gajek, Lesław& Kałuszka, Marek (2000)*Wnioskowanie Statystyczne (Statistical Reasining)* (in Polish), Warsaw.
[15] Fan, Jianqing& Gijbels, Irene (1995)"Data-driven bandwidth selection in local polynomial regression:variable bandwidth selection and spatial adaptation", *Journal of the Royal StatisticalAssociation, Series B*, Vol. 57, pp. 371 – 394.

AUTHOR

Marcin Michalak was born in Poland in 1981. He received his M.Sc. Eng. in computer science from the Silesian University of Technology in 2005 and Ph.D. degree in 2009 from the same university. His scientific interests is in machine learning, data mining, rough sets and biclustering. He is an author and coauthor of over 50 scientific papers.

INTENTIONAL BLANK

TOWARDS MODELING DISEASE OUTBREAK NOTIFICATION SYSTEMS

Farag Azzedin, Jaweed Yazdani, Salahadin Adam, Mustafa Ghaleb

King Fahd University of Petroleum and Minerals
Dhahran 31261, Saudi Arabia
{fazzedin,jaweed, adam,g200905270}@kfupm.edu.sa

ABSTRACT

Disease outbreak detection, monitoring and notification systems play an important role in assessing threats to public health since disease outbreaks are becoming increasingly common world-wide. There are several systems in use around the world, with coverage of national, international and global disease outbreaks. These systems use different taxonomies and classifications for the detection and prioritization of potential disease outbreaks. In this paper, we study and analyze the current disease outbreak systems. Subsequently, we extract features and functions of typical and generic disease outbreak systems. We then propose a generic model for disease outbreak notification systems. Our effort is directed towards standardizing the design process for typical disease outbreak systems.

KEYWORDS

Disease Outbreak Notification System, Taxonomy, Modeling, Health Systems. ETL, Databases

1. INTRODUCTION

Disease outbreak detection, monitoring and notification systems play an important role in assessing threats to public health since disease outbreaks are becoming increasingly common world-wide. There are several systems in use around the world, with coverage of national, international and global disease outbreaks. The prime purpose of these systems is ensuing quick detection of possible outbreaks and epidemics.

According to World Health Organization (WHO) [1], the widespread persistence of outbreaks such as MERS, SARS, H5N1 etc., poses immense risks for human life. An unforeseen disease outbreak could lead to a threat on a global scale. A number of factors contribute to this threat including the ability of the disease to mutate into new subtypes, the possibility of the disease to turn highly infectious for humans and the lack of timely response to develop immunity. As per WHO guidelines and recommendations, in the event of an outbreak, member countries are obliged to notify within 24 hours epidemiological information with regards to occurrence / reoccurrence of listed notifiable diseases, the occurrence of a new strain of a listed disease, a significant change in the epidemiology of a listed disease, or the detection of an emerging disease.

It is apparent that the increasing threat of disease outbreaks significantly increases the need to provide timely and accurate information to WHO and public health professionals across many jurisdictional and organizational boundaries. Also, the increasing frequency of biological crises, both accidental and intentional, further illustrates that Disease Outbreak Notification System (DONS) needs to be in place to meet the challenges faced by societies across the world. These systems should detect, monitor, prepare, and respond to a disease outbreak. There a large number of systems in existence that detect and prioritize potential disease outbreaks. Various DONS have

Natarajan Meghanathan et al. (Eds) : ACSIT, FCST, ITCA, SE, ICITE, SIPM, CMIT - 2014
pp. 55–69, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4606

been designed with different objectives, features and functions. There is no clear and standardized approach in the design and implementation of such systems.

Our effort in this paper is directed towards standardizing and proposing a model to be used as a reference when designing DONS in future efforts. We focus on various DONS that use different taxonomies and classifications for the detection and prioritization of potential disease outbreaks. We study and analyze the current systems. Subsequently, we extract features and functions of typical and generic systems under the DONS umbrella. We then propose a generic model for DONS.

2. OVERVIEW OF DISEASE OUTBREAK NOTIFICATION SYSTEMS

The evolution of DONS show a taxonomy that covers various geographies, functions and features based on system and user requirements. BioSense is an Internet-based software system that supports early detection of disease outbreaks by providing techniques for near real-time reporting, related analytics, implementation and automated outbreak detection on a national level. BioSense system collects and evaluates data from ambulatory, clinical laboratory test orders and results from Laboratory Corporation of America laboratory. It presents, summarizes, and visualizes data and analytical results through graphs, maps and tables based on day, source, state, disease type, and metropolitan area. The latest version of this application is BioSense 2.0 which provides data in a distributed cloud computing environment. [2], [3]

The Computer Assisted Search for Epidemics (CASE) is a framework for computer supported outbreak detection. The system developed and currently in use at the Swedish Institute for Communicable Disease Control (SMI) obtains data from SmiNet and performs daily surveillance. It is open source software that removes the personal identification and includes only the specific variables in the CASE database. The system performs outbreak detection in two steps: step 1 identifies different statistical algorithms that detect unexpected or unusual number of cases from collection of patient reports for a particular disease and step 2 initiates an investigation by an epidemiologist (a human expert). If CASE detects an outbreak, step 2 aids in determining whether the detected outbreak indicates an actual outbreak. In some cases, it might be able to detect outbreak diseases earlier than epidemiologists. Moreover, it might detect certain outbreaks that human experts would have overlooked. [4]

Another system, the National Notifiable Diseases Surveillance System (NNDSS) was launched in 1990 and managed by the Australian Government under the support of the Communicable Diseases Network Australia (CDNA). NNDSS collects, analyses, and disseminates data on communicable diseases. NNDSS is involved in national and international health practice and processes public health data through CDNA and other main stakeholders directly into the health system. It has been utilized to notify public health action, especially for the diseases that can be prevented through vaccination. Surveillance epidemiologists analyse the collected data from various jurisdictions every fortnight. In terms of reporting, various reports are placed on the CDA website to be available for public access. Data analysis and reports are also disseminated upon requests from the public, community groups or research organizations. [5]

HealthMap is another major system that facilitates the monitoring of global infectious diseases. As described by Freifeld et al. [6], this system uses a wide variety of online formal and informal information sources and channels such as Google News, ProMED, GeoSentinel etc. to collect and aggregate content in several languages which is then classified by infectious disease agents, geography and time. The system is based on open-source products, both for its development (Linux, Apache) and its continued use (Google Maps, Google Translate API, etc). The classification mechanism is entirely automated, and is based on algorithms that use factors such as the frequency and time frame of alerts as well as the number of sources reporting the

information to identify potential health events. An option is also provided for users to report outbreaks of infectious diseases in their region [6], [7].

Another novel approach for the detection of disease outbreaks and their forecasting is the INFERNO system, short for Integrated Forecasts and Early Enteric Outbreak. Nauvoma et al. [8] describe this system in their paper, explaining the use of a concept known as an "outbreak signature" that allows for the forecasting of outbreaks using existing knowledge of infectious disease epidemiology. The existence of the signature is based on the highly habitual nature of infectious disease events, and can be used to generate a long-term forecast. They further elaborate upon the four components of the system – training, warning and flagging, signature forecasting, and evaluation – each of which contribute to the observational knowledge about the nature and incidence of infection. The INFERNO system effectiveness in predicting an outbreak has been highlighted in the paper via the incidence of a substantial waterborne gastroenteritis disease outbreak in Milwaukee, Wisconsin in 1993. The system is slated to improve as information about infectious disease incidence and exposure increases, with improved predictions of outbreaks and greater accuracy.

Zelicoff et al. [9], in their paper, present the Rapid Syndrome Validation Project (RSVP), a system that allows for early detection of outbreaks and emerging biological threats. It is a collaboration project of several institutions: the Sandia and Los Alomos National Laboratories, the University Of New Mexico Department Of Emergency Medicine, and the NM Department of Health Office of Epidemiology. The system is built using Java and is platform-independent, and can be run with both a standalone Java database and a relational database such as Oracle for flexibility in deployment.

The RSVP system relies primarily on reports by medical personnel and provides an interface through which physicians quickly and easily enter clinical and demographic information about patients showing unusual or atypical symptoms and syndromes. This information is then used to assess potential emerging epidemics or outbreaks and send out early warning alerts to local departments of health for necessary investigation. The system also allows physicians to stay informed of new health alerts, as well as provides them feedback of any similarities between their reports and previous reported cases. The system facilitates the exchange of information between medical personnel and increases their involvement in disease and outbreak control [9].

In total, as part of our team effort, we identified and analysed 21 systems that were placed under our classification of DONS category. To match space requirements, only the study of a few systems has been presented in this section. However, in the next section we have included the taxonomy and analysis of all 21 systems.

3. ANALYSIS & TAXONOMY OF DONS

In developing the taxonomy covering the numerous implementations of DONS across the world, we identified major criteria for categorizing these systems. This section describes a taxonomy based on how, when, and where metrics of the 21 DONS we have analyzed. The "how" metric presents system features such as detection techniques, transparency, various layers of granularity, and disease coverage. The "when" metric looks at issues such as whether the systems, based on time frames, are static or dynamic, and whether the detection analysis is relayed in real time. The "where" metric is concerned with the geographical coverage and whether the system's detections and analysis are complete.

3.1 CATEGORIZATION OF CURRENT DONS

The study and analysis of DONS and related systems must provide a mechanism to categorize the functionalities, features and capabilities of these systems. The major criteria for inclusion of the

systems in our study were based on four categories that were developed to classify the systems. The four categories are: Collection & Analysis, Core Functions, System Features, and Support Functions. The Collection & Analysis category classifies the systems based on who owns the disease data, where such data comes from, what is the data format, and how the data is collected and analyzed. Table 1 shows the dimension and the description of each feature in this category.

Table 1 : Features of the "Collection & Analysis" category

Category	Dimension	Feature	Description
Collection & Analysis	Collection	Stakeholders	System owners & users including health-care providers, other DONS systems, experts, institutions, local & regional health authorities
		Data Sources	Health institutions such as hospitals, clinics, laboratories and pharmacy
		Collection Strategy	Surveillance disease list, collection objectives, collection methods, data cleansing and transformation strategy
		Data Formats	Standard data formats (e.g., XML, Standard Storage Format [SSF] for GPS files; Tagged Image File Format [TIFF] or Joint Photographic Experts Group [JPEG] for photographs; and Excel or InfoPath formats for electronic forms).
	Analysis	Computation Detection Algorithms	Detection algorithms capable of analyzing epidemiological and laboratory data
		Statistical Analysis	Statistical analysis (e.g. Threshold, SaTScan Poisson, SaTScan Space-Time permutation, and Farrington)

Table 2 : Features of the "Core Functions" category

Category	Dimension	Feature	Description
Core Functions	Case-Level	Detection	Defined process to identify individual cases as isolated cases or contributing to outbreaks
		Confirm & Register	Capacity to register/confirm as outbreak cases based on epidemiological and laboratory data
	Group-Level	Signal Extraction	Generate statistically significant signals based on grouped data for identifying outbreaks
		Interpretation	Interpret signals as outbreaks and map to appropriate alert and epidemic thresholds for public health action
		Reporting	Report confirmed outbreaks in a timely manner based on urgency
	Notifications	Communication Procedures	Appropriate and standard communication methods & medium to ensure delivery to identified stakeholders
		Target Entities	Appropriate entities such as health-care providers, other DONS systems, experts, institutions, local & regional health authorities
		Output/ Message Formats	XML, Standard Storage Format [SSF] for GPS files; TIFF or JPEG for photographs; and Excel or InfoPath formats for electronic forms; & pdf

The Core Functions category identifies how a disease outbreak is detected, confirmed, registered, and interpreted. In addition, this category also outlines how the detection is reported, formatted, communicated, and who the target entities are.

Table 2 shows the dimension and the description of each feature in this category. The System Features category consists of features such as completeness, timeliness, usefulness, mobility, reliability, flexibility, and others. These features are described in Table 3.

The last category Support Functions & System Interfaces includes features related to implementation, monitoring, reporting, evaluation standards, training, communication, evaluation, and administration. It also includes features related to how the system interacts with external actors. These features are described in Table 4.

Table 3 : Features of the "System Features" category

Category	Dimension	Feature	Description
System Features	Attributes	Completeness	Completeness of case reporting, reporting sites and notification form
		Timeliness	Data submitted in a timely manner e.g. (immediately, weekly and monthly reporting) to investigate and implement control measures.
		Usefulness	The value of data to use in detecting and responding to outbreak in appropriate time
		Real-time	Monitoring the case sources reporting and other functions in real time, helping to identify and investigate outbreaks in real time
		Coverage	The system can be used at various levels e.g. state, region, country, continent or global level
		Portability	The system can be run in multiple platforms and uses data standards such as XML
		Sensitivity	Sensitive to detect all possible cases/patterns for known and unknown diseases and report to the notification system
		Availability	Availability of system to be used at anytime, anywhere depending on its area of coverage
		Accessibility	Access the system through various platforms such as web browsers, smart devices etc. and by system staff to support and maintain the system
		Usability	The system can be used by predefined users to effectively and efficiently respond to outbreaks.
		Interface Design	System Interface is easy to use, easy to understand, and allow users to achieve their needs through intuitive interfaces .e.g. (GUI)
		Mobility	Native support for mobile devices such as smart phones, tablets etc.
		Flexibility	Ability to change and modify to detect new outbreak cases, modifying and redefining case definitions and threshold values
		Specificity	Ability to detect false positive and false negative cases

Table 4 : Features of the "Support Functions & System Interface" category

Category	Dimension	Feature	Description
Support Functions & System Interfaces	Support Functions	Standards	Use of standards for implementation, monitoring and evaluation, reporting (e.g. guidelines for priority diseases, action thresholds, data management tool, and guidelines for outbreak detection …)
		Training	Training users and staff to use the system e.g. (laboratory, epidemiology and health care persons, administrators)
		Communication	Effective and appropriate medium at each level to support reporting and feedback functions
		Monitoring & Evaluation	Monitor and evaluate the system to ensure that all planned activities for system are on track
		System Administration	Controlling and maintaining the system and ensuring that all activities are executed according to schedule
	Networking & Partnerships	Experts & Institutions	Support for an expert who is a physician or a health professional who is expert in an identified disease; when DONS detects a potential disease outbreak, it notifies a number of experts and/or institutions of the disease by sending messages to their mobile phones
		Coordination	Coordinating between the stakeholders and implementers for using the system in an effective way
		Feedback	Collection and processing of stakeholder response to newsletters, bulletins and supervisory visits

4. CRITICAL FEATURES IN DONS

As stated earlier, the functionalities and features listed in Tables 1 to 4 were identified based on various features and functionalities that exist in the current DONS. In order to be accurate, an initial scan of all the functionalities and features were collected and filtered to come with teh list of critical features required in a DONS.

The objective of this exercise was to include, through a rigorous process, all functionalities and features critical to a DONS system and then analyse the current DONS based on their adherence to the entries in the Tables 1 to 4. The justification for inclusion of the features as critical was based on several criteria including functional, technical, geographical and system factors.

We analyzed and classified 21 systems through this process. The result of our analysis is presented in Tables 5 and 6. This analysis was done through a detailed study of all the system by analyzing information from multiple sources. In order to ensure that the data collected during analysis is accurate, we carefully validated each feature mentioned in Tables 5 & 6 from more than one source. The primary references are listed in the first column of the two tables. Additional secondary references were also used in validating the contents of Tables 5 & 6.

Table 5 : Main features of existing major DONS

		Collection & Analysis						**Core Functions**							
		Collection				Analysis		Case-Level		Group-Level			Notifications		
Reference	**System Name**	Stakeholders	Data Sources	Collection Strategy	Data Formats	Computation Detection Algorithms	Statistical Analysis	Detection	Confirm & Register	Signal Extraction	Interpretation	Reporting	Communication Procedures	Target Entities	Output/Message Formats
[10]	Aegis		✓		✓		✓	✓			✓	✓		✓	
[11]	Argus					✓		✓	✓			✓			
[12]	BioAlirt		✓				✓	✓		✓	✓	✓		✓	
[13]	BioDefend							✓				✓			
[14]	BioPortal					✓	✓	✓		✓	✓	✓		✓	✓
[2]	BioSense	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓		✓	✓
[15]	BioStorm		✓	✓	✓	✓	✓			✓	✓	✓		✓	✓
[16]	BTsurveillance		✓				✓	✓				✓		✓	✓
[4]	Case	✓	✓		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
[17]	Distribute	✓	✓	✓	✓			✓				✓	✓	✓	
[18]	EARS					✓	✓	✓	✓			✓		✓	
[19]	Essence II	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
[20]	EWRS										✓	✓	✓	✓	
[6]	HealthMap		✓	✓				✓		✓		✓		✓	✓
[21]	TESSy	✓	✓	✓	✓	✓		✓	✓				✓	✓	✓
[8]	Inferno					✓		✓	✓	✓		✓			
[22]	NEDSS	✓			✓							✓			
[5]	NNDSS										✓	✓		✓	
[42]	RODS	✓	✓	✓	✓	✓	✓	✓	✓				✓	✓	✓
[9]	RSVP	✓	✓		✓			✓				✓	✓	✓	✓
[23]	SmiNet	✓	✓					✓	✓			✓			✓

Table 6 : More main features of existing DONS

		System Features														**Support Functions & System Interfaces**							
		Attributes														Support Functions					Networking & Partnerships		
Reference		Completeness	Timeliness	Usefulness	Real-time	Coverage	Portability	Sensitivity	Availability	Accessibility	Usability	Interface Design	Mobility	Flexibility	Specificity	Standards	Training	Communication	Monitoring & Evaluation	System Administration	Experts & Institutions	Coordination	Feedback
[10]	Aegis	✓			✓				✓		✓	✓				✓			✓		✓		✓
[11]	Argus					✓		✓		✓					✓		✓		✓		✓		
[12]	BioAlirt		✓	✓			✓	✓			✓				✓				✓		✓		
[13]	BioDefend		✓		✓							✓									✓		
[14]	BioPortal	✓		✓	✓	✓		✓		✓	✓	✓				✓		✓	✓		✓	✓	
[2]	BioSense	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[15]	BioStorm	✓	✓	✓		✓				✓	✓	✓		✓				✓	✓	✓	✓		
[16]	BTsurveillance			✓			✓		✓		✓	✓		✓		✓		✓			✓		
[4]	Case	✓	✓	✓	✓	✓	✓	✓			✓	✓			✓	✓		✓	✓	✓	✓	✓	✓
[17]	Distribute		✓	✓	✓	✓				✓		✓				✓			✓		✓		✓
[18]	EARS		✓	✓				✓	✓	✓					✓				✓				
[19]	Essence II	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[20]	EWRS		✓	✓	✓				✓	✓			✓					✓	✓		✓	✓	
[6]	HealthMap		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓			✓		✓	✓		✓		✓
[21]	TESSy	✓	✓	✓		✓			✓	✓	✓					✓	✓	✓	✓	✓	✓	✓	
[8]	Inferno							✓							✓		✓		✓				
[22]	NEDSS	✓	✓	✓				✓		✓		✓			✓	✓			✓				✓
[5]	NNDSS	✓														✓		✓	✓		✓	✓	
[42]	RODS	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓	✓		✓	✓	✓	✓	✓	✓
[9]	RSVP	✓	✓		✓		✓	✓	✓	✓		✓		✓	✓			✓	✓			✓	✓
[23]	SmiNet	✓	✓		✓					✓		✓				✓		✓	✓	✓	✓	✓	

As discussed before, DONS are classified as belonging to various categories including collection, analysis or reporting systems. Some systems are hybrid systems since they combine features, not necessarily all, from each of these categories. A system that has all the features of all the

categories is referred to as a complete system otherwise it is classified as a partial system. Based on the analysis of the systems in our study, Figure 1 shows examples of partial and complete systems. For example, the Australian NNDSS is a hybrid system where as Swedish SmiNet is only a data collection system.

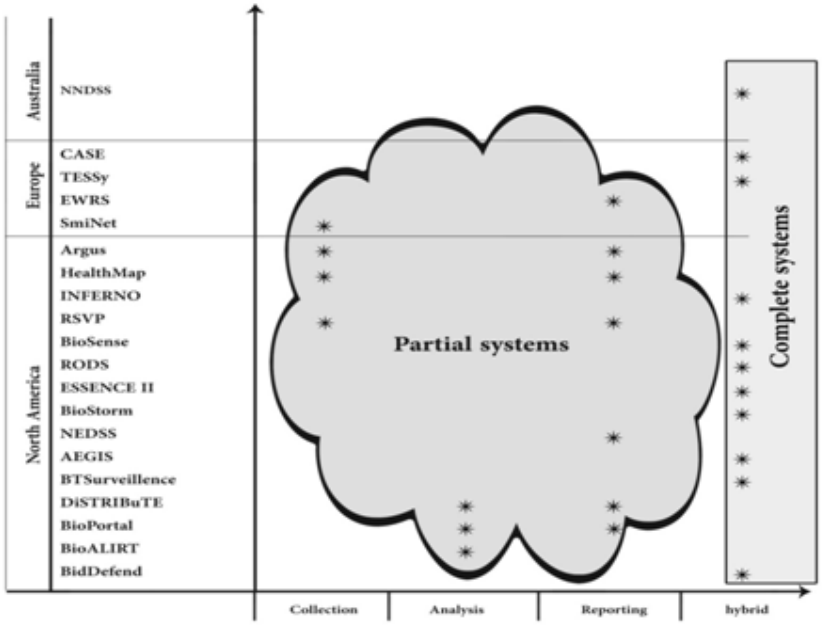


Figure 1 : Examples of partial and hybrid systems

5. MODEL OF A DONS

We propose that a DONS must detect potential disease outbreaks and notifies preregistered experts about the outbreak as a mandatory requirement. Based on the taxonomy proposed in the previous section, we have identified the critical features that the DONS should have to satisfy the mandatory requirements, referred to as Core Features in Table 7.

Table 7 : Core features of the DONS

	Collection & Analysis						**Core Functions**							
	Collection				Analysis		Case-Level		Group-Level			Notifications		
Rrequirements	Stakeholders	Data Sources	Collection Strategy	Data Formats	Computation Detection Algorithms	Statistical Analysis	Detection	Confirm & Register	Signal Extraction	Interpretation	Reporting	Communication Procedures	Target Entities	Output/Message Formats
DONS	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

We also propose that a Disease Outbreak Notification System (DONS), support the additional features listed in Table 8.

Table 8 : Additional features of the existing DONS

	System Features														Support Functions & System Interfaces							
	Attributes														Support Functions					Networking & Partnerships		
Rrequirements	Completeness	Timeliness	Usefulness	Real-time	Coverage	Portability	Sensitivity	Availability	Accessibility	Usability	Interface Design	Mobility	Flexibility	Specificity	Standards	Training	Communication	Monitoring & Evaluation	System Administration	Experts & Institutions	Coordination	Feedback
DONS	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓

5.1 High-Level Design of DONS

DONS interacts with three groups of external actors, namely, case sources, institutions, and experts. Figure 2 shows the context diagram of the DONS.

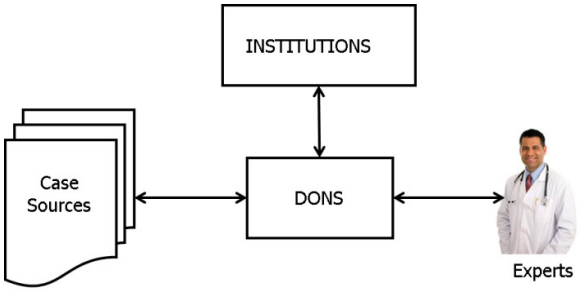


Figure 2 : The DONS context diagram

A case source is a health unit such as a hospital, a clinic, or a pharmacy which reports cases of notifiable diseases to DONS. DONS maintains a list of notifiable diseases. A case source reports diseases in the list of notifiable diseases, but it may also report unknown diseases. The interface between the case sources and DONS is web-based. To report a case, a case reporter must visit the secure DONS web page, where he logs in and fills the online form of the disease that he wants to report.

DONS can share information with local institutions, such as health institutions, research institutions, universities, and government institutions. DONS can also share information with international health institutions, such as WHO, and other disease outbreak notification systems. An example of information that can be shared with international health organizations includes how to detect a potential outbreak of a certain disease. DONS can also share information about new outbreak diseases in the Kingdom with selected international health institutions. The level of information sharing with each institution is different and will be decided by both parties. The information shared is usually via import and the export of files.

DONS has a list of experts to whom it notifies potential disease outbreaks. An expert is a physician, an epidemiologists or a medical professional who is specialized in one or more of the notifiable diseases. When DONS detects a potential disease outbreak, it notifies the appropriate

experts by sending messages to their preferred way of notification (SMS, email or phone call). The expert can then login to DONS website and gets more information about the outbreak. He can also write his feedback into DONS. DONS uses expert feedbacks to improve its accuracy of detecting potential disease outbreaks and to adapt to new situations.

We propose that a generic DONS must consists of at least eight main modules. These modules are the Extraction, Transformation and Loading (ETL) module, the DONS database, the detection module, the notification module, the feedback module, the export and import module, the operation module, and the learning module. Figure 3 shows the block diagram of DONS.

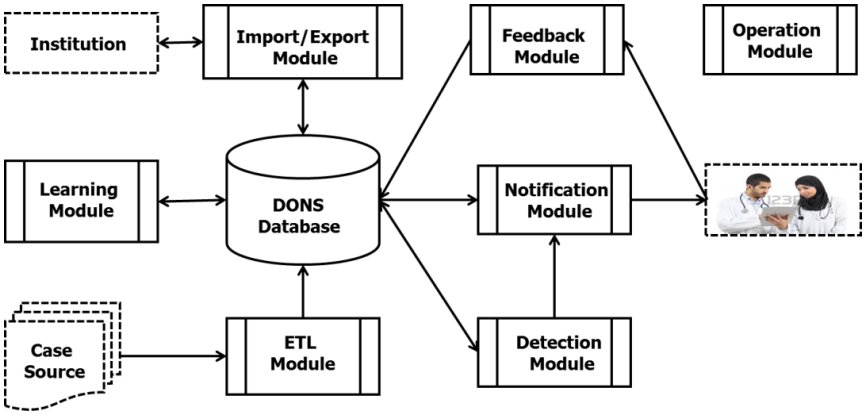


Figure 3 : DONS Block Diagram

The ETL Module receives new cases form the case sources, transforms them and loads them to the DONS database. New cases are reported by the case sources in one of two ways. In the first way, the case reporter goes to the DONS login web page, logs in to the DONS and fills the online form associated with the disease. In the second way, an ETL client process runs regularly in the case sources machine, scans the source database, and sends new cases of notifiable diseases to the DONS system. The new case is then stored in the DONS database. If the new disease is not in the DONS database, its symptoms are inserted into the database and they are also sent to experts. Figure 4 shows the activity diagram of the ETL module.

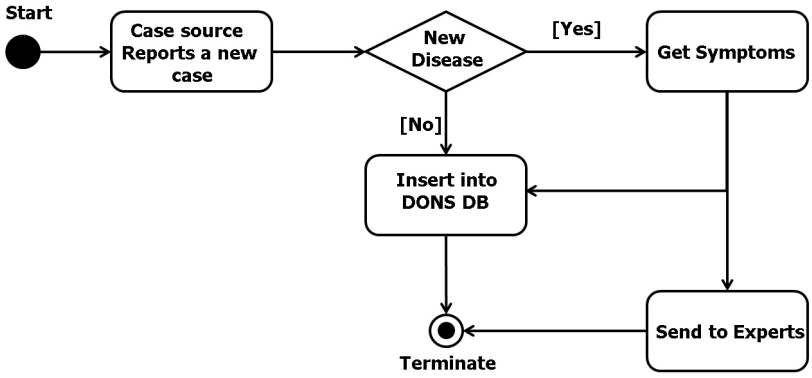


Figure 4 : The ETL module activity diagram

The DONS Database Module consists of many tables. Some of its main tables, we propose, are are: Disease, Symptom, Case, Patient, DetectionAlgorithm, DetectionAlgorithmParameter, Institution, Expert, HealthUnit, and Feedback. The Detection Module uses a number of statistical

algorithms to detect a potential disease outbreak. Some of these algorithms are: SaTScan Poisson, SaTScan Space-Time permutation [24]–[26], and Farrington [27]. The detection module uses different algorithms to detect different disease outbreaks [28]–[31]. Also the actual parameters of each detection algorithm are different for different diseases. DONS keeps the relationship between a disease, a detection algorithm, and the corresponding actual parameter in its database which makes it dynamic. The detection module contains a daemon which regularly checks the Event table. The Event table contains the events that trigger a particular detection algorithm to run. The detection algorithm scans the DONS database for any potential outbreaks. If it detects no potential outbreaks it terminates; and if it detects a potential disease outbreak it send the information to the notification module, its updates the database, and it terminates. Figure 5 shows the activity diagram of the Detection Algorithm.

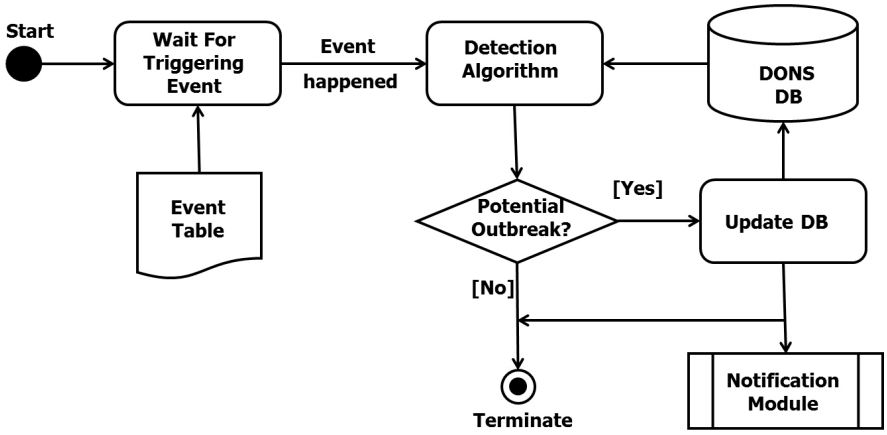


Figure 5 : The Detection module activity diagram

When the detection module detects a potential disease outbreak it passes that information to the notification module. The notification module then reads the appropriate experts from the database and notifies them about the potential disease outbreak. The expert can then login to the DONS system and read detailed information about the diseases from the notification module.
After the Notification Module notifies an expert about a potential disease outbreak, the expert evaluates the accuracy of the notification and gives his feedback. The feedback is mainly given through an online questionnaire and is stored in the DONS database.

The Learning Module enables DONS to handle new diseases and to tune the detection algorithm actual parameters. It first builds a clustering model by learning the symptoms of the notifiable diseases. It then uses the model to predict the cluster of a new disease. From the cluster of the new disease it associates the new disease to a particular detection algorithm and actual parameters which can then be modified from the feedback of the experts.

The Import and Export Module is the interface between DONS and the institutions that share information with it. Typical examples of information that are shared by these institutions are information about new diseases, detection algorithms, and detection algorithm parameter values. This module consists of a query builder, query processor and information formatter. The query builder builds the queries that are sent to the institutions. The query processor answers queries received from the institutions. The formatter formats the information received from the institutions before they are stored in to the DONS database. The kind of information it shares with each institution is stored in the DONS database. Figure 6 shows the Import/Export module.

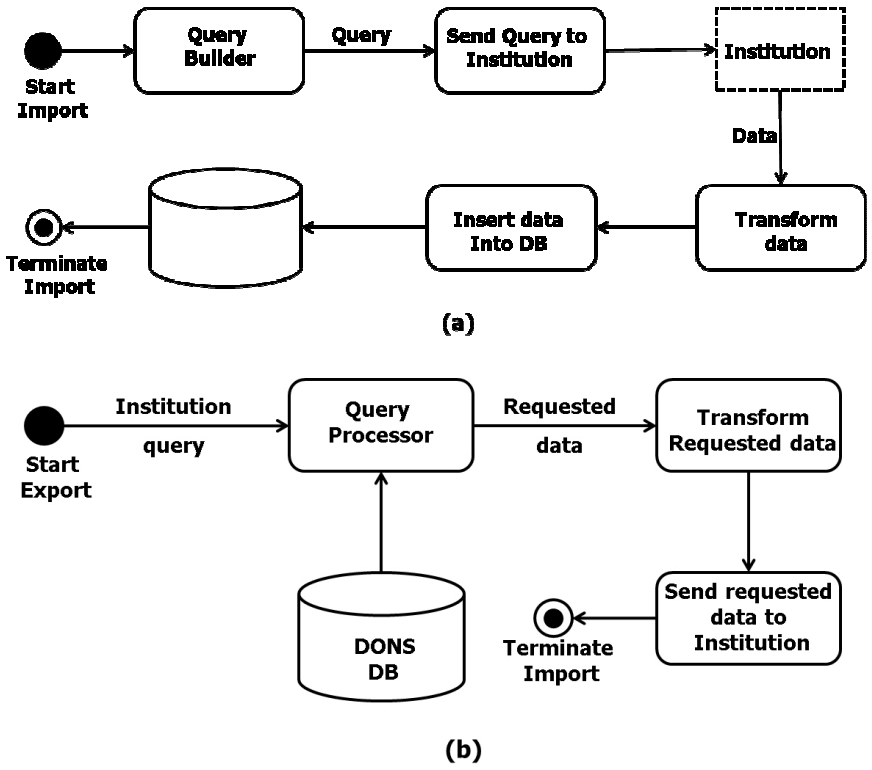


Figure 6 : (a) The Import module (b) the Export module

The Operation Module contains the tools and the personnel that manage, develop, and maintain the DONS. They personnel include system analysts, developers, database administrators, network administrators, system administrators, disease outbreak experts, managers, operators and others.

6. CONCLUSIONS

We have presented a study of disease outbreak detection, monitoring and notification systems that play an important role in assessing threats to public health. The current systems world-wide use different taxonomies and classifications for the detection and prioritization of potential disease outbreaks. Based on our study and analysis of the current disease outbreak systems, we extracted features and functions of typical and generic disease outbreak systems. The paper proposes the generic model for disease outbreak systems that we believe would be ideal for such systems design, development and implementation. The entire effort was also directed towards standardizing the design process for typical disease outbreak systems.

ACKNOWLEDGMENT

The authors would like to acknowledge the support provided by the Deanship of Scientific Research at King Fahd University of Petroleum & Minerals (KFUPM). This project is funded by King Abdulaziz City for Science and Technology (KACST) under the National Science, Technology, and Innovation Plan (project number 11-INF1657-04).

REFERENCES

[1] "WHO | Influenza at the Human-Animal Interface (HAI)." [Online]. Available: http://www.who.int/influenza/human_animal_interface/en/. [Accessed: 11-Feb-2014].

[2] C. A. Bradley, H. Rolka, D. Walker, J. Loonsk, and others, "BioSense: implementation of a national early event detection and situational awareness system," *MMWR Morb Mortal Wkly Rep*, vol. 54, no. Suppl, pp. 11–19, 2005.

[3] J. Gibson, B. T. Karras, and G. S. Gordon, "BioSense 2.0 Governance: Surveying Users and Stakeholders for Continued Development," *Online J. Public Health Inform.*, vol. 6, no. 1, 2014.

[4] B. Cakici, K. Hebing, M. Grünewald, P. Saretok, and A. Hulth, "CASE: a framework for computer supported outbreak detection," *BMC Med. Inform. Decis. Mak.*, vol. 10, no. 1, p. 14, 2010.

[5] "National Notifiable Diseases Surveillance (NNDSS)." [Online]. Available: http://www.health.gov.au/internet/main/publishing.nsf/content/cda-surveil-nndss-nndssintro.htm.

[6] J. S. Brownstein, C. C. Freifeld, B. Y. Reis, and K. D. Mandl, "Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project," *PLoS Med.*, vol. 5, no. 7, p. e151, 2008.

[7] C. Alexander, "Healthmap," *Ref. Rev.*, vol. 28, no. 1, pp. 30–31, 2014.

[8] E. N. Naumova, E. O'Neil, and I. MacNeill, "INFERNO: a system for early outbreak detection and signature forecasting," *MMWR Morb Mortal Wkly Rep*, vol. 54, pp. 77–83, 2005.

[9] A. Zelicoff, J. Brillman, D. W. Forslund, J. E. George, S. Zink, S. Koenig, T. Staab, G. Simpson, E. Umland, and K. Bersell, "The Rapid Syndrome Validation Project (RSVP).," in *Proceedings of the AMIA Symposium*, 2001, p. 771.

[10] B. Y. Reis, C. Kirby, L. E. Hadden, K. Olson, A. J. McMurry, J. B. Daniel, and K. D. Mandl, "AEGIS: a robust and scalable real-time public health surveillance system," *J. Am. Med. Informatics Assoc.*, vol. 14, no. 5, pp. 581–588, 2007.

[11] J. M. Wilson, "Argus: a global detection and tracking system for biological events," *Adv. Dis. Surveill.*, vol. 4, p. 21, 2007.

[12] D. Siegrist and J. Pavlin, "Bio-ALIRT biosurveillance detection algorithm evaluation," *MMWR Morb Mortal Wkly Rep*, vol. 53, no. Suppl, pp. 152–158, 2004.

[13] V. M. S. Zaheer S. Winn J. Perry, "Implementation of the BioDefend? syndromic surveillance system: electronic format versus web-base data entry," *Adv. Dis. Surveill.*, 2007.

[14] M. G. Baker and D. P. Fidler, "Global public health surveillance under new international health regulations.," *Emerg. Infect. Dis.*, vol. 12, no. 7, 2006.

[15] M. J. O'Connor, D. L. Buckeridge, M. Choy, M. Crubezy, Z. Pincus, and M. A. Musen, "BioSTORM: a system for automated surveillance of diverse data sources," in *AMIA Annual Symposium Proceedings*, 2003, vol. 2003, p. 1071.

[16] W. K. Yih, B. Caldwell, R. Harmon, K. Kleinman, R. Lazarus, A. Nelson, J. Nordin, B. Rehm, B. Richter, D. Ritzwoller, and others, "National bioterrorism syndromic surveillance demonstration program," *MMWR Morb Mortal Wkly Rep*, vol. 53, no. 43, p. 9, 2004.

[17] C. C. Diamond, F. Mostashari, and C. Shirky, "Collecting and sharing data for population health: a new paradigm," *Health Aff.*, vol. 28, no. 2, pp. 454–466, 2009.

[18] M. L. Hutwagner, M. W. Thompson, G. M. Seeman, and T. Treadwell, "The bioterrorism preparedness and response early aberration reporting system (EARS)," *J. Urban Heal.*, vol. 80, no. 1, pp. i89–i96, 2003.

[19] J. S. Lombardo, H. Burkom, and J. Pavlin, "ESSENCE II and the framework for evaluating syndromic surveillance systems," *MMWR Morb Mortal Wkly Rep*, vol. 53, no. Suppl, pp. 159–165, 2004.

[20] P. Guglielmetti, D. Coulombier, G. Thinus, F. Van Loock, and S. Schreck, "The early warning and response system for communicable diseases in the EU: an overview from 1999 to 2005.," *Euro Surveill. Bull. Eur. sur les Mal. Transm. Eur. Commun. Dis. Bull.*, vol. 11, no. 12, pp. 215–220, 2005.

[21] "The European Surveillance System (TESSy)." [Online]. Available: http://www.ecdc.europa.eu/en/activities/surveillance/TESSy/Pages/TESSy.aspx. [Accessed: 04-Mar-2013].

[22] N. E. D. S. S. W. Group and others, "National Electronic Disease Surveillance System (NEDSS): a standards-based approach to connect public health and clinical medicine.," *J. public Heal. Manag. Pract. JPHMP*, vol. 7, no. 6, p. 43, 2001.

[23] P. Rolfhamre, A. Jansson, M. Arneborn, and K. Ekdahl, "SmiNet-2: Description of an internet-based surveillance system for communicable diseases in Sweden.," *Euro Surveill. Bull. Eur. sur les Mal. Transm. Eur. Commun. Dis. Bull.*, vol. 11, no. 5, pp. 103–107, 2005.
[24] "SaTScan software for the spatial, temporal, and space-time scan statistics." [Online]. Available: http://www.satscan.org.
[25] "SaTScan Users Guide Report." [Online]. Available: http://www.satscan.org/techdoc.html.
[26] "SaTScan Version History Report." [Online]. Available: http://www.satscan.org/techdoc.html.
[27] C. P. Farrington, N. J. Andrews, A. D. Beale, and M. A. Catchpole, "A statistical algorithm for the early detection of outbreaks of infectious disease," *J. R. Stat. Soc. Ser. A (Statistics Soc.*, pp. 547–563, 1996.
[28] A. M. Pelecanos, P. A. Ryan, and M. L. Gatton, "Outbreak detection algorithms for seasonal disease data: a case study using ross river virus disease," *BMC Med. Inform. Decis. Mak.*, vol. 10, no. 1, p. 74, 2010.
[29] A. H. A.M. Kling M. Grünewald, "CASE User Manual (2012), Algorithms, parameter settings and Evaluation Module." 2012.
[30] M. Kulldorff, R. Heffernan, J. Hartman, R. Assunçao, and F. Mostashari, "A space--time permutation scan statistic for disease outbreak detection," *PLoS Med.*, vol. 2, no. 3, p. e59, 2005.
[31] A. M. Kling, K. Hebing, M. Grünewald, and A. Hulth, "Two Years of Computer Supported Outbreak Detection in Sweden: the User's Perspective.," *J. Heal. Med. Informatics*, vol. 3, no. 1, 2012.

AUTHORS

Farag Azzedin is an associate professor at the Department of Information and Computer Science, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia. He received his BSc degree in Computer Science from the University of Victoria, Canada. He received an MSc degree as well as a PhD degree in Computer Science from the Computer Science Department at the University of Manitoba, Canada.

Jaweed Yazdani is a faculty member at the Department of Information and Computer Science at King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia and a manager of Administrative Information Systems (ADIS) at KFUPM. He received his MS degree and a BS degree in Computer Science from King Fahd University of Petroleum & Minerals.

Salahadin Adam is an assistant professor at the Department of Information and Computer Science, King Fahd University of Petroleum and Minerals (KFUPM). He received his BS and MS degrees in Computer Science from KFUPM. He received a PhD degree in Computer Science from the Computer Science Department at the Monash University, Melbourne, Australia.

Mustafa Ghaleb earned his BS in computer science at King Khalid University (KKU), Abha, KSA in 2007. At the moment, he is pursuing his MS degree in Information & Computer Sciences at King Fahd University of Petroleum & Minerals (KFUPM).

INTENTIONAL BLANK

A COMPONENT MODEL WITH DYNAMIC PROTOTYPE TO TYPE TRANSFORMATION

Efim Grinkrug

Department of Software Engineering, National Research University Higher School of Economics, Moscow, Russia[1]
egrinkrug@hse.ru

ABSTRACT

The paper presents an extension for the JavaBeans component model that enables creating composed components dynamically, at runtime, without code generation. The composed components created can be used immediately for instantiation having their instances used for execution or for further components composition. The dynamic abilities are supported by extended type implementation based on additional superstructure provided with its Java API implementation and corresponding JavaBeans components. Using the component model and base components it provides, the new component composition is performed by building the composed prototype object that can be dynamically transformed into the new composed instantiable type. The component model can be used when implementing user defined types in declarative languages for event-driven models description.

KEYWORDS

Software Components, Component Model, JavaBeans, 3D Modeling

1. INTRODUCTION

The component-oriented programming (COP) is a promising approach for software development in many application areas. It provides many advantages from various points of view in software development process and constitutes the main idea of component-based software engineering (CBSE).

The idea of component-oriented programming is to create software products from composing parts – the idea that is at the base of the vast majority of technologies in other engineering areas. Composing parts of software products named components are created and used in accordance with a component model that defines what a component is, and what and how can be composed with that component [1].

In this paper we discuss some considerations about enhancing JavaBeans-component model that is widely used in Java software development. Java platform has become the most widely used object-oriented environment for software development starting from the time it was introduced [2]. The JavaBeans component model [3] initially was claimed as "the only component model for

[1] This work is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE).

Natarajan Meghanathan et al. (Eds) : ACSIT, FCST, ITCA, SE, ICITE, SIPM, CMIT - 2014
pp. 71–87, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4607

the Java machine" and it is widely used in Java-programming up to now, while currently there are many other, popular enough component models for Java-platform – both universal and domain specific (e.g. [4]).

This work is rooted in the practical experience earned from using JavaBeans-components to implement virtual reality modeling system with 3D-graphics support entirely in Java [5], and from developing instrumentation systems for wireless sensor networks modeling and commissioning (ZigBee Standard [6]). While the two kinds of applications are very different in nature, they both share the event-driven behavior model and benefit from component-based system implementation. We expect therefore that our conclusions may have some general application.

We begin with the problem statement, discussing how components are used and what the shortcomings of the JavaBeans component model are in Section 2, explaining our approach to the component model enhancements. Then we describe our component model with dynamic prototype to type transformation and its implementation principles in Section 3. After that in Section 4 we discuss how that component model can be used in application areas mentioned above. We find its place in the wide variety of component-based software technologies in Section 5 and conclude with future work direction.

2. COMPONENTS AND COMPONENT COMPOSITIONS

Since the time when component-oriented programming was recognized as base of software engineering [7], many definitions of a component were introduced. The most popular are following:

- "A software component is a unit of composition with contractually specified interfaces and explicit context dependencies only. A software component can be deployed independently and is subject to composition by third parties" [8];

- "A component is a software element (modular unit) satisfying the following conditions:

    1. It can be used by other software elements, its 'clients'.
    2. It possesses an official usage description which is sufficient for a client author to use it.
    3. It is not tied to any fixed set of clients."[9];

- "A [component is a] software element that conforms to a component model and can be independently deployed and composed without modification according to a composition standard."[10].

A software component model mentioned in the last definition, should define the syntax of components (how they are constructed and represented), the semantics of components (what components are meant to be), and the composition of components (how they are composed or assembled) [1].

We are specifically targeting the Java-platform and JavaBeans component model because this is the most popular object-oriented software development platform and component model.

2.1. JavaBeans components and their usage

The popularity of the JavaBeans component model is based on its relative simplicity, wide range of abilities it provides and corresponding tools to demonstrate that abilities. The first sentence of the specification is: "The goal of the JavaBeans APIs is to define a software component model for Java, so that third party ISVs can create and ship Java components that can be composed together into applications by end users" [3].

The visual composition of pre-existing components is at the origin of the JavaBeans component model and it is stated in the initial definition: "A JavaBean is a reusable software component that can be manipulated visually in a builder tool." And the abilities of JavaBeans components to be manipulated visually make JavaBeans component model attractive to use when developing modeling and visualizing applications with interactive abilities support. While JavaBeans components can be used as class libraries in traditional software development process, we are mainly interested in their dynamic composition abilities and corresponding tools support.

From an external view, JaveBeans components can communicate using four kinds of ports: methods, properties, event sources and event listeners.

The notion of method of a component is directly bound to the notion of method in the component implementation language: callable method must be present in the component implementation class.

The notion of property, from a client point of view, can be used for getting its value, setting a new value, or for binding with it to be notified with events whenever it changes its values.

Event sources generate events of certain type, while event listeners receive the events. Event sources provide operations to connect and disconnect listeners, supporting event-driven behaviors in applications composed of components.

In CBSE, composition is a central issue, since components are supposed to be used as building blocks from some repository and assembled or plugged together into larger blocks or systems.
JavaBeans-component model can be considered from a general, idealized component model point of view [1, 11, 12], that is expected to have three stage lifecycle: 1) design stage, when components are designed and developed at source code level of their implementation language (i.e. in Java, in our case) and possibly compiled into binaries; 2) deployment stage, when binary component representations are supplied into composition environment, and 3) runtime stage, when the components are instantiated and executed in a running program. (Actually, what we are going to do is remove borders between the stages.)

At the design stage, JavaBeans components are implemented by Java classes that satisfy simple rules (JavaBeans design patterns). Essentially, JavaBeans component is a Java class instantiable in any context (public class having public constructor with no arguments) and having support for persistence (to save and restore states of its instances). These classes are distributable for reuse in binaries (Java byte-codes) along with other classes and resources used to implement them. (A JavaBeans component, therefore, is a Java class that implements, possible using other Java classes, a type of objects, or instances, it creates. These notions are often erroneously mixed in literature hiding class-based object-oriented nature of the component model.) At the design stage, components composition can be performed using Java programming technology chain, resulting in components byte-codes produced by compiler that actually uses the composing components as classes from class libraries.

At the deployment stage, there is some composition environment that supports JavaBeans composition visually. The JavaBeans API Specification was supplied with Bean Development Kit (BDK [13]) that contained the composition environment prototype, the BeanBox, to illustrate interactive and visual composition support for JavaBeans components. That approach has been integrated in various IDEs, but the BeanBox from BDK is still used having more dynamic abilities than the IDEs provide. At that stage, JavaBeans component are instantiated in the composition environment and their instances are combined together to provide the composite functionality required. In the BeanBox components instances can be combined interactively and dynamically using all kind of ports to connect them together. In contrast with the BeanBox, IDEs use more static approach: they do not support direct component instances interactions, but help generating the source code for it, that gets compiled. In that sense, we can consider that IDEs as design stage tools that use composition ability to automate some code generation. When component instances are composed in the BeanBox tool the whole composition can be saved (serialized using binary or some other format), and restored (de-serialized) later.

At the runtime stage, components that were created at the design stage are instantiated and executed. But some of them, acting as the containers can use serialized composite objects stored at the design stage by de-serializing them inside their instances. In any case, we see that our composition abilities could not create a new component without compiling it. We cannot produce a new component in BeanBox like interactive tool: components must be classes, and classes may be created by their byte-codes generation only.

2.2. Shortcomings of the JavaBeans component model

An advantage of the JavaBeans component model is in its simplicity and usability: it is not based on a set of specific interfaces to be implemented by the components to work, as e.g. OSGi does [4]. (The later enhancement of the JavaBean component model that introduced BeanContext related features in that style is much less popular.)

There are many other engineering areas that use component-based technologies. In most of them instruments used to create their basic components differ from instruments used to build composed components from them. Usually, composing technologies are simpler (e.g., so called “screwdriver production” for computers).

In software, XML syntax is often used to declaratively define and create a composite object from components instances; that XML-based instance composition is much more simple technology than the compiler used to create the components (classes) themselves. We used that technology in [5] to compose 3D-scene from JavaBeans-components instances in the same way that is used, e.g. in XAML [14], but we used the VRML-parser of our own. That composite object composing instruments are significantly simpler than Java compiler and other tools required to produce JavaBeans-components. Moreover, building a composite object can be done interactively, as the Bean Box demonstrates for JavaBeans-components.

When we build a composite object from JavaBeans-components using some declarative language, or when we build the composite object from JavaBeans-components in a composer tool (similar to the Bean Box), we discover that the component technology we use is not “self-closed”.

Our source components are types of objects implemented as classes (compiled and supplied in form of instantiable types libraries.) As a result of composition made with a parser or interactively in a tool we get some composite object composed from supplied components instances or an instance of some predefined container type filled with our composite object.

A composite object we have created may be a workable object (e.g. it can be some GUI implementation or a scene made from VRML subset implemented in [5]). We can save and restore that composite object, we can clone it (either directly or by serialization/deserialization), but we always deal with it as an instance of some (predefined) container type, but not as a new (component) type, that can be used the same way as the components it was composed from. Having the components set in hand initially, we cannot produce some new composed component that can be instantiated like a type, instead of cloning it like composed object.

It is graphically visible while manipulating in BeanBox. First, a library with compiled JavaBeans-components classes is loaded into BeanBox repository (ToolBox). In the BeanBox container instance we instantiate components (types) that are dragged & dropped into it from the ToolBox. The instances just created are depicted inside BeanBox container instance, having their property values depicted in the Properties panel, where some values can be edited. Further on, we can link the instances together by their references, assigning one instance as a property value for another, or bind them by events. All that BeanBox provided abilities correspond to the JavaBeans Specification [3], that the tool was aimed to illustrate. We just want to highlight the lack of ability to enrich the ToolBox, filled with components we used initially, with the result of our manipulations with them during the composition.

By simple manipulations with components, we cannot define composed component that can be manipulated the same way. It is important to note that we would like to have that ability by simple means, i.e. by means other than the means that were used to create initial components. (Compare: electronic LSI components are created by more sophisticated technology than putting them on a PCB together.)

Initially, basic JavaBeans-components are represented as compiled classes that were conformed to the JavaBeans-component definition and the JavaBeans design patterns while they were coded. Creating JavaBeans-components is usually done statically as the result of standard software developing process, with packaging their byte-codes into a Jar-archive. We say “usually” because some dynamic code generation is possible (either by creating source code with its compilation on the fly, or by mean of immediate code generation using specific libraries for that purpose that are available). Both variants can be used at runtime (dynamically), but we do not consider them as “simple”.

Basic JavaBeans-components are created statically, by “hand-made” programming. When we use “hand-made” (or even automated to some extent) programming while creating new composed JavaBeans-component, using existing JavaBeans-components just as class libraries, then we apply the same technology that was used to produce basic components initially.
While that approach can be (and is) widely used (and, being sophisticated, can provide more effective result), it seems that a way to create composed component by means other than that of basic ones, can give some advantages – both technical and ideological.

From the technical point of view, the ability to define composed component interactively (and without compiling, dynamically) is attractive by simplicity of its usage – just as putting LSI circuits on the PCB (while we can have VLSI circuit for all of them later).

From the ideological point of view, we can talk about creating some “higher virtualization level” with its new type system, incorporating the lower level abilities to create and support base components functioning (i.e. on top, above the Java-platform). Base JavaBeans-components are represented as instantiable types (implemented as Java classes). In case we want “by means of simple manipulations” (i.e. without code generation) to get new composed component, we need to generalize a component (type) notion so that we can, when creating new composed types, use basic and composed components (types) in the same manner (equally). Along this way, compiled

(or hardcoded) components and composed components are just two kinds of type implementations at our "new virtualization level". An idea of that higher level implementation is to use JavaBeans-components to add a superstructure for a component type system.

JavaBeans-components model is not "logically closed" in ideology point of view because of the following consideration. The components are supplied in form of classes (implementing abstract data types) to be instantiated, and the classes were designed in accordance with class-based object-oriented programming paradigm. When composing their instances in some composing environment (e.g. in the BeanBox), a composite object is created that is not an instance of some composed type; it is just a content of the pre-existing container instance it was built inside (i.e. the content of a BeanBox container instance). In JVM, types may come into existence only by loading byte-codes of the corresponding classes. That composite object can be cloned (serialized / de-serialized, etc.), but its usage in that way corresponds rather to prototype-based programming paradigm (we have no class created for it during the composition.) When performing a components composition without its code generation we have to use the composite object as a prototype, thus substituting initial programming paradigm by another. We cannot produce the composite entity of the same nature as we had initially to compose it (without having to use the same technology as we used to produce initial components.)

Note though that in electrical engineering, for example, we can build a functional unit from its components using much more simple technology than technology used to produce them (and it is a matter of integration density.) We can draw its scheme as the composed unit type description and put it into production for future reuse. In case we could be able to use microcircuits with more density, we could implement the composed unit in one chip using chip manufacturing technology that we used before for our components manufacturing. But meanwhile we just can use soldering-iron with wires. Roughly to say: that's why Intel develops sophisticated chips while others use to wire them together in more simple manner, placing them wired on PCBs for different purposes.

What we are looking for is a relatively lightweight, dynamic, interactive composition technology to create new composed components without their codes generation.

3. TYPES AND COMPONENTS

Generally speaking, when coding in Java we can only write data type definitions. These type definitions are compiled into class-files with byte-codes that appear to be the runtime types in JVM upon class loading. Inside JVM, the types are represented by objects of type Class that are produced by class-loaders. All objects classes are immutable (static fields in classes can be mutable, but that style of programming it is not considered as good one), and they are not JavaBeans-component instances: they cannot be created by default-constructor of their class (Class); they are created by some ClassLoader instances having their byte-codes as input.
If we want to cross the boundaries of the runtime type creation ideology of the JVM, we need to define some superstructure over the JVM that has its own notion of object type. Since we are going to have that superstructure as a kind of JavaBeans component model extension and to implement it using our JavaBeans components (in component-based manner), we name it as BeanVM. The type notion in BeanVM should allow different type implementations: both types created from JVM classes loaded by class loaders and types created by our composition procedure designed specifically for that purpose. It means that BeanVM types can be classified as hardcoded-types and composed-types; the former comes into BeanVM from loaded JVM classes, the latter is produced inside BeanVM itself by composition.

Some of the BeanVM types are components (in BeanVM perspective, no matter how they are implemented). For now, we define the types that can be instantiated without any information provided from outside (i.e. from their instantiation context) as components (and we'll try not to mix them with the component instances, as it often happens in JavaBeans related texts). That definition recalls JavaBeans-component definition for JVM that must be instantiable using its default-constructor.

Like JavaBeans-components, BeanVM components can have named property sets with typed property values (i.e. having property value types in terms of BeanVM types). All BeanVM types are instances of a Type type (like all Java classes are instances of the Class class). Type type is implemented in Java by (abstract) class Type providing all type related operations for BeanVM. All BeanVM types are implemented as immutable objects – the information they contain does not change after they have been created. We are intentionally following the class-based object-oriented principles and trying to retain them when preforming components composition (in contrast with JavaBeans component model, as it was mentioned above).

To access BeanVM functionality implemented in Java we provide the BeanVM API that we will discuss below when needed.

3.1. Type representation

Any BeanVM type is represented by an immutable instance of that type implementation java class that is inherited from the abstract class Type and exposes the following type information:

$$\text{type} = \{\text{typeName}, \text{interfaceType}, \text{implementationType}\}, \quad (1)$$

where typeName provides the name of the type, interfaceType and implementationType reflect the type interface and implementation, correspondingly.

We expect to deal with types that were not compiled and, therefore, have no type specific methods to be invoked by any method calling mechanism of JVM. The only ports that can be used to communicate with our BeanVM component instances are properties that are supported by special BeanVM API for component instance property access.

An instance of the InterfaceType class is a set of PropertyType objects describing the interface properties:

$$\text{propertyType} = \{\text{propertyName}, \text{valueType}, \text{accessType}, \text{defaultValue}\}, \quad (2)$$

where popertyName is the name of a property. The valueType is the fixed BeanVM type that is used for type control when performing new value assignments: each BeanVM type can verify whether the given object belongs to its domain of values. The accessType defines operations to be applicable for the given property. A property can be readable (R), writable (W), bound (B). Indexed property can be indexed-readable (Ir) and/or, indexed-writable (Iw) as well. The defaultValue for a property is available for component properties only.

The implementationType part of the type information provides information on internal type implementation that is different for hardcoded and composed types.

Any object our BeanVM can deal with when it is functioning has its BeanVM type, and each BeanVM type can verify whether the given object belongs to the set of values of this type. That is used to implement value type control for property assignments. If an object was instantiated by BeanVM type, it knows that type upon creation. If an object was created by JVM class

instantiation and that class defines a component property value type, then that JVM class has been wrapped by the corresponding BeanVM type that delegates the object type check back to its implementations class. BeanVM supports BeanVM type instantiations and property access for their instances; all other activities are performed beyond the scope of its responsibilities (i.e. inside components behavior logic).

3.2. Hardcoded Types and Components

The hardcoded types in BeanVM are types implemented by loaded JVM classes. Hardcoded components are hardcoded types that are components. Here we consider how to represent JVM classes as BeanVM types and how to represent JavaBeans components as BeanVM components. The source information for hardcoded type definition in BeanVM is provided by corresponding class-object, loaded in JVM. We implement a primitive to map Java class to type by implementing the following method in Type type implementation class (as part of BeanVM API):

```
public static Type Type.forClass(Class someJavaClass);                    (3)
```

and we have our TypeLoader's hierarchy that mimics that of ClassLoaders (since each Java class is identified by its' name and the class-loader instance that the loaded class can provide). When implementing our TypeLoaders, we enhance the possibilities to create types in BeanVM: we can create types not only from loaded classes, as the primitive above does, but in some other ways described later.

We create a type for a given Java class once only, at the first attempt to get it; all subsequent calls will return existing hardcoded type from the type-loaders' (HashMap) table.

The hardcoded type creation procedure is based on reflection mechanism and standard JavaBeans introspection. The set of PropertyType objects is created based on an array of PropertyDescriptor objects that are provided by JavaBeans introspection procedure when introspecting the class (JavaBeans introspection can deal with any Java class, not only with JavaBeans-components). The PropertyType object (see (2)) gets the propertyName extracted from PropertyDescriptor object according to the JavaBeans design patterns [3]. The valueType is a BeanVM type created for a class of the property value by means of Type.forClass()-primitive (3). The property value class and the information to define the property accessType are available in the PropertyDescriptor object as well.

The InterfaceType object with the PropertyType objects array inside can be obtained from any Java class, but we are interesting in JavaBeans-components classes and their property value classes (that we wrap by our types). BeanVM-types for other Java classes are out of interest for the BeanVM.

The ImplementationType object for the hardcoded type is just a wrapper of the source class that implements the type in BeanVM (it reflects the old idea that any problem in Software Engineering can be solved by additional indirection).

The defaultValue in the propertyType (2) is needed in interface type for components only, and it is obtained only from our JavaBeans-components having our specific implementation. For all other (third party) JavaBeans-components we provide our hardcoded component-adaptor (that wraps any extraneous JavaBeans-component to be used in BeanVM environment).

Our hardcoded components are JavaBeans components having their implementation class inherited from our Bean class (directly or indirectly). The Bean class provides BeanVM API to

the internal implementation of the Bean component instance and its properties. All our hardcoded components are JavaBeans components that have BeanVM type instance implementation wrapped inside. The property access methods of our JavaBeans component implementation use BeanVM API and delegate to the wrapped instance implementation.

Here is a code snippet of the Bean class:

```
 public class Bean extends BeanVMObject {

      final Instance thisInstance;

      public Bean()   { thisInstance = Type.implementBean(this); }
      Bean(Type type) { thisInstance = type.createInstance(this); }
      // …
      protected final void initPropertyValue(String propertyName, Object
initValue) {…}
      public final Object getPropertyValue(String propertyName) {…}
      public final void setPropertyValue(String propertyName, Object newValue)
{…}
      // …
 }
```

The BeanVMObject is an abstract class that forces all BeanVM objects classes to provide their type getter method. Bean class has two constructors: public default constructor and package private constructor with a type argument. The default constructor, that is unavoidably executed when instantiating any Bean class ancestor (any our JavaBeans component), passes this component instance to its implementation factory method - Type.implementBean(this), that returns the instance implementation. The factory method, first, gets the type for this class and, second, lets the type to create the instance implementation. Hence, any Bean class ancestor it its constructor context is ensured that its implementation instance is already created and all its properties are implemented in it as appropriate. The concrete ancestor component is able to initialize its properties with their initial values using initPropertyValue() – method, and use the property value access methods (setPropertyvalue(), getPropertyValue(), etc.), that all delegate to the implementation instance.

Note, that initPropertyValue()-method works only once for a given property during the component type creation, when the component is instantiated for the very first time to collect property default values only (and store them in PropertyType objects (2)). When hardcoded type is created, the initPropertyValue()-method for its instances has no effect: they are already initialized with their default values when they are created in the implementation instance. Hardcoded component instance implementation contains only its mutable properties; immutable property values are stored in the PropertyType objects and shared by all instances of the type. The mutability of the property is determined by its accesType: the property is mutable if it is writable or bound (i.e. can change its value externally or internally, notifying about that).

Internal BeanVM API implementations of the property value access methods control the property accessType. All setPropertyValue()-methods control the property value type. All that control is implemented dynamically, and RuntimeException is thrown when violation occurs. To speed up the property access we provide methods in BeanVM API that translate the propertyName into an index of its PropertyType object in the component interface type along with the variants of property access methods that use the index instead of the propertyName. The concrete hardcoded component implementation can get its property indices in static initializer of its implementation class and use them when implementing its property access methods. In JVM perspective, all property values are stored internally in the BeanVM memory cells allocated for them and declared using the root of Java types hierarchy – java.lang.Object. That requires the explicit cast

to the property value type to be added when coding the concrete property getter using our getPropertyValue()-methods that return Object (some overhead with Java primitive types could be minimized by providing specific BeanAPI methods for them, but we omit these details here).

Note that our hardcoded components are JavaBeans components that can be manipulated visually in a builder tool (by definition) - like the BeanBox. It means that we keep all advantages of the JavaBeans component model, and we can use our hardcoded components, e.g. like we did before when implementing our 3D modeling framework by means of JavaBeans [5]. We could build a composite model from JavaBeans component instances and observe the model behavior, having placed it inside some predefined component container instance. But without generating byte-codes and loading them into the JVM, we could not create new composed component and place it into the ToolBox of the BeanBox (in visual terms), while all the components we used to create it are already there.

Now our goal is to make it possible: we have to deal with composed types and extend the builder tool (like the BeanBox), accordingly, to deal with them as well.

3.3. Composed Components

We have to define composed type in our BeanVM so that it does not correspond to some java class created and loaded especially for it, and can be defined dynamically, at runtime. The composed type that is instantiable without any arguments provided for it is our composed component. It has its interface definition and internal (hidden) implementation that we have to define using other components as the building blocks.

When defining composed types we comply with the class-based object-oriented approach. Our type definition is immutable structure that is able to create instances of that type instead of cloning them.

Like any type in BeanVM, composed type has its type name, the interfaceType and the implementationType (1). The interface type for the composed type is represented the same way as for a hardcoded type and is described by the array of PropertyType objects (2). But the implementationType is entirely different. It is composed by its composing types.

External communication with a composed component instance is performed using its interface properties access by means of BeanVM API mentioned above: any instance of a composed component is implemented as our Bean class ancestor instance that gets its composed type as an argument of the package private constructor (see Bean class code snippet) called when the composed type instantiation is performed. So, public BeanVM API for property access support works equally for any instance of any BeanVM type (no matter whether it is a hardcoded type or a composed type).

When creating an instance of a composed type, we create the instance internal implementation that, in this case, includes not only the mutable properties cells, but the implementationType instance. That implementationType instance is created as the result of composing graph traversal procedure where composing types are used as context-dependent refinements of composing components that are met during the traversal and instantiated to provide the composed type instance implementation.

The composed type interface defines the set of properties for the type instances. For each instance of the composed type, the instance internal implementation should be able to communicate with the instance external environment though the interface properties. The composed type instance

behavior is implemented by its composing components instances that express their behavior by their property value changes – as any our component instance does. Hence, to link the interface with internal implementation of the composed type instance, we need to link some interface properties to some properties of internal composing components instances. That composed type interface and implementation properties link can be organized basically in two different ways: by bidirectional property bindings and by interface property sharing. The second way is less expensive from the efficiency point of view, since there is no need to transfer property values.

Since we are supporting class-based object-oriented approach, we delegate our BeanAPI operations to the corresponding type object that implements them using the concrete instance reference (an internal instance implementation). When implementing an interface property access, we delegate to the interface property type that knows how to find and access the property value having the internal instance implementation. To implement that, we have PropertyType class sub-classed according property implementation categories: Immutable, Mutable, Bound and External. Each category (the PropertyType subclass) is responsible for corresponding property implementation and access. All mutable (and bound) property values are stored in an array of objects that is allocated by internal instance factory. All immutable property values are stored in their types and shared by design. PropertyType.External is used to share the enclosing type interface property with the given one. The PropertyType.External delegates the property access to the given property of the enclosing instance of the composed type. That enclosing instance is known as the composing type instantiation context.

We have mentioned previously that our components are instantiated without passing any arguments for their default constructor, and that was stated in the JavaBeans component definition. All our hardcoded components are JavaBeans components that are instantiated by their default constructors (without passing any context-related information). We do not use the BeanContext-related extensions of the JavaBeans Specification as well. While JVM provides some tricky way to get some information about constructor calling context using invocation stack trace, we do not rely on its usage. We use the fact that we are working in BeanVM (instead of JVM), that implements its own component instantiation primitive: createInstance(), that calls Java default constructors internally, and that default constructor delegates the internal instance implementation back to the BeanVM implemented instance factory method (see Bean class code snippet above). That factory method is able to use BeanVM primitive invocation stack to get the instantiation context (if it exists) and pass it to the internal instance implementation.

All the information needed to implement and access composed component instances is contained in their composed types implemented as immutable data structures that expose property value getters for reflection purposes (as implemented by abstract Type class and its type-specific subclasses). Now we'll describe how that structure can be created dynamically.

3.4. Prototype to Type Transformation

The natural way to create an instance of immutable data structure is to use its builder object that is mutable and can collect all relevant information to be provided for immutable objects initialization (i.e. to use the Builder design pattern) [15].

Naturally, we create the builder object in component-based manner. We use our specific hardcoded components to compose a prototype object that can be used as a source for creating an immutable composed type. These specific hardcoded components can be considered as meta-components (since they are components to construct components).

The prototype object is an instance of the Prototype component having the bound properties “name”,”interfacePrototype” and “implementationPrototype”. The type of property “name” is String. The “interfacePrototype” property can refer to an instance of InterfacePrototype component that is used to build the interfacePrototype object. The “implementationPrototype” property refers to an instance of ImplementationPrototype component to compose the implementation object.

The prototype to type transformation is performed by BeanVM primitive (here, for short, we omit the details concerning nested types that are enclosed in outer type implementation):

public static Type createFromPrototype(Prototype prototype), (4)

that returns new type in case when nothing prevents that from being happened (i.e. there are no namespace conflicts and other validation faults).

Generally speaking, we can create BeanVM type just having set the name only (i.e. having empty interface and empty implementation), that is similar to an empty JVM class (e.g. class Classname{}). Type with some interface part present, but with an empty implementation part can be used to instantiate its property set without any internal behavior linked with them (like a structure or record). Type with no interface part represents a separate executable entity type (e.g., “a scene” with its separate behavior).

The instance of the InterfacePrototype component exposes the set of property prototypes. The property prototypes serve two goals: 1) they are used as exposed handles to control the prototype implementation behavior, and 2) they provide source information to create the property types during the prototype to type transformation.

The instance of the ImplementationPrototype component exposes the set of component instances that compose the prototype object implementation – the set of composing prototypes. The compound prototype object can be built by defining graphs of two kinds: reference graph and events graph. The reference graph is defined by using some object as a property value (or as an element of indexed property value) of another. When an object is used as a property value of (i.e., is referenced by) several other objects, it is shared by them. The events graph is defined by component instances that bind the source bound property to the target to propagate the property change events (in correspondence with JavaBeans design patterns).

One of the main principal issues when defining composed components is defining a way to separate the interface of the component from its implementation while providing the way for them to intercommunicate. We define the interface in terms of properties that are prototyped using typed variables, i.e. objects having “value’-property with a given value type. We can use any BeanVM type as a value type for our typed variables (and as property value type after the prototype to type transformation). In fact, that is one of the two existing use cases for BeanVM types; another is BeanVM type instantiation. Having a type to be used as a value type, BeanVM creates (synthesizes) the BeanVM type for the typed variable automatically, assigning the synthesized name for that synthetic type and granting the full access rights to operate with its instances (i.e. read, write, bind their property “value”). In case the value type is an array type, we create BeanVM type for indexed typed variables that will serve as indexed property prototypes (having indexed access provided). That approach is similar with JVM array classes’ support (based on a given class of elements).

Property prototypes, implemented using typed variables and exposed through the interface prototype, should be linked with property prototypes of some composing prototypes inside the

prototype implementation. It can be done either by means of event routing graph (using bidirectional event routing for each link), or by sharing the property prototypes by means of reference graph. We use the latter approach (the more effective one) and provide support for the property prototypes sharing.

When a component is instantiated inside a prototype container, it gets its prototype-oriented internal implementation from the instance implementation factory. We use that instance as a composing prototype. The prototype-oriented implementation works in prototype-based style: it does not delegate the property access methods to the type of the instance (since it does not exist yet), but implements them using the prototype-oriented instance implementation itself. In that internal implementation, a composing prototype instance is represented with an array of property prototypes (in contrast with an array of property values, as it is implemented for component instances created outside the prototype container context). That additional indirection supports sharing the property prototypes of the interface prototype by composing prototypes instances in the prototype implementation (in case the value types are compatible). In principle, that sharing is similar to reusing the component instances inside the reference graph of the prototype implementation.

Each property prototype is supplied with access prototype instance that can be used to narrow the property prototype access rights by denying some of them for the given usage context. After the prototype to type transformation, narrowed access rights will be stored as accessType in the correspondent property type, as was mentioned in (2).

The composed prototype object is tunable and operational. Its' composition can be performed by visual manipulations with the correspondent tools support. When it is done, it can be used to produce the immutable BeanVM type. During the prototype to type transformation all prototypes are used as sources to create the corresponding types: property prototypes are transformed into property types with their access types created from the access prototypes, interface prototype is transformed into the interface type, composing prototypes are transformed into composing types, altogether transformed into the implementation type, and finally resulting in the composed type – or component - produced.

That newly created composed component can be instantiated as any other component –either using more efficient class-based object-oriented internal implementation, or using the more flexible prototype-oriented internal implementation (having been instantiated as a composing prototype in the prototype container).

Composed component can be serialized/deserialized (e.g., using some text format). To read the composed type from a text file by TypeLoader, we provide Type.forName(String typeName) primitive that loads the type by its name like JVM Class.forName(String classsName) loads classes. When looking for the source of type by name, we first try to load hardcoded type with the given name (using Class.forName()), if it exists, then the serialized type file to be parsed. The parser, essentially, reads into a prototype object and performs the prototype to type transformation.

4. SAMPLE APPLICATIONS

Both VRML [16] and X3D [17] Standards define a sets of elements that constitute a scene to be depicted using 3D and/or 2D graphics. The scene model in memory is represented by the directed acyclic graph (DAG), consisting of node instances whose types are predefined according to that standards. Node instances contain 'fields' whose values, in common, define the (scene) model state, and can vary, in event-based manner, while event propagations are performed among them.

For all the base predefined nodes, their field value types are statically known and defined by the Standards. Directed acyclic graph can be traversed with the result of earning some context-specific information, which, in combination with nodes field values, defines the visual presentation of the model that is rendered. The model behavior is defined by changes that happen in the node graph during events propagation, with events carrying data on their field value change from node source to the node target, and by reactions in receiving nodes, that, in their turn, can change their state and fire events. VRML and X3D Standards define base node types with their field types (and their meanings/semantics), the constructs to define event routing graph, standard scene access interfaces and so called 'profiles' – sets of independently distributable standardized functional components to support various (extendable) modeling abilities.

While the Standards do not expect their Java-implementation (and up to date they were not implemented in Java properly, despite of some attempts), their implementation using JavaBeans component model appears to be pretty natural; moreover, using JavaBeans-component model we can avoid some limitations of the Standards, that were not initially designed to use neither dynamic abilities of the Java platform nor JavaBeans-component model for it. JavaBeans components usage to implement subsets of the Standards with some additional features were discussed in [5], that shows how parallel compositions, behaviors and 3D-visualization of the models can be organized in standard JavaBeans-container (using BeanBox) along with standard JavaBeans-components.

Using the presented component model, we can implement user-defined nodes definitions that are described by the PROTO construction of VRML and X3D, in class-based object-oriented manner. It was not possible using JavaBeans component model, as we have discussed above.

The PROTO construction in VRML (or X3D) essentially provides the language features to describe new, user-defined node type, having supplied its name, its interface in terms of its fields (as that is done for the base, predefined node types) and internal implementation, that is similar to a scene graph. Both scene graph and PROTO implementation graph can contain nodes of any type – both predefined and user defined – provided they are defined prior to their usage in a graph by means of PROTO construction. Fields from the interface of a PROTO are bound to fields of its internal implementation nodes using special VRML syntax construction – "IS". (In our model that separate construction is not needed – it is essentially the same as "USE" construction to share interface property).

In practice the PROTO construction is implemented as a cloneable prototype, as its name implies, or just as macros definition to be substituted by parser. The whole scene description is not considered as a type definition to be instantiated; instead, it is just a prototype instance serialized in VRML or XML formats.

Using the proposed component model we expect to simplify and optimize software systems for the subject areas, where declarative languages (like VRML or X3D) are used to describe 3D-models with event-driven behaviors, and we will be able to support new abilities of these systems to benefit from.

The ability to build a flexible prototype object from components that can be transformed into composed instantiable component with context-dependent optimization may be useful in various applications. For example, that ability directly corresponds to the wireless sensor network commissioning task [6]. These networks can be described using a graph of nodes communicating by radio in event-driven manner, and are built from standard components to be tuned for a given application and environment. The concrete ad hoc network prototype (or a model) can be designed, then that typical network component can be created, and its instances, containing concrete network nodes settings, can be used for commissioning (e.g., over the air).

5. RELATED WORK

Component-oriented programming and its usage in software development formed a component-based software engineering – a branch of software engineering that has its long term history, valuable results and issues to be solved. There are many publications attempting to provide definitions for a component (we have referred to three most frequently cited [8, 9, 10]), investigating different component models with their taxonomies [1], and considering various aspects of component-based software development for different application areas [8, 10, 11]. Detailed overview of that works is beyond the scope of this paper.

We are concerned with issues we know from experience earned while developing component-based software system to implement 3D modeling and visualization [5]. This is a large area as well, and we will narrow our scope by platform in use, since there are popular platform dependent component models we have to leave aside, e.g. [18].

The most widely known 3D modeling software for Java platform is probably Java 3D [19] library, initially developed by Sun, then by the open source community. That heavy-weight library was not designed with component-oriented approach in mind, and attempts to implement VRML browser using Java 3D library were not finished. The library was used in X3D Standard development by web3d.org community [20], but currently Java 3D library is deprecated.

There are several 3D modeling and visualization tools for Java platform developed by commercial companies and universities, e.g. [21, 22], but they do not explicitly use a component-based approach or use their proprietary component model [23].

The PtolemyII project [24], having its long pre-Java history and covering, among others, the application area of our interest, has been moved to Java platform and use its component model reflected in Moml [25] specification. As far as it can be seen in publications, instantiation of declaratively defined composed type in implemented by cloning (i.e. in prototype-based manner).

At the time of highest VRML popularity, there were publications on extending it in object-oriented manner [26, 27]. Now we observe new wave of interest to these declarative languages that can be seen in new VRML/X3D compliant product – Instant Reality [28]. But that product does not extend the standards in component-oriented direction.

Both VRML [16] and X3D [17] standards specify Java authoring interfaces (for external model access and for internal scripting in Java). All that specifications have no relation with component models for Java-platform, while that functionality could be readily provided by component-oriented design.

There are some works on component-based software evolution [29, 30], but they are based on code generation. We can consider code generation as a feasible way to translate BeanVM components into JavaBeans components (that is similar to the Java just-in-time compiler translating from virtual machine level to executing machine level).

6. CONCLUSION

Both hardware and software architecture histories demonstrate the evolution with increasing dynamics abilities. Java platform, as the most popular software development environment, and JavaBeans component model, the most popular for that platform, provide the means to evolve in that direction as well. In this paper we have proposed a component model that demonstrates prototype-based and class-based programming paradigms interrelations. The prototype is a

flexible, mutable object sample that is built, lives and evolves in component-based manner; when it gets some desirable state of evolution, or just eventually by some stimulus, its genetic code is extracted into the type and saved for reuse by next generations (that reminds an everyman untaught view on genetics in the natural reality).

The component model proposed is a kind of dynamic extension for the JavaBeans component model, with the extension supported by specific JaveBeans components. Correspondingly, as for JavaBeans component model, we need the extended builder tool that utilize and demonstrate extended abilities of the component model. Developing that tool is the goal of our future work.

REFERENCES

[1] Kung-Kiu-Lau, Zheng Wang (2006) A Survey of Software Component Models (second edition), School of Computer Science, The University of Manchester, Preprint Series, CSPP-38.
[2] Tiobe community index. www.tiobe.com
[3] Sun Microsystems. JavaBeans API Specification. 1996.
[4] Open Services Gateway initiative – OSGi Alliance. http://www.osgi.org/
[5] E.Grinkrug. “3D Modeling by means of JavaBeans”, Proceedings of the 12th international workshop on computer science and information technologies, CSIT’2010, Moscow – Saint-Petersburg, Russia, 2010.
[6] E.Grinkrug, “Software Component Models in Wireless Sensor Network specification, deployment and control”, ZigBee Alliance, 1st European ZigBee Developers’ Conference, 2007, Munich, Germany.
[7] Douglas McIlroy. Mass-produced software components in: P. Naur and B. Randell, "Software Engineering, Report on a conference sponsored by the NATO Science Committee, Garmisch, Germany, 7th to 11th October 1968", Scientific Affairs Division, NATO, Brussels, 1969, 138-155. http://cm.bell-labs.com/cm/cs/who/doug/components.txt
[8] C.Szyperski, D.Gruntz, and S.Murer. Component Software:Beyond Object-Oriented Programming. Addison-Wesley, second edition, 2002.
[9] B. Meyer. The grand challenge of trusted components. In Proc. ICSE 2003, pages 660–667. IEEE, 2003.
[10] G. Heineman and W. Councill, editors. Component-Based Software Engineering: Putting the Pieces Together. Addison-Wesley, 2001.
[11] Christiansson, B., Jakobsson, L., Crnkovic, I.: CBD process. In: Crnkovic, I., Larsson, M.(eds.) Building Reliable Component-Based Software Systems, pp. 89–113. Artech House (2002).
[12] Lau, K.-K., Wang, Z., Software component models. IEEE Transactions on Software Engineering 33 (10), 2007, pp.709–724.
[13] The Bean Development Kit (BDK 1.0, 1.1). http://bhiggs.x10hosting.com/Courses/HighOctaneJava/JavaBeans/bdk.htm
[14] Xaml Object Mapping Specification. http://www.microsoft.com/en-us/download/details.aspx?id=19600
[15] J.Bloch. Effective Java. 2nd Edition, 2008.
[16] The Virtual Reality Modeling Language. ISO/IEC 14772. www.web3d.org
[17] Extensible 3D (X3D). ISO/IEC 19775. www.web3d.org
[18] Eddon, G., Eddon, H., Inside COM+ Base Services, Redmond, WA: Microsoft Press, 2000.
[19] Java3D. http://java3d.java.net/
[20] Xj3D. http://www.xj3d.org/
[21] jReality. http://www3.math.tu-berlin.de/jreality/
[22] jMonkeyEngine. http://jmonkeyengine.com/
[23] Demicron WireFusion. http://www.demicron.com/index.html
[24] The Ptolemy Project. http://ptolemy.eecs.berkeley.edu/index.html
[25] E.A.Lee, S.Neuendorffer, Technical Memorandum UCB/ERL M00/12, Dept. EECS, University of California Berkeley, CA 94720, USA. http://ptolemy.eecs.berkeley.edu/publications/papers/00/moml/
[26] Curtis Beeson. An Object-Oriented Approach to VRML Development. VRML '97 Proceedings of the second symposium on Virtual reality modeling language, p.17-24.

[27] Stephan Diehl, VRML++: A Language for Object-Oriented Virtual-RealityModels. Proceedings of the 24th International Conferenceon Technology of Object-Oriented Languages and Systems TOOLS Asia, Bejing, 1997, p. 141 – 150.
[28] Instant Reality. http://www.instantreality.org/
[29] A.McVeigh, A Rigorous, Architectural Approach to Extensible Applications, PhD thesis, Imperial College London, Department of Computing, 2009.
[30] A.McVeigh, Creating, Reusing and Executing Components in Evolve, 2010. http://www.intrinsarc.com/evolve

INTENTIONAL BLANK

SMP-BASED CLONE DETECTION

Hosam AlHakami, Feng Chen and Helge Janicke

Software Technology Research Laboratory,
De Montfort University, Leicester, UK
hosam.alhakami@myemail.dmu.ac.uk
{fengchen,heljanic}@dmu.ac.uk

ABSTRACT

Code cloning is a severe problem that negatively affects industrial software and threatens intellectual property. This paper presents a novel approach to detecting cloned software by using a bijection matching technique. The proposed approach focuses on increasing the range of similarity measures and thus enhancing the recall and precision of the detection. This is achieved by extending a well-known stable-marriage problem (SMP) and demonstrating how matches between code fragments of different files can be expressed. A prototype of our approach is provided using a proper scenario, which shows a noticeable improvement in several features such as scalability and accuracy.

KEYWORDS

Clone Detection, Stable Marriage Problem

1. INTRODUCTION

The Stable Marriage Problem (SMP) is a well-known problem that has been defined by Gale and Shapley in 1962 [1]. An example of the SMP is allocating the right jobs to their most suitable jobseekers (one-one). Similarly framed problems with differing cardinality are also considered to be instances of the SMP, such as matching graduated medical students to hospitals (one-many) [2]. The SMP grantees the stable match between the candidates.

Clone detection has intensively investigated due to the need of tackling code issues in the maintenance process. Current detection algorithms are more or less search based algorithms that do not consider the preferences of both candidates (code portions) in the process. In this paper, a variant of the stable marriage problem algorithm to clone detection is investigated to find clones of different source files. The extended algorithm introduces the preferences of code segments based on the values of predefined metrics, e.g. the number of calls from or to a method, cyclomatic complexity. The clone detection process should therefore consider the values of both parties.

The remainder of this paper is structured as follows. In Section 2 we introduce the background to the SMP research. In Section 3 we provide the context for the SMP algorithms to be applied to Clone Detection. In Section 4 we present an adapted SMP algorithm that is suitable to generate fair and stable matches between similar code fragments of different source files. In Section 5 we conclude the paper and outline future work in this area.

Natarajan Meghanathan et al. (Eds) : ACSIT, FCST, ITCA, SE, ICITE, SIPM, CMIT - 2014
pp. 89–101, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4608

2. SMP ALGORITHM

2.1. OVERVIEW

In 1962, David Gale and Lloyd Shapley published their paper College admissions and the stability of marriage [1]. This paper was the first to formally define the Stable Marriage Problem (SMP), and provide an algorithm for its solution. The SMP is a mechanism that is used to match two sets of the same size, considering preference lists in which each element expresses its preference over the participants of the element in the opposite set [1]. Thus, the output has to be stable, which means that the matched pair is satisfied and both candidates have no incentive to disconnect. A matching *M* in the original SMP algorithm is a one-to-one correspondence between the men and women. If man *m* and woman *w* are matched in *M*, then *m* and *w* are called partner in *M*, and written as *m = PM(w)* (which is the M-partner of *w*), *w = PM(m)* (the M-partner of *m*). A man *m* and a woman *w* are said to block a matching *M*, or called a blocking pairs for *M* if *m* and *w* are not partners in *M*, but *m* prefers *w* to *PM(m)* and *w* prefers *m* to *m = PM(w)* [2]. Therefore, a matching *M* is stable when all participants have acceptable partners and there is no possibility of forming blocking pairs. This problem is in interest of a lot of researchers in many different areas from several aspects. Matching problems on bipartite sets where the entities on one side may have different sizes are intimately related to the scheduling problems with processing set restrictions [3].

An instance *I* of SM involves *n* men and *m* women, each of whom ranks all *n* members of the opposite sex in strict order of preference. In *I* we denote the set of men by *m = m1, m2, mn* and the set of women by *w = w1, w2,….wn*. In SM the preference lists are said to be complete, that is each member of *I* ranks every member of the opposite sex as depicted in figure 1.

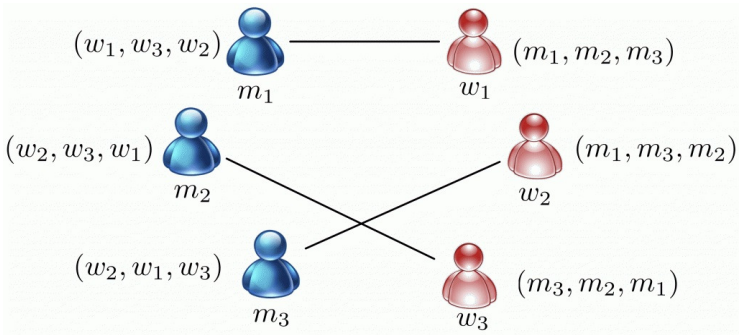


Figure1. General view of SMP. [4]

2.2. GALE SHAPLEY EXTENDED ALGORITHM

The algorithm presented by Gale and Shapley for finding a stable matching uses a simple deferred acceptance strategy, comprising proposals and rejections. There are two possible orientations, depending on who makes the proposals, namely the man-oriented algorithm and the woman-oriented algorithm.

In the man-oriented algorithm, each man *m* proposes in turn to the first woman *w* on his list to whom he has not previously proposed. If *w* is free, then she becomes engaged to *m*. Otherwise, if *w* prefers *m* to her current fiancé *m*, she rejects *m*, who becomes free, and *w* becomes engaged to *m*. Otherwise *w* prefers her current fiancé to *m*, in which case *w* rejects *m*, and *m* remains free. This process is repeated while some man remains free. For the woman-oriented algorithm the process is similar, only here the proposals are made by the women.

The man-oriented and woman-oriented algorithms return the man-optimal and woman-optimal stable matching respectively. The man-optimal stable matching has the property that each man obtains his best possible partner in any stable matching. However, while each man obtains his best possible partner, each woman simultaneously obtains her worst possible partner in any stable matching. Correspondingly, when the woman-oriented algorithm is applied, each woman gets her best possible partner while each man get his worst possible partner in any stable matching.

Theorem 1 All possible execution of the Gale-Shapley algorithm (with the men as proposers) yields the same stable matching, and in this stable matching, each man has the best partner that he can have in any stable matching [2].

According to the previous theorem if each man has given his best stable partner, then the result is a stable matching. The stable matching generated by the man-oriented version of the Gale-Shapely algorithm is called man-optimal. However, in the man-optimal stable matching, each woman has the worst partner that she can have in any stable matching, leading to the terms of man-optimal is also woman-pessimal. This results in the next theorem.

Theorem 2 In the man-optimal stable matching, each woman has the worst partner that she can have in any stable matching [2].

The following example in figure 2 gives the same output for both man-optimal and woman-optimal, the instance formed out of 4 elements.

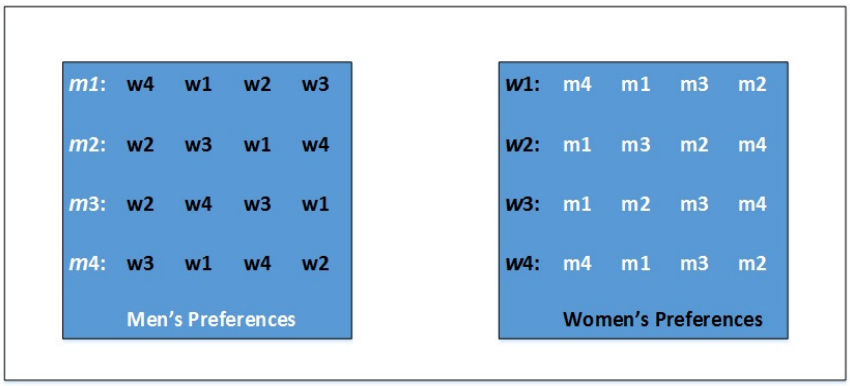


Figure 2. A stable marriage instance of size 4.

The results of different cases differ from man-oriented version to woman-oriented version. The stable matching generated by both man-oriented and women-oriented versions is *M0 = Mz = (1, 4), (2, 3), (3, 2), (4, 1).*

An extended version of Gale-Shapley algorithm has been designed to improve the basic algorithm. The extended version reduces the preference list by eliminating specific pairs that can be clearly identified as unrelated to any stable matching. The deletion process of such pair is performed by deleting each other from the preference lists.

Algorithm 1 Extended Gale-Shapley algorithm [2]

```
1: assign each person to be free
2: while some man m is free do
3: begin
4: w:= first woman on m's list;
5: if some man p is engaged to w then
6: assign p to be free;
7: assign m and w to be engaged to each other;
8: for each successor m' of m on w' list do
9: delete the pair(m',w)
10: end;
```

2.3. HOSPITALS/RESIDENTS PROBLEM

The hospitals/residents problem (also called Colleges/Students problem, and by many other names) reflects a cardinality of many-to-one of the stable marriage problem. This cardinality touches a wide range of large-scale applications that require stable matching such as students/colleges problem. Therefore, it has interested the researchers in different aspects for instance recruitment in which uses schemes to match a group and employers to a group of employees. The National Resident Match Program [5] is a real example in the US which annually matches hospitals to about 30,000 medical residents. An instance of the hospitals/residents (HR) problem consists of a set *R* of *n* residents and a set *H* of *m* hospitals, where each hospital *h* has capacity *ch*, the maximum number of positions available in *h*. Each resident ranks the hospitals in *H* that are acceptable to her in strict order of preference; likewise, each hospital ranks the residents in *R* that are acceptable to it in strict order of preference. A matching *M* for the instance is a set of resident-hospital pairs where in every pair the resident and the hospital are mutually acceptable to each other, every resident appears in at most one pair, and every hospital *h* appears in at most *ch* pairs. A pair forms a $(r, h) \notin M$ blocking pair with respect to *M* if

i) *r* is unmatched and finds *h* acceptable or *r* prefers *h* to the hospital she is assigned to and, simultaneously,

ii) *h* is not filled to capacity and finds *r* acceptable or *h* prefers *r* to one of the residents assigned to it.

Algorithm 2 Hospital-oriented algorithm

```
1: assign each resident to be free
2: assign each hospital to be totally unsubsidised;
3: while (some hospital h is unsubsidised) and (h's list
contains a resident r not provisionally assigned to h) do
4: begin
5: r:= first such resident on h's list;
6: if r is already assigned, say to h', then
7: break the provisional assignment of r to h';
8: provisionally assign r to h;
9: for each successor h' of h on r's list do
10: remove h' and r from each other's lists
11: end;
```

Intuitively, if *(r, h)* forms a blocking pair in *M* then *r* and *h* are likely to break their assignments under *M*, causing the matching to unravel. Thus, the goal of the HR problem is to find a matching

that is stable and has no blocking pairs. In their seminal paper [1], Gale and Shapley first tackled the problem in the simplier stable marriage (SM) setting where residents and hospitals are replaced by men and women. Every participant has a complete preference list (i.e., every man ranks all the women and every woman ranks all the men), and a capacity of one (i.e., every individual can have at most one assigned partner). They introduced the deferred-acceptance algorithm to find a stable matching, and showed that the algorithm can be extended to the more general HR setting. Consequently, they proved that every HR instance has a stable matching which can be computed in *O (nm)* time. In [6] Cheng et al. examined the structure of the set of all stable matchings of an HR instance and introduce the notion of meta-rotations in this setting. Also, they discuss the problem of finding feasible stable matchings.

Theorem 3

(i) The matching specified by the provisional assignments after the execution of the hospital-oriented algorithm is stable.

(ii) In this matching, a hospital h with q available places is assigned either its best q stable partners, or a set of fewer than q residents; in the latter case no other resident is assigned to h in any stable matching.

(iii) Each resident is assigned in this matching to his worst stable partner [2].

We will build on this background in section 4 where we use the extended Gale-Shapley algorithm in our application to the problem of clone detection.

3. CLONE DETECTION

3.1. OVERVIEW

Clone detection is a crucial field that has been intensively conducted by researchers and practitioners for the last two decades to enhance a software systems work and therefore, improves the maintainability for the future lifespan of the software system. Although the clone detection is a wide spread research problem over many years; is considered as a fuzzy terminology, since the researchers have differently defined it according to variants situations and criteria. Thus, it is essential to understand the meaning of clones and its uses to know how to deal with it properly? In this section, we provide different definitions and types of clones.

3.2. CLONE RELATION TERMS

Clone usually detected as a form that terms as one of either clone pair or clone classes. These two terms focus on the similarity relation between two or more pieces of cloned code. Kamiya et al. in [7] describe this relation as an equivalence relation (i.e., a reflexive, transitive, and symmetric relation). It can be said that there is a clone-relation between two fragments of code if (and only if) they have the same sequences (original characters strings, strings without whitespaces, token type etc.) From figure 3 below we can express the meaning of clone pair and clone classes based on the clone relation:

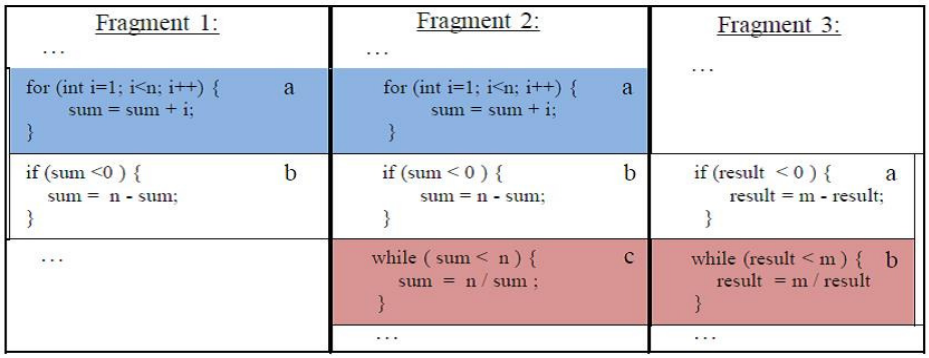


Figure 3. Clone pair and Clone class. [8]

- **Clone Pair:** two fragments of code are considered to form a clone pair when they have a clone-relation between them. That means these two portions are either identical or similar to each other. As seen in the figure 3 for the three code fragments, Fragment 1 *(F1)*, Fragment 2 *(F2)* and Fragment 3 *(F3)*, we can get five clone pairs, *(F1(a), F2(a))*, *(F1(b), F2(b)), (F2(b), F3(a)), (F2(c), F3(b))* and *(F1(b), F3(a))*. If we assume to extend the granularity size of cloned fragments, we get basically two clone pairs, *(F1(a + b), F2(a + b))* and *(F2(b + c), F3(a + b))*. And if we consider the granularity not to be fixed, we get seven clone pairs, *(F1(a), F2(a)), (F1(b), F2(b)), (F2(b), F3(a)), (F2(c), F3(b)), (F1(b), F3(a)), (F1(a+b), F2(a+b))* and *(F2(b + c), F3(a + b))*; each of these fragments is termed as a simple clone [9].

- **Clone Class:** is a maximal set of related portions of code that contains a clone pairs. It can be seen that the three code fragments of Figure 3, we get a clone class of *(F1(b), F2(b), F3(a))* where the three code portions *F1(b), F2(b)* and *F3(a)* form clone pairs with each other *(F1(b), F2(b)), (F2(b), F3(a))* and *(F1(b), F3(a))* result in three clone pairs. Consequently, a clone class is the union of all clone pairs which have portions of code in common [10,11].

- **Clone Communities:** as termed in [12] , it is another name of the Clone classes that reflecting the aggregation of related code fragments which form a clone pairs.

- **Clone Class Family:** researchers in [10] revealed the term of clone class family to group or aggregation of all clone classes that have the same domain.

- **Super Clone:** as have been outlined by [13] multiple clone classes between the same source entities (subsystems or clone classes) are aggregated into one large super clone which is the same as the clone class family.

- **Structural Clones:** it is an aggregation of similar simple clones that spread in different clone classes in the whole system [9] . Therefore, it can be classified as both a class clone (in early stage of clustering similar fragments of code) and super clone.

3.3. DEFINITION OF CODE CLONING

As aforementioned there is no original or specific definition of cloned code and therefore, all anticipated clone detection methods have their own definition for code clone [14,15] (Lakhotia, Li et al. 2003, Kontogiannis 1997). However, a fragments of code that have identical or similar

code fragments in the source code, considered to be a code clones. Regardless the changes that have been applied on a certain code clone, if still in the thresholds of the copied portion, then both the original and the copied fragments term as code clones and they form a clone pair.

Some researchers based their definition of clone code on some definition of Similarity whereas there is no specified definition of detection independent clone similarity. [16] (Baxter, Yahin et al. 1998) defined code clones as the fragments of code that are similar based on definition of similarity and they provide a threshold-based definition of tree similarity for near-miss clones. However, there is a fuzziness of the term similarity; what is meant by similar? , and to what extend are they similar? The definition provided by (Kamiya, Kusumoto et al.2002) zooms in this terminology as they define the clones as the segments of source files that are identical or similar to each other. Another ambiguous definition is proposed by [11,17] (Burd, Bailey 2002, Roy, Cordy 2007) in which fragment of code called clone when there is more existences of that fragment in the source code with or without minor modifications. However, a number of researchers [15,18, 19] (Kontogiannis 1997, Li, Lu et al. 2006, Kapser, Godfrey 2004) tried to control and specify their own detection dependent threshold based definition of the term similarity. Therefore, after several comparisons that run-out by [11,15,20] (Roy, Cordy 2007, Kontogiannis 1997, Bellon, Koschke et al. 2007) they attempt to automatically unify the result sets of multiple detectors, trying to solve the differential detector-based output.

4. EXTENDED SMP ALGORITHM (DUAL-MULTI-ALLOCATION) FOR CLONE DETECTION

4.1. OVERVIEW

SMP has solved several similar optimisation issues in different fields such as matching jobs to the most suitable jobseekers. Since the original SMP algorithm allows only the candidates of the first set (Men) to propose to their first choices, this research devotes to increase the fairness of SMP by allowing the candidates of the second set (Women) to make their own choices i.e. proposes to the best of their choices of the opposite set. The proposed approach considers a dual multi allocation technique that allows the candidates of both first and second set to enter the competition and propose again for a certain times to their preferences. So, each candidate of the first set may have more than one matched participants of the second set and vice versa. This adaption has enhanced the precision of the matching process; it is illustrated in figure 4 below. In the main SMP algorithm the desire is not controlled by the similarity, thus the assigned candidates are not meant that they are similar to each other. However, in clone detection the concept of similarity is essential. Therefore, aforementioned extension of the current state of SMP is necessary to be effectively applied in such applications. A novel matching scheme is needed to achieve smart interaction between the code fragments of the matched source files. This widens the spot to detecting every possible clone.

Practically, this process gives more than one stable matched pairs; respectively Hospital-Oriented-man and Hospital-Oriented-woman. Thus, we enclose a novel way of assigning the related code portions by adding a choosy strategy. This strategy helps to choose the pairs which form similar code clones to a certain threshold.

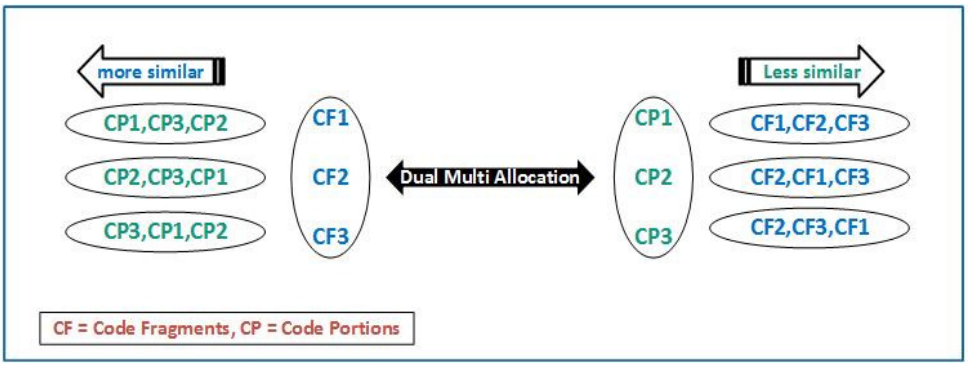


Figure 4. Dual Multi Allocation.

4.2. DUAL MULTI ALLOCATION ALGORITHM

This algorithm results in several stable matching pairs with dissimilar allocated candidates based on love's degree, which can be controlled to reach a certain level of desires. However, the matching process can be fixed as default to retrieve candidates of the highest rank love's degree factor.

The algorithm of Dual Multi Allocation consists out of two phases followed by the Choosy Strategy as following:

- **Phase 1 Hospital-Oriented-Man algorithm.**
- **Phase 2 Hospital-Oriented-Woman algorithm.**
- **Apply Choosy Strategy.**

Algorithm 3 Dual Multi Allocation algorithm

1: assign each person to be free
2: **while** (some man m is unallocated) and (m's list contains a woman w not allocated to m) **do**
3: begin
4: w:= first woman on m's list;
5: **if** w is already allocated, say to m', **then**
6: break the allocation of w to m';
7: assign w to m;
8: **for each** successor m' of m on w's list **do**
9: remove m' and w from each other's lists
10: end;
11: assign $M1$ to the Hospital-oriented-man pair or set of pairs;
12: assign each person to be free
13: **while** (some woman w is unallocated) and (w's list contains a man m not allocated to w) **do**
14: begin
15: m:= first man on w's list;
16: **if** m is already allocated, say to w', **then**
17: break the allocation of m to w';
18: assign m to w;
19: **for each** successor w' of w on m's list **do**
20: remove w' and m from each other's lists
21: end;
22: assign $M2$ to the Hospital-oriented-woman pair or set of pairs;
23: apply Choosy Strategy on $M1$ and $M2$
24: end;

4.3. CHOOSY STRATEGY

In the current state of the SMP algorithms, there are no needs to judge between two pairs to be chosen as an optimal pair. However, if a strategy to choose the optimal pair is raised, a competitive Choosy Strategy to support the newly built extension to help choosing the optimal pair is needed.

The choosy strategy formed out of two main factors, respectively, love's degree and contrast's degree. Love's degree reflects the degree of love from the view of both involved candidates (code fragment). To converge these views, we add the degree of love for both of participated (in the same pair) candidates and divide the result by two. The final result is the love's degree of the pair. The contrast's degree reflects the difference between the actual love's degree of the involved candidates. Thus, the most preferable pair is that with small difference in its contrast's degree. This factor helps when two different pairs has the same love's degree. Also, when more than one candidate has the same love's degree with a certain candidate, then the right candidate will be chosen. Figure 5 depicts the choosy strategy scheme.

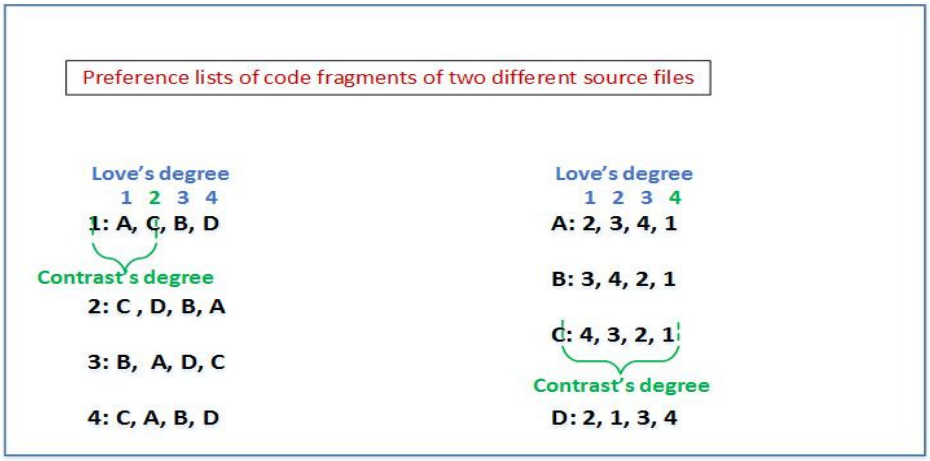


Figure 5. Choosy Strategy Scheme

4.4. SMP-BASED CLONE DETECTION

To apply the SMP algorithm in clone detection, it needs first to build the preference lists of both code fragments. This can be achieved using predefined metrics to specify the most similar related participants (code clone). Each code portion needs to strictly order the code fragments based on the similarity and vice versa. The traditional SMP algorithm performs a single assignment (one-to-one) for the involved candidates, which does not help especially in the case of allocating more than one code portion (method etc.) to the related code fragments of other source file. Multi Dual Allocation algorithm has been proposed to fulfil this requirement which widely needed in such fields. Figure 6 depicts a general prototype of code clones (method-based).

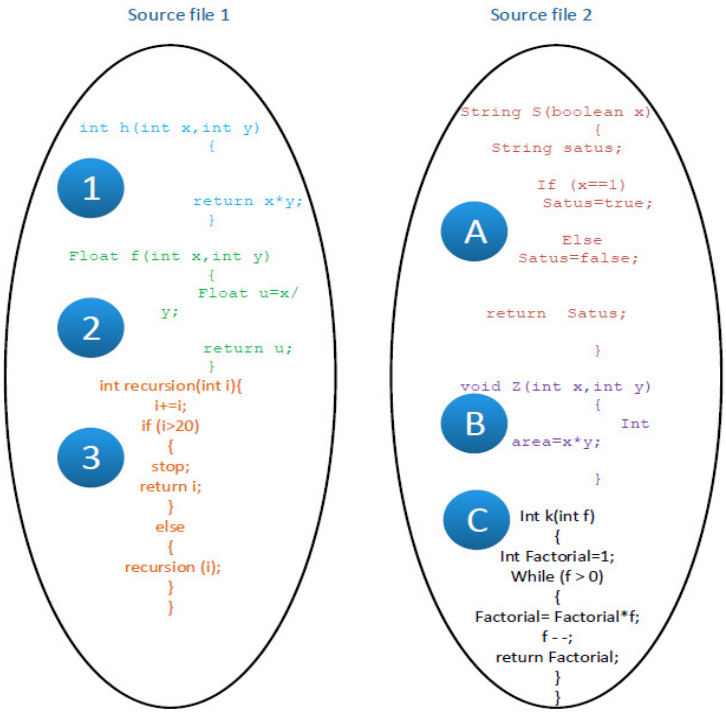


Figure 6. General Example of clones (method-based).

We have considered some metrics for fixed granularity of method based and calculate it using java plug-in with eclipse 1.3.3 (metrics 1.3.6). Figure 7 shows some of these metrics.

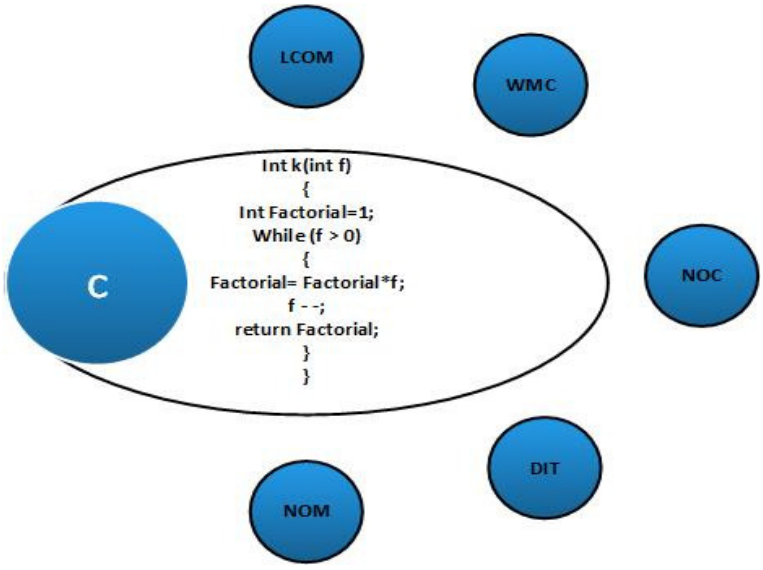


Figure 7. General view of metrics (method-based).

Table 1. Coupling Metrics

Abbreviations	Description
PROM	Number of protected methods
PUBM	Number of public methods
PRIM	Number of private methods
MCIN	Number of calls to a method
MCOUT	Number of calls from a method

Table 2. Method Metrics

Abbreviations	Description
LOC	Number lines of code
Nbp	Number of parameters
Nbv	Number of variables declared in the
Mca	Afferent coupling at method level
Mce	Efferent coupling at method level
CC	McCabe's Cyclomatic Complexity
NBD	Nested Block Depth

To apply the SMP algorithm we consider two main phases **Phase1**, building the preference list of each code fragment of first source file from the second source file's code portions, recording the most desired block and so on, repeats this process from second to first source files. **Phase2**, applying the adapted SMP algorithm based on the given metrics values. Figure 8 shows an assigned code fragment of the first source file to the most suitable (similar) code portions using the adapted SMP algorithm based on the metrics value.

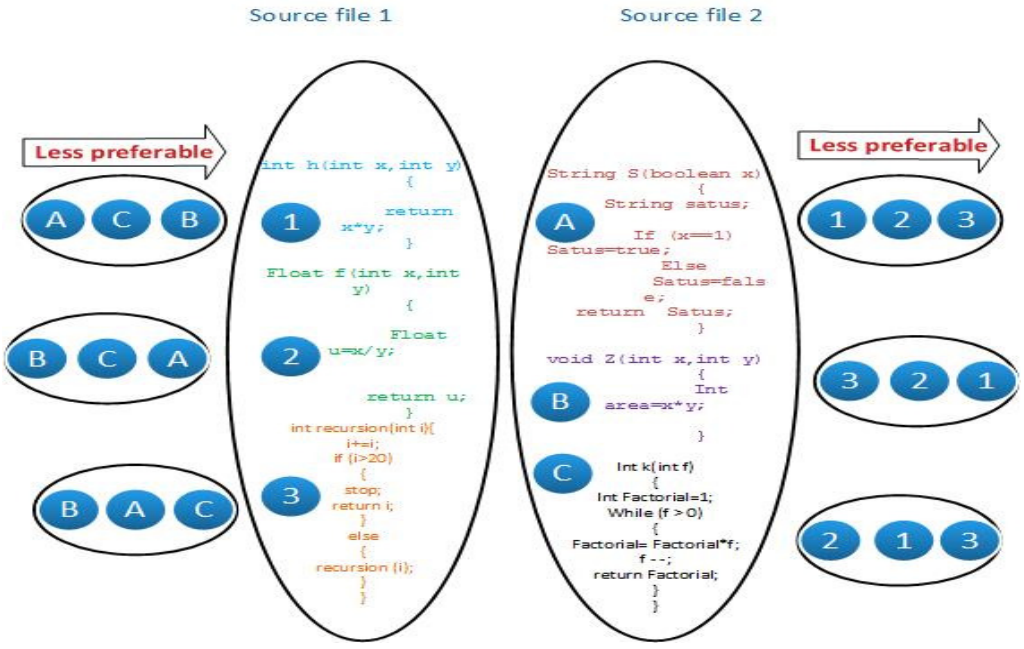


Figure 8. Example of clone detection SMP-based.

Table 3. Source file 1

Method	Metrics						
	LOC	Nbp	Nbv	Mca	Mce	CC	NBD
A	6	1	1	0	0	2	1
B	4	2	1	0	0	1	1
C	6	1	1	0	0	2	2

Table 4. Source file 2

Method	Metrics						
	LOC	Nbp	Nbv	Mca	Mce	CC	NBD
1	1	2	0	0	0	1	1
2	1	2	1	0	0	1	1
3	10	1	0	0	0	2	2

4.5. DISCUSSION

We proofed a remarkable efficiency of our approach by carrying out a case study on two middle size source files, each file with a minimum of 100 of specified blocks. Also, a set of metrics are predefined to determine the specs of each fragment of code, which help each candidate to build up its own preference list in order to apply the SMP algorithm. We observing some appointed features for the extended algorithm (e.g. performance) and the status of the detected clones (e.g. accuracy). This means that we are now able to develop match making code fragments that not only decide on the basis of the candidates' preferences of the first source file, but are actually trying to, within the current set of code fragments of both source files, to optimise the pairings from both perspectives fairly. Also, allowing the relation of many-to-many has increased the range of clones (high recall, high precision) that undetectable with most of previous clone detection approaches. However, the time complexity is challenging in this newly adapted algorithm, which still the same as the original SMP (polynomial time).

5. CONCLUSION

Stable marriage problem are well-known common matching algorithms, helped in several applications in different aspects of live for instance assigning medical schools graduates students to the most suitable hospitals. The paper presented a newly crucial extension which effectively touches a wide range of software engineering fields such as clone detection. Our contribution in this paper is the choosy strategy, which compromises between the preferences of the code fragments of two matched source files in clone detection process. Also, helped to increase the quality of retrieved code clones, through considering the desire of the matched candidates, which results in increased the satisfaction of the candidates in each pair. However, the proposed scheme has some limitations in terms of its complexity and would require longer time to reach the highly required stability. In our future work we would like to look at how dynamically scale the levels of metrics to consider different abstraction levels (e.g. package, class etc.).

REFERENCES

[1] Gale, David & Lloyd S. Shapley, (1962) "College Admissions and the Stability of Marriage", JSTOR, pp 9-15.

[2] Gusfield, Dan & Robert W. Irving (1989) The Stable Marriage Problem: Structure and Algorithms, MIT press Cambridge.

[3] Biró, Péter & Eric McDermid, (2011) "Matching with Sizes (Or Scheduling with Processing Set Restrictions)", Elsevier, .
[4] iQua, "Stable Matching in Networking." http://iqua.ece.toronto.edu/spotlights/matching/ (accessed 03/15, 2014).
[5] The match "National Resident Matching Program." http://www.nrmp.org/ (accessed 03/15, 2014).
[6] Cheng, Christine, Eric McDermid, & Ichiro Suzuki, (2008) "A Unified Approach to Finding Good Stable Matchings in the Hospitals/Residents Setting", Elsevier, Vol. 400, No. 1, pp 84-99.
[7] Kamiya, Toshihiro, Shinji Kusumoto, & Katsuro Inoue, (2002) "CCFinder: A Multilinguistic Token-Based Code Clone Detection System for Large Scale Source Code", IEEE, Vol. 28, No. 7, pp 654-670.
[8] Roy, Chanchal K. , (2009) "Detection and Analysis of Near-Miss Software Clones", IEEE, pp 447-450.
[9] Basit, Hamid Abdul & Stan Jarzabek, (2009) "A Data Mining Approach for Detecting Higher-Level Clones in Software", IEEE, Vol. 35, No. 4, pp 497-514.
[10] Rieger, Matthias, Stéphane Ducasse, & Michele Lanza, (2004) "Insights into System-Wide Code Duplication", IEEE, pp 100-109.
[11] Roy, Chanchal Kumar & James R. Cordy, (2007) Citeseer, Vol. 541, pp 115.
[12] Mayrand, Jean, Claude Leblanc, & Ettore M. Merlo, (1996) "Experiment on the Automatic Detection of Function Clones in a Software System using Metrics", IEEE, pp 244-253.
[13] Jiang, Zhen Ming, Ahmed E. Hassan, & Richard C. Holt, (2006) "Visualizing Clone Cohesion and Coupling", IEEE, pp 467-476.
[14] Lakhotia, Arun, Junwei Li, Andrew Walenstein, & Yun Yang, (2003) "Towards a Clone Detection Benchmark Suite and Results Archive", IEEE, pp 285-286.
[15] Kontogiannis, Kostas. , (1997) "Evaluation Experiments on the Detection of Programming Patterns using Software Metrics", IEEE, pp 44-54.
[16] Baxter, Ira D., Andrew Yahin, Leonardo Moura, Marcelo Sant'Anna, & Lorraine Bier, (1998) "Clone Detection using Abstract Syntax Trees", IEEE, pp 368-377.
[17] Burd, Elizabeth & John Bailey, (2002) "Evaluating Clone Detection Tools for use during Preventative Maintenance", IEEE, pp 36-43.
[18] Li, Zhenmin, Shan Lu, Suvda Myagmar, & Yuanyuan Zhou, (2006) "CP-Miner: Finding Copy-Paste and Related Bugs in Large-Scale Software Code", IEEE, Vol. 32, No. 3, pp 176-192.
[19] Kapser, Cory & Michael W. Godfrey, (2004) "Aiding Comprehension of Cloning through Categorization", IEEE, pp 85-94.
[20] Bellon, Stefan, Rainer Koschke, Giuliano Antoniol, Jens Krinke, & Ettore Merlo, (2007) "Comparison and Evaluation of Clone Detection Tools", IEEE, Vol. 33, No. 9, pp 577-591.

AUTHORS

Hosam Al Hakami received his B.Sc. degree in Computer Science from King Abdulaziz University, Saudi Arabia. He received his MSc degree in Internet Software Systems from Birmingham University, Birmingham, UK. He is studying now towards his PhD degree at the Faculty of Technology in De Montfort University, Leicester, UK.

Dr.Feng Chen was awarded his BSc, Mphil and PhD at Nankai University, Dalian University of Technology and De Montfort University in 1991, 1994 and 2007. As research outputs, he has published over 30 research papers in the area of software evolution and distributed computing.

Dr.Helge Janicke is heading the Software Technology Research Laboratory at De Montfort University, Leicester (UK). He is leading the research theme on Computer Security and Trust. His research interests are in area of software engineering where he is primarily looking at cyber security, in particular access control and policy-based system management.

INTENTIONAL BLANK

E-LEARNING: GENDER ANALYSIS IN HIGHER EDUCATION IN NORTH INDIA

Manu Sood[1], Virender Singh[2]

[1]Department of Computer Science, H.P. University, Shimla
[2]GGDSD College, Sector-32, Chandigarh
forviren@yahoo.co.in, soodm67@yahoo.com

ABSTRACT

E-Learning is an adaptable technology that can be used to cover different areas of interactive or live learning as well as training needs. It makes skills available through newer technologies and reduces the learning time even for complex topics. E-Learning is a changing trend in education that no longer limits the education to the four walls of a class room. Measuring the efforts to improve the learning skills with technology is utmost essential in order to effect any change in the education policies. This paper focuses on the analysis of the genders' interest in e-Learning in higher education in the northern part of India. A questionnaire survey designed for the purpose gathered information on students participation and opinions about the use of e-Learning in higher education. The analysis of the data thus collected shows that there is a definite change in trend among genders as far as the regular participation in e-Learning components are concerned.

KEYWORDS

e-learning, learning styles, attitudes of a learners, software engineering experience, software project management.

1. INTRODUCTION

E-Learning is a concept derived from the use of Information and Communication Technologies (ICT) to deliver teaching and learning both. A common definition states that e-Learning in higher education is a technique to enhance learning and teaching experiences and is used to educate students with or without their instructors through any type of digital media [1]. It can be used to replace traditional face-to-face teaching either completely or partially, as in some cases, the use of ICT is introduced as an additional resource along traditional teaching and learning methods. A major advantage of ICT is that accessing online learning resources is flexible and fast and has no geographical barriers [2][3].

The higher education sectors have been concentrating on increasing the use of online applications of e-Learning by using the internet to enhance education [4]. With the rapid growth of e-Learning, computers are now being used increasingly by students in many different educational processes and are considered to be valuable tools to enhance learning in higher education. The motivation behind this study is to try to determine if particular groups of students are learner or making sufficient use of online learning, so that these groups of students may be further encouraged to use online activities so as to enhance their overall learning experience.

Natarajan Meghanathan et al. (Eds) : ACSIT, FCST, ITCA, SE, ICITE, SIPM, CMIT - 2014
pp. 103–110, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4609

The aim of this study is to assess the usage of e-Learning activities/components in higher education by the students (male and female) of a Northern part of India.

The rest of the paper has been organized into different sections: section 2 presents a brief introduction to e-learning and section 3 provides an account of attitude of learners in e-learning. Section 4 presents the methodology used for deriving the results whereas section 5 provides a brief introduction about the questionnaires used for the collection of the data from the sample targets. Section 6 highlights how various statistical analyses techniques have been used to analyze the data collected through the questionnaires, most of the data having been presented in tables in summarized form. This section also sums up the finding and section 7 presents the conclusion of the work of the authors.

2. E-LEARNING

Latest trend in the education sector is E-learning [5], with the help of which a student can process more control over his learning process [6]. The definition of E-learning, however, is still in vague terms but the augmentation of the same, has been documented well [7]. A few attempts have been made to define e-learning like in [8], where it has been proposed that e-learning has four parts; learning with the help of computers, online learning, learning in classroom through virtual set-up and digital world. E-learning is basically nothing but a set of pre defined applications, and the services of same can be provided via internet, intranet, interactive TV and satellite.

Another definition has been given in [9], according to which, any learning guidance that is provided through a medium of extranets, or internets or even through mediums like audio/video tapes, interactive television, of CD's. All kinds of electronic medium are used to suit the students' needs and make them understand the concepts. Pollard and Hillage in [10] gave a lengthy explanation by saying that e-learning comprises of learning opportunities with the help of computers, networks and web-based technologies, that can help in the development of any individual and his performance. Sambrook in [11] and Homan & Macpherson in [7] suggest that e-learning basically comprises of electronic learning material which can be acquired in the form of CD's, that is PC friendly, or can even be downloaded via internet/ intranet. These are basically the interactive study material.

Any learning process carried out with the help of computers and web-based facility like intranet/ Internet is e-learning [12]. Torstein and Svein in [13] propose that e-learning is basically interactive learning, and all the study material is available online. This system gives an automatic feedback to its students as well.

3. ATTITUDES

E-learning is extremely helpful to the students stationed in remote locations especially in rising economies among developing countries like India. It has been found to be a promising alternative when compared to its counter-part [14]. Parker in [15] claims that students who have computer based know-how, have a positive approach towards e-learning and achieve more success in it. Shashaani in [16] suggests that there is a positive co-relation between experience in computers and attitude related to it. Woodrow in [17] states that it is very important to know the mind-set of students, related to the computers, as any computer based course and its development will be based on the attitude of the students. When learning becomes e-centric, a student is able to exercise greater control over his own learning process and will be empowered to learn more. It will provide him with time flexibility, place and mode flexibility, and will enhance the group

interaction among students, peers and teachers. It will also make institution like universities more accessible to students.

Link & Marz, in [18] and Hayashi et al. in [19] have pointed out that despite all the growth in the technology for higher education, there are many student who do not possess the skills required to handle technology. This group of student is handicapped, in terms of technology usage. Jones and Jones in [20] discuss about a study carried out by them, wherein a comparison between the outlook of learners and teachers on usage and effectiveness of the web-based management software has been presented. The results have been found to be in favor of teachers for being a greater supporter of web based learning as compared to students. To summarize, teachers believed in the positive help that technology brings in student communication and students did not. Guruajan and Low in [21] further highlighted that the learners took ICT as a convenient tool and not as a substitute tool. The learners still believed more in in-person learning through lectures given by teachers, and ICT was found to be helpful in the rare absence of a teacher's lectures. The main source of knowledge was still the text books and references, and web-based learning was analyzed to be second to text books.

4. METHODOLOGY

The authors of this paper have designed a questionnaire survey for gathering information on student experiences and opinions on the use of e-Learning components. The survey was conducted on a group of 392 people involved in higher education in Chandigarh (northern part of India) and surrounding areas. The group was a heterogeneous one and consisted of students in the fields related to Information Technology at the college and university level. The students targeted were the ones who were enrolled in different streams (related only to the field of IT) at under graduate and post graduate levels for studying in various colleges as well as in the university affiliating these colleges. The target groups were contacted during the months of October and November 2011 to participate in the survey. The survey responses collected over a period of five weeks were tabulated and then subjected to various statistical analysis techniques as described in section 4. The results thus obtained were analyzed further to find out the conclusion.

5. QUESTIONNAIRE

The first part of the questionnaire focused on collecting some demographic information about the respondents, such as age, gender, and some details regarding their streams of studies and type of degree etc. The questionnaire next, included a section on computer use, consisting of questions asking respondents to respond to the questions like whether they have access to a computer and a high-speed internet connection outside their college/ university. Next, the respondents were asked to provide an estimated number of hours spent per week on computer and internet. The next section on the questionnaire asked the respondents to provide details on how often they used a computer/ Internet for various tasks related to their studies, e.g. for preparing study material or using certain types of software, and how often they used the internet for contacting faculty members and tutors or to participate in online discussions etc. The responses for these questions were collected using a five-point scale having options never, occasionally, sometimes, quite often and regularly.

The results presented in this paper are part of a larger questionnaire study which also gathered information on informatics skills, satisfaction with college/university ICT provision, and opinions on the use of E-Learning etc.

6. STATISTICAL ANALYSES

Summary statistics has been presented in this section to depict the responses of respondents to the questions posed to them through the questionnaire. This summary includes breakdown of inputs from the respondents according to their gender i.e. male and female. The results are presented in the form of Frequency Distribution tables depicting numbers along with percentage differences among gender levels on various issues of research study.

Chi-Square test of proportionality has been applied on approach to E-Learning, Software Engineering Experiences in our case to investigate if any significant difference exist in this area among the proportion differences among male and female respondents.

Mann–Whitney non–parametric statistics has been used to test the significant differences in average score (median score) among male and female respondents on pre test and post test components.

Chi–Square test of associations has been used to statistically analyze the significant difference in number of male and female respondents by finding out the association between parameters of Knowledge about typical patterns observed in Software Projects & Software Project Management Literature.

6.1 RESULT ANALYSIS

The research was conducted on 392 respondents, among that 222 (56.6%) were male and 170 (43.4%) were female respondents. The majority of the respondent students i.e. 320 were in their third year of study and among them 57.2% were male and 42.8% were female respondents. Respondent student from 2nd year were 45 and among them majority were female respondents i.e. 33 (73.3%) and students from 1st year were all 27 male respondents.

6.1.1 PRACTICAL SOFTWARE ENGINEERING EXPERIENCE

Respondent students were further analyzed on their practical experiences in software engineering. Among 392 respondents 356 (90.8%) students have written software program and the proportion of the male (43.8%) and female (56.2%) respondent student differ significantly with $p = 0.002$. The practical experience in SE work of 369 respondents' trend was reversed among gender level in comparison to software program written as 53.9% (199) male respondents were having practical experiences as compared to 46.1% (170) female respondents and their proportion was statistically significant with $p = 0.003$. The major issue which came to the surface regarding big team work (team > 4 members) was quite unexpected, as 18.4% respondents had participated in bigger teams and among them 55.6% were male as compared to 44.4% were female respondents. Also, there was no significant difference (with $p = 0.78$) found among male (50.4%) and female (49.6%) respondents concerning their industrial exposure among total of 238 (60.7%) respondents.

6.1.2 SOFTWARE PROJECT MANAGEMENT LITERATURE

On literature issue respondents were first analyzed on number of references books respondents had used in their practical experience. The data collected from the respondents on this issue has been presented in a summarized form in table 1. The majority respondents (54.8%) preferred 3 – 5 books as compared to 24% respondents those preferred 1 – 2 books. Among male respondents 38.3% preferred 3 – 5 books and those who preferred to read 1 – 2 books or < 5 books were in the range of 21 – 24 % male respondents.

Table 1: Learning Styles

Books Read	0	1 – 2	3 – 5	> 5	Total
Male	36 (16.2%)	54 (24.3%)	85(38.3%)	47 (21.2%)	222
Female	0 (0%)	40 (23.5%)	130 (76.5%)	0 (0%)	170
Total	36 (9.2%)	94 (24.0%)	215 (54.8%)	47 (12.0%)	392
Learning Style	**Reading Text books**	**Class lecturers**	**Group Work**	**Web Based**	**Total**
Male	43 (19.4%)	16 (7.2%)	10 (38.3%)	153 (21.2%)	222
Female	2 (1.2%)	9 (5.3%)	9 (5.3%)	150 (88.2%)	170
Total	45 (11.5%)	25 (6.4%)	19 (4.8%)	303 (77.3%)	392

16.2% male respondents do not preferred to read any references book as E-Learning methodology. Female respondents were in majority (76.5%) reading 3 – 5 reference books while rest of the female respondents preferred 1 – 2 books. The chi square test with p = 0.0026 suggests that there were no significant association present among the gender level and the number of references books preferred.

In order to comment on the learning styles of the respondents, the questionnaire obtained the most preferred learning style among the choices from them. The most preferred style (77.3%) was web based learning style among both female (88.2%) and 2nd most for male respondents (21.2%) was the class lectures (6.4%). Group work (4.8%) was the least preferred style of learning overall but among those who preferred group work, 38.3% were male respondents. Reading textbooks (11.5%) was not significantly popular among females (1.2%) as compared to males (19.4%). The chi square test with p = 0.00 suggests that there was no significant association present among the gender level and the learning style.

6.1.3 PRE – TEST ANALYSIS: INTEREST IN SOFTWARE PROJECT MANAGEMENT

Respondents were pre test analyzed on likert scale about interest in software project management. The significant differences were statistically evaluates on p – values of Mann – Whitney test. Both male and female respondents were found to agree with no significant differences on Important to know about Software Project Management (SPM) (p = 0.284), Like to participate in Seminar on SPM (p = 0.807) & Important for software engineers to know about SPM (p = 0.259). Both genders with no significant differences were neutral on Like to get more information on SPM (p = 0.634) and Like to learn more about SPM (p = 0.969). Please refer to table 2 for the related data.

Table 2: Interest in Software Project Management (SPM)

	Male	Female	p-value
Important to know about SPM	2.03	2.10	0.284
Like to get more information on SPM	3.13	3.18	0.634
Like to participate in Seminar on SPM	2.00	2.04	0.807
Important for software engineers to know about SPM	2.00	2.13	0.259
Like to learn more about SPM	2.98	3.01	0.969

Respondents in this section were evaluated on the parameters of their knowledge about typical patterns observed in software projects. Please refer to the data given in table 3. 69.6% respondents (males 78.9%, females 57.6%) believed that finding and fixing any software problem after delivery was 5 times costly as compared to 30.4% respondents (males 21.2%, females 42.4%) believed that it was just 3 times. The addition of manpower compressed 25% of nominal schedule was assumed by 58.2% respondents (males 74.85, females 36.5%) as compared to 41.8% respondents (males 25.2%, females 63.5%) who assumed it to be at 10%.

6.1.4 KNOWLEDGE ABOUT TYPICAL PATTERNS OBSERVED IN SOFTWARE PROJECTS

Majority of respondents (76%) among both sexes (male 75.5 %, female 76.5%) believed that software development cost is primarily due to tool usage as compared to those 24 % respondents who believed it to be the product quality. On the issue of comparing software development process (SDP), majority of male respondents (69.8%) thought that it is the people skills that matters whereas female respondents thought that it is the programming language (47.6%) that matter. 'Programming style brings variations in software development process' was believed by 20.7% of respondents (male 24.3%, female 15.9%). 71.2% of male respondent thought that software inspection detects 60% of defects whereas 47.6% female respondents thought it to be 25% of detections.

Table 3: Knowledge about Typical Patterns Observed in Software Projects

Finding & Fixing a software problem after delivery is costlier by about		**3 times**	**5 times**
	Male	47(21.2%)	175(78.9%)
	Female	72(42.4%)	98(57.6%)
Nominal Schedule of typical SDP can be compressed up to		**10%**	**25%**
	Male	56(25.2%)	166(74.8%)
	Female	108(63.5%)	62(36.5%)
Software Development cost is primarily a function of		**Tool Usage**	**Product Quality**
	Male	168(75.7%)	54(24.3%)
	Female	130(76.5%)	40(23.5%)
On comparing SDP, variations between	**Programming Language**	**Programming Style**	**People Skills**
Male	13(5.9%)	54(24.3%)	155(69.8%)
Female	81(47.6%)	27(15.9%)	62(36.5%)
Software Inspection detects about	**25%**	**40%**	**60%**
Male	13(5.9%)	57(23.0%)	158(71.2%)
Female	81(47.6%)	32(18.8%)	57(33.5%)

7. CONCLUSION

As an effort to assess the gender participation in E-Learning activities among the students of higher education in the field of Information Technology, a questionnaire survey was conducted by the authors on a population of 392 students. This sample of students was drawn from different streams related to Information Technology to represent a cross section of the students enrolled in different courses at different undergraduate and postgraduate levels in various colleges and universities in the northern part of India. The analysis of the results clearly indicate that the E-Learning patterns are not gender sensitive as far as the web-base learning style is concerned. Similarly, it has been found that there is no significant gender sensitivity in the area of interests in software project management. But as far as the issues of 'knowledge about the typical patterns observed in software projects' is concerned, the various parameters have been found to be significantly gender sensitive.

REFERENCES

[1] Christie, M. F., & Ferdos, F. (2004). The Mutual Impact of Educational and Information Technologies: Building a Pedagogy of E-Learning. Journal of Information Technology Impact, 4(1), 15-26.

[2] Concannon, F., Flynn, A., & Campbell, M. (2005). What Campus-Based Students Think about the Quality and Benefits of E-Learning. British Journal of Educational Technology, 36(3), 501-512.

[3] Sivapalan S., & Cregan, P. (2005). Value of Online Resources for Learning by Distance Education. CAL-laborate, 14, 23-27.

[4] Arabasz, P., & Baker, M.B. (2003). Evolving Campus Support Models for E-Learning Courses. ECAR Edu-cause Center for Applied Research. Retrieved October 3, 2010, from http://net.educause.edu/ir/library/pdf/ecar_so/ers/ERS0303/EKF0303.pdf

[5] Spender, D. (2001). E-learning: Are Schools Prepared? Proceedings of the Annual Washington conference on E-Learning in a Borderless Market, 21-33.

[6] Acton, T. Scott, M. and Hill, S.(2005). E-Education – Keys to Success for Organisations 18th Bled eConference eIntegration in Action. Bled, Slovenia,

[7] Homan, G., and Macpherson, A. (2005). E-Learning in the Corporate University. Journal of European Industrial Training, 29(1), 75-99

[8] Beamish, N., Armistead, C., Watkinson, M., and Armfield, G. (2002). The Deployment of E-Learning in UK/European Corporate Organizations. European Business Journal. 14(3): 105-116.

[9] Govindasamy, T. (2001). Successful Implementation of E-Learning: Pedagogical Considerations. The Internet and Higher Education. 4(3-4): 287-299

[10] Pollard, E. and Hillage, J. (2001). Exploring E-Learning, Brighton: Institute for Employment Studies.

[11] Sambrook, S. (2003). E-learning in small Organizations, Education+Training, 45(8/9), 506-516.

[12] Hall, B., and Snider, A. (2000). Glossary: The Hottest Buzz Word in the Industry, Learning 44(4), 85-104.

[13] Torstein Rekkedal and Svein Qvist-Eriksen. (2003). Internet Based E-learning, Pedagogy and Support Systems NKI. Distance Education. March 2003.

[14] Gunasekaran, A., McNeil, R. and Shaul, D. (2002). E-Learning Research and Applications. Industrial and Commercial Training 34(2), 44-53

[15] Parker, M.(2003). Technology-Enhanced E-Learning: Perceptions of First Year Information Systems Students at the Cape Technicon. Proceedings of SAICSIT 2003, 316-319.

[16] Shashaani, L. (1994). Gender DIFFErences in Computer Experience and its Influence on Computer Attitudes. Journal of Educational Computing Research. 11(4) : 347-367

[17] Woodrow, J. E. (1991). A COMPARISon of Four Computer Attitude Scales. Journal of Educational Computing Research. 7(2): 165-187.

[18] Link, M .T. & Marz, R. (2006). Computer Literacy and Attitudes Towards E-Learning among First Year Medical Students. http://www.biomedcentral.com/1472-6920/6/34

[19] Hayashi, A., Chen,C., Ryan, T. & Wu, J. (2006). The Role of Social Presence and Moderating Role of Computer Self-Efficacy in Predicting the Continuance Usage of E-Learning Systems. http://findarticles.com/p/articles/mi_qa4041/is_200407/ai_n9437383 - Retrieved on 6 Aug 2008

[20] Jones, H.G. & Jones, H. B. (2005)A Comparison of Teacher and Student Attitudes Concerning Use and Effectiveness of Web-based Course Management Software. Educational Technology & Society, 8(2), pp. 125-135, http://www.ifets.info/journals/8_2/12.pdf - Retrieved on 6 Aug 2008.

[21] Guruajan, V. & Low, E. (2009). Using ICT Tools to Manage Knowledge: A Student Perspective in Determining the Quality of Education.

A CASE STUDY OF USING WEB-BASED TOOLS TO SUPPORT POSTGRADUATE STUDENTS' LEARNING IN A BLENDED LEARNING ENVIRONMENT

Xingmei Qiao

Sichuan Radio and TV University, Chengdu, China
qxm@scrtvu.net

ABSTRACT

This paper explores the implementation of Web-based tools in postgraduate students' courses and concentrates on students' participations and perceptions in such a blended learning environment. In this study, the focus is on the problem of how postgraduate students use web-based tools for learning support. Moreover, the writer discusses the factors that influence students' use of web-based tools in the learning process.

KEYWORDS

Web-based tools, Learning support, Blended learning

1. INTRODUCTION OF WEB-BASED TOOLS

The effective use of technological tools enables students to construct knowledge in an active way, shifting 'from didactic techniques to a unifying constructivist framework' [1] [2]. Due to the rapid pace of development in computer technology and the Internet, there are a wide range of web-based tools which provide multiple learning materials as well as a range of possible teaching formats in teaching and learning. Kirkley and Kirkley [3] argue that designing blended learning environments is a challenge since we simply incorporate a wide range of tools in the learning process. Considering the difference of design purpose, function and complexity in web-based tools, there are two categories: learning platform and personal learning tool.

A learning platform is a computer software (or computer system) integrating tools and services to support teaching and learning, particularly for building a virtual learning environment [4]. There are other terms used in previous articles, such as e-learning platform (or system), online learning platform, and virtual learning platform, which present similar meanings. There are various learning platforms, such as Blackboard, WebCT and Moodle, with different user interfaces and functions. This paper explores the WebCT, one of the most successful and widely used learning platforms (or systems). It has developed quickly in 15 years from 2 million users in 30 countries to over 10 million students in 80 countries for online learning.

Natarajan Meghanathan et al. (Eds) : ACSIT, FCST, ITCA, SE, ICITE, SIPM, CMIT - 2014
pp. 111–117, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4610

A Personal Learning System (or tool) is defined as computer software that mainly helps students manage their own learning, which can be viewed as a single and simplified learning platform [5]. Van Harmelen [6] points out that Personal Learning Systems (PLSs) may help students' set their own learning goals, manage their learning and communicate with others in the learning process. Actually, the personal learning tool is a new term which still has no unified definition.

The most frequently used personal learning tool in this study was PebblePad. PebblePad is an e-portfolio, providing electronic evidence such as text, images, multimedia, and hyperlinks. Actually, the designer indicates that PebblePad is a personal learning system, much more than an e-portfolio [7]. PebblePad supports personal learning, facilitating students' reflection on their own learning and needs. Compared with WebCT' preset framework interface, PebblePad is developed with a Flash user interface which is more customizable.

2. RESEARCH SAMPLINGS

For this research, a case study was performed within the context of the master courses at the University of Birmingham. The goal was to assess postgraduate students' perceptions of the use of Web-based tools in the learning process. The two sample courses, 'Online and blended learning' and 'ICT in society', both belong to the 'IT in Education' program.

A maximum of eleven students participated in the 'Online and blended learning' course, and eight students in 'ICT in society' course. Among these participants, four full-time students participated in both of the modules. They are all international students who speak English as a second language.

In the 'Online and blended learning', all the participants had previous experience of using ICTs in study. All of them replied that they had used e-mail, BBS, Chat and weblog before, and five of them had used WebCT. Relatively, four part-time students in 'ICT in society' were more familiar with these ICT tools because they also work in primary and high schools and have much more work experience on education and ICT, thus helping them to better understand tutors' instructional design and requirements.

3. FACTORS INFLUENCING STUDENTS' USE OF WEB-BASED TOOLS

3.1 Previous Experience and Technological Skills

Students' previous experience and current technological skills impact on students' preferences in selecting web-based tools. Kennedy et al. [8] argue that previous positive experiences with technology and previous skills with technology are two main factors which affect students' use of technologies for educational purposes. Some past studies indicate that experience with technologies may enhance students' performance in the class and suggest teachers need to consider students' backgrounds and previous knowledge when choosing tools to form a learning environment [9]. People who spend more time online or are frequent net users or have been using the net for a long time will have better online skills such as searching for information or using online communication, thus they are more likely to acquire better knowledge. This idea is demonstrated in many previous studies [10].

In this study, all the participants are familiar with the Internet and computer and usually use them for learning and entertainment. Moreover, all the participants have basic experience and knowledge of CMC tools such as BBS, Chat, Email and Instant Messaging tools like MSN. These commonly used CMC tools are unappealing for participants. Compared with these tools, some innovative tools are more attractive to students. For example, participants give a high rating to the Whiteboard tool embedded in the WebCT. When students discuss online, they may use it to draw objects, enter text, import images and share these outcomes with others in real time.

3.2 Self-efficacy

Students' self-efficacy also influences their use of web-based tools in the blended learning environment. Self-efficacy is a measure of individual's confidence in his or her ability to 'perform the behaviour required to produce a specific outcome and it's thought to directly impact the choice to engage in a task, as well as the effort that will be expended and the persistence that will be exhibited' [11].Self-efficacy has been shown in several studies to influence students' choice of whether to engage in a task and performance in the process [12].

Computer self-efficacy is regarded as a specific self-efficacy, particularly related to one's capability to use computer technologies and the acquisition of computer skills [13]. Previous computer experience contributes to students' believe that web-based tools and computer applications are easy. High self-efficacy, in turn, increases students' persistence in learning web-based tools. In fact, self-efficacy and previous experience permeate and interact with each other, help each other forward, and affect in the whole learning process.

In this study, students' self-efficacy is as important as their previous experience and computer skills in facing the challenge of using new ICT tools. In the 'Online and blended learning' course, the lack of self-confidence becomes a barrier to participants' success. Some participants just audit this course and give up the final task which is to design a Web-based Instruction site. One participant told me designing a WBI site looks 'difficult' and 'I am afraid I can't finish it'. However, participants with high self-efficacy maintain a positive attitude to meeting challenges. As the students gain more self-confidence, they move into a more autonomous phase of collaborative learning [14].

3.3 Computer Anxiety

Computer anxiety is the third factor relating to students' perceptions about the use of web-based tools. Computer anxiety is defined as "the tendency of individuals to be uneasy, apprehensive, or fearful about current or future use of computers"[15].Computer anxiety is regarded as an individual characteristic that plays a critical role in shaping perceived ease of use about new systems and computer tools, particularly in the early stages of users' experience and adoption [16].

Buche, Davis and Vician [17] indicate that at least a change of high-anxiety individuals' computer anxiety to a lower level can change their performance. Matsumura and Hann [18] argue that students with low computer anxiety feel more comfortable and may get better achievement than their high-anxiety peers who always feel worry about using computer and tools.

In this study, participants showed low computer anxiety because most of them were familiar with computer technologies and some participants even taught this subject in schools. Participants in other majors without computer-related backgrounds were more likely to become anxious when learning new web-based tools.

When participants discussed new functions, the high-anxiety students would make more negative comments like 'hard to use' and 'I don't think it's useful'. Maurer [19] points out that the effect of computer anxiety affecting academic achievement can be overcome by other factors, such as students' motivation, good instructional methods and materials. In addition, students who practice and use web-based tools in a relaxed and fun-filled environment may reduce their computer anxiety [20].

Tutors in this study offered detailed instructional materials and set realistic goals for learning new systems. As the study went further, the high-anxiety students became more relaxed and held a more positive view of the new tools. The situation in this study seems to support the idea that computer anxiety can be significantly reduced in a long-term and sustainable study, especially in the second half of the term [21].

4. THE USE OF WEB-BASED TOOLS FOR LEARNING SUPPORT

4.1. Support Collaborative Learning

As one of the most popular learning platforms, the WebCT not only allows teachers to create and host courses on the Internet, but also offers a collaborative learning environment for students. Communication tools integrated in the WebCT are used for building both asynchronous and synchronous interactions. Our participants used these tools to exchange opinions, feedbacks, and comments and ask questions in and out of classroom. There were two frequently used communication tools in the WebCT:

(1) Chat room: The Chat room is a synchronous communication tool which allows students in the same course to chat with one another in real time. It also has the Whiteboard function which can be used to draw objects, enter text, and import images at the same time.

(2) Discussions: The Discussions is an asynchronous communication tool which allows students to post messages and replies on specific course topics or discuss tasks and group work with instructors, teaching assistants, and other students. Specifically, there is a type of Discussion called 'Class blog'. The 'Class blog' is a more collaborative space, where students may post a chronological series of entries on a particular topic.

Both of these communication tools can support collaborative learning. Groups discuss their work, share documents and resources, and receive feedback from the instructor and partners. In the meanwhile, the instructor can participate and monitor groups' work [22] via the discussion forum and chat room.

In this study, the use of web-based tools to support collaborative learning has already been acknowledged and accepted by all our participants and used in students' practice. Analysing students' assignments of the 'Online and blended learning' course, all the participants used at least one of the communication tools in their WBI sites, and three participants added both the

discussion board and chat room into WBI sites in order to 'support real time and non-real time collaborative communication'.

4.2. Support Personalised Learning

PebblePad as an e-portfolio is easy to share personal experience and opinion. The 'Webfolio' which is an evidence-based website tool is embedded in the PebblePad and used to create personal e-portfolio. Pages contained in the Webfolio can be added to, edited or deleted at any time, and may also contain links or other assets easily within the personal e-portfolio. In addition, PebblePad supports a number of export options which allow users to move their webfolios out as a zip archive file or a web site, or individual items as html, thus users may put their webfolios into any personal web space or disk for distribution. In this study, tutors made a coherent assignment which was creating individual e-journal about the course contents, personal views, questions and discussion in the PebblePad. Students constructed their own e-journals through the 'Webfolio'. In each session, they used 'Blog' to record their learning views and experience, and added to their own Webfolio. Participants used 'convenient', 'not complicated' and 'forming a learning community' to describe the PebblePad. Due to the practical function, friendly user interface and good scalability, most of participants regarded the PebblePad as a good web-based tool to record and reflect on personalised learning.

5. STUDENTS' PERSPECTIVES ON THE USE OF WEB-BASED TOOLS FOR LEARNING SUPPORT

Participants gave positive evaluations of using WebCT for collaborative learning and PebblePad for personalised learning.

When students chose their tools, they did not care more about whether it is a big system or small one. Comparatively, they considered more whether it is easy to use or whether it fits their learning styles. Through my study of students' use of web-based tools, I found that students preferred to use a mix of tools to support their learning, depending on the tools' utility and ease of use.

What was highlighted through my observation of the course is that what is most needed to support students in using web-based tools in learning is having an authentic task in the learning process. When students first use a new tool, they do not only want to know how to use it, but what it can do or why they should use it. Tutors need to design authentic tasks and guide students to select and use appropriate tools. For example, in this study, we learned a virtual world system called 'Second Life' in the 'ICT in Society' course. The Second Life is a multi-user virtual environment supporting role-play, exploration, simulation and interaction through avatars [23]. Users may create their virtual world in this 3-D virtual environment, build and experience various social life and activities including academic activity, entertainment, business, and etc, as same as in the real world. There are few investigations about using SL in Education, but some research shows that multi-user virtual environments allow real time learning with visual, interactive components, which has great potential for the benefit to build a constructivist learning environment, especially in distance education [24].

The Second Life, as a role play tool, is totally different from WebCT. The Second Life requires more time online, and needs all participants to login at the same time, in this way it can be used to

extend our classroom learning in a virtual environment via virtual roles. Back to my study, in our Second Life exercise, we just had time to register and login in the system, then change our avatars and try to control them to explorer the virtual world. Due to the lack of authentic tasks, participants did not have more opportunities to study and experience the Second Life in detail and in depth. However, when participants talked about the SL after course, they used more negative comments than the WebCT and PebblePad, such as 'It is complex', 'no use' and 'what should I do next?'

This case reflects the importance of authentic tasks in using web-based tools at an early stage. The lack of authentic tasks may restrict students' motivation for learning and using new tools, especially such a complex virtual environment.

6. CONCLUSION

This study focused on the general use of WebCT and PebblePad, providing an initial exploration of postgraduate students' use of web-based tools in a blended learning environment and the students' experiences within them. Future research can be more detailed in studying students' achievement and outcomes, especially the knowledge construction elements.

ACKNOWLEDGEMENTS

The author would like to thank Dr Rachel Pilkington for her advice and help in this research, and thanks all the participants of IT in Education in the University of Birmingham.

REFERENCES

[1] Matusevich, M. N. (1995) School reform: What role can technology play in a constructivist setting? Montgomery County Public Schools. [Online]. Available: http://pixel.cs.vt.edu/edu/fis/techcons.html

[2] Jonassen, D. H., Peck, K.L. and Wilson, B.G. (1999) Learning with technology: A Constructivist Perspective. Upper Saddle, NJ: Merrill, Prentice Hall.

[3] Kirkley, S. and Kirkley, J. (2004)"Creating next generation blended learning environments using mixed reality, video games, and simulations," Tech Trends. Vol.49 (3), pp.42-53.

[4] Dillenbourg, P., Schneider, D., Synteta,V., (2002) "Virtual Learning Environments," in Proceedings of the 3rd Congress on Information and Communication Technologies in Education, Rhodes, Kastaniotis Editions, Greece, pp.3-18.

[5] Sharples, M. (2000)"The design of personal mobile technologies for lifelong learning," Computers and Education, vol. 34, pp.177-193.

[6] van Harmelen, M. (2006)"Personal Learning Environments," in Proceedings of the 6th International Conference on Advanced Learning Technologies (ICALT'06), IEEE.

[7] Pebble Learning Ltd, (2010) [Online]. Available:http://www.pebblepad.co.uk/

[8] Kennedy, G., Judd, T., Churchward, A., Gray, K. and Krause, K. (2008)" First year students' experiences with technology: Are they really digital natives? "Australasian Journal of Educational Technology, vol. 24(1), pp.108-122.

[9] Mehlenbacher, B., Miller, C. R., Covington, D., and Larsen, J. (2000)"Active and interactive learning online: A comparison of web-based and conventional writing classes," IEEE Transactions on Professional Communication, vol.43 (2), pp.166-184.

[10] Hargittai, E. (2002)"Second-level digital divide: Differences in people's online skills," First Monday, vol.7 (4), pp.1-19.

[11] Kinzie, M. B., Delcourt, M. A. B., and Powers, S. M. (1994)"Computer technologies: Attitudes and self-efficacy across undergraduate disciplines," Research in Higher Education, vol.35 (6), pp.745-768.
[12] Bouffard-Bouchard, T. (1990)"Influence of self-efficacy on performance in a cognitive task," The Journal of Social Psychology, vol.130, pp.353-363.
[13] Compeau, D. R., and Higgins, C. A. (1995) "Computer self-efficacy: Development of a measure and initial test," MIS Quarterly, vol.19, pp.189-211.
[14] Brown, J. S., Collins, A., and Duguid, P. (1989). "Situated cognition and the culture of learning," Educational Researcher, vol.18 (1), pp.32-42.
[15] Igbaria, M, and Parasuraman, S. (1989)"A Path Analytic Study of Individual Characteristics, Computer Anxiety, and Attitudes toward Micro-computers," Journal of Management, vol. 15(3), pp.373-388.
[16] Venkatesh, V. (2000)"Determinants of Perceived Ease of Use Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model," Information Systems Research, vol.11 (4), pp.342-365.
[17] Buche, M.W. Davis, L.R. Vician, C.(2007)"A Longitudinal Investigation of the Effects of Computer Anxiety on Performance in a Computing-Intensive Environment, " Journal of Information Systems Education, vol.18(4), pp. 415-424.
[18] Matsumura, S. and Hann, G. (2004)"Computer anxiety and students' preferred feedback methods in EFL writing," The Modern Language Journal, vol.88 (3).
[19] Maurer, M. M. (1994)"Computer anxiety correlates and what they tell us: A literature review, "Computers in Human Behavior, vol.10 (3), pp. 369–376.
[20] Brown, I. T. J. (2002) "Individual and Technological Factors Affecting Perceived Ease of Use of Web-Based Learning Technologies in a Developing Country," Electronic Journal on Information Systems in Developing Countries. vol. (9) 5
[21] Honeyman, D. S., and White, W. J. (1987)"Computer anxiety in educators learning to use the computer: A preliminary report," Journal of Research on Computing in Education, vol. 20(2), pp.129-138.
[22] Dabbagh, N. (2002) Using a web-based course management tool to support face-to-face instruction, The Technology Source. [Online]. Available: http://technologysource.org/article/389/
[23] Clarke, J., Dede, C., Ketelhut, D., Nelson, B., and Bowman, C. (2006)"A design-based research strategy to promote scalability for educational innovations," Educational Technology, vol.46 (3), pp.27-36.
[24] Burgess, M.L., Slate, J.R., Rojas-LeBouef, A and LaPrairie, K. (2010)"Teaching and learning in Second Life: Using the Community of Inquiry (CoI) model to support online instruction with graduate students in instructional technology," The Internet and Higher Education, vol.13 (1), pp. 84-88.

AUTHORS

Xingmei Qiao is an associate professor in the Sichuan Radio and TV University, China. Her research interests include ICT in Education, Distance Education.

INTENTIONAL BLANK

CLASSIFICATION OF CONVECTIVE AND STRATIFORM CELLS IN METEOROLOGICAL RADAR IMAGES USING SVM BASED ON A TEXTURAL ANALYSIS

Abdenasser Djafri[1] and Boualem Haddad[2]

[1,2] Department of Telecommunication, University of Science and Technology Houari Boumediene, Algiers, Algeria
a.djafri.dz@gmail.dz, h_boualem@hotmail.com

ABSTRACT

This contribution deals with the discrimination between stratiform and convective cells in meteorological radar images. This study is based on a textural analysis of the latter and their classification using a Support Vector Machine (SVM). First, we applied different textural parameters such as energy, entropy, inertia and local homogeneity. Through this experience, we identified the different textural features of both the stratiform and convective cells. Then, we used an SVM to find the best discriminating parameter between the two types of clouds. The main goal of this work is to better apply the Palmer and Marshall Z-R relations specific to each type of precipitation.

KEYWORDS

Meteorology, clouds, stratiform, convective, texture, SVM.

1. INTRODUCTION

Clouds are formed according to two processes: the convection and the progressive uplifting of the air mass (stratification).

The convective uplifting is due to the air instability. It is often vigorous and abrupt. The produced clouds are characterized by a large vertical extension and a limited horizontal extension. These clouds are generally designated by the term “cumulus”. They could be developed in different troposphere levels, where the instability exists.

The synoptic uplifting is the result of dynamic processes in the stable atmosphere, in a stratified flow. This gradual uplifting, producing clouds systems with uniform texture, could cover thousands of square kilometers. These clouds are generally designated by the term “stratus”.

The discrimination between these two types of clouds (stratiform and convective) is very important for applying the specific Palmer and Marshall Z-R relations for weather forecasting. [1]

Numerous researchers have been interested to solve this problem. In 2000, Michael Biggerstaff and Steven Listemaa developed an improved algorithm for the partitioning of radar reflectivity

Natarajan Meghanathan et al. (Eds) : ACSIT, FCST, ITCA, SE, ICITE, SIPM, CMIT - 2014
pp. 119–125, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4611

into convective and stratiform rain classifications. Their algorithm starts with output from the current operational version of the Tropical Rainfall Measuring Mission (TRMM) convective/stratiform scheme for the ground–based validation sites and corrects the output based on physical characteristics of convective and stratiform rain diagnosed from the three-dimensional structure of the radar reflectivity field. The modified algorithm improved the performance of echo classification by correcting two main sources of error. Heavy stratiform rain, originally classified as convective, and the periphery of convective cores, originally classified as stratiform. [2]

In 2005, Maria Franco *et al.* used the vertical profile of reflectivity (VPR) characteristics of the stratiform and convective rain combined with the algorithms by Sanchez – Diezma et al (2000) and Steiner et al (1995) to discriminate between the two kinds of precipitation. [3]

In 2006, Capsoni *et al.* based on the knowledge of the local yearly cumulative distribution function P(R), they studied the space-time evolution and the impact on electromagnetic waves propagation through the atmosphere of the stratiform and convective precipitation. Moreover, they modified the EXCELL rain cell model. [4]

In 2012, Xu Wang *et al.* used fuzzy logic to discriminate between convective and stratiform precipitation in Doppler weather radar. Based on the differences of radar reflectivity distribution morphology between stratiform and convective precipitation, they selected four recognition parameters, which were maximum reflectivity factor, altitude of echo top, vertical reflectivity gradient and vertical reflectivity gradient. [5]

Our approach is to use SVM to discriminate between convective and stratiform precipitation using textural features. We considered four textural parameters, which were energy, entropy, inertia and local homogeneity. Then we used SVM to identify which parameter could distinguish the best between the two types of precipitation.

2. DATA

The data were collected in Dakar (Senegal), Setif (Algeria) and Bordeaux (France). The observed images contain convective and stratiform cells. Our database consists of 400 images collected during 1999 in Dakar, 2000 observed in Bordeaux (France) and 5000 recorded in Setif (Algeria). We present in figure 1 a sequence of 20 images collected on August 13, 1999 in Dakar

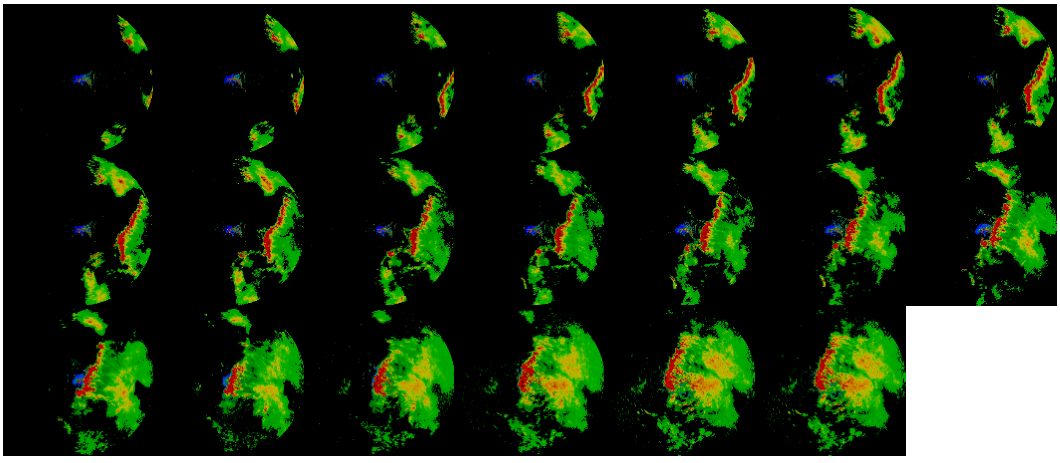


Figure 1. Sequence of meteorological radar images of Dakar.

3. TEXTURAL ANALYSIS OF CLOUDS

The textural features of the radar images of Setif, Bordeaux and Dakar have been calculated by using the histogram approach. The analysis of cloud cells is based on four parameters, namely, energy, entropy, inertia and local homogeneity. Let (x_{ij}) be the grey level of each pixel of the radar image (with i=1,…,N and j = 1,…,N), n_x, the number of pixels at grey level (x) and N_T, the total number of pixels in the image. Its relative frequency is given by:

$$P(x) = {}^{n_x}/_{N_T} \quad (1)$$

The textural parameters are then defined as:

Table 1. Mathematical formulas of textural parameters

Parameter	Mathematical formula
Energy	$E = \sum_i \sum_j x_{ij}^2$
Entropy	$S = \sum_i \sum_j x_{ij} \cdot \log(x_{ij})$
Inertia	$I = \sum_i \sum_j (i-j)^2 \cdot x_{ij}$
Local homogeneity	$HL = \sum_i \sum_j \left[\frac{1}{1+(i-j)^2}\right] \cdot x_{ij}$

First, we made up a database. To do, we considered that the convective cells are represented by a reflectivity factor superior to 42 dBZ, and the stratiform cells are described by a reflectivity factor between 5 dBZ and 40 dBZ. We constituted cloud cells with 4×4 pixels.

Next, we calculated the four textural parameters and we classified them depending on their reflectivity factor. The figures 2.a, 2.b, 2.c and 2.d show respectively the histograms of the four textural parameters, energy, entropy, inertia and local homogeneity.

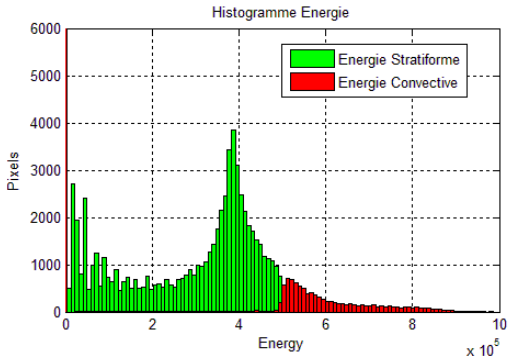


Figure 2.a. Energy histogram

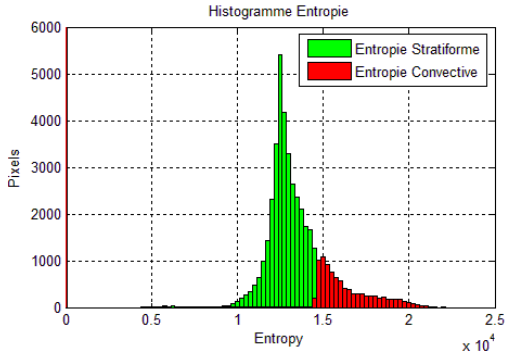


Figure 2.b. Entropy histogram

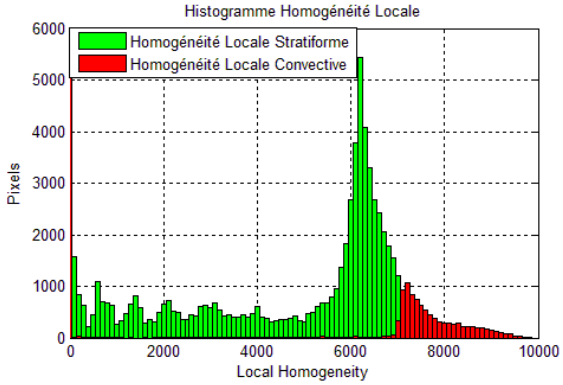


Figure 2.c. Inertia histogram

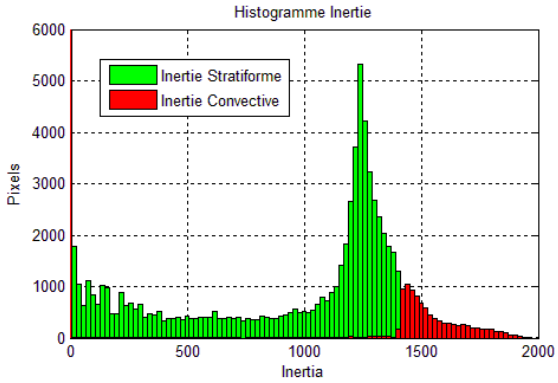


Figure 2.d. Local homogeneity Histogram

4. SVM CLASSIFICATION

Support Vector Machines (Commonly known as SVM) are a set of supervised learning techniques intended to solve discrimination and regression issues. They treat non-linear discrimination and reformulate the classification problem as a quadratic optimization problem.

They could be used to solve discrimination problems (i.e. to decide to which class should a

sample belong), or a regression problems (i.e. to predict the numerical value of a variable).

The concept of the SVM is to find the best hyperplan between the two classes (stratiform and convective). The hyperplan is a separator line that separates the stratiform class from the convective class. The figure 3 shows an example of an hyperplan.

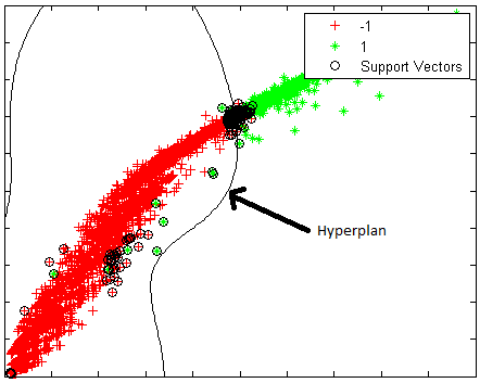


Figure 3. Hyperplan between two classes

The second step of our approach is to use an SVM to find out which one of the four textural parameters distinguishes the best between the convective and stratiform cells of a meteorological radar image. We used an SVM with "RBF" kernel function with default parameters

$$(C = 1 \; and \; sigma = 1)$$

The table 2 gives the correlation factor between the train base and the test base of the SVM for the four textural parameters (energy, entropy, inertia and local homogeneity).

Table 2. Correlation factor between the train base and the test base of the SVM.

Parameter	Energy	Entropy	Inertia	Local homogeneity
Correlation factor	97.01%	76.29%	96.02%	88.58%

According to the table 2, the two parameters (energy and inertia) are the most pertinent for the discrimination between the convective and the stratiform cells of meteorological radar images. The combination of the two parameters in an SVM gives 95.24% of good identification.

For example, we present in figures 3.a and 3.b the classification result of one image recorded in Dakar.

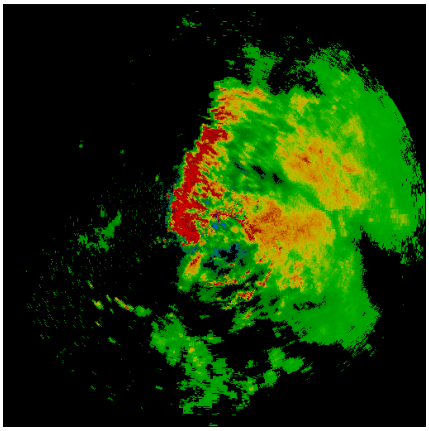


Figure 3.a. None-classified image

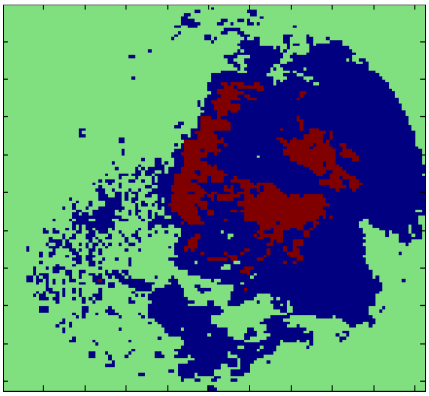


Figure 3.b. Classified image

The output image of the classifier represents the convective cells in red and the stratiform cells in blue. The ratio of good identification is 97%. This shows the efficiency of the proposed approach in the discrimination between the precipitation cells. Note that this approach has not been previously proposed.

5. SVM CLASSIFICATION

The aim of our study is to discriminate between stratiform and convective cells in radar images. This discrimination is very important to optimize the choice of the best Z-R Palmer relation to be used for quantitative rainfall estimation. The texture-based approach presented in this paper allow the classification of the stratiform and convective cells in a meteorological radar image without ambiguity. The proposed approach correctly identifies isolated convective and stratiform cells. The two parameters energy and inertia show the efficiency of this approach by a ratio of good identification of around 97%. It is also proven that the combination of good parameters in an SVM doesn't give necessarily a better result.

Next, the objective will be to see how this method can help to improve the precipitation estimation. This method will be tested on other sites where different climates prevail.

REFERENCES

[1] M. Pidwirny, Fundamentals of physical geography, 2nd ed. McGraw-Hill, 2006, ch. 8.
[2] M. I. Biggerstaff, and S. A. Listemaa, "An improved scheme for convective/stratiform echo classification using radar reflectivity," Journal of applied meteorology, vol. 39, pp. 2129-2150, Dec. 2000.
[3] M. Franco, R. Sanchez-Diezma, D. Sempere-Torres, and I. Zawadzki "Wind profilers and vertical profiles of reflectivity," Presented in the 32nd on radar meteorology, Alvarado, 2005.
[4] C. Capsoni, L. Luini, A. Paraboni, and C. Riva, "Stratiform and convective rain discrimination deduced from local P(R)," IEEE Trans. Antennas and propagation, vol. 45-11, pp. 3566-3569, Dec. 2006.
[5] X. Wang, X. Zheng, J. He, X. Li, "Using fuzzy logic to discriminate convective and stratiform precipitation in Doppler weather radar," in Proc. 9th International conference on fuzzy systems and knowledge discovery (FSKD 2012), 2012, pp. 623-627.

AUTHORS

Prof. Boualem HADDAD **got his degree** in **Telecommunications Engineer** from the national polytechnic school (Algeria) in 1982, his Magister degree in applied electronics in 1991 and his PhD degree in Atmospheric radiation from the University of Sciences and Technology of Algiers in 2000. His areas of interest are electromagnetic radiation, weather radar and instrumentation and atmospheric modeling. He is currently a professor at the University of Sciences and Technology of Algiers since 2001, and member of Radiation and image processing laboratory.

Mr. Abdenasser DJAFRI got his bachelor's degree in Electrical Engineering from the University of Science and Technology of Houari Boumediene (Algeria) in 2008, his Master's degree in telecommunication systems in 2010 and a PhD student till now in the same university. His areas of interest are radiocommunication, weather instrumentation and technology. He served as a member state delegate in the World Radiocommunication Conference of the International Union of Telecommunications (ITU-R WRC Geneva 2012). He also participates in the working party 5D of IMT Systems in the International nion of Telecommunications – Radiocommunication's Sector.

INTENTIONAL BLANK

HYPERSPECTRAL IMAGING WITH LIQUID-CRYSTAL TUNABLE FILTER FOR TISSUES CHARACTERIZATION

Jong-Ha Lee[1*] and Jeonghun Ku[1]

[1] Keimyung University, School of Medicine, Daegu, South Korea
segeberg@kmu.ac.kr

ABSTRACT

We developed and characterized a new near-infrared hyperspectral imaging system. The system consists of a charged coupled device and liquid crystal tunable filter, which is continuously tunable in the near-infrared spectral range of 650-1100 nm with a mean bandwidth of 5 nm. Experiments are conducted to quantitatively determine normal tissues characterization. In the first experiment, hyperspectral images are acquired from normal lung tissue phantom and analyzed. The data shows obvious peak absorption intensity at four different near infrared wavelengths (760, 805, 905, and 970 nm). In the second experiment, the simulated malign lung tissue phantom is used to compare the spectrum between normal and malign tissues. The experimental result indicates that for different type of tissues, the absorption intensity of the spectrum integrated over near-infrared spectral region was considerably different in normal tissues and simulated malign tissues. This difference provides the basis for the detection and delineation of the malign tissue margins from the normal tissues.

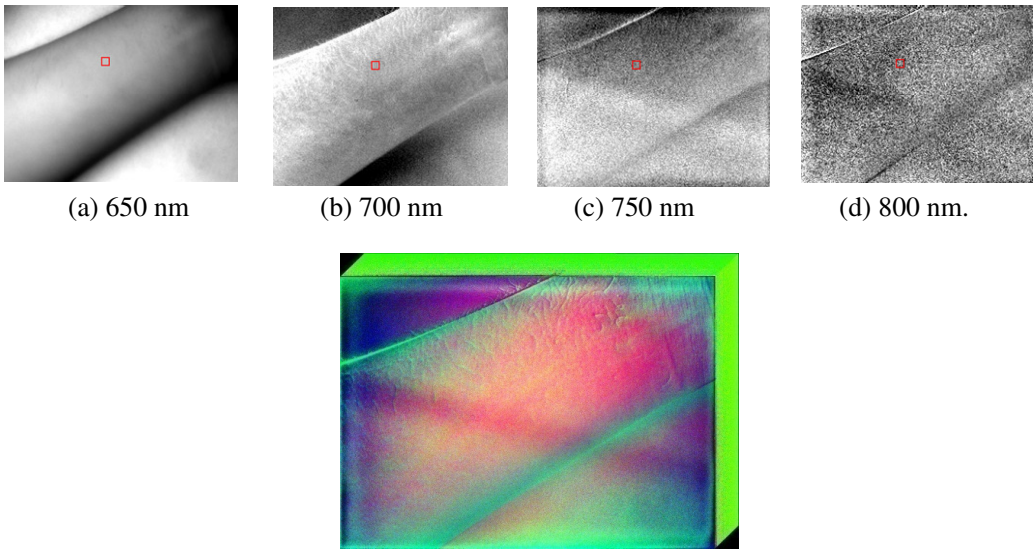
KEYWORDS

Hyperspectral, Tissue, Imager, Non-Invasive Method, Near-Infrared

1. INTRODUCTION

Hyperspectral imaging system (HIS) is a novel method to generate a spectral characteristic map of region of interest (ROI) based on the chemical composition. Previously, HIS has been used in non-medical applications including satellite investigation to find minerals on the ground or to access the condition of agriculture fields. Recently, HIS has been applied to the investigation of pathological changes in living tissue of animal and human. It has proven that HIS can provide valuable information as to the health or disease of tissue that sometimes other modalities are unavailable. HIS is a remote sensing technology to create 2-dimensional image having spectral information in each pixel. This information can be interpreted as the gradient map of species. It means HIS is a method of imaging spectroscopy combining the chemical specificity of spectroscopy with spatial information of imaging. The general concept of hyperspectral imaging is shown clearly in Fig. 1.

Natarajan Meghanathan et al. (Eds) : ACSIT, FCST, ITCA, SE, ICITE, SIPM, CMIT - 2014
pp. 127–137, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4612



(a) 650 nm (b) 700 nm (c) 750 nm (d) 800 nm.

(e) 3D cube from 650 nm to 1100 nm

Figure 1. Near infrared hyperspectral imaging of human arm in the wavelength of (a) 650 nm (b) 700 nm (c) 750 nm (d) 800 nm. (e) 3D hyperspectral cube reconstruction using the image from 650 nm to 1100 nm with 5 nm interval.

Biological tissues have optical signatures that reflect their chemical characteristics. The primary compositions in human tissue are oxyhemoglobin (HbO2) and deoxyhemoglobin (Hb). The Hb further breaks down into melanin, lipids and water (oxygenation). Total hemoglobin (HbT) indicates the combined quantities of HbO2 and Hb. In near infrared (NIR) region, HbO2 and Hb are major sensitive spectrum absorber. Since many diseases have specific effects on tissue oxygen and blood supply, tissue oxygenation and total hemoglobin concentration are major indicators of viability and tissue health. Thus in this paper, we manly focus on the NIR region, particularly the short wavelength NIR of 650 to 1100 nm. By comparing the acquired spectrum absorption measurement in NIR region, information about type, location and relatively concentration of chemical decomposition about the tissue can be quantified. As far as we know, non-invasive, real time, local measurement of tissue oxygenation and total hemoglobin is still not commercially available. In this paper, we construct and characterize of HIS capable of detecting photons in the NIR wavelength region to characterize the tissue condition. HIS we propose here has several advantages compare to other modalities such as CT and MRI. It penetrated into biological tissue deeply without radiation exposure such as CT, thus functional imaging with non-invasive and non-radioactive in real time is available. In addition it is portable and low cost compared to MRI.

The object of this work is to design HIS and investigate the ability of detecting normal tissues and distinguishing normal tissues from malign tissues. HIS is integrated with charged coupled device (CCD) and liquid crystal tunable filter (LCTF) to automatically capturing the spectrum information. LCTF controller is tuned to scanning from 650 nm to 1100 nm bands with 5 nm steps. From the data, the absorption spectrum of HbO2, Hb, lipids and water of normal and malign tissues are characterized and compared. The experimental study demonstrates the system capability of characterization of normal tissues and discriminate it from malign tissues by identifying key wavelengths. Identifying key wavelengths for tissue characteristic provides crucial in reducing the amount of data collected in subsequent specimen studies, thus allowing for rapid, optical, and clinical diagnosis.

2. EXPERIMENTAL SETUP

2.1. System Description

The portable hyperspectral tunable imaging system consists of 1.4 megapixel 12 bit digital imager (Qimaging Inc., Surrey, British Columbia), Liquid Crystal Tunable Filter (LCTF, Cambridge Research & Instrumentation Inc., Woburn, Massachusetts), and LCTF controller. The digital imager is a mono-cooled Charge Coupled Device (CCD) with 1392 (V) × 1040 (H) spatial resolution with 6.45 μm (V) × 6.45 μm (H) individual pixel size. LCTF is placed in front of the digital imager and filters bands in the short wavelength NIR range from 650 nm to 1100 nm. The filter is set to 5 nm full width at half waximum (FWHM). The FWHM is measured as the spectral separation between two points where the filter's transmission attains 50 % of the peak value. The LCTF controller synchronizes between digital imager and LCTF and switches the programmed sequential bands of filter. The tuning speed of the filter is between 50 ms to 150 ms.

All images were captured with 91 bands with center wavelengths separated by 5 nm. Data produced by HIS can be represented by a 3-D cube of image $I(x, y, \lambda_k)$, where (x, y) indicates the spatial coordinate of a pixel in the size, $x = 1, 2, ..., 1,392$, $y = 1, 2, ..., 1,040$, and λ_k denotes the wavelength of the k^{th} spectral band. Each value of $I(x, y, \lambda_k)$ is quantified by a grey scale level and has a minimum value of 0 and has a maximum value of $2^{12} = 4,096$. Individual 3-D cube of images are stored in a 12 bit binary format along a header file containing image parameter information. The data size of one image is approximately 1,392 pixels × 1,040 pixels × 91 bands × 12 bits = 197 megabytes. LCTF tuning, image acquisition, and storage are managed by a software compile by C++. A high-end laptop computer (Apple MacPro 2.53 GHz, Cupertino, CA) manages the instrument control, spectral image acquisition and synchronization. Image visualization is performed using ENVI software environment (Ver. 4.5, ITT Visual information solutions, Boulder, CO). The total scanning of LCTF is about 23 seconds including data transferring to computer and image rendering to screen. Fig. 2. shows HIS system in detail. HIS consists of three part, computer to LCTF controller, LCTF controller to camera, and camera to computer. The synchronization of this system is performed in the computer. Fig. 3. represents HIS transmission characteristics. Total bands from 650 nm – 1100 nm perform the similar transmission characteristics.

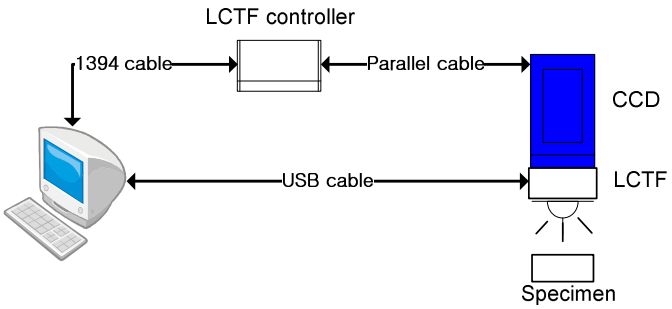


Figure. 2. Schematic diagram of the hyperspectral imaging system.

The connection consists of three parts. From computer to LCTF controller, 1394 cable is connected. From LCTF controller to camera, parallel cable is used. From camera to computer, USB cable is connected. The synchronization of this connection is handled in the computer.

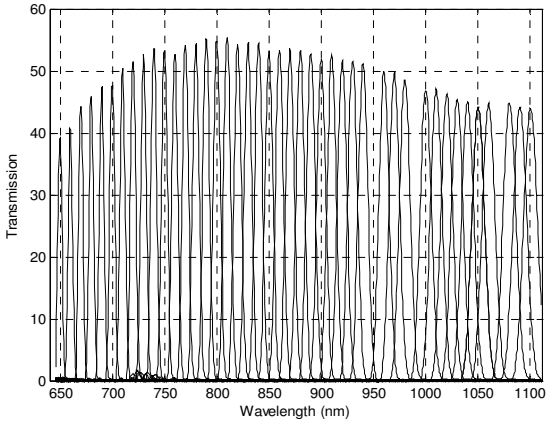


Figure. 3. Hyperspectral imaging system transmission characteristic at NIR 650 nm – 1100 nm. The transmission gives similar transmission characteristic.

2.2. Excitation Light Subsystem

To cover the short wavelength spectral range, two types of light sources are used for illumination. In the preliminary experiments, dual 500 W white quartz tungsten halogen lamps (QTH) range from 360 nm to 2500 nm was used. However, poor luminous efficiency (10 ~ 20 lm/W), non uniform illumination, and the overheating problem, it was not appropriate to use our purpose. To increase the luminous efficiency and prevent the overheating of the sample tissue, low working temperatures and high luminous efficiency Light Emitted Diode (LED) illuminator is considered with the following design. First a custom matching box with 99% reflection coating is machined such that its 10 mm diameter end hole fit into the nosepiece of fiber bundle. Inside a matching box, there is a LED array panel for illumination and a cooling fan to regulate the temperature of inside the box caused by LED array. The LED array panel consists of LR W5AP Osram Diamond Dragon LEDs. The luminous efficiency of this LED is 45 lm/W with 140 degree viewing angle. To make a same performance of 500 W QTH, paralleled 110 LEDs is integrated onto the single Printed Circuit Board (PCB). Since the illuminator needs to focus on the 10 mm diameter hole, the round shape of PCB with 30^{o} is used. Regarding the light guide, an optical fiber bundle consists of Corning SMF-28e optical fibers with 8.2 μm core diameters are customized and 6.35 mm stainless steel houses an optical fiber bundle. The connector between a matching box and a nosepiece of an optical fiber bundle is sealed with aluminum tape to eliminate light leakage. The light is then transmitted through an optical fiber bundle towards a light reflector and illuminates a sample. The illumination angle is within 5^{o} to minimize the shadows and directional scattering caused by the rough surface of the subjects. Fig. 4a and 4b indicates the proposed layout of QTH and LED array excitation light subsystem.

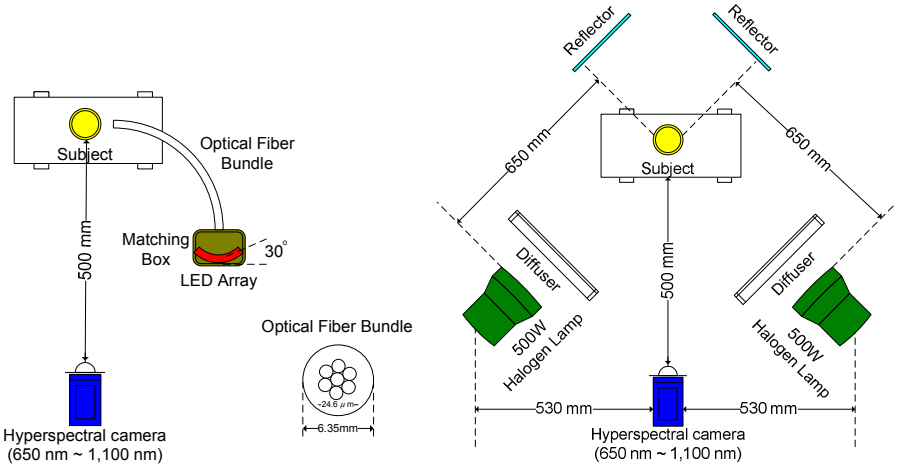


Fig. 4. (a) The layout of a hyperspectral imaging system with dual 500W Halogen Lamp Illuminator. (b) The layout of a hyperspectral imaging system with LED Array Illuminator

2.3. Optics Calibration

Since image intensity data from digital imager have unknown system offsets and gains and may vary over time, an image intensity converted to the reflectance data. According to [1], image intensity at spatial coordinate (x, y) at λ_k band can be modeled as

$$I(x, y, \lambda_k) = L(x, y, \lambda_k)S(x, y, \lambda_k)R(x, y, \lambda_k) + O(x, y, \lambda_k) \quad (1)$$

where $x = 1, 2, ..., 1392$, $y = 1, 2, ..., 1040$, and $k = 1, 2, ..., 91$. $L(x, y, \lambda_k)$ refers to the illumination, $S(x, y, \lambda_k)$ refers to the system spectral response, $R(x, y, \lambda_k)$ refers to the reflectance of the viewed surface, and $O(x, y, \lambda_k)$ is the offset due to the stray of the light. To compensate unknown system offsets and gains, Spectralon diffuse reflectance standards SRS-99 for an approximately 99 % reflectance and SRS-02 for an approximately 2 % reflectance are used (Labsphere, Sutton, NH). These standards used in the calibration were directly traceable to the US National Institute of Standards and Technology (NIST). For the image intensity of SRS-99 spectralon, we have

$$I_{SRS-99}(x, y, \lambda_k) = L(x, y, \lambda_k)S(x, y, \lambda_k)R_{SRS-99}(\lambda_k) + O(x, y, \lambda_k) \quad (2)$$

And for the image intensity of SRS-02 spectralon, we have

$$I_{SRS-02}(x, y, \lambda_k) = L(x, y, \lambda_k)S(x, y, \lambda_k)R_{SRS-02}(\lambda_k) + O(x, y, \lambda_k) \quad (3)$$

where $R_{SRS-99}(\lambda_k)$ and $R_{SRS-02}(\lambda_k)$ are reflectance functions for these two images and theoretically independent of (x, y) because the spectralon surface has the same reflectance property for all image pixels. By using the equations for $I_{SRS-99}(x, y, \lambda_k)$ and $I_{SRS-02}(x, y, \lambda_k)$, we can derive

$$L(x, y, \lambda_k)S(x, y, \lambda_k) = \frac{I_{SRS-02}(x, y, \lambda_k) - I_{SRS-99}(x, y, \lambda_k)}{R_{SRS-02}(\lambda_k) - R_{SRS-99}(\lambda_k)} \quad (4)$$

10 different spectralon images were obtained and averaged to estimate $I_{SRS-99}(u, v, \lambda)$ and $I_{SRS-02}(u, v, \lambda)$.

With this estimates, the final reflectance is given as below:

$$R(x, y, \lambda_k) = \frac{(I(x, y, \lambda_k) - I_{SRS-99}(x, y, \lambda_k))R_{SRS-02}(\lambda_k)}{I_{SRS-02}(x, y, \lambda_k) - I_{SRS-99}(x, y, \lambda_k)} + \frac{(I_{SRS-02}(x, y, \lambda_k) - I(x, y, \lambda_k))R_{SRS-99}(\lambda_k)}{I_{SRS-02}(x, y, \lambda_k) - I_{SRS-99}(x, y, \lambda_k)} \quad (5)$$

Finally the reflectance $R(x, y, \lambda_k)$ of samples are converted to the apparent absorbance, $A(x, y, \lambda_k)$, defines as the logarithm of the ratio between reflectance of the sample $R(x, y, \lambda_k)$, and the reflectance of certified 99 % standard, measured at the wavelength λ_k and the spatial coordinates x, y [2].

$$A(x, y, \lambda_k) = \log \frac{R_{SRS-99}(\lambda_k)}{R(x, y, \lambda_k)} \quad (6)$$

If the interest region is more than one pixel, the apparent absorbance vector of each ROI is averaging over $M + N$ pixels according to

$$A_{ave}(x, y, \lambda_k) = \frac{1}{M + N} \sum_{j=1}^{N} \sum_{i=1}^{M} A(x, y, \lambda_k) \quad (7)$$

This calibration step is performed at the beginning of every experiment.

3. IMAGE ACQUISITION AND ANALYSIS

In this section, we discuss the results of our experimental observations. The aim of these experiments is to demonstrate the potential of HIS for the normal tissue characterization and discriminate it from malign tissues.

3.1. Normal Lung Tissue Characterization

For the normal tissue characteristic experiment, a fresh pig's lung is collected from Temple University Hospital (3401 N. Broad Street. Philadelphia, PA 19140) in April 2008. Totally 91 spectral band images, each image of 1392×1040 pixels in size was obtained within 23 seconds. The 500 W QTH is used for illumination. Samples were kept in iced bags to minimize dehydration and then placed in a tray without ice while hyperspectral images are acquired. To demonstrate the capability of HIS to produce spectral contrast between different regions on the normal lung, four ROI was initially identified based on the visual sense and recorded as normal left lung, normal right lung, normal cardiac notch and normal trachea. The ROI is shown in Fig. 5.

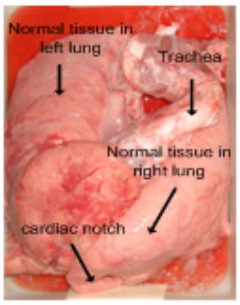


Fig. 5. Four ROI depending on the left lung, right lung, cardiac notch, and trachea.

The primary absorbers of the NIR spectrum in the tissue are deoxyhemoglobin (Hb), oxyhemoglobin (HbO2), bilirubin and methemoglobin, melanin, and water. Previously reported NIR spectrum characterization of biological tissues show Hb having an absorbance peak at 760 nm, HbO2absorbs broadly beyond 800 nm. Bulk lipids have an absorbance peak at 930 nm and water typically peaking at 970 nm [3]. Fig. 6. shows the measured spectrum of ROI of each category. The graph is drawn based on the average of 10×10 pixels apparent absorbance $A(x, y, \lambda_k)$. The range of $A(x, y, \lambda_k)$ is from 0 to 8.307. Within each 10 × 10 pixels ROI, 100 apparent absorption spectrum of each pixel are calculated and averaged. Since Hb and HbO2 blood vessels lying over the surface of the lung, a spectrum has a peak of Hb and HbO2 at 760 nm and 805 nm. The spectrum also indicates significant lipids and water absorption at 905 nm and 970 nm. The wavelength of peak of HbO2 and bulk lipids we acquired is slightly different from the literature [3]. We expect that this is due to the uncelebrated light source and sample condition. In addition to four peaks, the spectrum has a peak value of 1005 nm, 1035 nm, and 1070 nm due to the mixture of constituents in the lung.

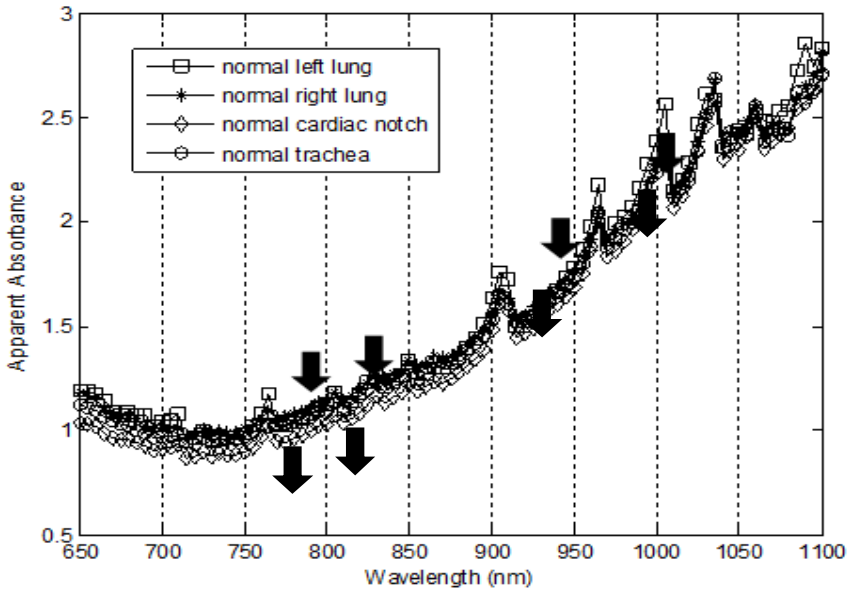


Fig.6. Reconstructed apparent absorbance spectrum of ROI (10 × 10 pixels) from normal left lung, normal right lung, normal cardiac notch, and normal tracheas. Measured spectra within the area were averaged and plotted. Spectrum contains an absorption peak at 760 nm, characteristic of Hb followed 800 nm peak typical of HbO2, with lipids at 930 nm and 970 nm for water peak. The peak beyond 970 nm is for a molecular mixture consistent with known constituents contained within lung.

3.2. Simulated Malign Lung Characteristic

For this experiment, the dehydration of lung sample is prepared to mimic unhealthy tissues with regard to reduced scattering, absorption, and autofluorescence. Same as the previous experiment, 91 spectral band images with each image of 1392 × 1040 pixels in size are obtained. The malign left lung, malign right lung, malign trachea, and malign cardiac notch are determined by 10 × 10 pixels on each category and 100 apparent absorption spectrum of each pixel are calculated and averaged. Fig. 7. shows the spectrum characteristic of malign lung tissues. It reveals the similar peak of 760 nm, 800 nm, 930 nm 970 nm as normal lung tissues, however the absolute value of apparent absorbance is lower than the normal tissue sample. This is because the malign tissue contains less HbT and water to absorb NIR spectrum. In addition, we notice that from 1050 nm, the spectrum is decreasing whereas the spectrum is still increasing in normal tissue sample.

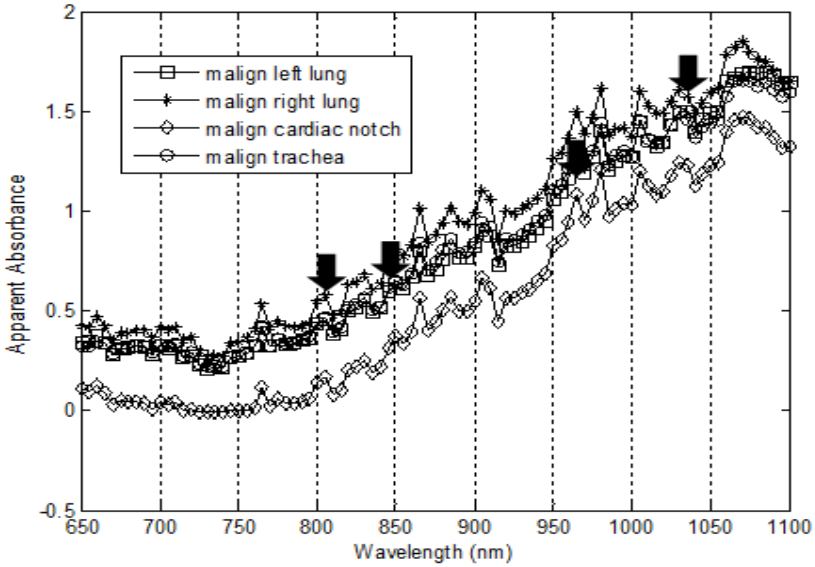


Fig. 7. Reconstructed apparent absorbance spectra of ROI (10 × 10 pixels) from malign left lung, malign right lung, malign cardiac notch, and malign tracheas. The spectrum has similar peak at 760 nm, 805 nm, 905 nm 970 nm.

3.3. Comparison of Normal and Malign Tissue Characteristics

The complete comparison for the wavelength from 650 nm to 1100 nm, is described in the following discussion. An average intensity value for an area of 10 × 10 pixels was obtained for normal left lung, normal trachea, malign left lung, and malign trachea. This average absorbance is plotted in Fig. 8. Notice that the major peak in the normal and malignant data occurs at the same wavelength. However, the intensity of malign tissues is significantly less than normal tissues, thus we can easily discriminate between two types of tissues.

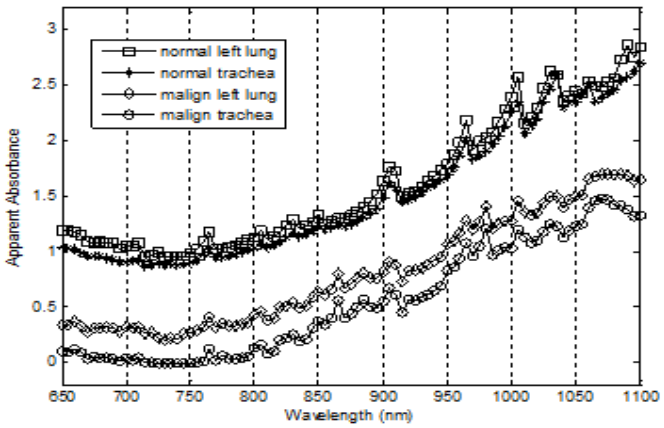


Fig. 8. Comparison of reconstructed apparent absorbance spectra of ROI (10 × 10 pixels) between normal tissue and malign tissues. Two types of tissue has similar wavelength of peak. However the absorbance intensity of malign samples is much lower than normal samples.

It is obvious that an acquisition time of 23s is unacceptable for a clinical instrument, therefore the key wavelength should be chosen. We already find the four key wavelengths of 760 nm, 805 nm, 905 nm, and 970 nm for total hemoglobin, lipids and water. In addition to these wavelengths the ratio between wavelengths are calculated and plotted.

$$r = \lambda_k / \lambda_{k+1},\ k = 1, 2, ..., 90 \tag{8}$$

Since we have totally 91 bands, the number of ratio r would be 8281. Fig. 9. presents the results for the wavelength ratio analysis and Fig. 10. shows the difference of wavelength ratio r between normal and malign tissues. From the results, we notice that when ratio r is 5256, the difference is the biggest. It means the ratio r between 1050 nm and 1100 nm provides the largest difference absorption characteristics between normal and malign tissues. Choosing the key wavelength in this way gives the instrument an acquisition time of approximately one second and differentiates normal and malign tissues simply and effectively.

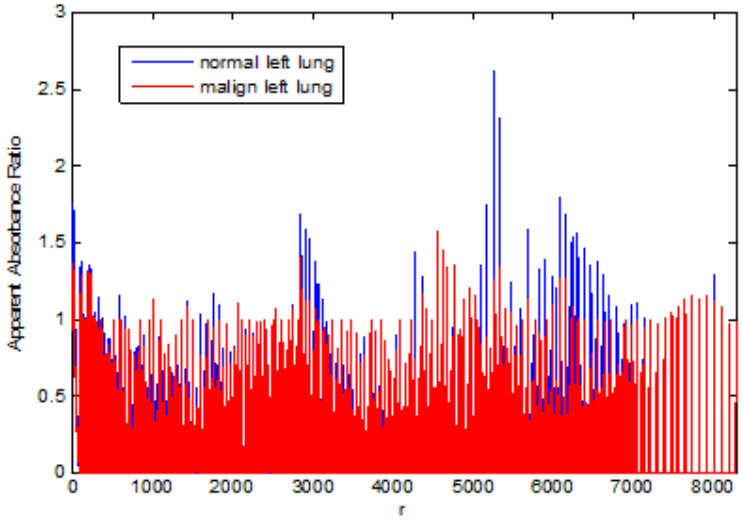


Fig. 9. Wavelength ratio of normal and malign tissues.

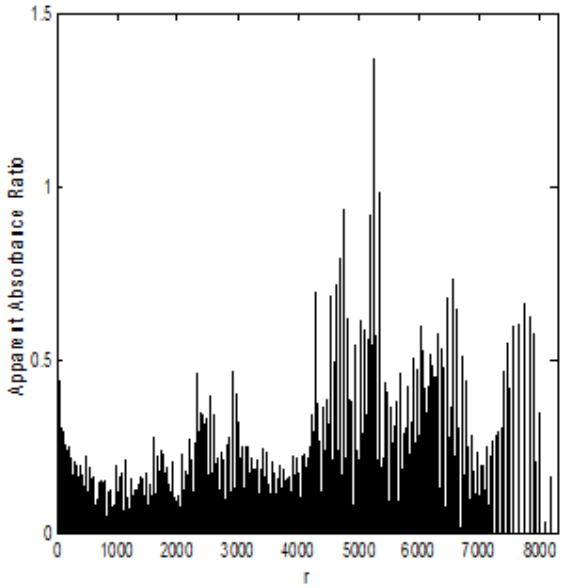


Fig. 10. The difference of wavelength ratio between normal and malign tissues. We observe that when the ratio is 5256, the difference is the biggest.

4. CONCLUSIONS

In this paper, we describe the development of a hyperspectral imaging system that combines several advances in photonics technologies, including CCD and LCTF. The LCTF is tunable over the spectral range of 650 – 1100 nm. The capability of the system has been proven through phantom lung studies. The apparent absorption differences between normal tissue and malignant tissue have been readily seen using this instrument. Throughout the experiment, we find that NIR imaging technology can provide a new modality for measuring changes in total hemoglobin concentration and oxygenation saturation between normal and malign human tissue. A key wavelength is also chosen that provides differentiation between normal and malignant samples. This key wavelength reduces the amount of data collected in subsequent work. The system we propose has obvious applications as a medical diagnostic tool. The modality of hyperspectral imaging combined with other data such as CT or MRI may prove useful in the characterization of normal tissues and detection of malignancies.

REFERENCES

[1] A. Chong, “Digital Near-Infrared Camera for 3D Spatial Data Capture,” The 16th Annual Colloquium of the Spatial Information Research Centre,p2004.

[2] J. Lammertyn, A, Peris, J. Baerdemaeker, and B. Nicolai, “Light Penetration Properties of NIR Radiation in Fruit with respect to Non-destructuctive Quality Assessment ,” Journal of the Postharvest and Technology, vol. 18, pp. 121-132, Feb. 2000.

[3] P. Bargo, T. Goodell, R. Sleven, GL Kovall, G. Blair, and S. Jacques, “Optical Measurements for Quality Control During Photodynamic Therapy,” Plenary Talk Int’l Photodynamic Assoc. Meeting, Jun. 2001.

[4] P. Bargo, “Optical Measurements for Quality Control in Photodynamic Therapy,” Ph. D. Dissertation of Electrical and Computer Engineering, Oregon Health & Science University, Jul. 2003.

[5]P. Taroni, A. Pifferi, A. Torriceclli, and R. Cubeddu, “Time-Resolved Optical Spectroscopy and Imaging of Breast,” Opto-Electronics Review , vol. 12, pp. 249-253, 2004.

[6] Z. Pan, G. Healey, M. Prasad, and B. Trombery, “Face Recognition in Hyperspectral Images,” IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 25, no. 12, pp. 1552-1560, 2003.

[7] K. Zuzak, S. Naik, G. Alexandrakis, D. Hawkins, K. Behbehani, and E. Livingston, "Characterization of a Near Infrared Laparoscopic Hyperspectral Imaging System for Minimally Invasive Surgery," Analytical Chemistry, vol. 79, no. 12, pp. 4709-4713, 2007.

ACKNOWLEDGEMENTS

This research was supported by the MOTIE(Ministry of Trade, Industry and Energy), Korea, under the Inter-Economic Regional Cooperation program (R0002625) supervised by the KIAT(Korea Institute for Advancement of Technology) and this work was supported by Business for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration in 2013

AUTHOR

Jong-Ha Lee

Jong-Ha Lee received the B.S. degree in electronics engineering in 2000 from Inha University, Incheon, Korea, the M.S. degree in electrical engineering in 2005 from Polytechnic Institute of New York University, Brooklyn, NY., and the Ph.D. degree in electrical engineering from Temple University, Philadelphia, PA. He worked at Samsung advanced institute of Technology (SAIT) as a research staff member. Currently, he is an assistant professor with the Department of Biomedical Engineering at Keimyung University, School of Medicine. His current research interests include tactile sensation imaging for tissue characterization, computer-aided diagnosis, medical image analysis, pattern recognition, and machine learning.

AUTHOR INDEX