



International Journal of Advanced Computer Science and Applications

Volume 5 Issue 8

August 2014



ISSN 2156-5570(Online)

ISSN 2158-107X(Print)



www.ijacsa.thesai.org



INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS



THE SCIENCE AND INFORMATION ORGANIZATION
www.thesai.org | info@thesai.org

OAlster

getCITED

Google
Scholar BETA

BASE
Bielefeld Academic Search Engine

ULRICHSWEB™
GLOBAL SERIALS DIRECTORY

arXiv.org

DOAJ DIRECTORY OF
OPEN ACCESS
JOURNALS

IET InspecDirect

INDEX COPERNICUS
INTERNATIONAL

WorldCat
Window to the world's libraries

Microsoft Academic
Search

EBSCO
HOST
Research
Databases

Editorial Preface

From the Desk of Managing Editor...

It is our pleasure to present to you the August 2014 Issue of International Journal of Advanced Computer Science and Applications.

Today, it is incredible to consider that in 1969 men landed on the moon using a computer with a 32-kilobyte memory that was only programmable by the use of punch cards. In 1973, Astronaut Alan Shepherd participated in the first computer "hack" while orbiting the moon in his landing vehicle, as two programmers back on Earth attempted to "hack" into the duplicate computer, to find a way for Shepherd to convince his computer that a catastrophe requiring a mission abort was not happening; the successful hack took 45 minutes to accomplish, and Shepherd went on to hit his golf ball on the moon. Today, the average computer sitting on the desk of a suburban home office has more computing power than the entire U.S. space program that put humans on another world!!

Computer science has affected the human condition in many radical ways. Throughout its history, its developers have striven to make calculation and computation easier, as well as to offer new means by which the other sciences can be advanced. Modern massively-paralleled super-computers help scientists with previously unfeasible problems such as fluid dynamics, complex function convergence, finite element analysis and real-time weather dynamics.

At IJACSA we believe in spreading the subject knowledge with effectiveness in all classes of audience. Nevertheless, the promise of increased engagement requires that we consider how this might be accomplished, delivering up-to-date and authoritative coverage of advanced computer science and applications.

Throughout our archives, new ideas and technologies have been welcomed, carefully critiqued, and discarded or accepted by qualified reviewers and associate editors. Our efforts to improve the quality of the articles published and expand their reach to the interested audience will continue, and these efforts will require critical minds and careful consideration to assess the quality, relevance, and readability of individual articles.

To summarise, the journal has offered its readership thought provoking theoretical, philosophical, and empirical ideas from some of the finest minds worldwide. We thank all our readers for their continued support and goodwill for IJACSA. We will keep you posted on updates about the new programmes launched in collaboration.

Lastly, we would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations.

We hope that materials contained in this volume will satisfy your expectations and entice you to submit your own contributions in upcoming issues of IJACSA

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 5 Issue 8 August 2014
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)
©2013 The Science and Information (SAI) Organization

Editorial Board

Editor-in-Chief

Dr. Kohei Arai - Saga University

Domains of Research: Human-Computer Interaction, Networking, Information Retrievals, Optimization Theory, Modelling and Simulation, Satellite Remote Sensing, Computer Vision, Decision Making Methodology

Associate Editors

Chao-Tung Yang

Department of Computer Science, Tunghai University, Taiwan

Domain of Research: Cloud Computing

Elena SCUTELNICU

"Dunarea de Jos" University of Galati, Romania

Domain of Research: e-Learning Tools, Modelling and Simulation of Welding Processes

Krassen Stefanov

Professor at Sofia University St. Kliment Ohridski, Bulgaria

Domains of Research: Digital Libraries

Maria-Angeles Grado-Caffaro

Scientific Consultant, Italy

Domain of Research: Sensing and Sensor Networks

Mohd Helmy Abd Wahab

Universiti Tun Hussein Onn Malaysia

Domain of Research: Data Mining, Database, Web-based Application, Mobile Computing

T. V. Prasad

Lingaya's University, India

Domain of Research: Bioinformatics, Natural Language Processing, Image Processing, Robotics, Knowledge Representation

Reviewer Board Members

- **Abbas Karimi**
Islamic Azad University Arak Branch
- **Abdel-Hameed Badawy**
Arkansas Tech University
- **Abdelghni Lakehal**
Fsdm Sidi Mohammed Ben Abdellah University
- **Abeer Elkorny**
Faculty of computers and information, Cairo University
- **ADEMOLA ADESINA**
University of the Western Cape, South Africa
- **Ahmed Boutejdar**
- **Dr. Ahmed Nabih Zaki Rashed**
Menoufia University
- **Aderemi A. Atayero**
Covenant University
- **Akbar Hossin**
- **Akram Belghith**
University Of California, San Diego
- **Albert S**
Kongu Engineering College
- **Alcinia Zita Sampaio**
Technical University of Lisbon
- **Ali Ismail Awad**
Luleå University of Technology
- **Alexandre Bouënard**
- **Amitava Biswas**
Cisco Systems
- **Anand Nayyar**
KCL Institute of Management and Technology, Jalandhar
- **Andi Wahju Rahardjo Emanuel**
Maranatha Christian University, INDONESIA
- **Anirban Sarkar**
National Institute of Technology, Durgapur, India
- **Anuranjan misra**
Bhagwant Institute of Technology, Ghaziabad, India
- **Andrews Samraj**
Mahendra Engineering College
- **Arash Habibi Lashakri**
University Technology Malaysia (UTM)
- **Aris Skander**
Constantine University
- **Ashraf Mohammed Iqbal**
Dalhousie University and Capital Health
- **Ashok Matani**
- **Ashraf Owis**
Cairo University
- **Asoke Nath**
St. Xaviers College
- **Ayad Ismaeel**
Department of Information Systems Engineering- Technical Engineering College-Erbil / Hawler Polytechnic University, Erbil-Kurdistan Region- IRAQ
- **Babatunde Opeoluwa Akinkunmi**
University of Ibadan
- **Badre Bossoufi**
University of Liege
- **Basil Hamed**
Islamic University of Gaza
- **Bharti Waman Gawali**
Department of Computer Science & information
- **Bhanu Prasad Pinnamaneni**
Rajalakshmi Engineering College; Matrix Vision GmbH
- **Bilian Song**
LinkedIn
- **Brahim Raouyane**
FSAC
- **Brij Gupta**
University of New Brunswick
- **Bright Keswani**
Suresh Gyan Vihar University, Jaipur (Rajasthan) INDIA
- **Constantin Filote**
Stefan cel Mare University of Suceava
- **Constantin Popescu**
Department of Mathematics and Computer Science, University of Oradea
- **Chandrashekhar Meshram**
Chhattisgarh Swami Vivekananda Technical University
- **Chao Wang**

- **Chi-Hua Chen**
National Chiao-Tung University
- **Ciprian Dobre**
University Politehnica of Bucharest
- **Chien-Pheg Ho**
Information and Communications Research
Laboratories, Industrial Technology Research
Institute of Taiwan
- **Charlie Obimbo**
University of Guelph
- **Chao-Tung Yang**
Department of Computer Science, Tunghai
University
- **Dana PETCU**
West University of Timisoara
- **Deepak Garg**
Thapar University
- **Dewi Nasien**
Universiti Teknologi Malaysia
- **Dheyaa Kadhim**
University of Baghdad
- **Dong-Han Ham**
Chonnam National University
- **Dragana Becejski-Vujaklija**
University of Belgrade, Faculty of organizational
sciences
- **Driss EL OUADGHIRI**
- **Duck Hee Lee**
Medical Engineering R&D Center/Asan Institute for
Life Sciences/Asan Medical Center
- **Dr. Santosh Kumar**
Graphic Era University, Dehradun, India
- **Elena Camossi**
Joint Research Centre
- **Eui Lee**
- **Elena SCUTELNICU**
"Dunarea de Jos" University of Galati
- **Firkhan Ali Hamid Ali**
UTHM
- **Fokrul Alom Mazarbhuiya**
King Khalid University
- **Frank Ibikunle**
Covenant University
- **Fu-Chien Kao**
Da-Y eh University
- **Faris Al-Salem**
- **GCET**
- **gamil Abdel Azim**
Associate prof - Suez Canal University
- **Ganesh Sahoo**
RMRIMS
- **Gaurav Kumar**
Manav Bharti University, Solan Himachal Pradesh
- **Ghalem Belalem**
University of Oran (Es Senia)
- **Giri Babu**
Indian Space Research Organisation
- **Giacomo Veneri**
University of Siena
- **Giri Babu**
Indian Space Research Organisation
- **Gerard Dumancas**
Oklahoma Medical Research Foundation
- **Georgios Galatas**
- **George Mastorakis**
Technological Educational Institute of Crete
- **Gunaseelan Devaraj**
Jazan University, Kingdom of Saudi Arabia
- **Gavril Grebenisan**
University of Oradea
- **Hadj Tadjine**
IAV GmbH
- **Hamid Mukhtar**
National University of Sciences and Technology
- **Hamid Alinejad-Rokny**
University of Newcastle
- **Harco Leslie Hendric Spits Warnars**
Budi Luhur University
- **Harish Garg**
Thapar University Patiala
- **Hamez I. El Shekh Ahmed**
Pure mathematics
- **Hesham Ibrahim**
Chemical Engineering Department, Faculty of
Engineering, Al-Mergheb University
- **Dr. Himanshu Aggarwal**
Punjabi University, India
- **Huda K. AL-Jobori**
Ahlia University
- **Iwan Setyawan**
Satya Wacana Christian University

- **Dr. Jamaiah Haji Yahaya**
Northern University of Malaysia (UUM), Malaysia
- **James Coleman**
Edge Hill University
- **Jim Wang**
The State University of New York at Buffalo,
Buffalo, NY
- **John Salin**
George Washington University
- **Jyoti Chaudary**
High performance computing research lab
- **Jatinderkumar R. Saini**
S.P.College of Engineering, Gujarat
- **K Ramani**
K.S.Rangasamy College of Technology,
Tiruchengode
- **K V.L.N.Acharyulu**
Bapatla Engineering college
- **Kashif Nisar**
Universiti Utara Malaysia
- **Kayhan Zrar Ghafoor**
University Technology Malaysia
- **Kitimaporn Choochote**
Prince of Songkla University, Phuket Campus
- **Kunal Patel**
Ingenuity Systems, USA
- **Krasimir Yordzhev**
South-West University, Faculty of Mathematics and
Natural Sciences, Blagoevgrad, Bulgaria
- **Krassen Stefanov**
Professor at Sofia University St. Kliment Ohridski
- **Labib Francis Gergis**
Misr Academy for Engineering and Technology
- **Lai Khin Wee**
Biomedical Engineering Department, University
Malaya
- **Lazar Stosic**
Collegefor professional studies educators Aleksinac,
Serbia
- **Lijian Sun**
Chinese Academy of Surveying and Mapping, China
- **Leandors Maglaras**
- **Leon Abdillah**
Bina Darma University
- **Ljubomir Jerinic**
University of Novi Sad, Faculty of Sciences,
Department of Mathematics and Computer Science
- **Lokesh Sharma**
Indian Council of Medical Research
- **Long Chen**
Qualcomm Incorporated
- **M. Reza Mashinchi**
- **M. Tariq Banday**
University of Kashmir
- **MAMTA BAHETI**
SNJBS KBJ COLLEGE OF ENGINEERING, CHANDWAD,
NASHIK, M.S. INDIA
- **Mazin Al-Hakeem**
Research and Development Directorate - Iraqi
Ministry of Higher Education and Research
- **Md Rana**
University of Sydney
- **Miriampally Venkata Raghavendera**
Adama Science & Technology University, Ethiopia
- **Mirjana Popvic**
School of Electrical Engineering, Belgrade University
- **Manas deep**
Masters in Cyber Law & Information Security
- **Manpreet Singh Manna**
SLIET University, Govt. of India
- **Manuj Darbari**
BBD University
- **Md. Zia Ur Rahman**
Narasaraopeta Engg. College, Narasaraopeta
- **Messaouda AZZOUZI**
Ziane AChour University of Djelfa
- **Dr. Michael Watts**
University of Adelaide
- **Milena Bogdanovic**
University of Nis, Teacher Training Faculty in Vranje
- **Miroslav Baca**
University of Zagreb, Faculty of organization and
informatics / Center for biomet
- **Mohamed Ali Mahjoub**
Preparatory Institute of Engineer of Monastir
- **Mohamed El-Sayed**
Faculty of Science, Fayoum University, Egypt
- **Mohammad Yamin**
- **Mohammad Ali Badamchizadeh**
University of Tabriz
- **Mohamed Najeh Lakhoua**
ESTI, University of Carthage

- **Mohammad Alomari**
Applied Science University
- **Mohammad Kaiser**
Institute of Information Technology
- **Mohammed Al-Shabi**
Assistant Prof.
- **Mohammed Sadgal**
- **Mourad Amad**
Laboratory LAMOS, Bejaia University
- **Mohammed Ali Hussain**
Sri Sai Madhavi Institute of Science & Technology
- **Mohd Helmy Abd Wahab**
Universiti Tun Hussein Onn Malaysia
- **Mueen Uddin**
Universiti Teknologi Malaysia UTM
- **Mona Elshinawy**
Howard University
- **Maria-Angeles Grado-Caffaro**
Scientific Consultant
- **Mehdi Bahrami**
University of California, Merced
- **Miriampally Venkata Raghavendra**
Adama Science & Technology University, Ethiopia
- **Murthy Dasika**
SreeNidhi Institute of Science and Technology
- **Mostafa Ezziyani**
FSTT
- **Marcellin Julius Nkenlifack**
University of Dschang
- **Natarajan Subramanyam**
PES Institute of Technology
- **Noura Aknin**
University Abdelamlek Essaadi
- **Nidhi Arora**
M.C.A. Institute, Ganpat University
- **Nazeeruddin Mohammad**
Prince Mohammad Bin Fahd University
- **Najib Kofahi**
Yarmouk University
- **NEERAJ SHUKLA**
ITM UNiversity, Gurgaon, (Haryana) Inida
- **N.Ch. Iyengar**
VIT University
- **Om Sangwan**
- **Oliviu Matel**
Technical University of Cluj-Napoca
- **Osama Omer**
Aswan University
- **Ousmane Thiare**
Associate Professor University Gaston Berger of
Saint-Louis SENEGAL
- **Omaima Al-Allaf**
Assistant Professor
- **Paresh V Virparia**
Sardar Patel University
- **Dr. Poonam Garg**
Institute of Management Technology, Ghaziabad
- **Professor Ajantha Herath**
- **Prabhat K Mahanti**
UNIVERSITY OF NEW BRUNSWICK
- **Qufeng Qiao**
University of Virginia
- **Rachid Saadane**
EE departement EHTP
- **raed Kanaan**
Amman Arab University
- **Raja boddu**
LENORA COLLEGE OF ENGINEERNG
- **Ravisankar Hari**
SENIOR SCIENTIST, CTRI, RAJAHMUNDRY
- **Raghuraj Singh**
- **Rajesh Kumar**
National University of Singapore
- **Rakesh Balabantaray**
IIIT Bhubaneswar
- **RashadAl-Jawfi**
Ibb university
- **Rashid Sheikh**
Shri Venkateshwar Institute of Technology , Indore
- **Ravi Prakash**
University of Mumbai
- **Rawya Rizk**
Port Said University
- **Reshmy Krishnan**
Muscat College affiliated to stirling University.U
- **Ricardo Vardasca**
Faculty of Engineering of University of Porto
- **Ritaban Dutta**
ISSL, CSIRO, Tasmaniia, Australia
- **Rowayda Sadek**
- **Ruchika Malhotra**
Delhi Technoogical University
- **Saadi Slami**
University of Djelfa

- **Sachin Kumar Agrawal**
University of Limerick
- **Dr.Sagarmay Deb**
University Lecturer, Central Queensland University,
Australia
- **Said Ghoniemy**
Taif University
- **Sasan Adibi**
Research In Motion (RIM)
- **Sérgio Ferreira**
School of Education and Psychology, Portuguese
Catholic University
- **Sebastian Marius Rosu**
Special Telecommunications Service
- **Selem charfi**
University of Valenciennes and Hainaut Cambresis,
France.
- **Seema Shah**
Vidyalankar Institute of Technology Mumbai,
- **Sengottuvelan P**
Anna University, Chennai
- **Senol Piskin**
Istanbul Technical University, Informatics Institute
- **Seyed Hamidreza Mohades Kasaei**
University of Isfahan
- **Shafiqul Abidin**
G GS I P University
- **Shahanawaj Ahamad**
The University of Al-Kharj
- **Shawkl Al-Dubae**
Assistant Professor
- **Shriram Vasudevan**
Amrita University
- **Sherif Hussain**
Mansoura University
- **Siddhartha Jonnalagadda**
Mayo Clinic
- **Sivakumar Poruran**
SKP ENGINEERING COLLEGE
- **Sim-Hui Tee**
Multimedia University
- **Simon Ewedafe**
Baze University
- **SUKUMAR SETHILKUMAR**
Universiti Sains Malaysia
- **Slim Ben Saoud**
- **Sudarson Jena**
GITAM University, Hyderabad
- **Sumit Goyal**
- **Sumazly Sulaiman**
Institute of Space Science (ANGKASA), Universiti
Kebangsaan Malaysia
- **Sohail Jabb**
Bahria University
- **Suhas J Manangi**
Microsoft
- **Suresh Sankaranarayanan**
Institut Teknologi Brunei
- **Susarla Sastry**
J.N.T.U., Kakinada
- **Syed Ali**
SMI University Karachi Pakistan
- **T C. Manjunath**
HKBK College of Engg
- **T V Narayana Rao**
Hyderabad Institute of Technology and
Management
- **T. V. Prasad**
Lingaya's University
- **Taiwo Ayodele**
Infonetmedia/University of Portsmouth
- **Tarek Gharib**
- **THABET SLIMANI**
College of Computer Science and Information
Technology
- **Totok R. Biyanto**
Engineering Physics, ITS Surabaya
- **TOUATI YOUCEF**
Computer sce Lab LIASD - University of Paris 8
- **VINAYAK BAIRAGI**
Sinhgad Academy of engineering, Pune
- **VISHNU MISHRA**
SVNIT, Surat
- **Vitus S.W. Lam**
The University of Hong Kong
- **Vuda SREENIVASARAO**
School of Computing and Electrical
Engineering,BAHIR DAR UNIVERSITY, BAHIR
DAR,ETHIOPA
- **Vaka MOHAN**
TRR COLLEGE OF ENGINEERING
- **Wei Wei**
- **Xiaojing Xiang**
AT&T Labs

- **YASSER ATTIA ALBAGORY**
College of Computers and Information Technology,
Taif University, Saudi Arabia
- **YI FEI WANG**
The University of British Columbia
- **Yilun Shang**
University of Texas at San Antonio
- **YU QI**
Mesh Capital LLC
- **Zacchaeus Omogbadegun**
Covenant University
- **ZAIRI ISMAEL RIZMAN**

- UiTM (Terengganu) Dungun Campus
- **ZENZO POLITE NCUBE**
North West University
 - **ZHAO ZHANG**
Deptment of EE, City University of Hong Kong
 - **ZHIXIN CHEN**
ILX Lightwave Corporation
 - **ZLATKO STAPIC**
University of Zagreb
 - **Ziyue Xu**
 - **ZURAINI ISMAIL**
Universiti Teknologi Malaysia

CONTENTS

Paper 1: An Information Hiding Scheme Based on Pixel-Value-Ordering and Prediction-Error Expansion with Reversibility
Authors: Ching-Chiuan Lin, Shih-Chieh Chen, Kuo Feng Hwang

PAGE 1 – 7

Paper 2: Semantic Similarity Calculation of Chinese Word
Authors: Liqiang Pan, Pu Zhang, Anping Xiong

PAGE 8 – 12

Paper 3: The Impact and Challenges of Cloud Computing Adoption on Public Universities in Southwestern Nigeria
Authors: Oyeleye Christopher Akin, Fagbola Temitayo Matthew, Daramola Comfort Y.

PAGE 13 – 19

Paper 4: WOLF: a Research Platform to Write NFC Secure Applications on Top of Multiple Secure Elements
(With an Original SQL-Like Interface)

Authors: Anne-Marie Lesas, Benjamin Renaut, Pr. Serge Miranda, Amosse Edouard

PAGE 20 – 31

Paper 5: Optimal Network Design for Consensus Formation: Wisdom of Networked Agents

Authors: Eugene S. Kitamura, Akira Namatame

PAGE 32 – 39

Paper 6: Managing Open Educational Resources on the Web of Data

Authors: Gilbert Paquette, Alexis Miara

PAGE 40 – 47

Paper 7: Development of Duck Diseases Expert System with Applying Alliance Method at Bali Provincial Livestock Office

Authors: Dewa Gede Hendra Divayana

PAGE 48 – 54

Paper 8: Toward Accurate Feature Selection Based on BSS-GRF

Authors: S.M. ELseuofi, Samy Abd El -Hafeez, Wael Awad, R. M. El-Awady

PAGE 55 – 59

Paper 9: A New Efficient Method for Calculating Similarity Between Web Services

Authors: T. RACHAD, J.Boutahar, S.El ghazi

PAGE 60 – 67

Paper 10: Role of Knowledge Reusability in Technological Environment During Learning

Authors: O. K. Harsh

PAGE 68 – 74

Paper 11: Neural Network Based Lna Design for Mobile Satellite Receiver

Authors: Abhijeet Upadhya, Prof. P. K. Chopra

PAGE 75 – 80

Paper 12: Multi-Agent Architecture for Implementation of ITIL Processes: Case of Incident Management Process

Authors: Youssef SEKHARA, Hicham MEDROMI, Adil SAYOUTI

PAGE 81 – 85

Paper 13: Path Planning in a Dynamic Environment

Authors: Mohamed EL KHAILI

PAGE 86 – 92

Paper 14: Online Monitoring System Design of Intelligent Circuit Breaker Based on DSP and ARM

Authors: Meng Song, Liping Zhang, Yuchen Chen, Weijin Zheng

PAGE 93 – 96

Paper 15: Evaluating Usability of E-Learning Systems in Universities

Authors: Nicholas Kipkurui Kiget, Professor G. Wanyembi, Anselemo Ikoha Peters

PAGE 97 – 102

Paper 16: Agent Based Personalized Semantic Web Information Retrieval System

Authors: Dr.M.Thangaraj, Mrs.Mchamundeeswari

PAGE 103 – 110

Paper 17: Social Learners' Profiles in a Distance Learning System Powered by a Social Network

Authors: HROR Naoual, OUMAIRA Ilham/MESSOUSSI Rochdi

PAGE 111 – 116

Paper 18: Regression Testing Cost Reduction Suite

Authors: Mohamed Alaa El-Din, Ismail Abd El-Hamid Taha, Hesham El-Deeb

PAGE 117 – 122

Paper 19: Discovering a Secure Path in MANET by Avoiding Black Hole Attack

Authors: Hicham Zougagh, Ahmed Toumanari, Rachid Latif, Y. Elmourabit, Noureddine.Idboufker

PAGE 123 – 130

Paper 20: Secure Deletion of Data from SSD

Authors: Akli Fundo, Aitenka Hysi, Igli Tafa

PAGE 131 – 134

Paper 21: Cost-Effective Smart Metering System for the Power Consumption Analysis of Household

Authors: Michal Kovalčík, Peter Fecilak, František Jakab, Jozef Dudiak, Michal Kolcun

PAGE 135 – 144

Paper 22: Information Communication Technology Adoption in Higher Education Sector of Botswana: a Case of Botho University

Authors: Clifford Matsoga Lekopanye, Alpheus Mogwe

PAGE 145 – 148

Paper 23: Improving TCP Throughput Using Modified Packet Reordering Technique (MPRT) Over Manets

Authors: Prakash B. Khelage, Dr. Uttam D. Kolekar

PAGE 149 – 156

Paper 24: Local and Semi-Global Feature-Correlative Techniques for Face Recognition

Authors: Asaad Noori Hashim, Zahir M. Hussain

PAGE 157 – 167

Paper 25: Activity Based Learning Kits for Children in a Disadvantaged Community According to the Project “Vocational Teachers Teach Children to Create Virtuous Robots from Garbage”

Authors: Kuntida Thamwipat, Pornpapatsorn Princhankol, Thanakarn Khumphai, Vitsanu Sudsangket

PAGE 168 – 172

Paper 26: Reconsideration of Potential Problems of Applying EMIM for Text Analysis

Authors: D. Cai

PAGE 173 – 181

Paper 27: Multilabel Learning for Automatic Web Services Tagging

Authors: Mustapha AZNAG, Mohamed QUAFAROU, Zahi JARIR

PAGE 182 – 191

Paper 28: cFireworks: a Tool for Measuring the Communication Costs in Collective I/O

Authors: Kwangho Cha

PAGE 192 – 197

Paper 29: Measuring Term Specificity Information for Assessing Sentiment Orientation of Documents in a Bayesian Learning Framework

Authors: D. Cai

PAGE 198 – 206

An Information Hiding Scheme Based on Pixel-Value-Ordering and Prediction-Error Expansion with Reversibility

Ching-Chiuan Lin

Department of Information
Management
Overseas Chinese University
Taichung, Taiwan

Shih-Chieh Chen

Department of Information
Management
Overseas Chinese University
Taichung, Taiwan

Kuo Feng Hwang

Department of Information
Technology
Overseas Chinese University
Taichung, Taiwan

Abstract—This paper proposes a data hiding scheme based on pixel-value-ordering and prediction-error expansion. In a natural image, most neighboring pixels have similar pixel values, i.e. the difference between neighboring pixels is small. Based on the observation, we may predict a pixel's value according to its neighboring pixels. The proposed scheme divides an image into non-overlapping blocks each of which consists of three pixels, and pixels in a block are sorted in a descending order. Messages are embedded into two difference values, where one is between the largest and medium pixels and the other is between the smallest and medium ones. In the embedding process, difference values equal to 0 or greater than 1 are unchanged or increased by 1, respectively, and those equal to 1 are also unchanged or increased by 1 if the message bit to be embedded is equal to 0 or 1, respectively. Calculating the difference value, one may extract a message bit of 0 or 1 if it is equal to 1 or 2, respectively. Recovering pixels is done by decreasing those difference values by 1 if they are equal to or larger than 2. Experimental results demonstrate that the proposed scheme may provide much larger embedding capacity, comparing to existing study, and a satisfied image quality.

Keywords—Reversible data hiding; Pixel-value-ordering; Prediction-error expansion

I. INTRODUCTION

Digital image is a digitized medium stored in an electronic file for presenting objects to people. If someone would like to know more about the image, a separated voice or text file is required, which is an inconvenient way. Alternatively, the owner of the image may type texts on the image for giving more information about the image to a viewer. This may damage or distort the image and the amount of added texts is limited. In addition, distorting an important image, e.g. a medical image, is unaccepted, since a distorted medical image may result in an incorrect diagnosis. Data hiding is a way of imperceptibly embedding important information into a medium, which may provide a way for annotating or watermarking an image, or secret communication. An image with embedded information is called a stego-image. Usually, the stego-image is visually the same as its original image so that people may not perceive the embedded objects.

When data are embedded into an image, the image may be distorted. In general, the more data we embed, the more the

image would be distorted. Embedding capacity and image distortion is a tradeoff. Therefore, a good data hiding scheme should be able to embed as many messages as possible and distort the cover image as slightly as it could. How to embed a large amount of data into an image and achieve a slightly distorted image is an important issue for data hiding applications. The issue is more important if we want the distorted image to be recoverable.

A famous scheme for reversibly embedding messages into an image, based on difference expansion, was proposed by Tian [1] in 2003. Based on differences between neighboring pixels in an image are small, his scheme expands a difference value between two neighboring pixels by increasing or decreasing their pixel values. Specifically, if a difference value d between pixels x and y is calculated as $d = x - y$, the difference is expanded to $2d = x' - y'$, where $x' = x + \lfloor d/2 \rfloor$ and $y' = y - \lfloor (d + 1)/2 \rfloor$ if $x \geq y$. Then, a message bit m is embedded into the expanded difference by setting $x'' = x' + m$, if d is expandable, i.e. x'' and y' are not over or under saturated. Later, for an expandable difference, the embedded message may be extracted by calculating $m = (x'' - y') \bmod 2 = (x' + m - y') \bmod 2 = (2d + m) \bmod 2$, and x and y are recovered by $x = (x'' - m) - \lfloor d'/2 \rfloor$ and $y = y' + \lfloor (d' + 1)/2 \rfloor$, where $d' = (x'' - m - y')/2$. Since we cannot guarantee that every difference is expandable, a location map recording whether a difference is expandable or not is required. Fortunately, most differences are expandable and the location map may be compressed in a satisfied compression ratio so that it consumes only a small part of embedding space. Later, a number of studies [2-5] inspired by Tian's scheme were proposed.

In 2006, a novel reversible data hiding scheme, based on shifting pixel histogram, was proposed by Ni et al. [6]. They calculated the number of pixels with the same pixel value and obtained a pixel histogram where the peak point was selected for embedding messages. Since most images contain few pixels with very small or large pixel values, the histogram on the right or left of peak point may be shifted one to the right or left side, respectively, so that there would be available space for embedding messages. A number of studies based on shifting histogram were proposed. References [7-10] improved Ni et al.'s scheme by shifting difference histogram, instead of

pixel histogram. Since difference between neighboring pixels usually is small, the peak point of difference histogram would be much higher than that of pixel histogram. Generally, these schemes have a higher embedding capacity comparing to those based on shifting pixel histogram.

A group of studies [11-16] explored neighboring pixels in an image and predicted a pixel value by its neighboring pixels. Then they adopted the histogram of predicted error for embedding messages. In general, the peak point of prediction error histogram is higher than difference histogram. Nevertheless, the embedding capacity of this kind of approach depends on its predictive method.

This paper proposes a data hiding scheme based on pixel-value-ordering and predication-error expansion. In a natural image, most neighboring pixels have similar pixel values, i.e. the difference between neighboring pixels is small. Based on the observation, we may predict a pixel's value according its neighboring pixels. The proposed scheme divides an image into non-overlapping blocks each of which consists of three pixels, and pixels in a block are sorted in a descending order. Messages are embedded into two difference values, where one is between the largest and medium pixels and the other is between the smallest and medium ones. In the embedding process, difference values equal to 0 or greater than 1 are unchanged or increased by 1, respectively, and difference values equal to 1 are also unchanged or increased by 1 if the message bit to be embedded is equal to 0 or 1, respectively. Calculating the difference value, one may extract a message bit of 0 or 1 if it is equal to 1 or 2, respectively. Recovering pixels is done by decreasing those difference values by 1 if they are equal to or larger than 2. Experimental results demonstrate that the proposed scheme may provide much larger embedding capacity, comparing to existing study, and a satisfied image quality.

The rest of this paper is organized as follows. Section II briefly reviews Li et al.'s scheme [15]. The proposed scheme is introduced in Section III. Section IV demonstrates our experimental results and compares the performance of the proposed scheme with that of Li et al.'s. Finally, conclusions are given in Section V.

II. RELATED WORK

Li et al. [15] proposed a data hiding scheme based on pixel-value-ordering and predication-error expansion. First, their scheme divides an image into non-overlapping blocks each of which consists of four pixels x_0, x_1, x_2 and x_3 as shown in Fig. 1(a). The four pixel values in a block are sorted in an ascending order as shown in Fig. 1(b) where p_0, p_1, p_2 and p_3 denote the sorted pixels. Then calculate

$$p'_3 = \begin{cases} p_3 & \text{if } p_3 - p_2 = 0, \\ p_3 + m & \text{if } p_3 - p_2 = 1, \\ p_3 + 1 & \text{otherwise,} \end{cases}$$

where p'_3 and m are the stego-pixel of p_3 and the message bit to be embedded, respectively. In Fig. 1(c), the stego-pixel is $p'_3 = 26$ (i.e. $x_1 = 26$) and this block may not embed a message bit since $p_3 - p_2 = 2$. Finally, the stego-pixels would

be $(x_0, x_1, x_2, x_3) = (20, 26, 22, 23)$. If $p_3 = 23$ (i.e. $x_1 = 23$), the embedding result would be unchanged, i.e. $(x_0, x_1, x_2, x_3) = (20, 23, 22, 23)$. In case of $p_3 = 24$, the embedding result would be $(x_0, x_1, x_2, x_3) = (20, 24, 22, 23)$ or $(x_0, x_1, x_2, x_3) = (20, 25, 22, 23)$ if the message bit to be embedded is $m = 0$ or $m = 1$, respectively.

x_0	x_1	x_2	x_3
20	25	22	23

(a)

p_0	p_1	p_2	p_3
x_0	x_2	x_3	x_1
20	22	23	25

(b)

p_0	p_1	p_2	p'_3
x_0	x_2	x_3	x_1
20	22	23	26

(c)

Fig. 1. An embedding process example of Li et al.'s scheme

The recovery process is the reverse of its embedding process. If one would like to extract an embedded message bit from a stego-block and recover a stego-pixel to its original pixel, he/she may calculate

$$m = \begin{cases} 0 & \text{if } p'_3 - p_2 = 1, \\ 1 & \text{if } p'_3 - p_2 = 2, \\ \text{null} & \text{otherwise,} \end{cases}$$

and

$$p_3 = \begin{cases} p'_3 & \text{if } p'_3 - p_2 \in \{0, 1\}, \\ p'_3 - 1 & \text{otherwise.} \end{cases}$$

The rationale of Li et al.'s scheme is based on the concept of pixel values in a block are similar for a natural image. Namely, the difference of pixel values in a block is small and they predicted a difference value (denoted by δ) of one. If the difference is larger than one (i.e. $\delta > 1$), they increased δ by one so that there would be an available space of $\delta = 1$ for embedding a message bit into a block. Then if the message bit m to be embedded is 0 or 1, δ is unchanged or set to 2, respectively.

III. PROPOSED SCHEME

The proposed scheme includes the embedding and extraction processes. The former embeds secret messages into a cover image and obtains a stego-image, and the latter extracts the embedded secret messages from the stego-image and completely recovers it to its original image. Most studies avoid the problem of saturated conditions (i.e. pixel value equal to 0 or 255 for a 256-gray-level image). It is worth mentioning that the proposed scheme also includes a solution for solving the

problem of saturated conditions. The two processes are presented in the following, respectively.

A. Embedding process

This section shows the embedding process for a cover image I with n pixels. Let the secret message, with r bits, to be embedded be a bit string $M = m_0 m_1 \dots m_j \dots m_{r-1}$, where $m_j \in \{0,1\}$ and $0 \leq j \leq r-1$. The embedding process is shown as follows.

1) For a gray-level image I , divide it into non-overlapping blocks each of which contains three neighboring pixels denoted by x_i, x_{i+1} , and x_{i+2} , respectively, where $0 \leq x_i, x_{i+1}, x_{i+2} \leq 255$.

2) For each block, sort their pixel values in a descending order denoted by $p_{i,max}$, $p_{i,med}$, and $p_{i,min}$, where $p_{i,max}$, $p_{i,med}$, and $p_{i,min}$, are the largest, medium, and smallest values, respectively, in block i .

3) Embed m_j by setting stego-pixel values $p'_{i,max}$ and $p'_{i,min}$ of $p_{i,max}$ and $p_{i,min}$, respectively, as

$$p'_{i,max} = \begin{cases} p_{i,max} & \text{if } E_{i,max} = 0, \\ p_{i,max} + m_j & \text{if } E_{i,max} = 1, \\ p_{i,max} + 1 & \text{otherwise,} \end{cases} \quad (1)$$

$$p'_{i,min} = \begin{cases} p_{i,min} & \text{if } E_{i,min} = 0, \\ p_{i,min} - m_{j+1} & \text{if } E_{i,min} = 1, \\ p_{i,min} - 1 & \text{otherwise,} \end{cases} \quad (2)$$

if $p_{i,min} = 1$ or $p_{i,max} = 254$, where $E_{i,max} = p_{i,max} - p_{i,med}$ and $E_{i,min} = p_{i,med} - p_{i,min}$ are the prediction errors of $p_{i,max}$ and $p_{i,min}$, respectively. Mark a block with $p'_{i,min} = 0$ or $p'_{i,max} = 255$ as overhead information. Let $H = h_0 h_1 \dots h_k \dots h_{t-1}$ be the required overhead information for extracting the embedded messages, where t is its length, $h_k \in \{0,1\}$, and $0 \leq k \leq t-1$.

4) Let M_0 be the sub-message embedded in step 3 and $M = M_0 || M_1$, where $||$ denotes concatenating. For each block i with $2 \leq p_{i,min}, p_{i,max} \leq 253$, embed H and M_1 by performing (1) and (2). Overhead information would not be generated in this step since $0 < p'_{i,min}, p'_{i,max} < 255$ in this step.

5) Finally, the stego-image I' is obtained.

In step 2, if the pixels in block i are sorted in an ascending, the embedding process in steps 3–4 would still be workable. However, for simplicity, the embedding process selects the descending order.

The proposed scheme applies the medium value $p_{i,med}$ in a block i to predict its neighboring pixel values $p_{i,max}$ and $p_{i,min}$. The rationale is that, in a natural image, most image blocks are smooth. Thereby we may expect that pixel values of $p_{i,max}$ and $p_{i,min}$ are similar to that of $p_{i,med}$ and the prediction error would be small. This implies that we would obtain more embedding space for embedding a message than without prediction.

The problem of recording saturated blocks with pixels modified is taken into account in step 3, and it is recorded by the overhead information embedded in the blocks mentioned in step 4. Note that we would not encounter the problem of saturated blocks in step 4.

B. Extraction process

Whenever a decoder gets the setgo-image I' , he/she may follow the following process to extract the embedded message and completely recover image I' to its original image I . In the process, the decoder could determine whether a saturated block needs to be recovered or not, from the extracted overhead information. The extraction process is presented as follows.

1) Divide setgo-image I' as it was divided by the encoder in the embedding process.

2) As in the embedding process, for each block i , sort their pixel values x'_i, x'_{i+1} , and x'_{i+2} in a descending order denoted by $p'_{i,max}, p'_{i,med}$, and $p'_{i,min}$, where $p'_{i,max}, p'_{i,med}$, and $p'_{i,min}$ are the largest, medium, and smallest stego-pixel values, respectively, in block i .

3) For each block i with $1 \leq p'_{i,min}, p'_{i,max} \leq 254$, calculate $E'_{i,max} = p'_{i,max} - p'_{i,med}$ and $E'_{i,min} = p'_{i,med} - p'_{i,min}$ and extract

$$HM_v = \begin{cases} 0 & \text{if } E'_{i,max} = 1, \\ 1 & \text{if } E'_{i,max} = 2, \\ \text{Null} & \text{otherwise,} \end{cases} \quad (3)$$

$$HM_{v+1} = \begin{cases} 0 & \text{if } E'_{i,min} = 1, \\ 1 & \text{if } E'_{i,min} = 2, \\ \text{Null} & \text{otherwise,} \end{cases} \quad (4)$$

where HM_v is a bit of mixed messages H and M . Then perform

$$p_{i,max} = \begin{cases} p'_{i,max} & \text{if } E'_{i,max} \leq 1, \\ p'_{i,max} - 1 & \text{if } E'_{i,max} \geq 2, \end{cases} \quad (5)$$

$$p_{i,min} = \begin{cases} p'_{i,min} & \text{if } E'_{i,min} \leq 1, \\ p'_{i,min} - 1 & \text{if } E'_{i,min} \geq 2. \end{cases} \quad (6)$$

Extract H from those blocks with $2 \leq p_{i,min}, p_{i,max} \leq 253$. Note that the bit order of h_k is determined by the block index. Specifically, if h_k is extracted from block a , h_{k+1} would be extracted from block b where $b > a$.

4) Given H in step 3, recover stego-pixels and extract message bits from those blocks with $p'_{i,min} = 0$ or $p'_{i,max} = 255$.

5) According to the block index, rearrange the message bits extracted from blocks, in steps 3 and 4, with $p_{i,min} = 1$ or $p_{i,max} = 254$ to get the sub-message M_0 . Another sub-message M_1 may be extracted from blocks, in step 3, with $2 \leq p_{i,min}, p_{i,max} \leq 253$. Finally, $M = M_0 || M_1$ is obtained and the original cover image I is completely recovered.

Block	x_i	x_{i+1}	x_{i+2}	p_{i_max}	p_{i_med}	p_{i_min}	H/M	p'_{i_max}	p'_{i_med}	p'_{i_min}	x'_i	x'_{i+1}	x'_{i+2}
0	145	148	147	148	147	145	1	149	147	144	144	149	147
1	147	148	149	149	148	147	01	149	148	146	146	148	149
2	146	146	145	146	146	145	1	146	146	144	146	146	144
3	254	250	254	254	254	250		254	254	249	254	249	254
4	1	2	4	4	2	1	0	5	2	1	1	2	5
5	1	2	4	4	2	1	1	5	2	0	0	2	5
6	0	2	3	3	2	0		3	2	0	0	2	3

Fig. 2. An example of the proposed scheme

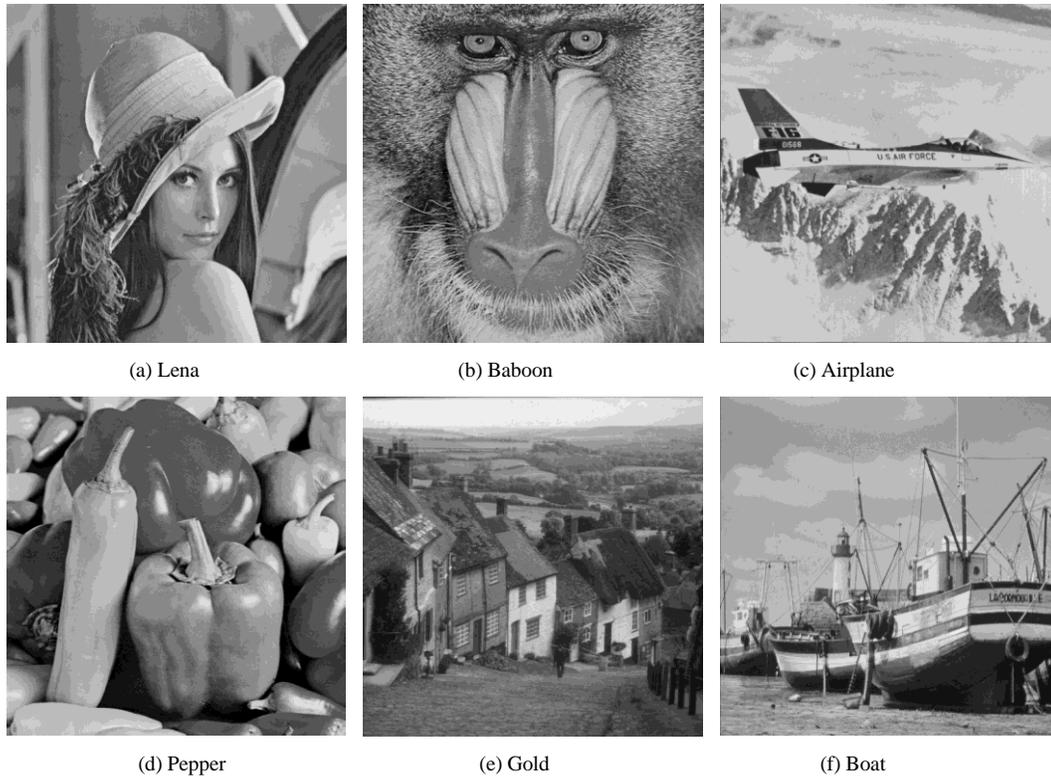


Fig. 3. Test images

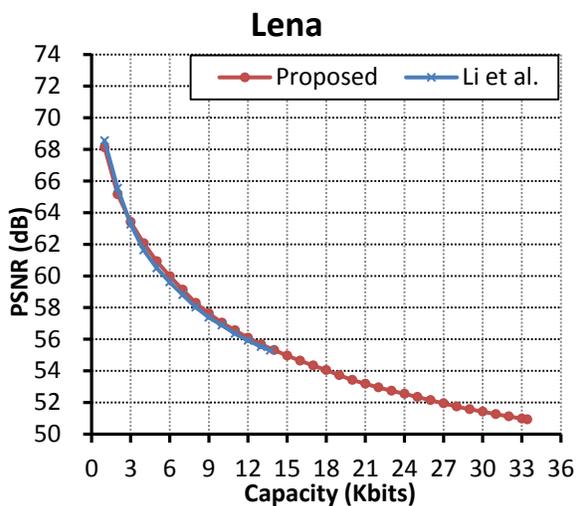
C. An example of the proposed scheme

An example with 7 blocks is given in this section to illustrate the proposed scheme. The example image is a gray-level one with pixel values between 0 and 255.

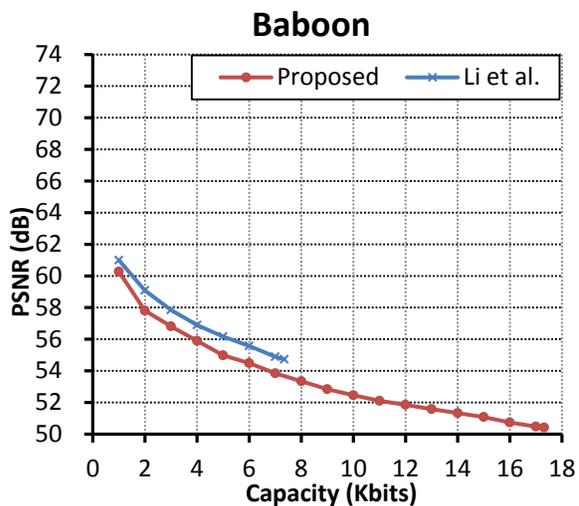
In the embedding process, the divided blocks are shown in Fig. 2, where column H/M denotes both overhead information and message bits to be embedded. Pixel values in a block are sorted in a descending order and the largest and smallest pixel values in a block are marked by red and blue color, respectively. Let the message to be embedded be $M = 0111$. The first processed block is, in step 3, block 3 which embeds nothing, and $m_0 = 0$ is embedded into block 4 by setting $p'_{i_max} = 5$ and $p'_{i_min} = 1$. Then $m_1 = 1$ is embedded into block 5. Block 6 is a saturated block and its pixels is remained unchanged, and $M_0 = 01$ in this step. For simplicity, assume the overhead information is $H = 10$.

In step 4, H is embedded into blocks 0 and 1, and $M_1 = 11$ is embedded into blocks 1 and 2. Note that block 1 embeds two bits, one is from the second bit of H and the other is from the first bit of M_1 . After M_1 is embedded, the embedding process is completed.

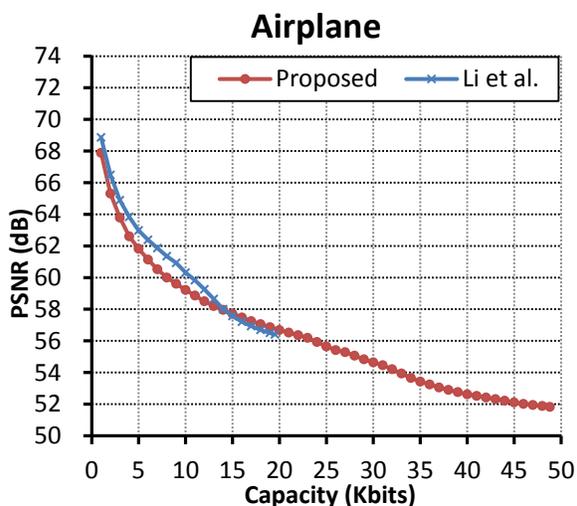
Moving to the extraction process, we divide setgo-image I' into non-overlapping blocks as it was divided by the encoder in the embedding process and sort pixel values in a descending order for each block. First, in step 3, mixed messages of H and M , 10110, are extracted from blocks 0–4 by performing (3) and (4) and these blocks are recovered by performing (5) and (6). Next, the overhead information $H = 10$ is extracted from blocks 0–3, since $2 \leq p_{i_min}, p_{i_max} \leq 253$ in these blocks. For simplicity, assume that $H = 10$ means that block 5 needs to be recovered. Then a bit of 1 is extracted from block 5, and the block is recovered. After all messages are extracted and



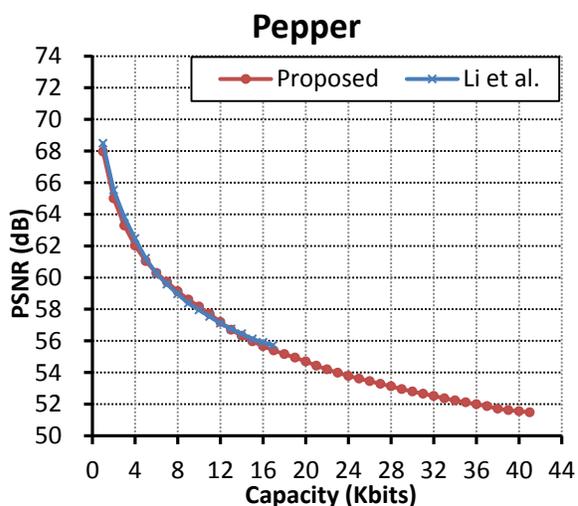
(a) Lena



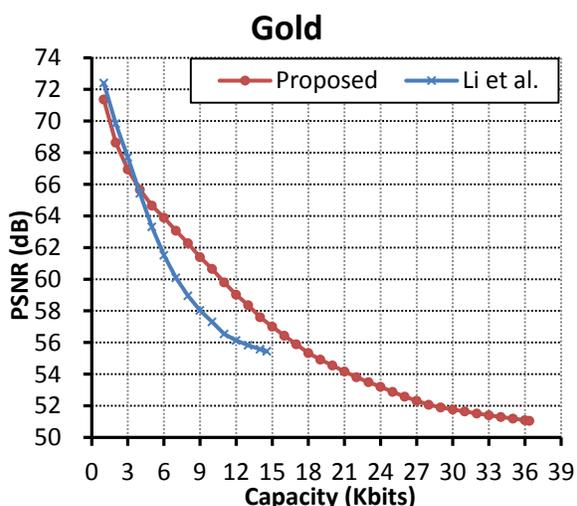
(b) Baboon



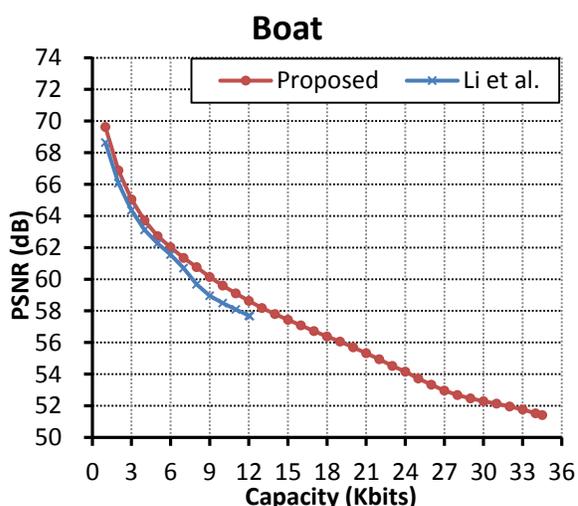
(c) Airplane



(d) Pepper



(e) Gold



(f) Boat

Fig. 4. Comparing performance of the proposed scheme with Li et al.'s scheme

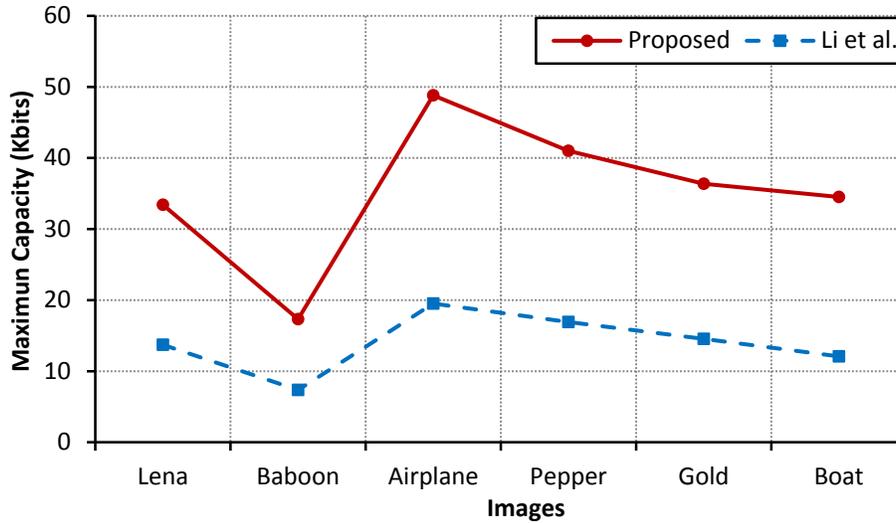


Fig. 5. Comparison of maximum embedding capacity

stego-pixels are recovered, $M_0 = 01$ is obtained from blocks with $p_{i_min} = 1$ or $p_{i_max} = 254$ by rearranging the extracted bit string according to their block indexes. Similarly, $M_1 = 11$ is obtained from the above mixed messages 10110 but removing $H = 10$ and the last bit (i.e. the first bit of M_0). Finally, $M = M_0 || M_1 = 0111$ is extracted and the cover image is completely recovered.

From the above example, we observe that both stego-blocks 5 and 6 are saturated blocks. However, stego-block 5 needs to be recovered and, in the extraction process, the block must be distinguished from stego-block 6 which was not changed in the embedding process. The problem about whether a saturated block needs to be recovered or not may be solved by the embedded overhead information. In step 3 in the extraction process, we first extract messages from stego-blocks with $p'_{i_min} = 1$ or $p'_{i_max} = 254$, since these blocks would not be saturated ones. Then we decode overhead information from blocks with $2 \leq p_{i_min}, p_{i_max} \leq 253$. As long as the overhead information was decoded, we can recognize which saturated blocks need to be recovered. Note that blocks with $p_{i_min} = 1$ or $p_{i_max} = 254$ embed user's messages, if any, instead of overhead information in step 3 in the embedding process. The example in the section has illustrated how the overhead information is embedded into and extracted from a stego-image.

IV. EXPERIMENTAL RESULTS

To evaluate the performance of proposed scheme, we implemented the proposed scheme in Java on a personal computer and embedded randomly generated secret messages into cover images, as shown in Fig. 3, which were downloaded from [17]. All cover images are grayscale with 256 levels and the dimension is 512×512 . A test image may be divided into $\lceil 512 \times 512 / 3 \rceil = 87381$ non-overlapping blocks. In the extreme condition, if a block embeds two bits, the embedding capacity of an image may be up to 174762 bits. Each p_{i_max} and p_{i_min} in a block may be modified no more than one. For a 256-gray-level image with n pixels, the image quality, or the

similarity between a stego-image and its cover image, is evaluated by peak signal to noise ratio (PSNR) calculated as

$$\text{PSNR} = 10 \times \log_{10}(255 \times 255 / \text{MSE}),$$

In the above equation, MSE is mean square error calculated as

$$\text{MSE} = \sum_{j=0}^{n-1} (x_i - x'_i)^2 / n$$

where x_i and x'_i denote cover and stego-pixel values, respectively.

We also implemented Li et al.'s scheme, on the same platform, to compare the performance of our scheme with their scheme's. An image applying the proposed scheme may obtain a larger number of blocks comparing to applying Li et al.'s. Therefore, the image may provide a larger embedding capacity if it applies the proposed scheme.

A smooth image (e.g. Airplane) may contain a larger number of smooth blocks comparing to a complex image (e.g. Baboon). Here a smooth block is an image block with similar pixel values, and it may result in a smaller prediction error. The proposed scheme embeds a message bit into a block with prediction error equal to one which is a smaller prediction error. Since Baboon and Airplane are smooth and complex images, respectively, their prediction errors are usually larger and smaller, respectively, than the other test images.

Fig. 4 illustrates the comparison of performance between the proposed and Li et al.'s schemes in terms of image quality (PSNR) and embedding capacity. The figure shows, in a low embedding capacity, the two schemes have similar performance. We can observe that the PSNR is higher than 50 dB for each test images in Fig. 4, and the proposed scheme may provide image quality similar to Li et al.'s scheme but a much higher embedding capacity than their scheme.

In the worst condition, if p_{i_max} is increased by one and p_{i_min} is decreased by one, we have $\text{MSE} = (1^2 + 1^2) / 3 =$

0.667 and $PSNR = 10 \times \log_{10}(255 \times 255/0.667) = 49.89$ dB. This means the proposed scheme may guarantee the image quality higher than 49.89 dB for a 256-gray-level image. Even in this condition, the difference between a cover image and its stego-image would not be detected by human eye.

Fig. 5 shows the comparison of maximum embedding capacity between the proposed and Li et al.'s schemes. For the test images in Fig. 3, the embedding capacity of the proposed scheme is more than twice as high as Li et al.'s. A reason is that we may embed, at most, two message bits into a block, whereas Li et al.'s scheme may embed, also at most, only one message bit into a block. In addition, the proposed scheme could provide more blocks than Li et al.'s. The proposed scheme may satisfy more applications' requirement if they need a higher embedding capacity and satisfied image quality.

V. CONCLUSIONS

We have introduced an information hiding scheme, with reversibility, based on pixel-value-ordering and prediction-error expansion. The proposed scheme divides an image into non-overlapping blocks each of which contains three pixels and sorts pixels in a block in a descending order. After embedding, the property of pixel-value-ordering in a block is invariant so that the image can be recovered. Comparing to Li et al.'s scheme, the proposed scheme can achieve more blocks for embedding. In addition, a block may embed up to two message bits. Consequently, the proposed scheme can obtain a higher embedding capacity and satisfied image quality. Experimental results show that the proposed scheme achieves an embedding capacity more than twice as high as Li et al.'s scheme on the same level of image quality. The proposed scheme is a good candidate for reversible data-hiding applications which need a high embedding capacity and low distortion.

REFERENCES

- [1] J. Tian, "Reversible data embedding using a difference expansion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, pp. 890–896, August 2003.
- [2] A. M. Alattar, "Reversible watermarking using the difference expansion of a generalized integer transform," *IEEE Transactions on Image Processing*, vol. 13, pp. 1047–1156, August 2004.
- [3] C. C. Chang and T. C. Lu, "A difference expansion oriented data hiding scheme for restoring the original host images," *Journal of Systems and Software*, vol. 79, pp. 1754–1766, December 2006.
- [4] D. M. Thodi and J. J. Rodríguez, "Expansion embedding techniques for reversible watermarking," *IEEE Transactions on Image Processing*, vol. 16, pp. 721–730, March 2007.
- [5] O. M. Al-Qershi and B. E. Khoo, "High capacity data hiding schemes for medical images based on difference expansion," *Journal of Systems and Software*, vol. 84, pp. 105–112, January 2011.
- [6] Z. Ni, Y. Q. Shi, N. Ansari, and W. Su, "Reversible data hiding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, pp. 354–362, March 2006.
- [7] C.-C. Lin and N.-L. Hsueh, "A lossless data hiding scheme based on three-pixel block differences," *Pattern Recognition* vol. 41, pp. 1415–1425, April 2008.
- [8] C.-F. Lee and H.-L. Chen, "Adjustable prediction-based reversible data hiding," *Digital Signal Processing*, vol. 22, pp. 941–953, December 2012.
- [9] W.-L. Tai, C.-M. Yeh, and C.-C. Chang, "Reversible data hiding based on histogram modification of pixel differences," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, pp. 906–910, June 2009.
- [10] C.-C. Lin, W.-L. Tai, and C.-C. Chang, "Multilevel reversible data hiding based on histogram modification of difference images," *Pattern Recognition*, vol. 41, pp. 3582–3591, December 2008.
- [11] P. Tsai, Y.-C. Hu, and H.-L. Yeh, "Reversible image hidings cheme using predictive coding and histogram shifting," *Signal Processing*, vol. 89, pp. 1129–1143, June 2009.
- [12] V. Sachnev, H. J. Kim, J. Nam, S. Suresh, and Y. Q. Shi, "Reversible watermarking algorithm using sorting and prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, pp. 989–999, July 2009.
- [13] W. Hong and T.-S. Chen, "A local variance-controlled reversible data hiding method using prediction and histogram-shifting," *Journal of Systems and Software*, vol. 83, pp. 2653–2663, December 2010.
- [14] W. Hong, T.-S. Chen, Y.-P. Chang, and C.-W. Shiu, "A high capacity reversible data hiding scheme using orthogonal projection and prediction error modification," *Signal Processing*, vol. 90, pp. 2911–2922, November, 2010.
- [15] X. Li, J. Li, B. Li, and B. Yang, "High-fidelity reversible data hiding scheme based on pixel-value-ordering and prediction-error expansion," *Signal Processing*, vol. 93, pp. 198–205, January 2013.
- [16] B. Ou, X. Li, Y. Zhao, and R. Ni, "Reversible data hiding using invariant pixel-value-ordering and prediction-error expansion," *Signal Processing: Image Communication*, in press.
- [17] <http://sipi.usc.edu/database/>

Semantic Similarity Calculation of Chinese Word

Liqliang Pan, Pu Zhang, Anping Xiong

Institute of technology of computer communication
Chongqing University of Posts and Telecommunications
Chongqing, China

Abstract— This paper puts forward a two layers computing method to calculate semantic similarity of Chinese word. Firstly, using Latent Dirichlet Allocation (LDA) subject model to generate subject spatial domain. Then mapping word into topic space and forming topic distribution which is used to calculate semantic similarity of word(the first layer computing). Finally, using semantic dictionary "HowNet" to deeply excavate semantic similarity of word(the second layer computing). This method not only overcomes the problem that it's not specific enough merely using LDA to calculate semantic similarity of word, but also solves the problems such as new words(haven't been added in dictionary) and without considering specific context when calculating semantic similarity based on semantic dictionary "HowNet". By experimental comparison, this thesis proves feasibility, availability and advantages of the calculation method.

Keywords— semantic similarity; LDA; subject model; HowNet

I. INTRODUCTION

The semantic similarity calculation methods of word have been widely used in question-answering system, information retrieval, machine translation, etc. Different application Background have different definition of semantic similarity. In question-answering system and information retrieval, semantic similarity of word mainly focuses on the approximate degree of synonymy or same-meaning. While in machine translation it focuses on the approximate degree of mutual substitution in different contexts. The application background of this paper is Chinese question-answering system. So the understanding of word semantic similarity is approximate degree of synonymy of two words without caring about contexts. Semantic similarity of two words is higher if they are more synonymy in different contexts, otherwise the similarity is lower.

There are mainly two semantic similarity computing methods of word[1]. One is counting word information in documents, the other is constructing knowledge of "world". The first method, using statistical information of word to calculate word semantic similarity, is based on aggregation phenomenon of the analogue. The method is objective and specific, so it can reflect similarity and difference of word in syntactic, semantic, pragmatic, etc. However, the method is dependent on training corpus and counting algorithm. In addition, this method is easily interfered by data sparsity and noise. Sometimes there are some obvious errors. For example, using LDA(Latent Dirichlet Allocation) subject model[2] to generate distribution of subject-word and document-subject. Words are aggregated according to topics, so words in the same topic have semantic similarity. The second method, using knowledge of "world", is based on the fact that everything is interrelated. Generally it describes the characteristic of word

and relation of word using special description-language and building a structure like dictionary. For example semantic dictionary "HowNet" describes the connections of word through relationship of "sememe" and reflects synonymy of word through the approximate degree of similarity of sememe [1]. The method accurately reflects semantic similarities and differences of word, but the result obtained by this method is greatly influenced by subjective consciousness. From the perspective of development of things, construction dictionary can't be completed and can't keep pace with the times, thus it can not accurately reflect objective facts.

Above all, The two kinds of semantic similarity computing methods both have advantages and disadvantages. The thesis puts forward a new semantic similarity computing method (two layers computing method) by combining the two methods and redefining similarity calculation method of word. Firstly, The method uses LDA subject model to excavate topic-word distribution. Using LDA topic model reflects the objective existence of word. Then thesis uses semantic dictionary "HowNet" to further excavate the semantic similarity of word which reflects the objective substantiality of word. The new method lays foundation for similarity judgment of question sentence in Chinese question-answering system.

II. THE FIRST LAYER SEMANTIC SIMILARITY CALCULATION

A. Problem description

Sentence C1: What is the fastest search engine in search field?

Sentence C2: In Chinese retrieval, Baidu is more efficient than Google.

We can see that there are no common words between C1 and C2, but they are still similar. The reason is that Google and Baidu are two specific examples of Search engine. In fact, we often encounter those problems such as correlation and similarity of word and sentence in Search engine algorithm and question-answering system. In traditional information retrieval field, there have been a lot of methods to measure sentence similarity, such as the classical VSM model. However, those methods are often based on a assumption that the more repetition of words between sentences, the more similar they are. Through the example above, we can see that it does not conform to the reality. Most of the time, the approximate degree of synonymy of sentences depends on semantic relations behind words rather than repetition of words, especially suitable for short texts and questions with few words. Therefore, we need to adopt LDA topic model to find subject distribution behind words and judge semantic similarity of word.

B. Brief introduction of Latent Dirichlet Allocation subject model

LDA subject model, proposed by Blei and etc, is a three layers Bayesian generative model—text-topic-word [2]. The essence of LDA is to find topic structure of text using feature of words co-occurrence in text. In generation process, each text is represented as mixture distribution of subjects, and each subject is a probability distribution over words. Based on pLSA[4], leading a hyper-parameter α into the model's document-topic probability distribution, thus the new model obeys Dirichlet distribution. Then Griffiths and etc apply Dirichlet prior distribution to another parameter β , which makes the LDA subject model come into being a completed model. The model is represented by Fig. 1, with the meanings of symbols shown in table 1.

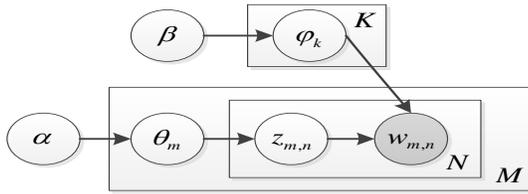


Fig.1. LDA probability graph model

TABLE I. SYMBOL IN LDA MODEL

Symbol	Meaning	Symbol	Meaning
α	Hyper-parameter of θ	$w_{m,n}$	word
β	Hyper-parameter of φ	M	Text No.
θ_m	Text-topic probability distribution	N	word No.
φ_k	Topic-word probability distribution	K	Topic No.
$z_{m,n}$	Distribution of words in a topic		

According to Fig. 1, the Joint probability distribution of LDA is:

$$p(w_m, z_m, \theta_m, \Phi | \alpha, \beta) = \prod_{n=1}^N p(w_{m,n} | \varphi_{z_{m,n}}) p(z_{m,n} | \theta_m) p(\theta_m | \alpha) \bullet p(\Phi | \beta) \quad (1)$$

We often set Hyper-parameter $\alpha = 50/K, \beta = 0.1$, K is number of topics. Seeing [4] for detailed information of choosing α and β values.

we can estimate the parameters using:

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^N n_k^{(t)} + \beta_t} \quad (2)$$

$$\theta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^K n_m^{(k)} + \alpha_k}$$

Where $n_k^{(t)}$ denotes to the number of times that word t has been observed with topic k, $n_m^{(k)}$ denotes to the number of times that topic k has been observed with a word of document m. If

you want more detailed information, you can see the paper of Blei [4].

C. Semantic similarity Calculation method of word in subject spatial domain

Running LDA topic model and doing Gibbs sampling on the document corpus D, we get K topics hidden in the documents and topic-word probability distribution Φ . The element $\varphi_{s_k w_i}$ of Φ shows the probability of word w_i belongs to topic s_k ($1 \leq k \leq K$).

K Topics build a feature space: $V = (s_1, s_2, s_3, \dots, s_k)$

(3) So the word w_1 and w_2 distribution vector in K topics feature space is:

$$V_{w_1} = (\varphi_{s_1 w_1}, \varphi_{s_2 w_1}, \varphi_{s_3 w_1}, \dots, \varphi_{s_k w_1})$$

$$V_{w_2} = (\varphi_{s_1 w_2}, \varphi_{s_2 w_2}, \varphi_{s_3 w_2}, \dots, \varphi_{s_k w_2})$$

The semantic similarity calculation of two words w_1 and w_2 is:

$$Sim(w_1, w_2) = \cos ine(V_{w_1}, V_{w_2}) = \frac{V_{w_1} \bullet V_{w_2}}{|V_{w_1}| |V_{w_2}|} \quad (4)$$

The value of (4) is higher, the similarity of two words w_1, w_2 is more approximate, vice versa.

III. THE SECOND LAYER SEMANTIC SIMILARITY CALCULATION

A. Problem description

Sentence C1: What is the fastest search engine in search field?

Sentence C2: In Chinese retrieval, Baidu is more efficient than Google.

Sentence C3: The search result on Google is more accurate than on Baidu.

By constructing topic spacial, we find that Search engine, Google and Baidu have semantic similarity by calculating their subject distribution cosine (4). Concluding that C1 has similarity with C2 and C3. But after doing further analysis, we find that C1 describes search speed, C2 describes efficiency of retrieval, and C3 describes search accuracy. In other words, searching C1 on Search Engine, we expect that the feedback is more about performance information of search engine or not. So we need further judge synonymy of other words. As we all know, there have synonymy among speed, efficiency and accuracy, but the semantic similarity between speed and efficiency is higher than between speed and accuracy. Of course, we also see that the topic spatial domain created by LDA topic model can judge the correlation between words through calculating their topic distribution cosine (4), but for further specific semantic information of words can not be presented. In order to make up this shortcoming, we use the

following method based on semantic dictionary "HowNet" to analyze specific semantic similarity between words.

B. Brief introduction of "HowNet"

"HowNet" is a common sense knowledge bases, of which description objects are concepts and semantic items, and can describe Chinese and English word using description objects [1]. Using the basic content of "HowNet" to compute the relationship of words or phrase. As the meaning of Chinese words are very complex, its semantic meanings are different in different contexts. So one word are described as the collection of several semantic items and concepts in "HowNet". "HowNet" use "sememe" to future describe semantic items. Special word "sememe" is the smallest unit of semantic meaning and does not vary with the contexts.

Sememes are the most basic unit of describing the meaning item and exiting complicated relations[1]. In "HowNet", there are eight relations of sememe: hyponymy, synonymy, relative, antonymy, part-whole, attribute-host, event-role, materials-production. Hyponymy is the most important sememe relation. It is a kind of hierarchy system, which is described through tree structure which is easy to operate by computer. The top describe abstract concepts and the bottom describe specific concepts. As follows, we will use the hyponymy relation of sememe to compute semantic similarity of words. If you want more concreteness calculation, you can take other relations of sememe into account .

C. Similarity computing method of word based on "HowNet"

There are two Chinese words: w_1 and w_2 . Assume w_1 has n semantic items, w_2 has m semantic items. And the similarity of w_1 and w_2 is the biggest similarity of their semantic items.

Thus, the similarity between two words is transformed into the similarity of two semantic items. Of course, the specific context of two words is not considered here. Actually it is best to use sentence context to disambiguate words first. In other words, designating the word for a particular semantic item. then computing similarity of corresponding semantic items, which is more accurate and will be further researched in future.

By observing semantic dictionary "HowNet", Finding that semantic items are divided into function semantic items and notional semantic items. So the description of semantic items is different with different classes in "HowNet". Function semantic item is described in {relation sememe} or {syntactic sememe}. So, function semantic item only needs to compute the similarity of corresponding relation sememe or syntactic sememe. However, descriptions of notional semantic item are more complex and are divided into four parts:

- 1) *The first independent sememe Description: The first sememe of independent sememes (without special symbols or relation symbol in front of sememe).*
- 2) *Other independent sememes Description: Specific words and Independent sememes except the first sememe.*
- 3) *The relation sememe Description: Sememe Described in relation symbol.*

4) *The symbol sememe Description: Sememe Described in special symbol.*

So, Notional semantic items S_1, S_2 similarity calculation are divided into four parts and each parts similarity marked as $Sim_i (1 \leq i \leq 4)$. Different parts have different weight β_i . The first part present the main semantic of word, so it have the highest weight. In order to lower the weight of other parts. The calculation formula is as follows :

$$Sim(S_1, S_2) = \sum_{i=1}^4 \beta_i \prod_{j=1}^i Sim_j(S_1, S_2) \quad (5)$$

In(5),the $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$ and $\beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$.reflecting the latter parts have lower significance to the overall similarity. You can adjustable the parameters β_i .

In computing similarity between function word and notional word, we know that the possibility of same semantic they both express is very small in actual application. So we think the similarity of function word and notional word is always zero in the thesis.

Finally, all of similarity calculation of semantic items are ultimately attributed to similarity calculation of sememe. We use the hyponymy relation of sememe to compute semantic similarity of sememe. Obtained by experimental analysis:

$$Sim(p_1, p_2) = \frac{\alpha}{Dis\ tan\ ce(p_1, p_2) + \alpha} \quad (6)$$

p_i present sememe, $Dis\ tan\ ce(p_1, p_2)$ is the path length of p_1, p_2 in hierarchy tree. α is a parameter can be adjusted according to the practical application. know more information about "HowNet" [1].

IV. THE TWO LAYERS SEMANTIC SIMILARITY CALCULATION METHOD

The similarity calculation method of word based on LDA subject model Sim_1 embodies characteristic of words co-occurrence. The similarity calculation method of word based on semantic dictionary "HowNet" Sim_2 reflects the semantic connection of words. We combine the two algorithms to acquire a two layers semantic similarity calculation method Sim . If the words have similar subject distribution and semantic connection, the similarity of words should be high, Vice versa.

Computing similarity of words w_1 and w_2 use:

$$Sim(w_1, w_2) = \gamma_1 Sim_1(w_1, w_2) + \gamma_2 Sim_2(w_1, w_2) \quad (7)$$

The γ_1 and γ_2 can be adjusted according to actual application.

V. EXPERIMENTS AND RESULTS

A. Preparations of Latent Dirichlet Allocation subject model

- Experimental data

document number M Using the complete version of Chinese text classified corpus of Sougou laboratory(107M), The text sets have 10 categories, including automobile, finance, IT, health, sports, tourism, education, employment, culture, military(Each category has 8000 pieces, 80000 pieces of document in total).You can get this data sets from [8].

- Experimental setup

Preprocess Do preprocessing, word segmentation, erasing stop-word to original documents. Algorithm of Chinese word segmentation adopts ICTCLAS segmentation system of Chinese Academy of Sciences. Algorithm of delete stop-word adopts conventional removal method at the beginning and then repeatedly observing generating data, writing regular expressions to remove some words(for example name entities and no specific meaning words such as time, place) again. Erasing stop-word can lower the spatial dimension of word which is useful for computing semantic similarity of words. The final word dimension is 207499(N word number). As we know, Chinese word is a combination structure with single characters and the combination method is very complex. It leads to very high word dimension. Reducing word dimension should be further study features of Chinese words formation.

Topic number K Abstract 20000 documents from M (each categories have 2000 documents) to acquire the most suitable topic number. By observing perplexity-index to determine number of topic. The perplexity-index represents uncertainty when forecasting data. The lower value, the better performance. The calculation formula is as follows[10]:

$$Perplexity(D) = \exp \left\{ - \frac{\sum_{m=1}^M \log p(w_m)}{\sum_{m=1}^M N_m} \right\} \quad (8)$$

In (8), N_m denotes the length of document m , M denotes the documents sets. $p(w_m)$ denotes the possibility of word w in document m creating by LDA and It's calculation method as follows:

$$p(w_m) = \sum_d \prod_{n=1}^N \sum_{j=1}^K p(w_i | z_i = j) p(z_i = j | w_m) p(d) \quad (9)$$

Three experiments are made to set subject number. Each experiment as 10-100 (interval 10 add). The Fig.2 shows that topic number and perplexity-index present inverse relation. When topic number is about 97, Decline trend of perplexity-index is not obvious. Bigger topic number is, Calculation of

LDA subject model's parameters estimating is more complicated, so setting $K=100$.

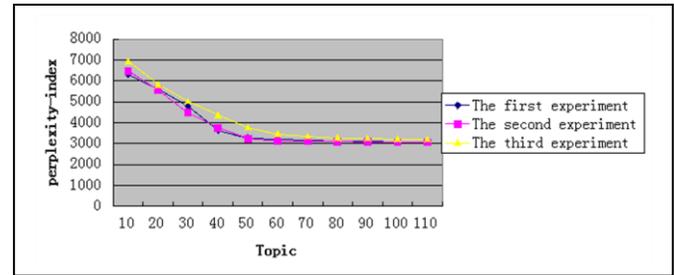


Fig.2. The relation of topic number and perplexity-index

Other parameters $\alpha = 50 / K, \beta = 0.1$.

B. Preparations of semantic dictionary "HowNet"

Data sets Quoting two data sets sorted out by Liuqun (gloss.dat, whole.dat). Gloss.dat stores description of semantic item of words(66142 records in total). Whole.dat stores hierarchy relation of sememe(1618 records in total). As gloss.dat data is massive and access frequency is high, gloss.dat is stored into mysql database. It makes search more faster.

Parameter settings

$$\alpha = 1.6, \beta_1 = 0.5, \beta_2 = 0.2, \beta_3 = 0.17, \beta_4 = 0.13$$

C. Preparation of computing method on two layers

Parameter setting Set γ_1 and γ_2 as 0.5, you can change the value according application.

D. Experimental

Table2 shows result of three semantic similarity computing methods of Chinese word. We choose seven groups word, detail information seeing experimental result.

Method 1: Based on LDA subject model described in the thesis

Method 2: Word's semantic similarity computation Based on the HowNet by Qun.Liu[1].

Method 3: The two layers Semantic similarity calculation method

TABLE II. THREE SEMANTIC SIMILARITY COMPUTING METHODS OF CHINESE WORD

No.	Phrases 1	Phrases 2	Method 1	Method 2	Method 3
1	Search engine	Google	0.999994	0.000000	0.499997
	Search engine	Baidu	0.999999	0.000000	0.499995
	Google	Baidu	0.999986	0.000000	0.499993
2	Speed	Efficiency	0.304053	0.557143	0.430598
	Speed	Accuracy	0.132498	0.557143	0.344821
	Efficiency	Accuracy	0.183966	0.588889	0.386428
	Patient	sick person	0.989468	0.500000	0.744734

3	Patient	Doctor	0.760043	0.588889	0.674466
	Patient	Disease	0.818214	0.093023	0.455619
4	Red	Pink	0.935126	0.700000	0.817563
	Red	Light red	0.942354	0.700000	0.821177
	Red	Blood red	0.039634	0.700000	0.369847
5	Like	Love	0.640461	0.500000	0.570321
	Like	Hobby	0.627409	0.500000	0.563704
	Like	Hate	0.469384	0.142870	0.306121
6	Strike	Attack	0.615560	0.500000	0.557780
	Strike	Assault	0.574370	0.500000	0.537185
	Strike	Fondle	0.027581	0.222222	0.111249
7	Apple	Computer	0.984101	0.093023	0.538562
	Apple	Jobs	0.988315	0.000000	0.494157
	Compute	HP	0.990974	0.000000	0.495487
	Compute	Google	0.039012	0.000000	0.019506
	Compute	Keyboard	0.949653	0.083333	0.516493
	Compute	Main engine	0.762508	0.222222	0.492365

E. Analysis experimental results

Words in group 1 and 2 are keywords extracted from previous examples.

From group one, we find that those new specific words(Search engine, Baidu and Google) are not included into the semantic dictionary "HowNet". So we cannot use the semantic dictionary "HowNet" to calculate their semantic similarity. The result from group one embodies the "limitations" of application scope of "HowNet". However, LDA topic model uses statistical approach (training from large scale corpus, then generating potential theme and assembling words according to their subject distribution) property to break through the limitations of new word. Therefore, as long as training corpus is wide enough and updated, the application field of LDA subject model can be extended without limit. The extensibility of LDA subject model can make up the limitation of "HowNet " very well.

In group two, our purpose is to find the most similar word to "speed" from "efficiency" and "accuracy". We know that speed reveal the degree of fast or slow, and accuracy refers to the degree of precision or recall rate in search field, while efficiency is a comprehensive noun which can express both speed and accuracy. Through experimental data, we can see that if we only use the semantic dictionary "HowNet", speed, efficiency, accuracy are consistent in similarity, without any differences. However, the calculation method of semantic similarity on two layers can reflect the differences between words very well.

Analysis the third group of phrases,we want to find the highest similarity with "patient" from "sick person", "doctors", "diseases".From experiment result, Only using LDA to compute phrases similarity, we will find that doctors,sick person and disease both have high similarity with patient. This

method has some certain distinction, but can not reach the aim of our application. Because Our application is ultimately used in Chinese question answering system, and the feature of our phrase similarity is synonymity. LDA topic model guarantee phrases similarity difference by sampling and complicated calculation, but is also a probabilistic model which reflect word co-occurrence. As we know, word patient frequently appear in a document,which is very likely to be have sick person, doctor, diseases and etc. That is the reason sick person, doctor, diseases have high similarity with patient when only using LDA to compute similarity. Therefore, in order to further distinguish difference of phrases semantic, we use the semantic dictionary to further mining phrase semantic. As the table shows that the two layer of the semantic similarity calculation method can reflect the greatest similarity(patient and sick person). At the same time, distinguishing doctor and disease with patient. Introducing the semantic dictionary to refine similarity of phrases.

In group four, the keywords are about colors. In semantic dictionary "HowNet", semantic items do not reflect approximation degree of color attribute value (red and pink are similar, while red and white look much different). It's very difficult to describe the approximation degree of colors using objective language, but it is able to tell differences to some extent by the method in this thesis. But the effect will be not consistency using different training corpus.

Words in group five are about affection tendency. Words in group six are about the degree of action. Words in group seven are computer vocabulary.Through analysis of the experimental data, the method in this thesis is also able to distinguish similarity between phrases to some extent, showing the most intuitive feeling and proving the feasibility of the method.

VI. CONCLUSION

The paper presents a two layers semantic similarity calculation method to excavate semantic similarity of Chinese words. Through lots of experiments, this method is feasible and applicable.

ACKNOWLEDGMENT

This work is supported by National Social Science Foundation Project of P.R. China (No.14BFX156).

REFERENCES

- [1] Q.Liu and S.J.li, "Word's semantic similarity computation Based on the HowNet", The 3rd Chinese lexical and semantic proseminar, Taipei, China, 2002.
- [2] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation", Journal of Machine Learning Research. 3:993-1022, January 2003.
- [3] B.Ge,F.F.Li,S.L.Guo, "Word's semantic similarity computation method based on HowNet", Application Reserach of Computers, Vol.27, No.9, pp.3329-3333, Sep.2010.
- [4] T. Griffiths, "Gibbs sampling in the generative model of Latent Dirichlet Allocation", Technical Report, 2003.
- [5] T.Griffiths and M.Steyvers, "Finding scientific topics", Proc of the National Academy of Sciences, 2004.
- [6] Hu, Feng Song, Guo, Yong, "An improved algorithm of word similarity computation based on HowNet", Computer Science and Automation Engineering, IEEE International Conference, Vol.3, May 2012.
- [7] Z.Dong and Q.Dong,HowNet,http://www.keenage.com.
- [8] Text categorization corpus of sougou. http://www.sogou.com/labs/dl

The Impact and Challenges of Cloud Computing Adoption on Public Universities in Southwestern Nigeria

Oyeleye Christopher Akin¹

Department of Computer Science and Engineering,
Ladoke Akintola University of Technology,
Ogbomoso, Oyo State, Nigeria

Fagbola Temitayo Matthew², Daramola Comfort Y.³

Department of Computer Science^{2,3},
Federal University, Oye-Ekiti^{2,3},
Ekiti State, Nigeria

Abstract—This study investigates the impact and challenges of the adoption of cloud computing by public universities in the Southwestern part of Nigeria. A sample size of 100 IT staff, 50 para-IT staff and 50 students were selected in each university using stratified sampling techniques with the aid of well-structured questionnaires. Microsoft excel was used to capture the data while frequency and percentage distributions were used to analyze it. In all, 2, 000 copies of the questionnaire were administered to the ten (10) public universities in the southwestern part of Nigeria while 1742 copies were returned which represents a respondent rate of 87.1%. The result of the findings revealed that the adoption of cloud computing has a significant impact on cost effectiveness, enhanced availability, low environmental impact, reduced IT complexities, mobility, scalability, increased operability and reduced investment in physical asset. However, the major challenges confronting the adoption of cloud are data insecurity, regulatory compliance concerns, lock-in and privacy concerns. This paper concludes by recommending strategies to manage the identified challenges in the study area.

Keywords—cloud computing; cloud adoption; information-communication-technology; public-universities

I. INTRODUCTION

Information and Communication Technologies (ICT) are powerful enabling tools for educational change and reform introducing new methods of teaching and conducting research as well as provisioning of educational facilities for online learning, teaching and research collaboration. It thus represents a potentially equalizing strategy for developing countries. The great flexibility offered by ICT strongly facilitates the acquisition and use of available knowledge to expand access to education, strengthen the quality of education and improve the quality of the classroom teaching-learning processes via access to electronic active teaching and learning, research, training and development resources on the global collaborative network of internetworks and use of ICT tools in education. It can be said to be the bedrock for successful scientific research and development in education.

ICT is considered a critical tool in preparing and educating students with the required skills for the global workplace. It educates students so that they can continually adapt to a work world of continuous technological innovations [1]. The ability to become lifelong learners within a context of collaborative

environment and the ability to work and learn from experts and peers in a connected global community are major flexibilities offered by ICT [2]. Iwasokun, Alese, Thompson and Aranuwa [3] stressed that ICT is a versatile tool for running a smooth and efficient university system, giving support in areas such as lecture delivery, private studies, information disseminations, program (conferences and seminars) planning and execution, communication at different levels, crisis prevention and management.

Unfortunately, the recurring Global Economic Meltdown (GEM) and national financial hiccups currently embattling the developing countries continue to pose a serious threat to the survival of quality education as governmental institutions and University administrators helplessly fight the provision of unlimited fundamental ICT facilities and support tools, services and applications needed to facilitate effective teaching and Sustainable Educational Research and Development (SERD) activities in Universities. Furthermore, developing countries generally face challenges in terms human and financial resources needed to harness the potential of ICT successfully and effectively in education [1].

As much as the adoption of ICT in education becomes imperative, cost of owning of the required ICT infrastructures, licensing, standards requirement, cost of maintenance, electrical power supply and physical security of these facilities come at a great financial expense. Gerald and Eduan [4] stressed that availability and accessibility to ICT infrastructures and services by staff and students in Universities in most developing economies are limited or non-existent. Inadequate funding of universities by the government at all levels, erratic power supply, operational cost, high cost of equipment renewal, cost of maintenance and bandwidth, lack of maintenance practice and lack of ICT budget by the Universities are the major factors responsible for the failure of the survival of ICT in Universities.

Effective teaching-learning process, research and development activities have been hampered as a result of these menaces. For example, when power is rarely supplied, the admirable goals of transforming education with ICT and taking a paradigm shift in education is all a dream; having access to educational resources on demand, anytime, anyhow and anywhere is a story and e-learning would not be sustained either [5].

Sequel to these challenges, the adoption of cloud computing, a service-oriented alternative to ICT provisioning and deployment, with the potential to yield low cost, improved efficiency and availability become imperative in Universities. Mehmet and Serhat [6] identified some of the benefits offered by cloud computing in education to include on-demand access to online database repositories, e-learning platforms, digital archive, portals, research applications and tools, file storages, e-mails and other educational resources anywhere for faculty, administrators, staff, students and other users in university.

Therefore, in this paper, the impact and challenges of the adoption of cloud computing on public Universities in the Southwestern part of Nigeria is investigated. The impact assessment investigates the gains derived from the adoption of cloud computing in Nigerian Universities while the challenges assessment investigates the problems and constraining factors mitigating with the successful adoption and use of cloud computing in Nigerian Universities. It concludes by recommending strategies to manage the identified challenges in the study area.

Section II of this paper presents a literature review and conceptual underpinnings of cloud computing adoption in public Universities. The materials and method are presented in Section III. In Section IV, the results and interpretation are presented and discussed while the conclusion, recommendation and future research work are presented in Section V. Questionnaire for the evaluation of the impact and challenges of cloud computing adoption and use by Universities in Southwestern, Nigeria is provided at the appendix section after the references.

II. LITERATURE REVIEW AND CONCEPTUAL UNDERPINNINGS

The conceptual, logical and architectural development over Networking, Internet and Grid computing has given birth to the third (3rd) technological revolution after Personal Computer (PC) and the internet known as cloud computing [6]. Cloud computing can be described as a composite three-tier delivery, development and application platform [7-8]. As a delivery platform, it uses an on-demand cloud-based infrastructure to deploy an infrastructure or applications, for example, the Amazon Elastic Cloud.

The on-demand cloud-based development environment provides a general purpose programming language (for example, Bungee Labs, Coghead, google sites) as a development platform. As an application platform, it is used to develop and deploy end-user applications (for example, Salesforce.com, NetSuite, Cisco-WebEx and google docs).

Olabiyisi *et al.* [9] defined cloud computing as an elastic and scalable utility model that offers flexible, ubiquitous, on-demand network access to a shared pool of configurable computing resources (for example, servers, data centers, networks, applications and services) that can be rapidly provided and released with limited interaction of service provider or the management. It provides shared infrastructure, self-service, dynamic and virtualized pay-per-use platforms which put it on high demand. Cloud computing implies a level of dynamic, flexible resource sharing and allocation of assets.

Edtech [10] conducted interview with a panel of the world's top technologists discussing new technologies changing higher education especially "education in the cloud" trend. Shel Waggener, the senior vice president of Internet2 and former Chief Information Officer (CIO) at University of California in Berkeley, Ted Dodds, Chief Information Officer and Vice President at Cornell University, Ron Kraemer, the vice president and CIO at Notre Dame University and Bill Wroblewski, Director of infrastructure services for information and technology services at University of Michigan discussed issues relative to benefits of cloud adoption, risk factors and risk management practices. A great number of advantages of using cloud computing in education were highlighted and techniques for mitigating the risk of cloud adoption were explained.

Gerald and Eduan [4] conducted a survey on the adoption of cloud computing among public universities and FET colleges within South Africa. The authors argued that public universities and colleges share many similar operational processes such as course offerings, admissions, enrollments, bursaries, research and graduations that can be standardized across the higher education sector and offered as a set of services through cloud to the many colleges and universities in a more cost effective way than is currently the case. The results of the analysis shed some light on the current state of cloud computing adoption within the South African public higher education sector, the main factors that fuel its adoption, the main barriers that impede its adoption and the direction it may take in future as it matures.

Abdulsalam and Fatima [5] argued that cloud computing is the solution to ICT in higher education in Nigeria. The authors identified scarcity of ICT infrastructure and lack of access, high cost of ownership, unsteady and inadequate electrical power supply as factors that are limiting the infusion of ICT in Nigeria higher education. They claimed that the prospect of a maturing cloud of on-demand infrastructure, application and support services is important as a possible means of driving down the capital and total costs of ICT in higher education, facilitating the transparent matching of IT demand, scaling ICT, fostering further ICT standardization and accelerating time to market by reducing ICT supply bottlenecks.

III. MATERIALS AND METHOD

The methodology and approach adopted in this paper are described below. In this section, the research questions are highlighted, the study area, sampled population and research techniques used are discussed.

A. Research Questions

To realize the purpose of this research study, three (3) research questions are formulated as follows:

- What is the level of adoption of cloud computing by Universities in the study area?
- What are the benefits associated with the adoption of cloud computing by Universities in the study area?
- What are the challenges and the constraining features to the successful adoption and use of cloud computing by Universities in the study area?

B. Data Source and Presentation

This study is an empirical research which investigates the level of adoption, benefits and challenges of cloud computing on universities in southwestern part of Nigeria. The instrument for data collection was a well-structured questionnaire titled, "The Evaluation of the Impact and Challenges of Cloud Adoption and Use on Universities in Southwestern, Nigeria" with three (3) parts. The first part provides vital biodata information about each respondent while the second part provides information on the assessment of the adoption of cloud computing in universities. The third part assesses the impacts of cloud computing on Universities in south western part of Nigeria while the fourth part investigates the challenges of using cloud computing in the study area.

The questionnaire was validated and tested for reliability using the Pearson Product Moment Correlation. A Cronbach alpha reliability coefficient (α) of 0.89 was obtained, an indication that the instrument was reliable for data collection. In all, 2,000 copies of the questionnaire were administered to the ten (10) public universities in the southwest geo-political zone of Nigeria while 1742 copies were returned which represents a respondent rate of 87.1%.

A total of hundred (100) IT staff, fifty (50) para-IT staff and fifty (50) students were surveyed in each university. Microsoft Excel was used to capture and analyze the data obtained from the duly-filled copies of questionnaire while frequency, mean and percentage distributions were the descriptive techniques used. The descriptive survey was adopted to obtain the opinion of a representative sample of the target population so as to be able to infer the perception of the entire population.

IV. RESULTS AND INTERPRETATION

The results of the research on the trends of adoption of cloud computing, the impacts and challenges associated with the adoption and use of cloud computing on Universities in Southwestern Nigeria are presented and discussed in this section.

A. Trends of Adoption of Cloud Computing by Universities in Southwest Nigeria

The analysis of the findings for research question 1 is presented in figures (1 and 2). The responses obtained from 1742 respondents in this research study indicated that out of the ten (10) universities in the study area, nine (9) universities have already adopted the technology and using it which represents 90% adoption rate. This confirms the report by Edudemic [11] that many higher education and research institutions have moved to the cloud for email and collaboration services. However, the primary reluctance of few other institutions to expand their use of cloud services has been based on concerns over privacy, security and the potential or perceived risks associated with intellectual contents.

As illustrated in figure 2, the responses obtained indicated that seven (7) universities use SaaS. Applications including google docs, Moodle, Google Mail, Yahoo Mail and NetSuite offered by SaaS in universities are universal, free and in high

demand by the entire university community which accounted for its widest adoption and use. PaaS in universities including Google sites, online databases, Microsoft Dynamics CRM online and integrated development environment are only used by a limited number of consumers including the developers, researchers and other technical personnel that are for research, training and development purposes. Thus, in the result obtained, only two (2) universities use PaaS. IaaS in education offers oracle coherence, educationERP.net, microsoft, virtual computing laboratories, servers and operating systems and adopted by one (1) university for e-learning and portal hosting. This result corroborates the report of Marinela and Anca [12] that the highest number of cloud consumers subscribes to SaaS.

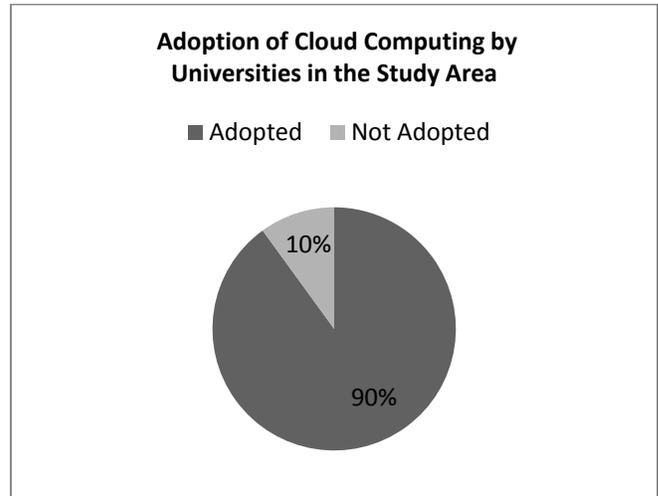


Fig. 1. Result of adoption of cloud computing

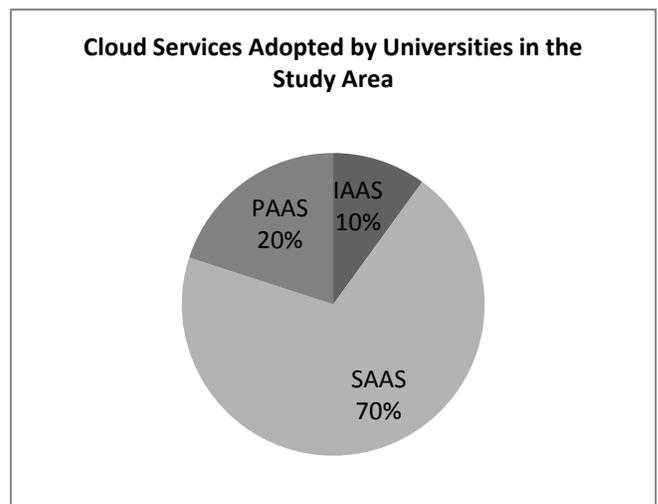


Fig. 2. Result of cloud services adopted in the study area.

The result of the findings obtained for research question 2 is presented in Table 1. The major benefits being derived by the use of cloud computing include cost efficiency which is the most important factor that drives most Universities in Nigeria to adopt cloud, followed by enhanced availability, low environmental impact, reduced IT complexities, mobility,

scalability, increased operability and reduced investment in physical asset in that order. This result is supported by Behrend, Wiebe, London and Johnson [13] who stated that cloud computing is a technological innovation with a major purpose of reducing IT costs for the college and eliminating many of the time-related constraints for students, making learning tools available and accessible to a larger number of students. EDUCAUSE [14] stressed that cloud computing offers to universities the possibility of concentrating more on teaching and research activities rather than on complex IT configuration and software systems.

Westmont College reports that after deploying six cloud-centric service platforms, it has achieved numerous benefits, including a 65 percent cost reduction up front (over more traditional deployments), and a 55 percent cost saving over the useful lifetime of the solutions. Beyond the cost savings, the college reports a significant increase in user satisfaction, as well as a significant decrease in the amount of IT management time required [8]. Sasikala and Prema [15] emphasized the Cloud Computing trend of replacing software traditionally installed on campus computers (and the computers themselves) with applications delivered via the internet is driven by aims of reducing universities' IT complexity and cost.

TABLE I. BENEFITS OF CLOUD COMPUTING IN NIGERIA UNIVERSITIES (N = 1742)

S/N	Benefits of Cloud Computing in the Study Area	% of Respondents
1	Enhanced Availability	99.3
2	Reduced Cost	100
3	Scalability	84.2
4	Low Environmental Impact	95.6
5	End-User Satisfaction	74.6
6	Mobility	85.4
7	Reduced IT Complexities	93.4
8	Reduced Physical Asset Investment	82.8
9	Increased Interoperability	83.5

Based on the analysis of the findings obtained for the research question 3 in this study as presented in Table 2, a number of challenges currently embattling Universities using cloud services in the study area have been identified. These constraining factors include data insecurity, regulatory compliance concerns, lock-in, privacy concerns, unsolicited advertising and reluctance to eliminate staff positions, reliability challenge and resistance to change in technology in that order. This result is supported by the work of Dan [16] who identified that approximately, 75% of Chief Information Officer and IT specialists consider security as being the number one risk.

IITE [17] explained that major concern of university administrators and CIO is around the security of data. Institutions may consider that their data is more secure if it is hosted within the Institution. Transferring data to a third party for hosting in a remote data centre, not under the control of the Institution and the location of which may not be known presents a risk. Another risk identified is that cloud providers target users with unsolicited email or advertising.

Lock-in is also of concern as some companies such as Google and Microsoft allow institutions to co-brand their cloud products. There may be a risk in associating an institution too closely with these companies whose popularity is variable with users [17]. Probably of greater risk is that an institution will become "locked-in" to the products of a particular provider. There are significant costs in migrating from any widely used system. Institutions which start to integrate educational processes with the cloud systems will find it even more difficult to migrate. If a better rival product emerges or the cloud provider decides to impose or increase charges on institutions it may be too late to change.

TABLE II. CHALLENGES OF CLOUD COMPUTING IN THE STUDY AREA (N = 1742)

S/N	Challenges of using Cloud Computing	% of Respondents
1	Data insecurity	89.3
2	Unsolicited Advertising	64.6
3	Lock-in	77.6
4	Reluctance to eliminate staff positions	64.6
5	Privacy Concerns	68.9
6	Reliability challenge	64.2
7	Regulatory compliance concerns / User control	80.0
8	Institutional culture / resistance to change in technology	59.2

V. CONCLUSION, RECOMMENDATION AND FUTURE WORK

The poor state of ICT in Nigerian universities has really limited its impact on socio-economic development, quality of graduates and research outputs. Cost has been identified as the major barrier to the survival of ICT in education which can be leveraged through the adoption of cloud computing. Cloud computing can actually help institutions reduce high expenditures on hardware, software and IT maintenance. It can also offer enhanced availability, low environmental impact, reduced IT complexities, mobility, scalability, increased operability and reduced investment in physical asset.

However, the constraining factors to successful adoption and use of cloud computing include data insecurity, regulatory compliance concerns, lock-in, privacy concerns, unsolicited advertising and reluctance to eliminate staff positions, reliability challenge and resistance to change in technology.

Based on the results obtained from this research work, the following recommendations are made. The cloud can help universities to:

- 1) Accommodate the rapid increase in mobile device dependency
- 2) Open their technology infrastructures to businesses and industries for research advancements.
- 3) Remain updated with the ever-growing resource requirements and energy costs.
- 4) Store expansive amounts of sensitive data and information that's easily accessible
- 5) Teach students in new, different ways and help them manage projects and massive workloads with the provisioning of a digital campus storage for class notes, papers and projects.
- 6) Acquire and implement the latest software and application updates
- 7) Streamline enrollment and admissions processes that are costly and time-consuming
- 8) Turn to subscriptions that are scalable and provide options
- 9) To use applications without installing them on their computers and also allows access to saved files from any computer with an Internet connection.

Future research work can investigate on how the constraining factors to the successful adoption of cloud computing in Nigeria Universities can be managed easily without incurring additional overheads. The readiness assessment of the Universities to the adoption of ICT in various services being offered can also be conducted.

REFERENCES

- [1] B. Neil and Associates, "ICT, Education, Development and the Knowledge Society". GeSCI African Leadership in ICT Program, 2011.
- [2] S. Osaat and L. Nsereka, "Impact of Information and Communication Technology on Distance Education: The Case of National Open University of Nigeria", African Research Review, An International Multidisciplinary Journal, Ethiopia Vol. 6 (1), Serial No. 24, January, 2012, Pp. 325-341.
- [3] G. B. Iwasokun, B. K. Alese, A. F. Thompson and F. O. Aranuwa, "Statistical evaluation of the impact of ICT on Nigerian universities", International Journal of Education and Development using Information and Communication Technology (IJEDICT), 2012, Vol. 8, Issue 1, pp. 104-120.
- [4] M. Gerald and K. Eduan, "Cloud Computing in Higher Education: Implications for South African Public Universities and FET Colleges", Annual Conference on WWW applications, 2012.
- [5] Y. G. Abdulsalam and U. Z. Fatima, "Cloud Computing: Solution to ICT in Higher Education in Nigeria", Advances in Applied Science Research, 2011, 2 (6): 364-369, Pelagia Research Library.
- [6] F. E. Mehmet and B. K. Serhat, "Cloud Computing for Distributed University Campus", International Conference on the Future of Education, Pixel Publishing International, 2011.
- [7] J. Anjali and U.S. Pandey, "Role of Cloud Computing in Higher Education", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, 2013, Pg 966-972.
- [8] N. Sclater, eLearning in the Cloud, International Journal of Virtual and Personal Learning Environments, Vol 1, No 1, 10-19, IGI Global, 2010.
- [9] SO Olabiyisi, TM Fagbola, RS Babatunde. An Exploratory Study of Cloud and Ubiquitous Computing Systems. World Journal of Engineering and Pure and Applied Sciences 2012; 2(5):148-155.

- [10] Edtech, "Education in the cloud", Education in the Cloud _ edtechdigest.com.htm, 2013.
- [11] Edudemic, "The Future of Higher Education and Cloud Computing", www.edudemic.com / The Future of Higher Education and Cloud Computing - Edudemic - Edudemic.htm, 2013.
- [12] M. Marinela and L. A. Anca, "Using Cloud Computing in Higher Education: A Strategy to Improve Agility in the Current Financial Crisis". IBIMA Publishing, Vol 20 (2010), Article ID 875547, DOI:10.5171/2011.875547.
- [13] T. S. Behrend, E. N. Wiebe, J. E. London and E. C. Johnson, Cloud Computing Adoption and Usage in Community Colleges. Behaviour & Information Technology, 30 (2), 2011, 231-240.
- [14] EDUCAUSE, "Cloud Computing Explained", http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolum/CloudComputingExplained/206526, 2012.
- [15] S. Sasikala & S. Prema, Massive Centralized Cloud Computing (MCCC) Exploration in Higher Education. Advances in Computational Sciences and Technology, 3 (2), 111-118, 2010.
- [16] M. Dan "Cloud Computing in Education", September 12, 2011, [Online]. Available: http://www.cloudave.com/14857/cloud-computing-in-education/
- [17] IITE, "Cloud Computing in Education", UNESCO Institute for Information Technologies in Education (IITE) Policy Brief, September 2010.

APPENDIX

Questionnaire for the Evaluation of Impact and Challenges of Cloud Computing Adoption and Use by Universities in Southwestern, Nigeria

The purpose of this questionnaire is to evaluate the impact and challenges of cloud adoption and use by universities in southwestern Nigeria. Your sincere contribution to the research by giving very accurate and honest responses is solicited as confidentiality of volunteered information is guaranteed.

Part I: Biodata of the Respondent

1. Status: IT Staff Para-IT Staff Student
2. Age:
3. Sex: Male Female
4. University:
5. Rank:

Part II: Assessment of the Trends (Awareness and Adoption) indices of Cloud Computing in Nigerian Universities (To be completed by IT member of Staff)

1. Are you aware of cloud computing in Education?
Yes No

If yes, what cloud services are being adopted and used in your University? Tick as appropriate:

- Infrastructure as a Service (IAAS):** for example, storage of educational multimedia resources, hosting of the E-library resources, Institutional Learning Management Systems (LMS) like Moodle and Blackboard, Computer laboratories, Telephony, University portal on cloud by cloud providers like Amazon EC2 Elastic Cloud, IBM, Terramark, GoDaddy and Intuit Quick Base among others who presented infrastructure components for rent.
- Software as a Service (SAAS):** for example, use of ERP, Identity Services, Google Apps which covers the following three main areas: messaging (Gmail, Calendar and Google Talk), collaboration (Google Docs, Video and Sites) and security (email security, encryption and archiving).
- Platform as a Service (PAAS):** offers Integrated Development Environments (IDE) / platform for rent. For example, use of Google Sites and cloud-based APIs and .NET platforms.

Part III: Assessment of the benefits of Using Cloud Computing in Nigerian Universities. Rate as applicable to your University on the likert scale of Strongly Agree, Agree, Not Sure, Disagree and Strongly Disagree (N = 1742).

Serial Nos	Benefits	Strongly Agree	Agree	Not Sure	Disagree	Strongly Disagree
1	Enhanced Availability	992	738	09	03	0
2	Cost effectiveness / Affordability	1156	586	0	0	0
3	Elasticity / Scalability	965	502	260	15	0
4	Lower environmental impact	869	797	67	9	0
5	End-User Satisfaction	967	333	345	97	0
6	Mobility	931	557	189	65	0
7	Reduction in IT Complexities	1005	622	112	03	0
8	Less investment in physical assets	1210	233	215	84	0
9	Increased Interoperability between disjointed technologies	851	604	202	85	0

Part IV: Assessment of the challenges of using Cloud Computing: Rate as applicable to your University on the likert scale of Strongly Agree, Agree, Not Sure, Disagree and Strongly Disagree (N = 1742).

Public Universities	Total Number of Questionnaires Returned by Respondents	Total Not Returned
Ladoke Akintola University of Technology, Ogbomoso	190	10
Adekunle Ajasin University, Akungba-Akoko	169	31
University of Ado-Ekiti, Ado-Ekiti	160	40
University of Agriculture, Abeokuta	177	23
University of Ibadan, Ibadan, Oyo State	189	11
Lagos State University, Ojoo, Lagos State	173	27
Federal University, Oye-Ekiti, Ekiti State	182	18
Obafemi Awolowo University, Ile-Ife	172	28
University of Lagos, Lagos State	168	32
Federal University of Technology, Akure	162	38

Serial Nos	Challenges	Strongly Agree	Agree	Not Sure	Disagree	Strongly Disagree
1	Data insecurity	728	828	108	78	0
2	Unsolicited Advertising	612	513	511	106	0
3	Lock-in	543	809	289	101	0
4	Reluctance to eliminate staff positions	690	434	467	151	0
5	Privacy Concerns	606	596	385	155	0
6	Reliability challenge	389	729	420	204	0
7	Regulatory compliance concerns / User control	765	752	244	104	0
8	Institutional culture / resistance to change in technology	432	599	480	231	0

Table Showing the Summary of the Public Universities in the Southwestern Part of Nigeria

WOLF: a Research Platform to Write NFC Secure Applications on Top of Multiple Secure Elements (With an Original SQL-Like Interface)

(July 2014, LSIS / MBDS Research Report)

Anne-Marie Lesas, PhD student with Gemalto and LSIS research lab of Aix-Marseille University, MBDS innovation lab at the University of Nice – Sophia-Antipolis, France

Pr. Serge Miranda
Director of MBDS Master degree and innovation lab (www.mbds-fr.org)
University of Nice Sophia Antipolis, France

Benjamin Renaut,
FIRST Research Engineer & Project manager
MBDS innovation lab
University of Nice – Sophia-Antipolis,
TOKIDEV, Nice, France

Amosse Edouard, PhD student
I3S Research laboratory
University of Nice – Sophia
Antipolis, France

Abstract—This article presents the WOLF (Wallet Open Library Framework) platform which supports an original interface for NFC developers called “SE-QL”. SE-QL is a SQL-like interface which eases and optimizes NFC secure application development in making the heterogeneity of the Secure Element (SE) transparent. SE implementation could be “embedded” (eSE) in the mobile device, or inside the SIM Card (UICC), or “on-host” software-based, or in the Cloud (e.g. through HCE); every SE implementation has its own interface(s) making NFC secure-application development extremely cumbersome and complex. Proposed SE-QL solves this problem. This article demonstrates the feasibility and attractiveness of our approach based upon an original high-level API.

Keywords—*Mobiquitous services; Near Field Communication (NFC); Secure Element (SE); Smart card; Structured (English as a) Query Language (SQL); Digital Wallet; TSM / OTA; UICC*

I. INTRODUCTION: WOLF AND SE-QL

Based upon both our Near Field Communication (NFC) know-how and our SQL (Structured Query Language for databases) expertise [1], we propose a generic innovative Framework called WOLF (Wallet Open Library Framework) for facilitating Service Providers (SPs) in the process of development and deployment of new NFC secure applications on a wide range of smartphones having different SE implementations.

In this article we give an overview of WOLF platform developed at MBDS innovation laboratory, which allows NFC developers to interact easily with an application based on NFC Card Emulation mode by using SQL-like language, “SE-QL”. WOLF framework allows SPs to reduce NFC development costs and the time-to-market, improve and ensure the quality of products applications for new secure NFC services.

WOLF extends “NFC Container” project [2], [3] previously developed at MBDS in 2008-2010 for J2ME cell phones. Its

SE-QL interface idea stems from a long-term research background on SQL and database (DB) systems; SE-QL simplifies NFC service development by abstracting the core software complexity associated with the management of multiple SE environments.

WOLF supports the secure ecosystem of FIRST project in India with Tata Consulting Services (TCS), Gemalto and the Indian Institute of Sciences (IISc) of Bangalore under research contract from IFCEPAR | CEFIPRA (www.cefipra.org). FIRST encompasses a portfolio of NFC financial and rural inclusion use cases for unbanked people in India (70% of them owning a cell phone) managed within a FIRST wallet developed by TCS on top of WOLF platform. FIRST primary goals were to demonstrate Financial Inclusion (FI) services for unbanked people based upon:

- Virtual “mobiquitous money” [4],
- The appeal of NFC standard both to strategic use cases (Financial Inclusion, Narega, Mobiquitous NFC Public Distribution System “M-PDS” [5], [6], [7]), and disruptive ones (e-coins, rural animal bank with crowdfunding “BARTER2.0” [8]).

WOLF has been successfully tested in the development of M-PDS use case prototyped at MBDS since 2012 and it is integrated into FIRST generic platform.

We have been working on the Android SE-QL interface with the delivery of WOLF API which is compliant with SIMAlliance Open Mobile API (OMAPI) with Gemalto SE-UICC, as well as Android Host-based Card Emulation (HCE) using WOLF generic applet; this work was presented at the WIMA’s research track conference in Monaco (April, 2014) [9], at the Indo-French Conference in New Delhi (October, 2013), and at the Indo-French Center for the Promotion of Advanced Research (IFCPAR | CEFIPRA) meeting in St Malo (May, 2014).

The remaining of this document is organized as follows: in section II, we do a synthetic state of the art around the NFC standard and NFC secure application development; we also discuss the NFC Container project, its context, scopes and benefits for WOLF contribution. In section III, we study the proposed SE-QL interface and the WOLF API based upon related works and existing technologies. WOLF and SE-QL implementation are also described in this section with some uses cases concerning the FIRST project. In conclusion, we summarize the benefits of this research platform and present some promising extensions

II. NFC ECOSYSTEM AND SECURE NFC APPLICATION DEVELOPMENT AROUND THE SE

NFC is a very fast establishment and very short-range (for security purposes) contactless communication technology and world standard since 2004. NFC uses the inductive coupling making a source device (acting as initiator) able to provide energy and exchange data with a target passive device (without need battery) by backward induction. NFC will be a standard in next generation smartphones.

Our future will be “ubiquitous” [2], [3], [10] around the convergence of mobility of cell phones becoming computers (smartphones) and ubiquity of Internet (becoming social and broadband); NFC is an underlying technology and standard supporting mobility. NFC connectivity induces five additional dimensions to enrich information services (the five “W”). “Who”: the identity of the end-user (with habits, preferences and POIs), “Where & When” space and time (“here and now”, when tapping), “Whereabout” the goal, expected result (information? transaction?) and the “What”, the final outcome (information, ticket transaction, appointments, service, triggering mechanisms, etc.) [2]. Unlike contactless smart cards services, NFC mobile services benefit from features of the smartphones: network high connectivity, embedded sensors, location-based ecosystem, camera, high-definition and touchscreen user interface bring NFC services at a higher level of expectation and innovation, extending SP information system to the end-user allowing screens, real time interactivity, personalization, traceability and live updates without storage limitation (and Cloud synchronization).

In terms of tracking the smartphone coupled with NFC at least enables to get 3-dimensional transaction identification: space, time and biometrics (as demonstrated in [3]).

NFC standard can be classified into two categories regarding NFC applications: (i) those that do not require security using NFC read / write mode (m-tourism, marketing 2.0... Similar to QR codes) or NFC P2P mode (e.g. initiate Bluetooth® pairing), and (ii) those that need to store confidential data and do transactions in the secure environment provided by the Secure Element (SE) using the NFC card emulation mode (digital identity, ticketing, couponing, electronic money / m-payment, access control, transactions etc.) where the NFC device acts as a smart card.

In our research we focus on the latter case, i.e. on NFC card emulation mode which requires the NFC service and sensitive data to be hosted in the SE.

A. NFC standard

The NFC standard is a very revealing technology of the expected convergence between the worlds of telecommunications, consumer electronics and computing. It is one of the numerous RFID standards (known technology since the 1940s) operating over the unlicensed 13.56 MHz frequency of up to ten centimeters (one to four in practice). It is a wireless technology which could be integrated into mobile phones allowing them to exchange information with other devices (mobile phone, printers, locks or any NFC card readers) and NFC tags (ISO / IEC 14443 - NXP MiFare - Type A or Type B, and Sony-FeliCa).

Since initial standard specification in 2006 by the NFC Forum (www.nfc-forum.org, funded in 2004 by Nokia, Sony, and Philips semiconductors, now NXP) of NFC standard, many specifications have increased the attractiveness of this world standard especially since it is widely available on Android devices.

NFC Forum specifications apply to the physical and data link layers of Open System Interconnection model (OSI) whereas GlobalPlatform (GP) widely contributes to the standardization of SE security architecture, internal and external mechanisms, etc., and Trusted Environment Execution (TEE). GP is the main reference for SE standardization; GSMA and ETSI also play an important role, as the SIM-centric model is the only one to be standardized end-to-end. Most popular SEs, at the moment, are SIM-based.

The NFC standard (ISO / IEC 14443) has three basic operating modes:

- Read / Write: The NFC cell phone acts as an active reader and can read and / or write data to or from a passive NFC tag.
- NFC Peer-to-Peer (P2P): Allows two NFC devices to be active and exchange information interactively.
- Card Emulation: The phone behaves as well as a passive contactless card (EMV, American Express, access card, Ticketing...) for a NFC reader (POS).

B. NFC card emulation mode with APDU messages

The card emulation mode is the extension of contact-based smart cards to contactless NFC cards; it inherits the smart card programming standards, especially the small data packets called APDUs (Application Protocol Data Unit) used to communicate with the services hosted and running in the smart card / SE (also called “applet” or “cardlet”). This requires a high-profile developer with strong expertise in communication protocols to handle low-level data structures at the byte level and integrate strong environmental constraints related to the execution environment of the SE.

APDU protocol was originally specified in the Java Specification Request (JSR) 177 (Security and Trust Services API for J2ME™) taken up by smart cards standard ISO / IEC 7816-4 (Identification Cards - Integrated Circuit Cards with Contacts: Organization, security and commands for interchange) now extended to contactless smart cards. In

addition, there are related standards (GP / OMAPI, EMV, EN 726-3 for prepaid memory cards, etc.).

Remark: APDU protocol does not manage the device connection or the channel opening.

The APDU commands (C-APDU) is the message sent by the client application (initiator) to the service running in the SE. The structure is composed of a mandatory header of 4 fields of one byte: (i) {CLA} (class field) defines the command type (standard, industry, using security or not), (ii) {INS} defines the command instruction, (iii) {P1} is the first parameter (0x00 if not defined), (iv) {P2} is the second parameter (0x00 if not defined), and a conditional body of 3 optional parameters of a variable length: (i) {Lc} is the conditional data length of 1 or 3 (extended APDU*) bytes if not empty, (ii) {Data} is the payload data of {Lc} length if not empty, (iii) {Le} is the expected length of the response data varying from 1 to 3 (extended APDU*) bytes if not empty.

The APDU response (R-APDU) is made of the optional response data (that cannot exceed the length defined in {Le} field provided in the C-APDU), and the 2 bytes response status words {SW1} and {SW2} giving the status of the C-APDU. When the command is successful, the service returns the status words 0x9000.

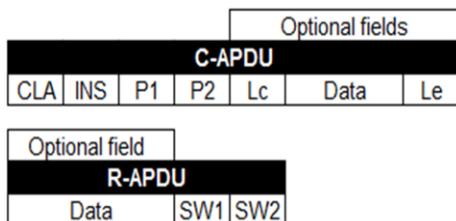


Fig. 1. APDU messages structure

TABLE I. EXAMPLE OF APDU ERROR STATUS WORDS

SW1, SW2	Meaning
0x6A82	File not found
0x6700	Incorrect data length
0x6981	Incorrect file type
0x6982	Security status not satisfied
0x6984	Invalid data
0x6985	Conditions not satisfied
0x6A86	Incorrect P1 and/or P2 parameter(s)
0x6D00	Unsupported command instruction

TABLE II. EXAMPLE OF CLA CODES

CLA byte	Command type
0x00	ISO standard command
0x04	ISO standard command with security
0xB0 to 0xCF	ISO standard INS instruction
0x80	GP standard command
0x84	GP standard command with security
0xFF	Commands for the reader

TABLE III. EXAMPLE OF INS CODES

INS byte	Instruction description
0xA4	ISO / IEC 7816-9 SELECT FILE used to initiate communication with a service identified by its AID (provided in the payload data field)
0x05	OMAPI SELECT SECURE STORAGE ENTRY
0xB0	ISO / IEC 7816-4 READ BINARY
0xD0	ISO / IEC 7816-4 WRITE BINARY
0xD6	ISO / IEC 7816-4 UPDATE BINARY
0xE0	ISO / IEC 7816-4 ERASE BINARY
0x82	ISO / IEC 7816-4 MUTUAL AUTHENTICATION

All standards combined, hundreds INS codes can be found (65536 possible combinations).

Card Query Language (CQL) initially designed by Pierre Paradinas in the 1990s [11], [12], with GEMPLUS (now Gemalto), was the first approach of SQL-like APDU instructions. Smart Card Query Language (SCQL) has been standardized in 1999 by ISO / IEC 7816-7 “Interindustry commands for Structured Card Query Language (SCQL)” [13]. But SCQL is limited to specific smart cards with embedded lite Relational DataBase Management System (RDBMS) whereas SE-QL is not....

TABLE IV. EXAMPLE OF SCQL INS CODES

INS byte	SCQL instructions in the payload
0x10	CREATE, DROP, INSERT, DELETE, DECLARE, FETCH
0x12	CREATE KEY, AUTHENTICATE, CHECK, BEGING TRANSACTION, COMMIT and ROLLBACK
0x14	CREATE USER, CHANGE PASSWORD, UNLOCK, DELETE USER, etc.

C. NFC mobile services basics (card emulation mode)

There are two situations involving the communication with the NFC mobile service: (i) the client is an NFC terminal (e.g. POS), or an NFC handset (acting as a terminal); communication is done via NFC when the handset is approached to the terminal and the terminal has successfully initiated the communication, (ii) the client is a mobile application (typically a user interface); communication is done via a bridge depending on the target SE and platform, through the Radio Interface Layer (RIL), or sometimes through NFC Controller using Single Wire Protocol (SWP)...

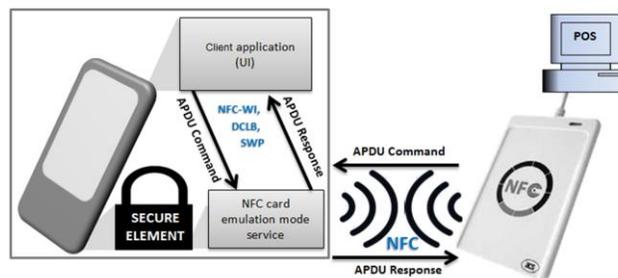


Fig. 2. NFC mobile application architecture (card emulation mode)

In the architecture shown on Fig. 2, the NFC service running inside the SE processes the received C-APDU and

returns R-APDU. However, mobile handset may also act as a reader and be client for an external NFC service: in that case, the mobile application will send C-APDUs to the external NFC device and will receive the R-APDUs.

Remark: A SE can host multiple services and the mobile application (user interface) can be client of a portfolio of NFC services; such an application is called a mobile wallet (m-wallet). A service hosted in the SE is identified by its AID (specified by ISO / IEC 7816-5 “Numbering system and registration procedure for application identifiers”); the AID is used to route the messages.

D. The Secure Element (SE)

SE is mostly an electronic chip with its own processor capable of running applications such as JavaCard (applets) and guaranteeing a certain level of security and functionality. Hardware-based SEs have the same characteristics as the smart cards: a minimalist computing environment on a single chip, complete with CPU, ROM, EEPROM, RAM and I/O ports, preprogrammed with a multi-execution environment OS and security domains separated by firewall that guarantees mutual isolation between running applications. Recent smart cards include coprocessors implementing cryptographic algorithms (such as DES, AES and RSA) and conform to TEE specifications.

SEs for mobile phones have all the capabilities of a smart card or even higher; they can theoretically be used for all types of applications using a smart card (prepaid cards, transportation cards, credit / debit cards, health, loyalty, couponing, storage of VPN access parameters, etc.).

1) Hardware-based SE: SIM-SE, eSE or Removable SE

a) The SIM-based SE handled by the Mobile Network Operator (MNO); Gemalto is providing such SE in our FIRST Consortium. The SIM card with its NFC interface is a Universal Integrated Circuit Card (UICC).

b) Embedded SE (eSE) outside the SIM into the terminal and handled by the device retailer (like Google, Nokia, Samsung).

c) Removable SE in an external SD card or a sticker under control of a SP (e.g. banks, retail companies) with or without the NFC chip.

The SE could encompass various applets with their own business models and access keys controlled by their owners. Theoretically the 3 hardware-based types of SE could work together depending on the Host Controller Interface (HCI) routing capabilities

2) Software-based (card emulation) and SE in the Cloud

The software emulation (of a smart card) is an approach to card emulation for NFC phones for services that do not need to be always available (i.e. when the mobile is off). It was introduced to mobile phones by Research In Motion (RIM) on the Blackberry platform. In addition to supporting different types of SE, the Blackberry 7 introduced the card emulation mode of NFC tags with a software application on the mobile phone. Host-based Card Emulation (HCE) is the software card emulation solution available on Android devices since the end of 2013 with Android KitKat. HCE services run in the host

processor as well as other services making this solution less secure and most vulnerable to malwares [14].

On the other hand, NFC software-based implementation is much lighter to develop and truly simplifies the deployment; SP can deploy its NFC services itself. Furthermore, even if software card emulation approach cannot be a solution “as is”, security could be strengthened by encryption mechanisms; such a “crypted soft-SE” is currently studied by IISc Bangalore in FIRST Consortium.

Another solution is to relocate the SE in the cloud (Cloud-SE) [15]. In this case, NFC services running on the mobile device relay the instructions to the remote Cloud-SE server (e.g. via Web services) and get back the returned responses. This approach reduces the complexity of the chain of deployment of new NFC services. It also has the major advantage to be hardware-independent and offers better safety (neither credentials nor sensible data are stored on the smartphone) and higher (unlimited?) storage capacity than other SEs, but increases the transactions latency. This approach which assumes always-on air reachability was discarded from our FIRST project, but it will also be studied for WOLF API.

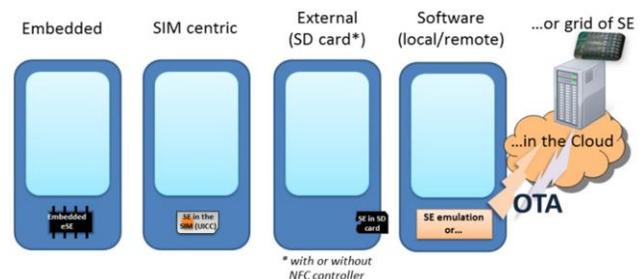


Fig. 3. Several SE architectures

E. Trusted service manager (TSM) and OTA interactions

NFC ecosystem incorporates several actors: computer information systems developers, MNOs, SPs, handsets retailers and chip cards manufacturers, etc. This implies an interoperability of the bodies that govern many heterogeneous domains (telecom, banking, technology, security and cryptography, rights to privacy, government, and other consecutive intermediates). This drives SPs outside their business domain. The TSM standards have emerged as the “split TSM” (SP TSM and MNO TSM, see Fig. 4) proposed by GP in 2011 which is basically provided as Web services.

NFC services hosted in the SE can be initially installed and built-in by MNOs or could dynamically be loaded on demand (by touching an NFC tag or by scanning a 2D tag or through Internet, etc.) onto the cell phone. This is done “Over-The-Air” (OTA), under the responsibility or not of the MNO managing the SE and the applets. Remote dynamic download and provisioning of applications, content, services, tickets, coupons... are then possible in a secure way. Every applet could be managed remotely by the TSM.

When the SE is in the SIM, MNOs have the responsibility of creating security domains on the SIM card for the NFC SPs. TSMs have the responsibility of credentials and encryption keys management of the SP security domain, application

loading into SP security domain, mobile wallet management, and service lifecycle management services (activation, personalization, maintenance, deletion). TSMs implement OTA communication into the NFC-enabled mobile phone and into the SE.

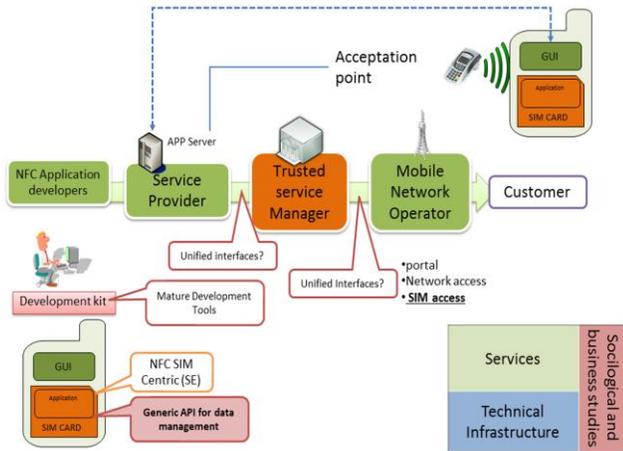


Fig. 4. NFC: a complex ecosystem [1]

F. NFC mobile payment and ubiquitous money

The payment industry seems to be naturally part of the NFC ecosystem by dematerializing payment cards in the SE to allow the phone to be used as a mean of payment.

The payment operators add contactless payment capabilities to their payment card (e.g. American Express ExpressPay, MasterCard PayPass, and Visa PayWave), to create an open loop approach to virtual local emerging money. These services enable end-users to tap and pay their purchase approaching their contactless card or NFC phone to an NFC payment terminal (POS).

In emerging countries, ubiquitous money could replace the cash playing a double attractive role for unbanked people: NO CASH OUT solutions (by accessing a digital wallet) and targeted help / donation through tagged products delivered to the beneficiary and biometric identification like in the prototype of M-PDS use case of FIRST project [5], [6], [7], [8], [9].

G. Related work: “NFC Container” project at MBDS with Gemalto

The project inherits of the know-how of MBDS innovation lab at University of Nice - Sophia Antipolis: in 2011, the university deployed an innovative project based on “NFC container” architecture called “Nice Future Campus” where the student’s phone plays the role of a virtual card with a portfolio of services running in the SIM-SE provided by Orange (m-payment for the university cafeteria, access to geotagged campus information, consulting jobs and internships for students, access to the library, etc.).

NFC container is a SE content management API proposed by MBDS in 2008 as a result of a research contract with the French DGE (“Direction Générale des Entreprises”) of the Ministry of Industry. This project has been driven in order to

reduce the development time and the “time-to-market” of new NFC services. It is partly described in [2], and [3].

This project aimed to provide a set of tools in order to:

- Develop secure applications using NFC c.
- Deploy these applications to a large panel of NFC enabled mobile phone customers.
- Interact with those applications within the SE.

NFC Container project led to a high level API for SPs in order to manage their NFC services more easily.

The proposed API was a generic JavaCard application that allows a dynamic content management onto the SE. The implementation was compliant with NFC Forum standard specifications and runnable on every GP compliant SE.

The project was based on a generic applet that can be preloaded into the SE or loaded “on demand” by the end-user (through TSM / OTA).

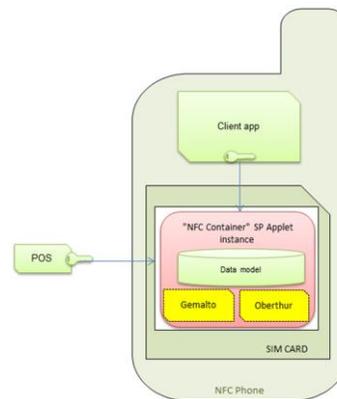


Fig. 5. NFC container instance [2]

The main idea was to access the “SE as a DB” where any privacy data from a client application can be managed. Each user (e.g. SP developer) is identified by a security key on the applet. NFC Container API provides developers some ways to implement additional layers for more complex data management. In that case, the SP develops its own application using the generic API which provides a set of functions (SQL-like) for data management on the card.

We illustrate the added value in terms of code simplification by the following Java coding example retrieved from [2]:

Source code 1. Example of NFC Container code simplification

```

//-----//
// Example of Java coding without using //
// NFC Container //
//-----//
short index = GetIdxToDo(tag);
if (index == (short) DO_NOT_FOUND) {
//DO not found-> create a new one//
short index2Free = GetIdxToFreeSpace();
short freesize= (short) (SIZE MEMORY -
index2free);
//calculate size of free space//
short DOsize = (short) (lc+LEN_TAG+LEN_LEN);

```

```
//calculate size of the new DO//
if (DOsize <= freesize) {
    //it is enough space for a new DO//
    memory[index2free] = (byte) tag;
    //set DO tag//
    memory[(short) (index2free + LEN_TAG)] =
        (byte) lc;
    //set DO length//
    //copy the DO atomic into the memory//
    Util.arrayCopy(cmd apdu,
        (short)((ISO7816.OFFSET CDATA) & 0x00FF),
        memory, (short) (index2free + LEN_TAG +
            LEN_LEN), lc);
//-----//
// Versus coding with NFC Container //
//-----//
void
insertRecord(byte[] record,
    shortrecordOffset,
    shortrecordLength)
```

NFC Container advantages:

Since access to the hardware-based SE is supervised by the NFC ecosystem owners, the deployment of new NFC services is necessarily related to them. To be able to test a new service to be installed in the SE, the developer has firstly to get the keys of the security domain. He has to deal with (physical) SE providers (MNOs, Gemalto, Oberthur...) and get some SIM cards. This is not simple as those access keys must remain secret in order to keep SE's security level.

Through the NFC Container, the developer has access to the generic pre-installed and fully customizable applet on the SE. Thus, the API allows developing new NFC applications without necessarily needing any high level security access and requires a smaller level of knowledge for NFC developers and services providers, allowing them to focus on their client application not in the applet management.

WOLF API is an extension of NFC Container project to the smartphone ecosystem with the formalization of SE-QL.

III. SE-QL AND WOLF API: A GENERIC SQL-LIKE INTERFACE TO COMMUNICATE WITH THE SE

The primary SE function is to store (sensitive and confidential) data: the majority of the (contact or contactless) card applications consist in storing and retrieving data (with or without pre-processing, with or without security mechanisms). The second fact is the building APDUs remains a cumbersome task for the developers: manipulating hexadecimal codes is not simple to human understanding; it is time consuming to be implemented and it is difficult to maintain.

In this section we illustrate the SE-QL foundation based on DataBase Management System (DBMS) concepts. We present our contribution to the project with the WOLF API at the development current stage, and we describe the implementation within use cases prototypes of the FIRST project.

A. Relational databases and SQL background

A database is a set of structured data associated with a schema derived from real world by applying a data model (relational, object). DBMS allow databases management through a standardized interface called SQL [1], [16], [17].

They provide TIPS (Transaction, Integrity, Persistence and Structuration / schema) services for development of information systems development [2].

A DBMS integrates 3 levels:

- Data Description Language (DDL): Defines a language allowing description of objects (tables, domains, databases, views, procedures...).
- Data Manipulation Language (DML): Defines a standard data manipulation language allowing interrogating and updating a DB without specifying any access algorithm (SQL3 being the current standard [1] for relational databases).
- Data Control Language (DCL): Defines some integrity and confidentiality constraints in order to manage user's rights and authorizations on objects.

The DBMS ensures consistency of data in databases and defines some mechanisms such as sharing, security, physical and logical independence of data, access performances in share or exclusive mode.

DBMS can handle several standard request mechanisms allowing manipulating data (read, write, delete, update, sort, etc.). The SQL language, created in 1974 and first normalized in 1986 is used to perform operations on relational databases. It allows developers to interact with a RDBMS without showing physical aspect of data and is compatible with DDL, DML and DCL. SQL's instructions syntax is pretty close to the human language in order to make it easier to learn and to read by a human user.

B. From SQL DBMS to SE's Applet

GP compliant SE is divided in a set of Security Domains (SDs) separated by firewalls allowing several services to be safely executed on the same card.

Each SD is dedicated to a type of business model. On Fig. 6, the first SD is reserved for the card issuer (MNO, manufacturer...); it contains maintenance and additional customer's (SP) services. Others are dedicated to the SPs services (transportation, payment card, loyalty card, coupons, etc.).

A DBMS manages a set of databases contains tables consisted of rows and columns / fields. On the other hand, the SE is composed of SDs hosting running applets managing their own data / fields (see Fig. 7).

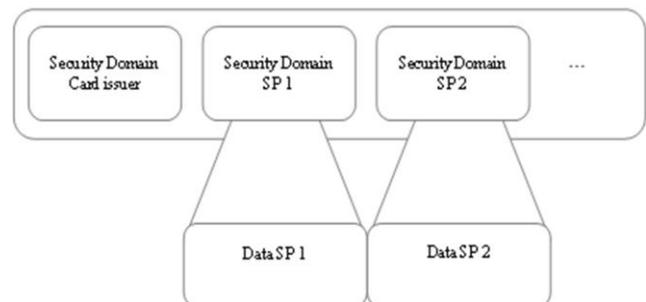


Fig. 6. Security Domains inside the SE

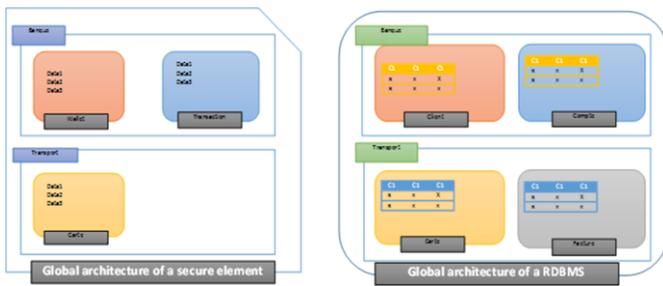


Fig. 7. Similarities between a SE and a DBMS

We can infer the following conclusions:

- The execution environment provided by the SE could be seen as equivalent to a DBMS: the OS manages the SDs and the DBMS manages databases;
- An NFC service hosted in the SE (applet) could be seen as a table of the DB;
- The data fields of an applet could be seen as the columns of the table.

C. DDL, DML, and DCL applied to the SE-QL

Note: at this stage, the instructions are not yet all been implemented in SE-QL.

1) Data Description Language (DDL)

These instructions are intended for the SE management with special maintenance privileges:

- *CREATE* – to create new services (applets) in the SE
- *ALTER* – to modify the structure of the SE data storage
- *DROP* – delete objects from the SE

2) Data Manipulation Language (DML)

These instructions are intended for the NFC service management with the SP privileges:

- *SELECT* – read data from the applets of the SE
- *INSERT* – write data fields into an applet of the SE
- *UPDATE* – update existing data fields within an applet of the SE
- *DELETE* – deletes the records from an applet of the SE

3) Data Control Language (DCL)

- *GRANT* – provide access privileges
- *REVOKE* – remove access privileges

D. Principle of SE-QL

First objective is to hide this low-level implementation of byte arrays by providing the same generic and developer-friendly interface regardless of the platform (initially developed on Java and Android devices, later on Windows Phone and other platforms like HTML5).

1) SE-QL algorithm

A SQL query is composed of two parts: a mandatory part containing the statement and the object affected by the

command, and an optional part using algebraic operators for projection, selection, junction and division.

For a given applet uniquely identified by its alias (matching its AID), SE-QL model makes a correspondence between an SQL-like instruction (CREATE, INSERT, SELECT, UPDATE, DELETE, etc.) and the assigned APDU instruction, whereas the data model gives the records SE-QL alias, their corresponding hexadecimal identification (ID) in the APDU instruction parameter, and the maximum length expected by the applet.

For example, we can illustrate a simplified use case in which the developer would check and update the balance stored by a portfolio applet. This is done in two steps: (i) read “balance” record from “portfolio” applet, (ii) update “balance” record from “portfolio” applet. This will be translated as following in SE-QL language: (1) “SELECT balance FROM portfolio”, (2) “UPDATE portfolio SET balance = {value}”. To be done, the metadata must provide portfolio applet AID, SELECT and UPDATE translation into APDU protocol, and the “balance” record byte identifier:

TABLE V. EXAMPLE OF SE-QL INSTRUCTIONS

Description	SE-QL meaning/ alias	APDU transcription
Applet / relation {used for the communication initialization}	portfolio	AID: F000000001
	Class standard & Security compliance	CLA: B0 {ISO / IEC 7816-4}
Instruction	SELECT	INS: B2 {READ RECORD}
	UPDATE	INS: DC {UPDATE RECORD}
Record projection	Balance	P1: 50
	Record value expected length	Le: 0E {14}

TABLE VI. EXAMPLE OF SE-QL TO APDU COMMAND

SE-QL command	APDU command	Return
SELECT <i>balance</i> FROM <i>portfolio</i>	B0B250000E	Fields values
UPDATE <i>portfolio</i> SET <i>balance</i> = 2000 {payload data is given according the field length 0x0E defined by the metadata}	B0DC50000E0000000002000	Number of rows updates

In more complex use cases, several APDU commands may be handled within a single SE-QL instruction, for example, a statement regarding multiple fields. The activity diagram shown on Fig. 8 illustrates the processing of several fields algorithm, where APDU command sending corresponds to the greyed task.

These examples illustrate the similarity between an APDU command and a SQL query; the rationale of our approach is to identify mechanisms between these two worlds. Moreover applets are not intended to manage large amounts of data like tables in relational databases; usually there are small pieces of data corresponding to tuples of information in a table in a DBMS. In addition, SE-QL language will be limited by the

already defined standards; SE-QL goal is only to ease NFC secure application development.

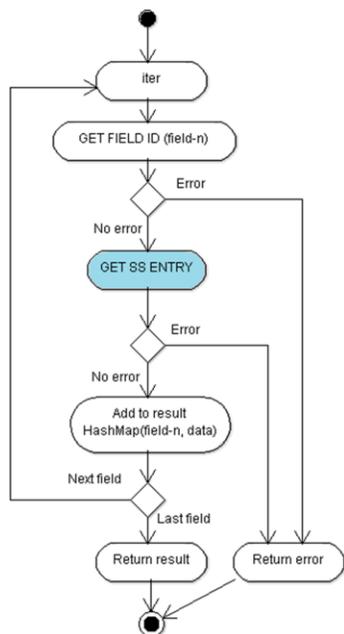


Fig. 8. SE-QL SELECT instruction algorithm

SE-QL is delivered as a Java package integrated in WOLF libraries that manages SE / reader connection and applet communication initialization. This is managed from SE-QL controller that is the entry point for the client applications of the SE. SE-QL controller is in charge to redirect instructions to the appropriate SE-QL interface according targeted platform, implementing the method “execute” that accepts the SE-QL instruction and the command ID. The result for a given command (identified by an ID) is forwarded through an event of the SE-QL callback currently in the form of a HashMap of returned data, concerned field name being the key.

2) Metadata

Metadata is actually provided as a XML (eXtensible Markup Language) file (which could be acquired on the fly during the installation of the application, for example from a web server); it contains one (or more) applet(s) configuration defining two models: the former describes the correspondence APDU / SE-QL commands, while the second describes the pattern of the data managed by the applet. The metadata are used by SE-QL parser on the side of the client application at the APDU commands building, as well as the generic applet for its data storage initialization (could also be provided as a script).

Source code 2. Example of XML metadata file

```

<seqlmetadata>
  <applets>
    <appletModel alias="pds_applet" AID="F0014144500002">
      <instructions>
        <seqlModel>
          <ins>SELECT</ins>
          <cla>80</cla>
          <value>B2</value>
        </seqlModel>
        <seqlModel>
          <ins>INSERT</ins>
        </seqlModel>
      </instructions>
    </appletModel>
  </applets>
</seqlmetadata>
  
```

```

<cla>80</cla>
<value>D2</value>
</seqlModel>
</instructions>
<tables>
  <tableModel>
    <name>username</name>
    <value>30</value>
    <length>9</length>
  </tableModel>
  <tableModel>
    <name>password</name>
    <value>40</value>
    <length>9</length>
  </tableModel>
  <tableModel>
    <name>key</name>
    <value>3i</value>
    <length>255</length>
  </tableModel>
  <tableModel>
    <name>pin</name>
    <value>50</value>
    <length>4</length>
  </tableModel>
</tables>
</appletModel>
<appletModel alias="wolf_hce" AID="F0014144500001">
  <instructions>
    ...
  </instructions>
</appletModel>
  
```

E. WOLF API

As mentioned, WOLF aims to be an ontology-driven Framework based on self-descriptive metadata (being currently XML format). This is our prerequisite to ensure maintainability and portability of the components. The architecture of WOLF is drawn around the central concept of SE-QL being the interface for a simplified and optimized handling of the data that is the same regardless of the mobile and whatever SE type and whatever covered platform. This interface allows developers to define their own configuration in the metadata file (mapping the APDU instructions, description and alias of the data, etc.), making the interface compliant with already existing (contact or contactless) smart card standards or proprietary instructions.

WOLF API encompasses a generic applet and a generic wallet each based on its own metadata that can be easily personalized for a rapid implementation (see Fig. 9).

1) WOLF current progress

Current API provides:

- SE-QL package which is common for all Java-based platforms contains the SE-QL generic interface, APDU's handling classes, SE-QL callback interface, SE-QL parser, and metadata objects.
- Android packages contains: (i) OMAPI SE-QL controller to access the SE with OMAPI API on Android SDK17 platform, (ii) HCE controller for Android SDK19 platform SE-QL, and (iii) the generic WOLF HCE service (that emulates the SE) able to communicate with (iv) WOLF generic applets. Android package will also contain the off-host APDU service** (to communicate with the SIM-SE), the generic Wallet UI, and Android reader for external devices (work in progress).

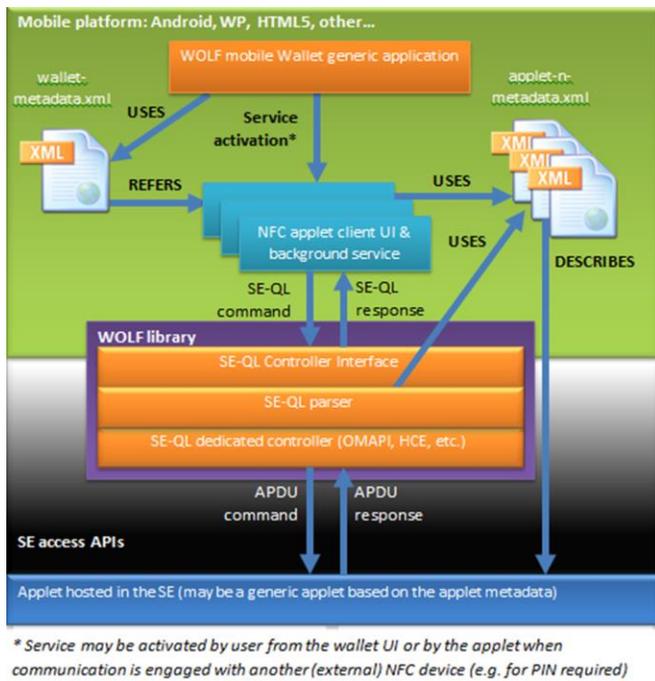


Fig. 9. WOLF architecture overview

- Smart card reader package provides the SE-QL controller for Java™ Smart Card I/O API compliant with PC / SC (NFC) readers.
- (4) HTTP package contains the HTTP requests helper for the sending of GET, POST, PUT, and DELETE requests (e.g. to access RESTful Web services), and the SOAP Web services helper.
- Cryptography helper (encode / decode, generate hash key) is provided in the tools package.
- TSM package contains tools dedicated to TSM (work in progress).

WOLF for Android and SE-QL have been tested for three use cases so far : (i) SIM-based NFC service on Android device compiled with SDK17 (Android Jelly Bean, build 4.2.2 or earlier on nonstandard build of CyanogenMod) using OMAPI, (ii) Android HCE generic service compiled with SDK19 (Android KitKat, build 4.4.2), and (iii) a Java application to test the communication between a USB NFC reader plugged to the PC (using standard PC / SC driver, see www.pcscworkgroup.com/) and an NFC device.

The beta version of WOLF plugin for Android was implemented last year for CyanogenMod with seek-for-Android [47] implementation of OMAPI (need the device bootloader to be unlocked and CyanogenMod build to be installed in replacement of standard OS on the phone). Afterward, OMAPI was formally integrated as an external API for Android standard build 4.2.2. But since Android KitKat, OMAPI is no more supported by the official Android build; it has been replaced by Android HCE, the software based card emulation mode. ***On Android KitKat, SIM-SE access must be routed by Android off-host APDU service...*

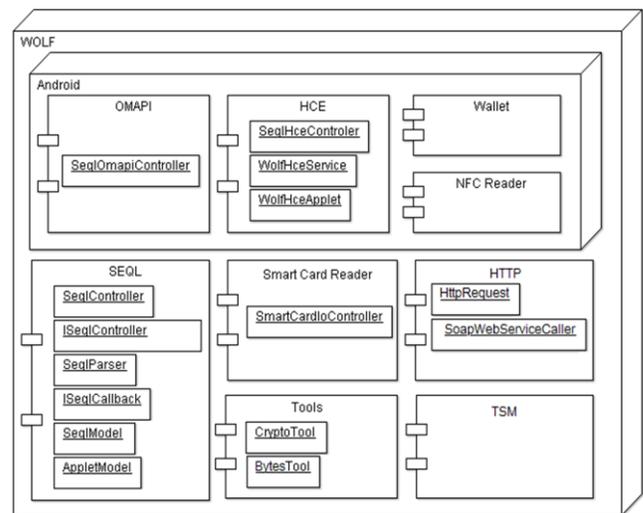


Fig. 10. Major WOLF current components

2) WOLF experimentation within FIRST project

The FIRST wallet for Financial Inclusion (FI) uses SE-QL commands and WOLF platform in the end-to-end supply chain NFC services for the secure traceability of aids delivery. WOLF was successfully tested in the development of two use cases prototyped at MBDS in 2012-2014 within FIRST project; M-PDS (Mobiquitous NFC Public Distribution System) [8] and BARTER 2.0 (Bank of Animal in Rural TERRitories 2.0 / Social Network of Donators) [8]: WOLF API and SE-QL are the kernel of mobile NFC services modules within use cases prototypes of FIRST wallet as shown in Fig. 11.

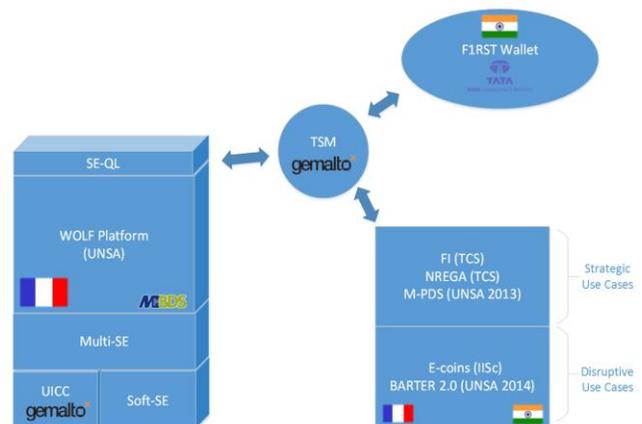


Fig. 11. Architecture of FIRST ecosystem [8]

3) WOLF / SE-QL implementation

On Android platform, WOLF library has to be referenced in the project properties. Once it is done, SE-QL controller class can be instantiated once with appropriate parameters (targeted platform, application context if Android, metadata file). Then, you can get the controller and use SE-QL to communicate with the SE, and receive events with the returned responses from the applet as shown in source code 3:

Source code 3. Android implementation of WOLF: SE-QL controller instantiation

```
1. import org.mbds.wolf.seql.ISeqlCallBack;
2. import org.mbds.wolf.seql.SeqlController;
3. import org.mbds.wolf.seql.exceptions.ApduError;
4.
5. public class MyApplication extends Application
6.     implements ISeqlCallBack {
7.     SeqlController ctrl = null;
8.     //.....
9.     protected void init() {
10.         boolean ok = false;
11.         try {
12.             SeqlController.OS os =
13.                 SeqlController.OS.ANDROID_HCE;
14.             if (android.os.Build.VERSION.SDK_INT<=19)
15.                 os = SeqlController.OS.ANDROID_OMAPI;
16.             ctrl = new SeqlController(
17.                 getApplicationContext(), os,
18.                 File metadata, this);
19.             ok = true;
20.         } catch (ClassNotFoundException e) {
21.             e.printStackTrace();
22.         } catch (NoSuchMethodException e) {
23.             e.printStackTrace();
24.         } catch (InstantiationException e) {
25.             e.printStackTrace();
26.         } catch (IllegalAccessException e) {
27.             e.printStackTrace();
28.         } catch (InvocationTargetException e) {
29.             e.printStackTrace();
30.         }
31.         if (!ok) {
32.             Toast.makeText(this, "SE-QL controller
33.                 could not be instantiated,
34.                 application will finish!",
35.                 Toast.LENGTH_LONG).show();
36.             quitApp();
37.         }
38.     }
39. }
```

Source code 4. Android implementation of WOLF: executing SE-QL instruction

```
1. public class MyActivity extends Activity
2.     implements ISeqlCallBack {
3.     private SeqlController ctrl;
4.     //...
5.     @Override
6.     protected void onCreate(Bundle
7.         savedInstanceState) {
8.         super.onCreate(savedInstanceState);
9.         MyApplication act =
10.             (MyApplication) getApplication();
11.         ctrl = act.getController();
12.     }
13.     @Override
14.     protected void onResume() {
15.         super.onResume();
16.         ctrl.setCallback(this);
17.         if(!ctrl.isServiceConnected()) {
18.             ctrl.initService();
19.         }
20.     }
21.     private boolean executeSeqlCommand(String
22.         statement, int commandId) {
23.         return ctrl.execute(statement,
24.             commandId);
25.     }
26. }
```

```
25. @Override
26.     public void onPINRequired() {
27.         startActivityForResult(new
28.             Intent(this,
29.                 PinEntryView.class),
30.                 Constant.PIN_REQUEST);
31.     }
32.     @Override
33.     public void onResponse(Map<String, Object>
34.         results, int commandId) {
35.         //Process results
36.     }
37. }
```

Fig. 12 gives an overview of components interaction showing how the client module (service UI) requires the name and the password of the handset end-user, stored by the applet using SE-QL; then, WOLF plugin initiates the connection and the applet requests the user PIN entry. When user PIN entry has been successfully transmitted, the previous SE-QL instruction can be processed. This sequence derived from FIRST use cases prototyped at MBDS was a first proof-of-concept; the client module has been tested with the applet embedded in the SIM and the generic HCE applet of WOLF without requiring any code changes except for the manifest (and the Android host-apdu-service XML metadata), since HCE services must be declared in the client application.

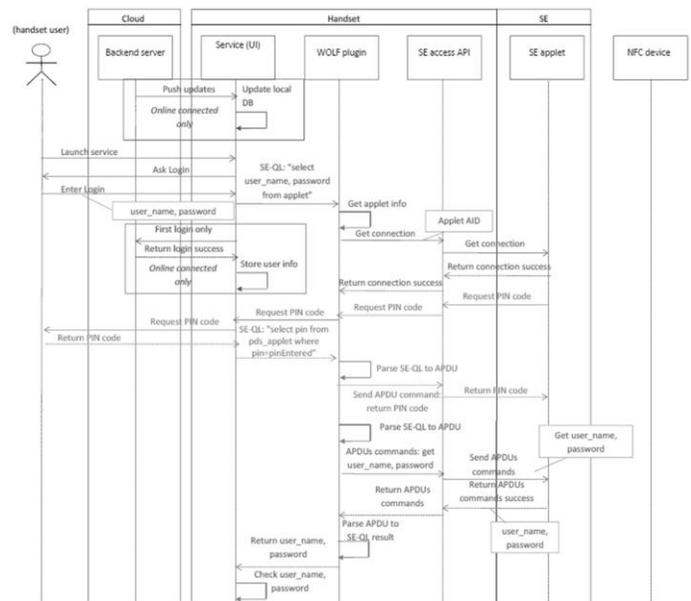


Fig. 12. Sequence diagram example using WOLF API

The WOLF SE-QL reader Java tester is intended to test the communication with the applet hosted into the SE of the mobile phone: the tester frame is shown in Fig. 13: user grabs the SE-QL instruction in the input text area and clicks on the “Execute” button. Then, he is asked to place the device on the reader and the SE-QL instruction is parsed into APDU(s) command(s); the result is shown in the log trace view.

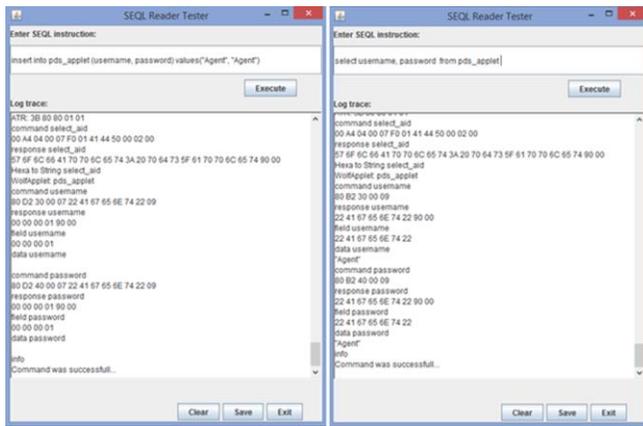


Fig. 13. SEQL tester for PC / SC Reader

IV. CONCLUSION AND FUTURE WORKS

In this report, we demonstrated SE-QL high added value making easy and fast the development of secure NFC mobile services using card emulation mode. The great advantage of SE-QL is that it is compatible with most APDU standards and can be used for already existing applets. SE-QL is a very good solution for applications that do not need complex transactions.

Withal, it does not already fully meet ACID (Atomicity, Consistency, Isolation, and Durability) properties that guarantee the transactions are processed reliably; in the next release, we are focusing on the transactions integrity by implementing the mechanisms for the handling of a set of transactions that need to be completely executed in order to maintain system consistency. This will be done using the "BEGIN transaction", "COMMIT", and "ROLLBACK" instructions (taking care of the limitation of RAM will require a temporary serialization before complete execution) and the management of a time-out. But, as we talk about contactless communication, we may face untimely interruptions transactions when the devices are removed prematurely.

Several researches have been done on database implementation in smart cards (that can apply to the SE) [11], [12], [13], [18]; this highlights the need for embedded DB system. Nowadays, lite databases have been adopted for low capacity systems like SQLite (www.sqlite.org) which is multiplatform compliant and it is provided as part of Android OS. This is why we will also explore other areas of investigation around embedded DB, and also other object-oriented structures in the payload such as RESTful (Representational State Transfer) services, and the ability to provide XML, JSON (JavaScript Object Notation), etc.

Another current important stuff is to manage security according standards specification. This will be done combining IISc Bangalore research on the crypted SE to ours, e.g. using cryptographic APIs (e.g. Cipher) and studying existing protocols, or by proposing new ones and Cloud-based solutions (studied at LSIS lab of Aix-Marseille University [11]).

ACKNOWLEDGMENT

Special Thanks to Gemalto (supporting MBDS apprentices and a CIFRE research Scholarship) and IFCEPAR | CEFIPRA

organization supporting FIRST (Financial Inclusion based upon Rural ubiquitous Services Technological platform) project (2012-2015) and more personally to Mr. Ilan Mahalal, Program Manager at Gemalto, Mr. Debi Pati, CEO lead at TATA CS, and Nicolas Pastorelly, Olfa Arfani, Mohamed Sidime, Guillaume Larroque, Pierrick Morizot from University of Nice Sophia Antipolis and MBDS for their technical contributions to NFC Container, WOLF platform and FIRST project.

REFERENCES

- [1] S. Miranda, "Relational Objects Databases (Bases de données objets relationnelles (SQL3 et ODMG))," Dunod, 2004.
- [2] S. Miranda et al., "Mobiquitous Information Systems (Systèmes d'Information mobiquitaires)," in *Ingénierie des systèmes d'information*, RTSI Série ISI, Vol. 16 no 4, Hermes Lavoisier, France, 2011.
- [3] S. Miranda, N. Pastorelly, V. Ishkina, D. Torre, V. Chaix, "Lessons inferred from NFC mobiquitous innovative information service prototyping at UNS", in [1], 2011, pp. 15-47.
- [4] M. Della Peruta, A. Atour, "Business models for mobiquitous (social) money: Application to M-PDS program in India" research report, to be published in *International Journal of Complementary Currency Systems (IJCCS)*, 2014.
- [5] D. Pati, S. Miranda, confidential document, "FIRST Project: Collaboration Agreement," Feb, 2012.
- [6] D. Pati, "Architecture of FIRST system," IFCEPAR | CEFIPRA Report, Nov., 2013
- [7] O. Arfani, M. Sidime "M-PDS (mobiquitous Public Distribution system) USE CASE for FIRST project," M.S. thesis, MBDS CS department, University of Nice – Sophia-Antipolis, France, Oct., 2013.
- [8] G. Larroque, P. Morizot, B. Renaut, S. Miranda: "Proposal of a disruptive Use Case for FIRST : Mobiquitous Bank of Animals – BARTER 2.0 project", draft Nov. 2013.
- [9] A.-M. Lesas, "FIRST Research project Mobiquitous NFC Financial services for unbanked people", presentation at WIMA Conf. in Monaco, NFC Research Track, April 22, 2014.
- [10] C. Papetti, K. Sok, S. Miranda, "Mobiquitous NFC Tourism (Une plateforme de gestion de Tags pour le tourisme mobiquitaire du Futur)," *Monde du Tourisme*, to be published, 2014.
- [11] P. Paradinas, J.-J. Vandewalle, "A personal and portable database server: the CQL card," 1994, available: <http://cedric.cnam.fr/~paradin/presentation/CQL.pdf>
- [12] P. Paradinas, J.-J. Vandewalle, "How to integrate Smart Cards in Standard Software without writing specific code?," 1994, available: <http://cedric.cnam.fr/~paradin/presentation/CTST.pdf>
- [13] 3GPP TSG-T3, "Phone book management with ISO 7816 part 7 (SCQL)," Document T3-99167, Source: Gemplus, Miami, June, 14-16th, 1999.
- [14] M. Roland, "Software Card Emulation in NFC-enabled Mobile Phones: Great Advantage or Security Nightmare?," NFC Research Lab Hagenberg, Univ. of Applied Sciences, Austria, IWSSI / SPMU, June, 2012, available: <http://www.medien.ifi.lmu.de/iwssi2012/papers/iwssi-spmu2012-roland.pdf>
- [15] L. Pesonen, "TSM Point of View and Issues Faced," EC/ETSI Workshop on Collaborative Ecosystem for M-Payment, Sophia Antipolis, France, July, 1, 2014, available: http://docbox.etsi.org/Workshop/2014/201407_MPAYMENTWORKSH OP/S02_ECOSYSTEM_and_ISSUES/S02_Pesonen_GD.pdf
- [16] C.-J. Date, "Introduction to data base systems," Addison-Wesley Educational Publishers Inc., U.S., 1975.
- [17] E.-F. Codd, "A Relational Model of Data for Large Shared Data Banks," IBM Research Report, San Jose, Aug. 19, 1968.
- [18] N. Anciaux, L. Bouganim, P. Pucheral, "Embedded RDBMS within a smart card feedback (SGBD embarqué dans une puce : retour d'expérience)," *Techniques et Sciences Informatiques*, vol. 27, no 1-2, Sept., 2008, pp. 141-180.

- [19] S. Miranda, A.-M. Lesas, " VAMP project: Mobiquitous NFC car (Architecture logicielle du projet Vamp. Plateforme mobiquitaire embarquée à bord de véhicules mobiles)," *Revue du Génie Logiciel*, No103, Dec., 2012, pp. 38- 48.
- [20] Li, Y., Boucelma, O., Provenance Monitoring in the Cloud, IEEE 6th International Conf. on Cloud Computing, June 27-July 2, 2013, Santa Clara, USA.
- [21] V.Coskun, K. Ok, B. Ozdenizci, "NFC application development for Android," Wrox Ed, John Wiley&Sons, UK, 2013.
- [22] E. Coleen Coolifge, P.Hourani, "Securing cloud and mobility," CRC Press, Auebach Publication, USA, 2013.
- [23] L. Francis, G.-P. Hancke, K.-E. Mayes, K. Markantonakis, "Practical Relay Attack on Contactless Transactions by Using NFC Mobile Phones," *Cryptology ePrint Archive*, Report 2011/618, 2011, available: <http://eprint.iacr.org/2011/618>.
- [24] V. Alimi, "An Ontology-based Framework to Model a GlobalPlatform Secure Element," presentation at WIMA Conf. in Monaco, NFC Research Track, April 11, 2012.
- [25] M. Roland, "Secure Element APIs and Practical Attacks on Secure Element-enabled Mobile Devices," presentation at WIMA Conf. in Monaco, NFC Research Track, April 11, 2012.
- [26] L. Francis, G.-P. Hancke, K.-E. Mayes, K. Markantonakis, "Practical NFC Peer-to-Peer Relay Attack Using Mobile Phones," *RFID Security and Privacy Issues*, LNCS vol. 6370/2010, Heidelberg, 2010, pp. 35-49.
- [27] G.-P. Hancke, K.-E. Mayes, K. Markantonakis, "Confidence in smart token proximity: Relay attacks revisited," *Computers & Security*, Elsevier Ltd., Springer Berlin, 2009, pp. 615-627.
- [28] H. Ailisto, T. Matinmikko, J. Häikiö, A. Ylisaukko-oja, E. Strömmer, M. Hillukkala, A. Wallin, E. Siira, A. Pöyry, V. Törmänen, T. Huomo, T. Tuikka, S. Leskinen, J. Salonen, "Physical browsing with NFC technology," *VTT Research Notes* 2400, 2007.
- [29] T. Tuikka, M. Isomursu, "Touch the Future with a Smart Touch," *VTT Research Notes* 2492, 2009.
- [30] V. Coskun, K. Ok, B. Ozdenizci, "Professional NFC Application Development for Android™," Wrox, John Wiley & Sons, Ltd., 2013.
- [31] D. Schall, "Service-Oriented Crowdsourcing: Architecture, Protocols and Algorithms," *SpringerBriefs in Computer Science*, 2012.
- [32] K. Finkenzeller, "RFID handbook," third edition, John Wiley & Sons, Ltd., 2010.
- [33] T. Igoe, D. Coleman, B. Jepson, "Beginning NFC," O'Reilly Media, Jan., 2014.
- [34] Urien, P., Piramuthu, S., "Towards a secure Cloud of Secure Elements concepts and experiments with NFC mobiles," *IEEE International Conf., CTS*, May 20-24, 2013, San Diego, USA.
- [35] P. Pourghomi, G. Ghinea, "Managing NFC payment applications through cloud computing," *IEEE, ICITST*, Dec. 10-12, 2012, London, UK.

Optimal Network Design for Consensus Formation: Wisdom of Networked Agents

Eugene S. Kitamura
Department of Computer Science,
National Defense Academy
Yokosuka, Japan

Akira Namatame
Department of Computer Science,
National Defense Academy
Yokosuka, Japan

Abstract—The wisdom of crowds refers to the phenomenon in which the collective knowledge of a community is greater than the knowledge of any individual. This paper proposes a network design for the fastest and slowest consensus formation under average node degree restrictions, which is one aspect of the wisdom of crowds concept. Consensus and synchronization problems are closely related to variety of issues such as collective behavior in nature, the interaction among agents as a matter of the robot control, and building efficient wireless sensor networks. However, designing networks with desirable properties is complex and it may pose a multi-constraint and multi-criterion optimization problem. For the purpose of realizing such efficient network topology, this paper presents an optimization approach to design networks for better consensus formation by focusing on the eigenvalue spectral of Laplacian matrix. In both the fastest and slowest networks presented, consensus is formed among local structures first, then on a global scale. This suggests that both local and global topology influence the networks dynamics. These findings are useful for those who seek to manage efficient consensus and synchronization in a setting that can be modeled as a multi-agent system.

Keywords—wisdom of crowds; consensus problem; Laplacian matrix

I. INTRODUCTION

There is a strong interest in many fields to answer the following questions. How do interacting individuals with micro-motives produce the desirable or undesirable outcomes at the aggregate level? How are interactions among agents that produce some regularities of interest at the macroscopic level identified? Most of our social activities are substantially free of centralized management, and although people may care about how it all comes out in the aggregate, people's own behaviors are typically motivated by self-interest. Therefore, in examining collective outcome, the observer shall draw heavily on the individual behaviors. It might be argued that understanding how individuals behave is sufficient to understand most parts of the social system. Although individual behaviors are important to understand, they are not sufficient to describe how a collection of agents generate desirable macroscopic outcome. To make the connection between microscopic behavior and macroscopic outcome of interests, the observer usually has to look at the system of interactions among agents described as the interaction network topology.

There are social systems for which it is difficult to understand how they work or to find better ways to make them

work. For instance, social systems often produce inefficient outcomes at the aggregate level in a way that the individuals who comprise the system cannot evaluate or are not even aware of. When the system results in some undesirable outcome, the cause is often thought of as the members who comprise the system. The resulting outcome is observed as corresponding to the intentions of the members who compromise the system. It is not easy to tell from emergent phenomena just what the motives are behind individuals and how strong they are.

Social systems often result in the features of *emergent properties*, which are properties of the system in which separated components by themselves do not have. Other social phenomena are also viewed as *emergence* that have arisen from billions of small-scale and short-term decisions of interacting agents. Billions of people make billions of decisions everyday about many things. It often appears that the aggregation of these unmanaged individual decisions leads to unpredictable outcomes. Unintended consequences and side effects are closely related to emergent properties. In other words, the global or macroscopic functionality of a system is the sum of all side effects of all emergent properties.

People constantly interact with each other in different ways and for different purposes. Somehow these individual interactions produce some coherence at the aggregate level, and therefore, aggregation may generate structure and regularity. The individuals involved may have a very limited view of some part of the whole system but their activities are coordinated extensively and produce a desirable outcome at the aggregate level. These emergent properties are the result of not only the behavior of individuals but the interactions between them as well.

In his book, titled *The Wisdom of Crowds*, Surowiecki explores an idea that has profound implications: a large collection of people are smarter than an elite few, no matter how they are brilliant and better at solving problems, fostering innovation, coming to wise decisions, even predicting the future [1].

He explains the wisdom of crowds emerges only under the right conditions: (1) diversity, (2) independence, (3) decentralization, and (4) aggregation. His counterintuitive notion, rather than the madness of crowd such as herding or cascade as traditionally understood, suggests new insights for the issue on how complex social and economic activities should be organized.

In contrast, Lorenza et al. [2] demonstrates by experimental evidence that even mild social influence can undermine the wisdom of crowd effect in simple estimation tasks. In the experiment, subjects could reconsider their response to factual questions after having received average or full information of the responses of other subjects. They compare subjects' convergence of estimates and improvements in accuracy over five consecutive estimation periods with a control condition in which no information about others' responses was provided. Although groups are initially wise, knowledge about the estimates of others narrows the diversity of opinions to such an extent that it undermines the wisdom of crowd effect. Especially the social influence effect diminishes the diversity of the crowd without improvements of its collective error.

This observation derives requirements for a more general model of network effects. Therefore a new area of research has emerged aiming at explaining the phenomena of strong positive or negative network effects in markets and their implications on market coordination and efficiency. However, the assumptions and simplifications implicitly used for modeling social interaction processes fail to explain the individual cognitive decision-making process as well as the network structure.

Consensus problems have a long history in computer science and control theory [3]. In networks of agents, consensus means to reach an agreement regarding a certain quantity of interest that depends on the state of all agents. A consensus algorithm is an interaction rule that specifies the information exchange between an agent and all of its neighbors on the network. The theoretical framework for solving consensus problems for networked systems was introduced by Olfati-Saber and colleagues [4].

Many of the essential features displayed by complex systems emerge from their underlying network structure. Whether optimization plays a key role in shaping the evolution of optimal networks is an important question [5]. One of the broadest areas of research, the optimization has a long history. Optimization is most often connected to a function that the system performs. In numerous cases the function is multifaceted. How network structure influences the global performance of such systems is probably the question that is posed most frequently in network research [6][7].

Here the following question is addressed: what topology fosters or dampens consensus or synchronization on networks? In this paper, such optimal topologies are constructed for any fixed number of nodes and links, by employing an evolutionary optimization procedure. The issue of the wisdom of networked agents in terms of convergence speed by formulating consensus problems is addressed. An evolutionary algorithm is used involving optimizing the eigenvalue ratio of the Laplacian matrix under the constraint of the average degree. Traditionally, optimization has a strict mathematical definition, which refers to obtaining the solutions that strictly optimize a well-defined objective function. Here a looser definition of the word is adopted by extending it to include a tendency of the system to improve its behavior as a result of a selection pressure based on artificially imposed fitness function. It comprises the variation principles or the survival-of-the-fittest principles that pervade

biology and engineering, the foundational hypotheses of numerous computer algorithms, and the frameworks for addressing the improvement of efficiency in various contexts. Optimization in complex networks has a broad significance, incorporating static and dynamical properties and serving as an instrument to analyze and control the evolution and function of both natural and engineered systems.

In this paper, the fast consensus topology is introduced that is not Ramanujan for average degree less than 2 that has not been previously reported, generated with the genetic algorithm. Additionally, the fastest consensus is contrasted with the slow consensus topology, based on a heuristic model which was originally inspired by the evolutionary algorithm. Finally, a consensus of two clique topology connected by a single link will be examined. All of these networks display a topological priority in the progression of consensus, where consensus is formed initially with locally oriented agents that form close proximity. Formation of such topology by an agent society may be referred to as their wisdom, resulting in varying aggregate performances from a global perspective.

II. NETWORK DESIGN FOR BETTER CONSENSUS

It is broadly recognized that most complex systems in nature are organized as intricate network patterns. This observation has triggered an intense research effort aimed at understanding the organizing principles of these networks, their structural properties, and the interplay between topology and dynamics. Understanding the network structure of individual systems has led to tremendous advances in the past decade [8][9].

Most of the complex systems seen in real life also have associated dynamics, and the structural properties of such networks have to be linked with their dynamical behavior. While most of the initial effort was put into understanding the topological properties of networks, the interest has gradually shifted towards the analysis of the interplay between a topology and the dynamics of network components. In general, each element (node) in a network undergoes a dynamical process while coupled to other nodes. The system's collective behavior depends strongly on the efficiency of communication paths, which is in turn dictated by the underlying network topology. In this way, the network structure determines to a large extent the possibility of a coherent response.

Thus, the topology of the network remains static while the states of the nodes change dynamically. In this respect, one of the questions of obvious significance is whether there is a relation between the stability of the dynamics against small perturbations in the dynamical variables and the specific arrangement of the network's connections. If the perturbation decays quickly, so that it is unable to spread to the rest of the network, the network is said to be stable. Such a property is necessary if networks are to survive the noisy environment that characterizes the real world.

It has sometimes been argued that, networks with larger number of nodes, links and stronger interconnections are more stable. Such assertions are partly based on empirical observations, e.g., in ecology, where it has been found that more diverse and strongly connected ecosystems are more

robust than their smaller, weakly connected counterparts. Some important processes studied within this framework include synchronization of the individual dynamical systems and consensus processes such as opinion formation. Studies like these have clarified that certain topological properties have strong impacts on the dynamics of the networks. In recent studies, the reason for the occurrence of synchronized networks became clear and the underlying network topology is important, however, little is known about what the best network topology is for synchronization [10][11].

Two principal approaches have contributed to understanding network structures so far. The first is an assembly mechanism that derives the structure of large complex networks from processes that describe the piecewise addition of nodes and links according to stochastic rules over time. Preferential attachment is an important mechanism in this category. The second approach is via optimization, thus assuming that a network structure observed in the real world represents the end point of some guided reorganization mechanism that aims at optimizing system performance during its evolution [12].

When modeling consensus, the underlying network restricts communication among agents where agents can exchange information only with the connected agents. There are many studies to analyze the influence of the network topology on the convergence speed of the iterative consensus algorithm.

Based on the concept of “The wisdom of crowds”, Golub et al. [13] developed a social network approach to model consensus phenomenon. Their study basically uses an important model of network influence largely due to DeGroot who studied consensus problems in groups of experts who originated in statistics [14]. In their paper, Golub et al. examined one aspect of this broad theme: for which social network structures will a society of agents who communicate and update naïvely come to aggregate decentralized information completely and correctly. They focus on situations where there is some true state of nature that agents are trying to learn and each agent’s initial belief is equal to the true state of nature plus some idiosyncratic zero-mean noise. The network structure of agents is described using a weighted network. Agents have beliefs about some common question of interest—for instance, the probability of some event. At each time step, agents communicate with their neighbors in the social network and update their beliefs. The updating process is simple. An agent’s new belief is the average of his or her neighbors’ beliefs from the previous period.

An outside observer who could aggregate all of the decentralized initial beliefs could develop an estimate of the true state that would be arbitrarily accurate in a large enough society. Golub et al. studied learning in a setting where agents receive independent noisy signals about the true value of a variable and then communicate it in a network. The agents naïvely update beliefs by repeatedly taking weighted averages of neighbors’ opinions. They show that all opinions in a large society converge to the truth if and only if the influence of the most influential agent vanishes as the society grows. They also identify obstructions to this, including prominent groups, and provide structural conditions on the network ensuring efficient

learning. Whether agents converge to the truth is unrelated to how quickly consensus is approached.

The consensus problem is also related to synchronization. Synchronization is the most prominent example of coherent behavior, and is a key phenomenon in systems of coupled oscillators as those characterizing most biological networks or physiological functions. Synchronous behavior is also affected by the network structure. The continuous range of stability of a synchronized state is a measure of the system’s ability to yield a coherent response and to distribute information efficiently among its elements, while a loss of stability fosters pattern formation.

In recent studies, the reason for the occurrence of synchronized networks became clear and the underlying network topology turned out to be important. However, synchronization often occurs unexpectedly and little is known what the best network topology is for synchronization.

III. OPTIMAL NETWORK DESIGN FOR BETTER CONSENSUS

The analysis of consensus problems relies heavily on matrix theory and spectral graph theory [15]. The interaction topology of a network of agents is represented using an undirected graph G with the set of nodes and edges. Neighbors of agent i is denoted as n_i .

Consider a network of agents with the following dynamics:

$$\dot{x}_i = \sum_{j \in N_i} a_{ij} (x_j(t) - x_i(t)), \quad (1)$$

where a_{ij} is the weight of agent i on agent j .

Here, reaching a consensus means asymptotically converging to the same internal state by way of an agreement characterized by the following equation:

$$x_1 = x_2 = \dots = x_n = \alpha. \quad (2)$$

Assuming that the underlying graph G is undirected ($a_{ij} = a_{ji}$ for all i, j), the collective dynamics converge to the average of the initial states of all agents:

$$\alpha = \frac{1}{n} \sum_{i=1}^n x_i(0). \quad (3)$$

The dynamics of system in (1) can be expressed as

$$\dot{x} = -Lx(t). \quad (4)$$

L is the graph Laplacian matrix of the network G ; the Laplacian matrix is defined as

$$L = D - A, \quad (5)$$

where $D = \text{diag}(d_1, d_2, \dots, d_n)$ is the diagonal matrix with elements $d_i = \sum_j a_{ij}$ and A is the binary adjacency matrix ($n \times n$ matrix) with elements a_{ij} for all i, j where a_{ij} is 1 if agent i and agent j is connected or 0 if they are disconnected.

Notice that because the networks in this paper are undirected, \mathbf{L} is a symmetric matrix with all real entries, and therefore a Hermitian matrix. In this research, \mathbf{L} being a Hermitian matrix is always met with equality since the diagonal entry of each row in \mathbf{L} is the degree of node i , and each link connected to i results in -1 in the same row. So the sum of all off diagonals in a row is d_i . Therefore \mathbf{L} is a positive semi-definite matrix. Since \mathbf{L} is semi-definite (and therefore also Hermitian), the following ordering convention for the eigenvalues will be adopted:

$$0 = \lambda_1 \leq \lambda_2 \leq \dots \lambda_n \quad (6)$$

The interval in which the synchronized state is stable is larger for a smaller ratio of the two eigenvalues $Q = \lambda_n/\lambda_2$, therefore a network has a more robust synchronized state if the eigenvalue ratio $Q = \lambda_n/\lambda_2$ is smaller. Focusing on the first part, the network optimization will be the guided evolution of networks subject to some constraints. Attempting to explain the formation of a small-scale size of networks, the constrained evolution of networks towards consensus or synchrony optimality will be investigated. Constraint optimizations are included via a fitness function that combines the desired goal such as synchronization properties of the networks optimized, or the propensity of a network to synchronize with an average degree requirement needed to connect the nodes.

Ramanujan graph is known in the literatures as the best networks for fast consensus. Ramanujan graphs [16] are k -random regular networks with the second minimum eigenvalue satisfying:

$$\lambda_2(L) \geq k - 2\sqrt{k-1}, \quad (7)$$

It is known that among the second minimum eigenvalue $\lambda_2(L)$ of various Laplacian matrices, Ramanujan graphs have the largest $\lambda_2(L)$ [17]. One class of Ramanujan graphs is a random regular network, which is easy to construct and used in many application for better consensus.

However, for sparse networks with lower average degree, the condition in (7) does not carry any information especially for $k=2$. In this case a Ramanujan graph is a ring network with the degree of 2 as shown in Fig.1(a). However, the second minimum eigenvalue $\lambda_2(L)$ of the ring network is very small, and consensus is very slow on the Ramanujan graph with the degree $k=2$.

Genetic algorithms have been extensively used in single objective optimization for various communication network related optimization problems. Optimizing complex networks usually involve multiple objectives such as the network size as well as various network properties. In this paper, an evolutionary algorithm involving minimization of the eigenvalue ratio of the Laplacian matrix with the constraint of the average degree is used in order to design the optimal network.

Now consider an optimization approach to design networks with better consensus. Although optimization requires finding an optimal solution for a well-defined objective function,

optimization as a selection pressure to minimize the following fitness function is considered [18]:

$$E(\omega) = \omega \frac{\lambda_n}{\lambda_2} + (1 - \omega)\langle k \rangle \quad (8)$$

where $\langle k \rangle$ is the average degree, and ω ($0 \leq \omega \leq 1$) is a parameter controlling two objects.

The eigenvalue ratio $Q = \lambda_n/\lambda_2$ decreases as the average degree increases and the convergence speed becomes much faster. Therefore an interesting question is how to design a sparse network with a few degrees that guarantees a certain convergence performance. However this is a very difficult combinatorial problem. Therefore, an evolutionary design method is an effective way to design such a sparse optimal network.

Initially 10 random networks with the Poisson degree distributions are generated and the genetic algorithm to obtain better networks in terms of improving the fitness function in (8) is used. The network is encoded as a binary adjacency matrix to perform the mutation and crossover. Next, the most suitable matrices among the parents and children matrices are chosen, and the others are eliminated.

The multi-point crossover was used. After the crossover, each element in the matrix switches to a reverse state with a specific probability. In this paper, the network is an undirected graph, and so, if one element is reversed, the symmetry element is reversed at the same time.

There is a possibility that an isolated network appears after crossover and mutation. In this paper, when an isolated node appears in a new network that the node has infinite distances to another node, the network is dumped. Therefore, non-isolated matrices can be used. After many generations have passed, an optimal network which minimizes the fitness function defined in (8) can be obtained. For dense networks with the average degree $\langle k \rangle$ is larger than 4, the evolutionary optimized networks are Ramanujan graphs.

However for a sparse network of the average degree $\langle k \rangle=2$, an evolutionary optimized network is a ring-trees type, where many modules networks with tree structures are combined by a ring network. Fig. 1 illustrates the difference of the network topology of an evolutionary optimized network (Fig. 1(b)) from a Ramanujan graph (ring network with $k=2$) (Fig. 1(a)).

IV. SIMULATION RESULTS ON CONSENSUS FORMATIONS

In this section, simulation results on fastest and slowest consensus formation on an evolutionary optimized network and a heuristically designed network are shown.

A. Fastest Consensus formation on a sparse network with the average degree $\langle k \rangle=2$

To examine how fast the process is for achieving consensus among agents in a network, the convergence speed is measured. In the following (9), various initial values are given to each node.

$$x_i(0) = i \quad (i = 1, 2, \dots, n) \quad (9)$$

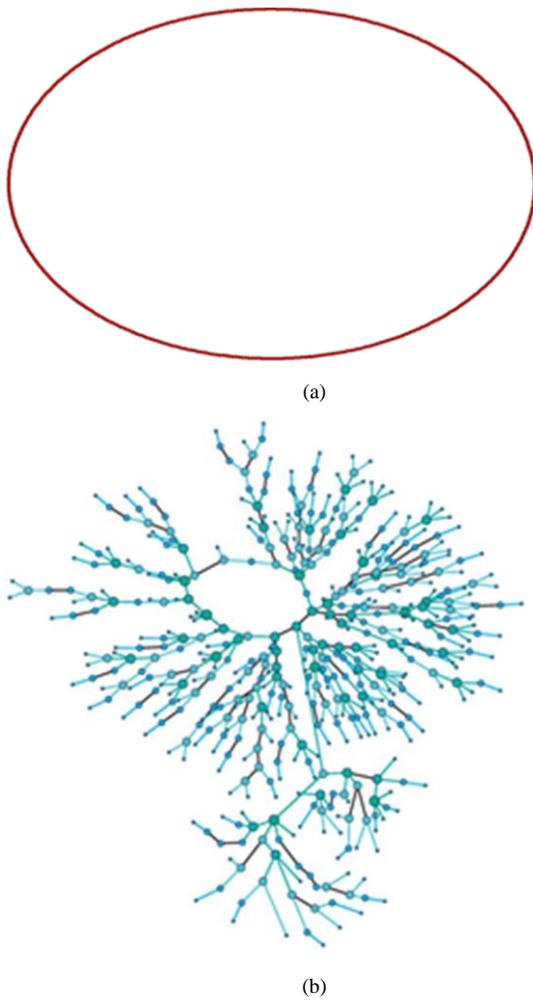


Fig.1. The convergences of the optimized network for (a) Ramanujan network with $\langle k \rangle = 2$ and (b) ring-trees network.

The state of each agent (node) x_i ($1 \leq i < n$) in the system converges to a constant value. The time required for achieving a consensus is compared, that is, until the states asymptotically converge to the same internal state by way of an agreement characterized by the following equation:

$$x_1 = x_2 = \dots = x_n \quad (10)$$

Fig. 2 shows the convergences of the optimized network. This means the collective dynamics converge to the average of the initial states of all agents. In general, the optimized network is faster than the Ramanujan network for $\langle k \rangle = 2$, the fastest among the previously network models.

Fig. 2 (a) is from Ramanujan network and Fig. 2 (b) is from ring-trees network as shown in Fig. 1 with 100 nodes and 100 links. Consensus dynamics on ring-trees network is achieved much faster than Ramanujan network, where the time of consensus is 1201 steps on ring-trees network and over 5000 steps on Ramanujan network. On the ring-trees network, where agents are divided into each tree network at first. After that, the global consensus is formed via the ring network. It is very interesting

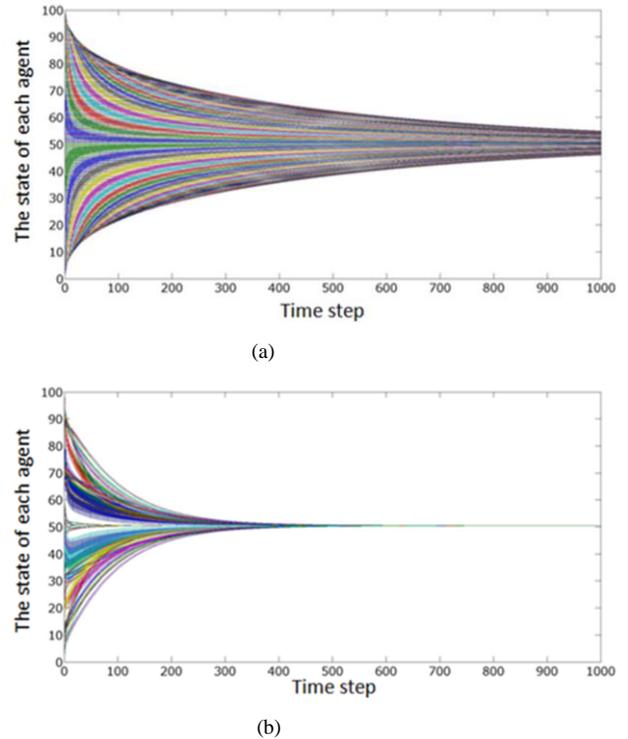


Fig.2. Network topologies of (a) a Ramanujan graph and (b) an evolutionary optimized network, a ring-trees network, both with 500 nodes and 500 links with an average degree $\langle k \rangle = 2$.

that this simple mechanism makes a big difference between these networks on the convergence time of consensus dynamics.

B. Network with the slowest consensus formation

In this section a heuristic network for slow consensus is proposed. The original topology resulted from the evolutionary design approach. For the slowest design, the inverse of the eigenvalue ratio λ_2/λ_n is minimized under the constraint of the average degree:

$$F(\omega) = \omega \frac{\lambda_2}{\lambda_n} + (1 - \omega)(k) \quad (11)$$

Then the network which minimizes the fitness function in (11) is obtained. Then the optimal network for the slowest consensus is a clique-with-line network as seen in Fig. 3.

The essence of the evolutionary optimized network is extracted and that lead to the following heuristic network design. The line network with a single clique has N_M nodes as the dense core and N_{Line} nodes in the line network. There are L_M links in the clique and L_{Line} links in the line network. If the total resource of building network is limited by n nodes and L links, the following relationships between variants are obtained as,

$$\begin{aligned} n &= N_M + N_{Line}, \\ L &= L_M + L_{Line}, \end{aligned}$$

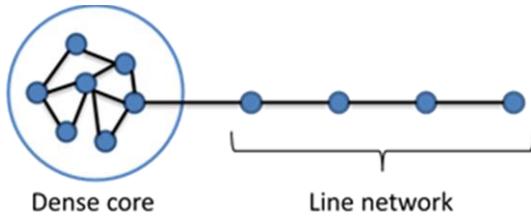


Fig.3. The interconnected network with a clique and a line network

$$L_{Line} = N_{Line}. \quad (12)$$

When the length of line network with N_{Line} nodes is given in addition to total nodes and total links (n and L), the N_{Line} should meet the following condition to avoid disconnected network and double links between same two nodes.

$$n - N_{Line} - 1 \leq L - N_{Line} \leq_{N-N_{Line}} C_2. \quad (13)$$

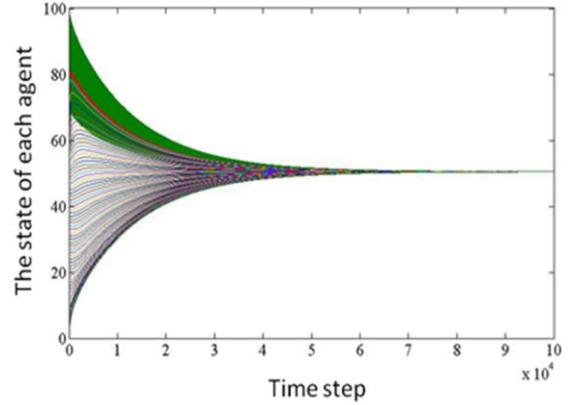
Now the consensus formation on this optimal network for the slowest consensus formation is evaluated. The internal state value of each agent is plotted over time in Fig. 4(a). Consensus formation is very slow and it takes steps 100 times more than a ring-tree network in Fig. 1(b). The changes of the agent states at the beginning are shown in Fig. 4(b). This figure shows that consensus is achieved among agents in the clique first, then among agents on the line, and globally in the end.

V. CONSENSUS FORMATION ON A SYMMETRIC DUMBBELL NETWORK

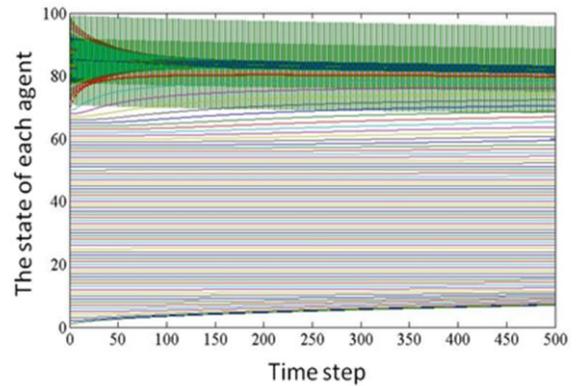
Now consider consensus formation in a network with community structures. As a typical example, a “dumbbell (or barbell) graph” with two cliques of identical network structure connected by a single link is used. Since the cliques used here are complete graphs, a Ramanujan [19], any node can be chosen to connect the two cliques. The bridging nodes on both cliques are strategically important since they are the only nodes that connect the two cliques. Additionally, these bridging nodes act as a bottleneck [20] since all nodes in the clique are indirectly influenced by the opposite clique only through their bridging node. Such model may be considered to represent two business teams working together for a common goal.

A graphical image of a dumbbell graph is shown in Fig. 5. Each clique has 50 nodes and 1,225 links therefore 100 nodes and 2,451 links in total. The consensus performance is shown in Fig. 6. The initial values assigned to one clique is ≤ 50 and to the other clique is $50 <$. The progression of consensus in Fig. 6 shows that consensus is formed within the clique before a global consensus is achieved through the bridging nodes and the bridging link. The time it took for consensus is 9,961 steps.

Since the graph used in each clique is a complete graph where all agents are connected to the rest of the agents, the



(a)



(b)

Fig.4. The diagram of consensus dynamics between agents on a clique-with-line network with $N = 100$ nodes and $L = 500$ links. The number of agents in a single core is 34. In this case the number i is assigned to agent i initially. The plot range of time step is $[0, 10^5]$ for (a) and $[0, 500]$ for (b).

consensus formation is instant if the cliques were on their own [21]. Although the fastest consensus is achieved in both graphs if they are separated, the consensus becomes very slow if they are connected only by a few links. Therefore the consensus



Fig.5. A diagram of a symmetric dumbbell graph. Each clique is composed of 50 agents. The graph contains a total of 100 nodes and 2,451 links. Since each clique is a complete graph, a Ramanujan, the bridging node may be chosen arbitrarily.

dynamics is an example to illustrate that the network dynamics will be affected by the local topological properties as well as the global properties of the network.

VI. CONCLUSION AND EXTENDED WORKS

Consensus and synchronization problems are associated with various issues, therefore, the optimal networks proposed prove themselves as one of the best network structures in terms of significance and effectiveness. Designing desirable networks is complex and it may pose a multi-constraint and multi-criterion optimization problem.

This paper presented a genetic optimization approach to design an optimal network for consensus formation and synchronization while simultaneously minimizing the eigenvalue ratio and the link density. The convergence speed of evolutionary optimized networks both for fastest convergence and slowest convergence for $\langle k \rangle = 2$ was investigated. Additionally, the priority of consensus formation within the network with community structures was observed.

Evolutionary optimized networks for faster consensus is made of a ring and many tree structures that are connected to various nodes on the ring. Although Ramanujan networks are known to have fast consensus characteristics, the optimized network is neither a Ramanujan graph nor random regular networks in which all nodes have the same degrees. Most of the nodes have the same degrees but some nodes have more links and other nodes have less links than the majority of nodes. For slow consensus, a core-with-line network was obtained where one end of a string of nodes is a dense clique with many nodes. Finally, two clique graphs were connected by a single link and its consensus property was observed. The consensus dynamics of the above networks show that consensus is formed among agents with denser links first, then a global consensus is formed. This demonstrates that both local and global topological properties contribute to the overall behavior of the network dynamics.

As for future development of this research based on the above results obtained, solving a network design problem of additional complexity by including larger number of objectives and constraints is an area of further investigation. The constraints may differ for networks in a geographic setting in addition to a topological layout, as in the case with real networks such as the internet and the brain cell network. For additional constraint, not only the wiring cost but also the maintenance of links in real distances may be used [22].

Many networks occurring in real life have modular structures that are arranged in a hierarchical structure. A more complex consensus model can also be considered where each module can represent a community of various interests [23] as the dumbbell network, or even a module as a single agent with distinct parameters represented by each node, with thresholds for each node in order to gain consensus within the module of nodes. Such a setting may be used to model an individual's psychology and their motivational level especially when there are many factors to be weighed in a decision making process.

Still further ahead, how to design ultra-large networks with the optimal principle will be investigated. Optimal networks (network modules) are achieved by treating each module as a node that has the same degree distribution as the rest of the optimal modules. The connecting nodes of two modules are selected stochastically in proportion to the node positions in

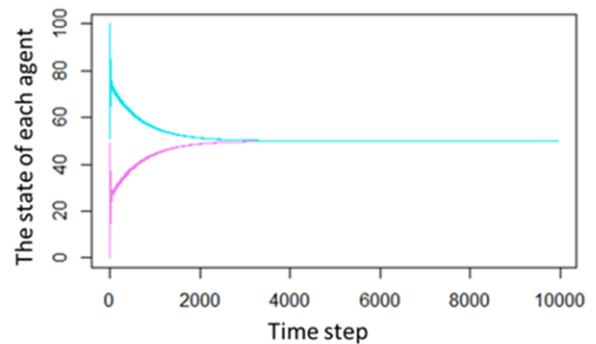


Fig.6. A consensus progression diagram for the dumbbell graph in Fig. 5. The initial values assigned to agents in each clique are first half of 100 to one clique and the second half of 100 to the other clique.

each module. The benefit of such modularization of a network is that the sparse connectivity between modules can prevent the contagion of risk spreading [24]. Observation of consensus dynamics and speed can reflect the robustness of large modularized optimal networks.

REFERENCES

- [1] J. Surowiecki, *The Wisdom of Crowds*, New York City, NY: Random House, 2004.
- [2] J. Lorenza, H. Rauhuth, F. Schweitzer, and D. Helbing, "How social influence can undermine the wisdom of crowd effect," *PNAS*, vol. 108, no. 22, pp.106–112, 2011.
- [3] A. Pikovsky, M. Rosenblum, and J. Kurths, *Synchronization: A Universal Concept in Nonlinear Sciences*, Cambridge, Cambridge University presses, 2003.
- [4] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems", *Proceedings of the IEEE*, vol. 95, no. 1, pp.215–233, 2007.
- [5] A. Adamatzky, *Dynamics of Crowd-Minds*, Singapore, World Scientific, 2005.
- [6] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, pp. 440–442, 1998.
- [7] A. L. Barabási, Z. Dezso, E. Ravasz, S. H. Yook, and S. Oltvai, "Scale-free and hierarchical structures in complex networks," *Modeling of Complex Systems: Seventh Granada Lectures*, vol. 661, pp.1–16, 2003.
- [8] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp.268–276, March 2001.
- [9] A. L. Barabási, *Linked: The New Science of Networks*, New York city, NY: Perseus Publishing, 2002.
- [10] L. Donetti, P. I. Hurtado, and M.A. Munoz, "Entangled networks, synchronization and optimal network topology," *Physical Review Letters*, vol. 95, 188701, 2005.
- [11] I. Belykh, M. Hasler, M. Lauret, and H. Nijmeijer, "Synchronization and graph topology," *International Journal of Bifurcation and Chaos*, vol. 15, no. 11, pp. 3423–3433, 2005.
- [12] H. Sato, O. Isao, and S. Kobayashi, "A new generation alternation model of genetic algorithms and its assessment," *Journal of Japanese Society for Artificial Intelligence*, vol. 12, no. 5, pp. 734–744, 1997.
- [13] B. Golub and M. Jackson, "Naïve learning in social networks and the wisdom of crowds," *American Economic Journal: Microeconomics*, vol. 2, pp.112–149, 2010.
- [14] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, iss. 345, pp. 118–121, 1974.
- [15] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems & Control Letters*, vol. 53, no. 1, pp. 65–78, 2004.
- [16] A. Lubotzky, R. Phillips, and P. Sarnak, "Ramanujan graphs," *Combinatorica*, vol. 8, no. 3, pp. 261–277, 1988.

- [17] S. Kar and J. M. F. Moura, "Topology for global average consensus," *Signals, Systems and Computers*, pp. 276-280, 2006.
- [18] T. Komatsu and A. Namatame, "Distributed Consensus and Mitigating Risk Propagation in Evolutionary Optimized Networks," J.-H. Kim, et al. (Eds.): *AsiaSim2011, PICT 4*, Springer, pp. 200–209, 2012.
- [19] M. R. Murty, "Ramanujan graphs," *J. Ramanujan Math. Soc.*, vol. 18, pp. 1-20, 2003.
- [20] S. Sreenivasan, R. Cohen, E. Lopez, Z. Toroczkai, and H. E. Stanley, "Structural bottlenecks for communication in networks," *Physical Review E*, vol. 75, 036105, 2005.
- [21] F. Fagnani, "Consensus dynamics over networks," 2014.
- [22] R. K. Pan and S. Sinha, "Modular networks with hierarchical organization: The dynamical implications of complex structure," *Pramana Journal of Physics*, vol. 71, no. 2, pp. 331–340, 2008.
- [23] J. Gao, S. Buldyrev, E. Stanley, and S. Havlin, "Networks formed from interdependent networks," *Nature Physics*, vol. 8, pp. 40–48, 2012.
- [24] T. Komatsu and A. Namatame, "An evolutionary optimal network design to mitigate risk contagion," in *Proc. The 7th International Conference on Natural Computation*, IEEE, pp. 1980-1985, 2011.

Managing Open Educational Resources on the Web of Data

Presenting COMETE, an RDF-based Resource Manager

Gilbert Paquette

CICE Research Chair, LICEF Research Center
Télé-université
Montreal, Canada

Alexis Miara

CICE Research Chair, LICEF Research Center
Télé-université
Montreal, Canada

Abstract—In the last few years, the international work on Massive Open On-line Courses (MOOCs) underlined new needs for open educational resources (OER) management within the context of the Web of Data. First, within MOOCs, all (or at least most) resources must be open and available on the Web through URIs, including the MOOCs themselves. Second the evolution of research and practice in the field of OER repositories, notably the focus in international e-learning standards, is moving recently from OER metadata stored in relational databases towards RDF-based descriptions of resources stored in triple stores. Third, new resource management tools like COMETE provide more intelligent search capabilities within the Web of data, both for designers who are building MOOCs, and also for students who should be equipped with friendly tools to personalize their environment. We will present some COMETE use cases to illustrate these new possibilities and advocate for their integration within MOOC platforms.

Keywords—resource management; resource repositories; OER; open resources; MOOCs; RDF; Web of data; Semantic Web; IEEE-LOM; ISO-MLR

I. INTRODUCTION

This paper results from a decade of research started in 2002-2004, within the eduSource canadian project, where a resource manager called PALOMA [1] was first produced and improved in the following years. Later on, the main research moved within the Learning Object Repository Network (LORNET), a large pan-canadian 5-year effort directed by the first author. This effort resulted in the production of TELOS [2], a design workbench for learning scenario or knowledge management workflows based on a new manager describing resources in the terms of an OWL technical ontology. In the last three years, our research on educational resource management moved to the use of semantic technologies for the Web of data [3,4,5], resulting in a mature tool, COMETE, that is being used in the colleges of Quebec for educational resource referencing and search.

The present paper summarizes this recent effort for the first time, describing the COMETE system and linking it to the ISO-MLR standard [6], proposing its use as a tool for the design and use of massive open on-line courses (MOOCs) [7,8].

The paper is organized into four more sections. In section 2, we introduce the notion of an educational resource repository and of a resource manager. Section 3 presents a

recent evolution of e-learning standards, norms and application profiles resulting in the publication of the ISO standard on Metadata for Learning Resources (ISO-MLR), which is based on the Resource Description Framework (RDF) enabling the use of the Web of data. In the fourth section we provide an overview and some insights on COMETE, our RDF-based OER manager. In the last section, we use COMETE to illustrate its use both for MOOC design by professors or designers, and for MOOC personalization by students and tutors.

II. OPEN EDUCATIONAL RESOURCES REPOSITORIES

We provide here some background information on open educational resources repositories and resource management.

A. Open educational resources repositories

The term “Open educational resources” was first coined at UNESCO’s 2002 Forum on Open Courseware and defined as “teaching, learning and research materials in any medium, digital or otherwise, that reside in the public domain or have been released under an open license that permits no-cost access, use, adaptation and redistribution by others with no or limited restrictions. Open licensing is built within the existing framework of intellectual property rights as defined by relevant international conventions and respects the authorship of the work.”[9]

Ten years after, UNESCO held in Paris an international OER congress on 20-22 June 2012 where the so-called Paris OER Declaration was issued. This declaration recommends that States, within their capacities and authority “foster awareness and use of OER” in many ways including “encourage research on OER”, “promote the understanding and use of open licensing frameworks”, and “facilitate finding, retrieving and sharing of OER.”

This last recommendation refers to the important on-going work in the last ten years on so-called “learning objects repositories (LOR)”. Many definitions have been given for “learning objects (LO)”. It is a more general concept than OER, since not all LOs are open, some being copyright protected. But still, LO repositories use the same methods and technologies than OER repositories.

B. First interoperability norms: Dublin Core and IEEE-LOM

The idea that educational contents could be seen as “objects” to be reused in multiple contexts dates back to the

late 60's but it started to become a reality only by the middle of the 90s with the generalization of the Internet [10].

In 1995, an international consensus arose around the necessity of e-learning standards to promote tools' interoperability and learning objects reusability. The aim was to insure the reuse of educational objects jeopardized by the diversity of referencing metadata schema around the world. This goal was shortly concretized by the Dublin Core (DC) metadata initiative [11] proposing a first set of standardized metadata, expressed in XML. Since then, the Dublin Core metadata schema has become one of the most used vocabularies on the Web of Data.

In 1996, the IEEE created the Learning Technology Standards Committee to integrate previous work on the concept of Learning Object Metadata (LOM) [12]. In June 2002, on the basis of a joint IMS-ARIADNE proposal, IEEE approved a LOM standard that was largely accepted internationally. From then on, major resource repository initiatives bloomed rapidly: ARIADNE in Europe, MERLOT in the USA, EdNA in Australia.. These and many other organizations, including our own LORNET, joined the GLOBE [13] consortium that operates actually a large repository of nearly one million resources.

A metadata record is a standardized set of properties of a learning object that makes its retrieval possible throughout the world using computer software as if the resources were in a unique reservoir, whatever their actual location.

A resource manager is a piece of software that provides the essential functionalities to make it work. Fig. 1 presents a view of the PALOMA resource manager [1]. Most LOM-based managers like this one have the following components.

- The *Harvester* or *Metadata Repository Builder* help find the location of interesting LOs on the Web or on a Local area network and creates a LOM record for each LO or resource. To make a resource more widely available, this component will sometimes transport it on some predefined server location. Harvesting is now the main method to add metadata records in a repository. It rests on the OAI-PMH protocol [14], based on a client-server architecture, in which "harvesters" request information on updated records from repositories. In this way metadata records already present in some repositories can be selected and grouped in larger or more specialized repositories.
- The *Metadata Editor* provides forms to enter all the metadata for any resource and stores the metadata

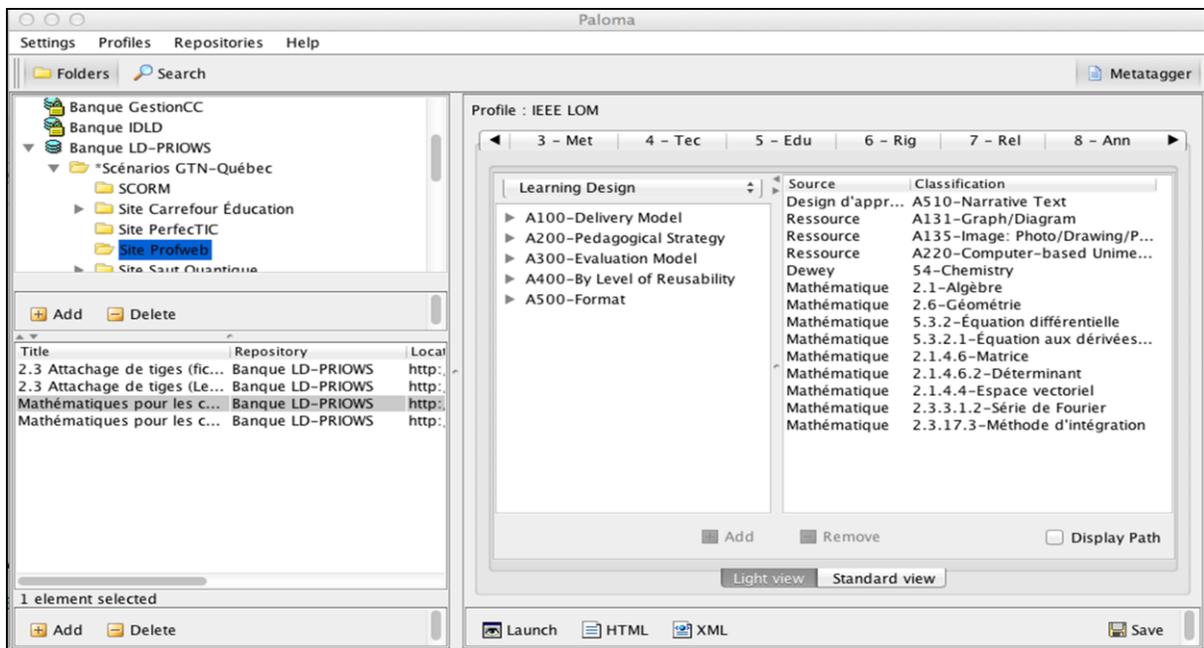


Fig. 1. Example of an early OER manager: PALOMA [1]

C. Resources managers based on Learning Object Metadata

The learning object (or OER) repositories are basically computer servers, databases and management software operating on the Web that can deliver the learning objects to any computer connected to the Internet. The learning objects are described by metadata stored in databases. In most configurations, the metadata is separated from the resources that can reside on one or more servers connected to the Web.

record in a permanent relational/XML database. Fig.1 shows on the right part some metadata entered for the selected resource within the selected repository/folder on the left. The Metadata shown here is selected in section 9 of LOM standard where terms in classifications can be chosen.

- The *Repository Search Agents* apply user-defined constraints involving metadata properties to find and display a set of the corresponding metadata records for a user to select and view the metadata and the resource.

- The *Access Manager* helps define a user's profile and its access rights to folders, metadata records and resources. User rights may include viewing, adding, modifying, deleting and assigning rights to other users.
- The *Repository Structure Manager* is a commodity for grouping resources' metadata records into folders. It can move a metadata record from one folder to the other, copy an alias in another folder, suppress a record or a folder, and duplicate a record to speed up the metatagging of a similar resource.
- The *Collaborative Annotator* provides message editing by users about a resource. It also provide ratings fonctionnalités that can be consulted by other users, sometimes offering resource display according to these ratings.

D. Potential and Limits of DC/LOM Resource Repositories

There is an large interest around the world for learning object or resource repositories as exemplified by the number of existing repositories, of organizations building and sustaining them, of contributors integrating learning objects in repositories and of the users of these learning objects. Most international conferences and journals on technology-based learning include scientific communications, some journals being specifically devoted to the subject¹.

The fundamental reasons for this interest are the growing educational demands in all countries, the limited capacity of face to face education to fulfill the demand in a timely manner, the important effort and cost involved to build online multimedia learning materials and the new possibilities offered by the Internet.

While it is a fact that millions of documents can be found on the Internet using search engines like Google, there is no guarantee that a query will lead to trustable material on which high quality education can be built. Learning object repositories offer a solution to this problem.

- First, resources repositories are maintained by educational institutions and professors that put their expertise and credibility in the balance, providing a certain trust in the quality of the referenced resources.
- Second, these resources are often peer-reviewed to ensure their quality; comments on the documents are made on the repository website to identify their actual use and their reusability capacity in different areas.
- Third, the metadata associated to the learning objects give precious information to the users, such as the name and location of the authors, the type of learning or teaching resource, the knowledge contained in it, the educational use that can be made of it, the languages in which it is delivered and the technical requirements for its proper use.
- Fourth, this metadata serves to make focused queries according to a user needs based on the properties of the

resource, not only on vague keywords that leads to thousands of references that one needs to read to understand what kind of content they provide.

- Finally, the vast majority of these learning objects are in the public domain. They are OER to be reused free of charge. The resources can be adapted or aggregated, and referenced back in a repository to extend the availability of good learning material.

After a decade of research and practice in this field, although they provide a solution to cope with the growing educational needs of the knowledge society, there are still a number of limitations to a larger use of OER repositories.

- *Cultural issues, author recognition.* Many authors will keep for themselves and their students the resources they build for their courses. While research papers are easily shared, educational resources are seen by many as private property that should be protected by copyright. Some repository like Merlot provide various kind of recognition and rewards for authors who share openly, but these methods not as widespread as in the case of research papers.
- *Rigid, closed institutional systems.* Many educational institutions keep their resources in house, integrated closely within online course stored in a learning content management system (LCMS) or online course platform, thus preventing a larger use of the ressources.
- *Slow standard adoption.* Without the adoption of an international standard, resources described by proprietary metadata remain local, unshared from one institution to another.
- *Multiplicity of IP holders.* Another severe drawback is the slow adoption of open licences like Creative Commons or GPL/LGPL, blocking the reusability of learning ressources, moreover if the resource is complex and subject to multiple intellectual property (IP) rights.
- *Heavy referencing process.* Even when the LOM standard is adopted and resources are submitted to open licences, the referencing process is complex. The LOM has 9 sections and 86 possible metadata entries so most institution will reduce the metadata set to a LOM profile covering only part of the standard and they will add specific vocabularies, thus facilitating the referencing process, but complicating search operation of resources coming from multiple repositories.

III. ISO-MLR AND THE WEB OF DATA

Many of these limitations will be overcome through more information to institutions and authors on standards and open licences. Still others require new approaches such as the Web of linked data.

A. ISO-MLR: an OER referencing standard based on RDF

Although the Dublin Cores and the IEEE-LOM are widely used to describe learning resources, interoperability among

¹ Please consult the IJKLO journal at www.ijklo.org and its successor at <http://www.informingscience.us/icarus/journals/ijello>

metadata sets from multiple repositories is still challenging, as best practices are only recommended.

For example, instead of using ISO 8601, a DC Date element can be written in plain language making impossible its processing by queries. Ambiguous definitions pose another challenge. For example a Date element can represent a resource creation time, a time of update or a time of publication. As mentioned above, LOM records can be based on a wide variety of Application profiles each defined in their own way by various agencies.

The ISO/IEC 19788 standard [7], in short ISO-MLR, is intended to provide optimal compatibility with both DC and the LOM. It presents the following advantages.

- Insuring the coherence and the non-duplication of concepts by proposing an RDF-based data model.
- Preventing the proliferation of non interoperable application profiles.
- Supporting the extension of description vocabularies in precise ways while preserving interoperability.
- Supporting multilingual and cultural adaptability requirements from a global perspective.
- Integrating resource referencing and search with other data sets in the Web of linked data.

The graph in Fig.2 shows part of the ISO-MLR RDF model. The ovals represent classes of resources, the rectangles are value types, properties are written on the links. This graph summarizes the RDF triples in the section 5 of the standard. Here are some of the triples present on Fig.2:

- (Learning resource, has learning activity, Learning activity)
- (Learning activity, learning method, *method value*)
- (Learning resource, has contribution, Contribution)
- (Contribution, has contributor, Person)
- (Learning resource, has annotation, Annotation)
- (Annotation, annotation date, *date value*)
- (Annotation, annotator, Person)

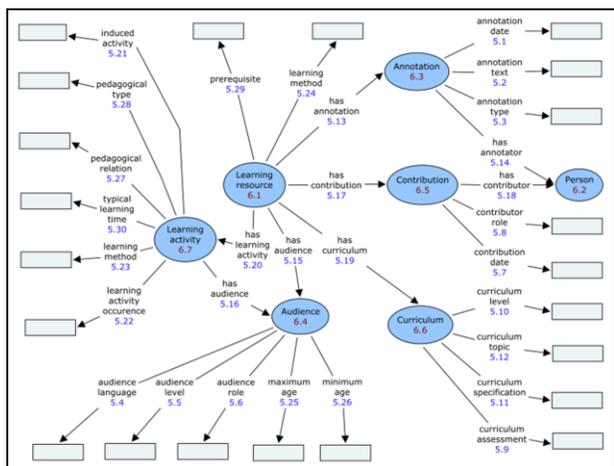


Fig. 2. Part of the ISO-MLR RDF model

B. A standard for the Web of data

The fundamental thing here is that ISO-MLR proceeds from a different vision than previous standards like the IEEE-LOM, where resources are seen only as documents. ISO-MLR, using technologies like RDF and RDF schema, integrates well to a Web of linked data, instead of simply a Web of documents.

The origin of the Web of data, also termed “Semantic Web”, dates back to 2001 when the actual director and founder of the Web, Tim Berners-Lee and his colleagues [15] proposed to integrate to Web pages information on the knowledge (concepts, properties) present in Web documents. The URLs who provide locations on the Web were to be generalized to URIs that could represent people, real-world objects, but also abstract concept and properties. These entities and the values of their properties would be linked together by declaring RDF triples.

It then becomes possible to describe the meaning, the semantic of Web pages beyond the syntax of natural languages and their inherent ambiguity. A Web of linked data enables computer agents to follow the links and perform more intelligent operations using the knowledge behind the words.

For this, the SPARQL Protocol and RDF Query Language [16] enables queries within the huge graph of RDF triples that constitutes the Web of linked data. Fig. 3 shows the state of this graph at the end of 2011 that grouped over 200 datasets, 26 billions RDF triples and 400,000 property links.

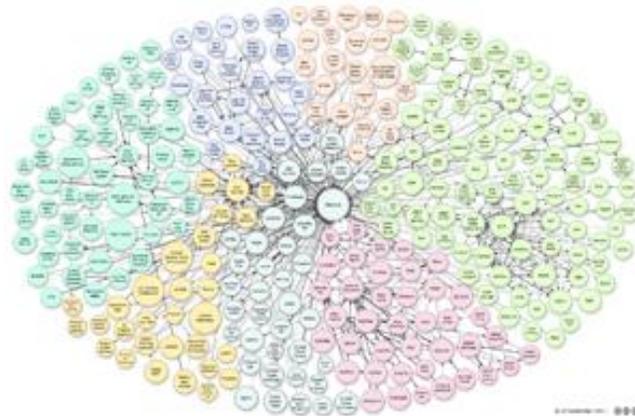


Fig. 3. The Linking Open Data (LOD) cloud diagram in 2011 [17]

Each node on the figure represents a data set. For example, the Dbpedia node at the center of the figure groups most of the information in Wikipédia, while the FOAF dataset groups information about persons having a URI on the Web. The links between two nodes means that the terms of the vocabulary in one node are linked with terms in the other node. For example, “persons” in DBpedia is related to “persons” in FOAF and their geographical localization can be found in another vocabulary such as GEONAMES.

In the same way, terms in ISO-MLR are linked to terms of other vocabularies on the Web of data.

For example, iso-mlr5:Person in the graph of Fig.2 has the same meaning as foaf:person or dcterms:person. This means

that a computer agent that would search for an iso-mlr:learning_resource can also retrieve its iso-mlr:Contributors, find these persons and retrieve their Wikipedia pages from Dbpedia, their email from FOAF and their localization from GEOBASE.

IV. COMETE, A RDF-BASED RESOURCE MANAGER

COMETE is a learning resource repository manager based on the RDF approach. It allows locating, aggregating and retrieving educational resources that constitute the heritage of an organization. Basically, it is a database containing metadata about learning resources on which users can perform queries to find and discover educational material that they can reuse for their various needs.

Fig.4 describes the technical architecture of a COMETE implantation instance. It's a 3-tiers client-server architecture developed in Java technology.

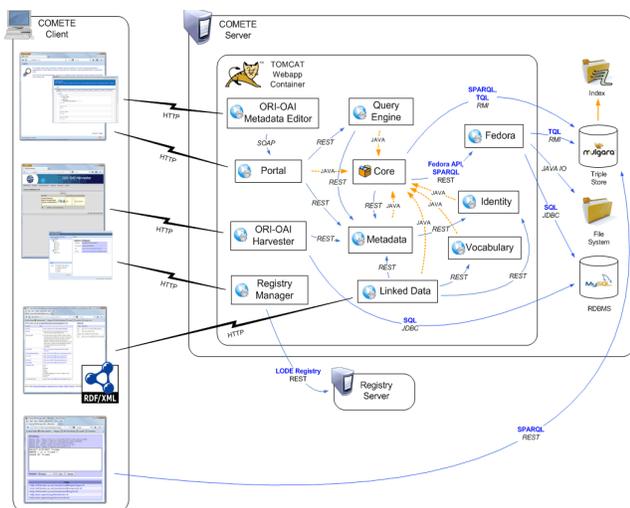


Fig. 4. COMETE architecture schema

Various web applications powered by an Apache Tomcat server provide specialized REST services that allow different types of clients to exploit the open data contained in the repository. Most of the clients use their favourite web browsers to access the system through a user-friendly web interface. A SPARQL endpoint is also available for advanced users who want to directly access the raw RDF triples to built various applications or Web services.

A. Integrating new resources in the triple store

The integration of resources inside a COMETE repository is done by imports of their metadata records. There are different ways to achieve this task. The metadata records can be imported manually by uploading an archive file containing a collection of metadata records. Most of the time, however, the metadata records will be harvested automatically by either an OAI-PMH Harvester or a HTML Spider. In such a case, a Harvest Definition will declare the technical information required to access the repository to be harvested. Using the facilities provided by the operating system on which COMETE runs, it's also possible to program harvest schedules so that the

process is executed periodically to make sure that new or updated metadata records are always imported to the system.

These records are ingested by the system and a XSL transformation extract data for generating all pertinent triples. COMETE enable data mining across multiple metadata schemas like Dublin Core, IEEE LOM and other application profiles. Fig. 5 illustrates the result this process, that is a homogeneous graph of data in accordance with COMETE's internal metamodel (Fig. 6).

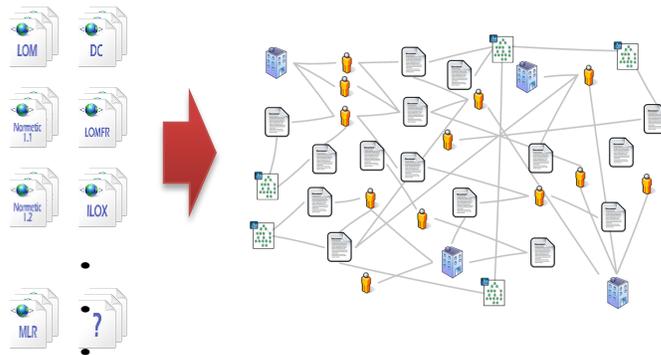


Fig. 5. Transformation from metadata to RDFgraph

All the generated triples are stored into Mulgara [18], an open source RDF triple store system, where data is organized around different RDF graphs. The default graph contains all the triples about learning resources whereas some other specialized graphs manage SKOS thesaurus and other different views of the system.

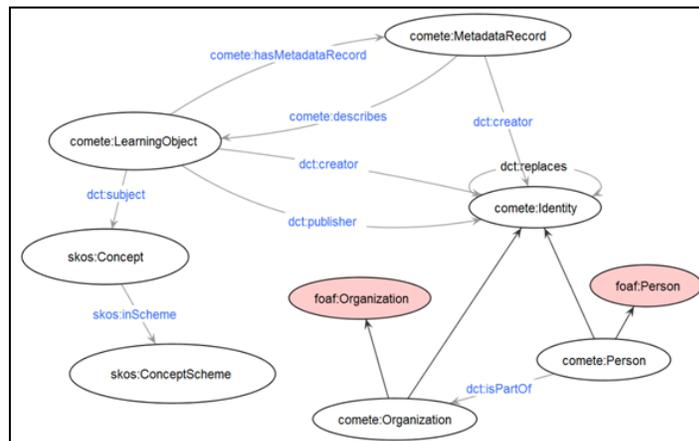


Fig. 6. COMETE metamodel (main classes)

As a semantic network, the RDF graph represents the model entities as nodes. Mains nodes are learning resources (LearningObject), persons and organizations (Identity) and element of vocabulary (SKOS Concept). By various techniques, the system tries to maximize the inner coherence of the graph.

The Identity module (on Fig. 4) implements the management of identities (metadata about persons or organizations). This includes importation of identities, identity resolution (identifying any identity in the database that represents the same person or organization, making sure it

stays unique, and completing it as new details are known), and representations of the identity (such as VCards and HTML code for the end users). Furthermore, manual merge of identities is also provided within a set of administrative tools for a better control of data integrity.

The Vocabulary module (on Fig. 4) implements the management of vocabularies and thesauri, which involve importing from VDEX or SKOS formats, unambiguously identifying the vocabulary that a term is from, and finding a computer readable representation of the whole vocabulary, updating from source automatically, transparently converting from one format to another, replacing a vocabulary when updates are available, publishing vocabularies automatically and providing user interface elements reusable by other modules, such as efficient vocabulary term choosers for queries to the repository.

This module manages also correspondences between taxonomies. Indeed, SKOS concept alignment between different ontologies (or vocabularies) can be taken into account by the query engine. A useful example of alignment is the mapping between different school-level taxonomies of different countries to promote the interoperability of resources between national repositories. For instance we can search resources which target audience is Junior High School in United States and the results may contain pertinent Secondary School I-III tutorials produced in Québec. We can imagine here a wide range of possibilities.

B. Querying the triple store

All of the previously presented modules provide rich graphs of data that allow doing more sophisticated searches in the repository. All nodes have many textual information (literal triples) where values are indexed and which can be used by fulltext search methods.

Nodes are also linked together and their graph can be traversed to perform more “intelligent” searches. Furthermore, all of the elements in the model are referenced with a unique identifier (URI).

The next figure represents the simplest COMETE search mode. As in a Google search, it only needs a field of keywords.

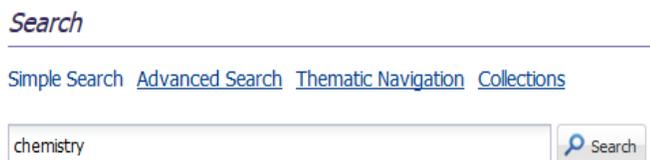


Fig. 7. COMETE Simple Search

Suppose now we seek the resources of an author X dealing with Organic Chemistry from Dewey classification. COMETE web interface offers different kinds of search interfaces, including advanced search. Fig 8 illustrates the easy way to fill the previous query.

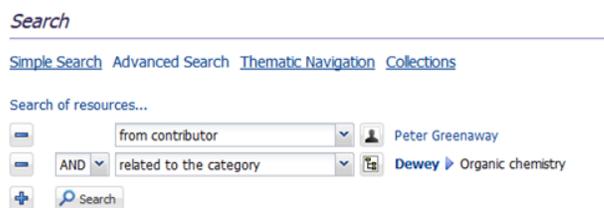


Fig. 8. COMETE Advanced Search

All of the queries will be translated in SPARQL language by the QueryEngine module of Fig.4 and then be run on the triple store. By combining different conditions, mixing keyword-based approach and by using negative prefixes, more complex queries can be performed.

A third way to navigate inside resources is to use the Thematic Navigation (Linked data on Fig.4) module that lets user discover resources from available thesaurus.

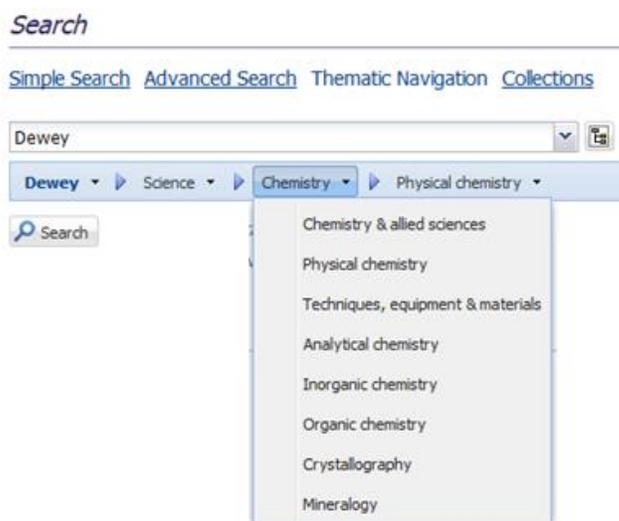


Fig. 9. COMETE Thematic Navigation

Results will be returned and displayed to the user interface. In connection with alignment of vocabularies seen previously, queries may be extended with option like: “include equivalent categories from Library of Congress”. Obviously, resources linked with that classification must be present to obtain more results.

Finally, a fourth search mode in COMETE user interface is the Collection mode. It offers to users a list of preset complex queries to avoid having to deal with. Example: the “Last month Algebra and Euclidean geometry resources from Montréal University” set.

Like we said previously, there is also a SPARQL endpoint for querying the triple store. To achieve this task, a Snorql module is deployed. It offers a user web interface with a text area field to enter and run SPARQL queries. Although it requires some technical knowledge, it’s still a simple way to explore data.

C. Linking with the Web of data

The link with the web of data, as a global data space, is done by the following facts:

COMETE respects the basic principles of Linked Data:

- ❖ All objects are described with HTTP URIs.
- ❖ URIs can be dereferenced over the HTTP protocol into a description of the identified object or concept. Moreover, the content negotiation mechanism (or 303 redirect strategy) makes possible to serve different versions depending on the context of the client; HTML page for web browser clients, RDF/XML for software agent.
- ❖ Links are included to other URIs as soon as possible to promote future discovering of resources.
- ❖ A unifying RDF data model provides a globally unique identification of entities and allows different schema to be used (and reused) in parallel to represent data.

This last bullet point is supported by a vocabulary reference [19] that details the COMETE meta-model in terms of class and property definitions and the reuse of existing vocabularies (Dublin Core, FOAF and SKOS).

The publishing of data via a SPARQL endpoint allows interaction with COMETE data by external systems.

V. USING COMETE WITHIN A MOOC PLATFORM

In this section, we present two use cases where COMETE is used to interoperate with a MOOC platform like OpenEdX. Within such a platform, the role of COMETE is twofold: 1) enable designers to search and reference OERs within a MOOC; 2) reference MOOC themselves to produce a searchable standardized MOOC portal.

A. The OpenEdX MOOC platform

OpenEdX is the open-source release of the edX platform developed by a non-profit organization founded by Harvard and MIT in the USA. In April 2013 Stanford and edX agreed to collaborate on future developments of the edX platform. In September 2013, Google committed to the development of OpenEdX. In France, France Université Numérique (FUN) uses a version of OpenEdX. Amongst many other institutions, Télé-université du Québec has also adopted OpenEdX for its MOOC initiative.

OpenEdX provides essentially two server-based applications. The first one, edX-STUDIO, is the application where designers build courses. Resources and activities are grouped in course modules and stored in Mongo no-sql XML files and in MySQL databases. Students interact at runtime with a second application, the Learning Management System (LMS) that performs learner authentication, runtime learning scenario support, forums and online group meetings, automatic and peer-assessed grading of learners and learning analytics operations [20].

B. Designing a MOOC using the COMETE OER Manager

Figure 10 present an example of a course structure. The course is subdivided in sections (e.g. modules) and each module in sub-sections (e.g. lessons). For each lesson, the upper bar provides access to the lessons' sequential components. It is can be composed of four kind of components: discussion components, HTML components, problem components, and video components. In principle, all these components should be open educational resources (OER).

All these OER components re found mainly on the Web. Actually. Designers use search engines like Google or Bing to find open resources to reuse or adapt for their course. As explained previously in section II-D, there are many adavantages in using a learning resource repository manager like COMETE to find suitable ressources.

Using REST web services, a call to COMETE from OpenEdX studio could start efficient search operations and facilitate the selection of ressources of the four categories proposed in STUDIO. Conversely, STUDIO could be upgraded to provide forms to edit metadata for the resources in a standardized DC, LOM or ISO-MLR application profile suitable for Studio. This would enable designers to automatically create a resource repository for a course, for a whole program or for all edX users. The creation of this local repository would produce a URI where the edX ressources can be harvested by COMETE or other OER Managers and integrated into larger repositories for future use.

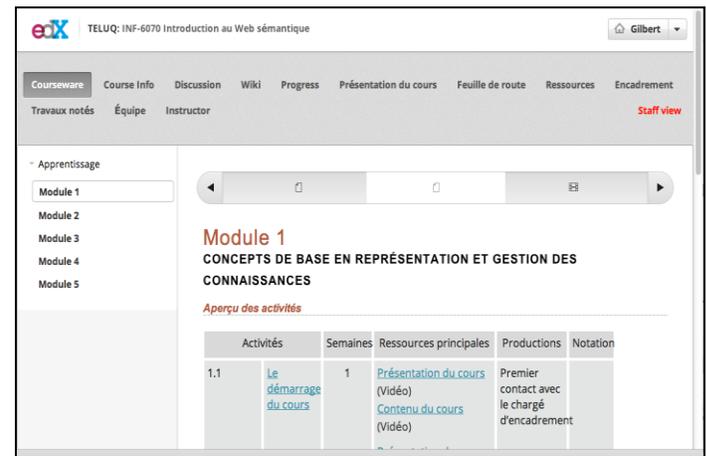


Fig. 10. A screen from OpenEdX LMS

C. Referencing a MOOC using COMETE

When a new MOOC is created in OpenEdX, a screen like the one on Fig. 11 is offered to the designer. Actually, only four metadata are asked: the course name, the organization that supports it, the course number and the periods when it will be offered.

The form shown on figure 11 could be easily extended to fields from a DC, LOM or ISO-MLR application profile that would take in account the differences between small resources within a MOOC, compared to large OERs like complete MOOCs.

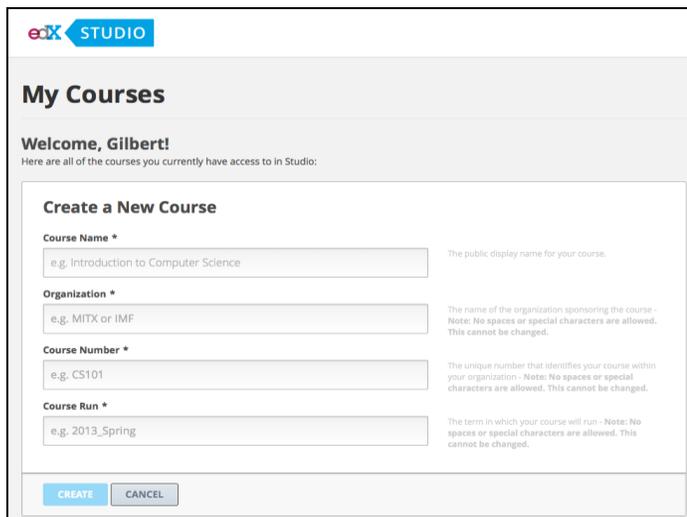


Fig. 11. Creating a new MOOC in edX STUDIO

Then, automatically, each time a new MOOC is created, it would have a URI on the Web of data together with its composing resource. COMETE could then provide a searching facility in MOOC repositories like Class Central [21], which is a free online MOOC aggregator from top universities like Stanford, MIT, Harvard, etc. offered via Coursera, Udacity, edX, NovoED, & others.

Actually, in a MOOC portal, courses are classified by subject, universities, level and provider, which are the only meta-data entries available to browse for a course. Most of the time one must open each course to know what's in it. With standardized metadata, COMETE could power a MOOC portal with various kinds of search and navigation within repositories containing hundreds of MOOC, including advanced search combining many metadata and navigating on the Web of data.

CONCLUSION

We have presented a solution to one of the main problems in Open Educational Resources repositories, which is the multiplicity of norms, standards and application profiles that preclude efficient search for resources within multiple repositories. We have built a first Linked data OER repository manager, COMETE, relying on semantic web techniques, largely complying to the new ISO-MLR encompassing and flexible standard. Also, its use for MOOC and MOOC components referencing using RDF triples will become an asset as massive online courses are growing rapidly in most countries. Our next work will be to investigate various integration of COMETE tools with MOOC platforms as indicated in the present contribution.

ACKNOWLEDGMENT

The authors wish to thank Frédéric Bergeron who has strongly contributed to the development of COMETE. Also, we must underline the implication of Marc-Antoine Parent, Benoit Grégoire, Gilles Gauthier and Pierre-Julien Guay who

have provided advice on some of the orientations of the tool. We also thank Claude Coulombe and Mohammed Ben Jemia for their implication with the OpenEdX MOOC platform.

Finally we wish to acknowledge the financial support of the Ministère de l'Enseignement Supérieur de la Recherche, de la Science et de la Technologie (MESRST), Télé-université (Centre LICEF) and the Collège Bois-de-Boulogne (Vitrine Technologie-Éducation).

REFERENCES

- [1] Paquette G, Lundgren-Cayrol K, Miara A, Guérette L (2004) The Explor@2 Learning Object Manager, in R. McGreal (ed), *Online education using learning objects*. pp 254-268. London: Routledge/Falmer.
- [2] Paquette, G. (2010). An ontology-driven System for e-learning and knowledge Management. In Paquette, G., *Visual Knowledge Modeling for Semantic Web Technologies: Models and Ontologies*. Hershey, PA: IGI Global, pp 302-324
- [3] Allemang D. and Hendler J. (2011) *Semantic Web for the Working Ontologist – Effective Modeling in RDFS and OWL*. 2nd Edition. Morgan-Kaufmann/Elsevier, Amsterdam.
- [4] Domingue, J., Fensel, D. et Hendler, J. A. (dir.). (2011). *Handbook of semantic web technologies*. Berlin, Allemagne : Springer-Verlag.
- [5] Heath, T. et Bizer, C. (2011). *Linked data: Evolving the web into a global data space*. In *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1), 1-136.
- [6] ISO-MLR (2013) ISO-IED 19788 Information technology – Learning, education and training – *Metadata for learning resources multipart standard*. http://en.wikipedia.org/wiki/ISO/IEC_19788.
- [7] Hollands, F. and Tirthali, D. (2014) *MOOCs: Expectations and Reality*. http://cbcse.org/wordpress/wp-content/uploads/2014/05/MOOCs_Expectations_and_Reality.pdf
- [8] Daniel, J. (2012) Making Sense of MOOCs: Musings in a Maze of Myth, Paradox and Possibility. *Journal of Interactive Media in Education*. <http://www-jime.open.ac.uk/article/2012-18/html>
- [9] Unesco (2012). 2012 Paris OER Declaration. http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/WPFD2009/English_Declaration.html.
- [10] Duval, E. and Robson, R. (2001) Duval, E. and Robson, R. Guest Editorial on Metadata. *Interactive Learning Environments, Special issue: Metadata*, Volume 9-3, December 2001, pp. 201-206
- [11] DC – Public Core Metadata initiative. <http://dublincore.org> .
- [12] IEEE-LOM – Learning Object Metadata, http://fr.wikipedia.org/wiki/Learning_Object_Metadata .
- [13] GLOBE – Global Learning Object Brokered Exchange, <http://globe-info.org> .
- [14] OAI-PMH – Open Archives Initiative Protocol for Metadata Harvesting. <http://en.wikipedia.org/wiki/OAI-PMH>.
- [15] Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. *Scientific American Magazine*, vol. 284, no. 5, mai 2001, pp. 29–37
- [16] SPARQL – SPARQL 1.1 Query Language, W3C Recommendation, 21 March 2013. <http://www.w3.org/TR/sparql11-query/>
- [17] LOD Cloud – The Linking Open ata cloud diagram. <http://lod-cloud.net>
- [18] Mulgara RDF Triple Store System. [http://en.wikipedia.org/wiki/Mulgara_\(software\)](http://en.wikipedia.org/wiki/Mulgara_(software))
- [19] Miara A. and Bergeron, F. *COMETE RDF Metamodel*. <http://comete.licef.ca/reference/>
- [20] Coulombe C. (2014) *Expérimentation de la plateforme OpenEdX*, rapport technique LICEF, Télé-université du Québec.
- [21] Class Central (2014) *Free Online Education Portal*, <https://www.class-central.com>

Development of Duck Diseases Expert System with Applying Alliance Method at Bali Provincial Livestock Office

Dewa Gede Hendra Divayana
Chairman of Information Technology Department
Faculty of Computer, Indonesia Technology University
Bali, Indonesia

Abstract—Farming is one of the activities that have a business opportunity. One is raising ducks. The main results can be obtained from the breeding duck is a duck meat and eggs for consumption and also means praying ceremony in Bali, as well as duck egg shells that can be used for jewelry. Since the outbreak of avian influenza began in 2008, have an impact on consumer demand of ducks decreased and consumers become more careful in choosing and consuming duck. The avian influenza virus not only spread across the country of China, Thailand and Vietnam, but also in Indonesia, Bali is no exception. This is evidenced by the discovery of cases of death due to bird flu virus in some areas in Bali, among others: the regency of Karangasem, Badung, Tabanan, Klungkung and Jembrana. From this, the Bali Provincial Livestock Office took steps to develop an expert system in the detection of diseases ducks. This expert system uses a alliance method is a combination of forward chaining, backward chaining and weighted product to search the physical symptoms and behavioral symptoms duck by the name of a known disease and to determine the percentage of disease attack level in ducks. In this study, the analytical techniques used to analyze the truth is a alliance method of duck disease expert system. Activity data collection and information to support research conducted by, among others, literature studies, interviews, and observations.

Keywords—Expert System; Forward Chaining; Backward Chaining; Weighted Product; Alliance Method; Duck Diseases

I. INTRODUCTION

Farming is one of the activities that have a business opportunity. One is raising ducks. Raising ducks is one business that can be used as a promising source of income for most people in Indonesia and Bali in particular. Besides the cheap price of the seed, maintenance of duck also not as difficult as raising pigs or cows. It is shown from the results of research conducted Bali Provincial Livestock Office that shows the average demand for duck meat increased by 30% each year and is followed by the rise of the merchant ducks in some areas in Bali.

The main results can be obtained from the breeding duck is a duck meat and eggs for consumption and also means praying ceremony in Bali, as well as duck egg shells that can be used for jewelry.

From some of the advantages and benefits gained from raising ducks, of course there are also the challenges or

obstacles faced duck breeders include overcoming disease in ducks. In fact, since the outbreak of avian influenza in the 2008-2012 year range ducks impact on demand from consumers has decreased and consumers become more careful in choosing and consuming duck.

The avian influenza virus not only spread across the country of China, Thailand and Vietnam, but also in Indonesia, Bali is no exception. This is evidenced by the discovery of cases of death due to avian influenza virus in some areas in Bali, among others: the regency of Karangasem, Badung, Tabanan, Klungkung and Jembrana.

Therefore, the community needs to know what are the types of diseases that can be contracted on a duck. With the importance of knowledge about the types of the duck disease, it is deemed necessary to provide a medium that can provide information about diseases in ducks. As for some of the media that may be obtained easily is through magazines, newspapers, television broadcasts, radio broadcasts, educational and training organized by the Bali Provincial Livestock Office, can also even through the computerized system.

The computerized system is an expert system to duck diseases detection. Duck diseases expert system have ability to ducks diseases detect and analyzed in detail.

II. LITERATURE REVIEW

A. Expert Systems

In [1], Expert Systems is a branch of Artificial Intelligence that makes extensive use of specialized knowledge to solve problems at the human expert level.

In [2], an expert system is the computer system that emulates the behaviour of human experts in a well-specified manner, and narrowly defines the domain of knowledge. It captures the knowledge and heuristics that an expert employs in a specific task. An overview of current technologies applied with an expert system that is developed for Database Management System, Decision Support System, and the other Intelligent Systems such as Neural Networks System, Genetic Algorithm, etc.

In [3], an Expert system is a software that simulates the performance of a human experts in a specific field. Today's

expert systems have been used in many areas where require decision making or predicting with expertise.

In [4], the Expert System (ES) is one of the well-known reasoning techniques that is utilized in diagnosis applications domain. In ES, human knowledge about a particular expertise to accomplish a particular task is represented as facts and rules in its knowledge base [4].

From the definitions of the above can be concluded in general that expert systems is an artificial intelligence system that combines knowledge base with inference engine so that it can adopt the ability of the experts into a computer, so the computer can solve problems such as the often performed by experts.

B. Forward Chaining

The inference engine contains the methodology used to perform reasoning on the information in the knowledge base and used to formulate conclusions. Inference engine is the part that contains the mechanism and function of thought patterns of reasoning systems that are used by an expert. The mechanism will analyze a specific problem and will seek answers, conclusions or decisions are best. Because the inference engine is the most important part of an expert system that plays a role in determining the effectiveness and efficiency of the system. There are several ways that can be done in performing inference, including the Forward Chaining. In [5], forward chaining is matching facts or statements starting from the left (first IF).

C. Backward Chaining

Also in [5], backward chaining is matching facts or statements starting from the right (first THEN). In other words, the reasoning starts from the first hypothesis, and to test the truth of this hypothesis to look for the facts that exist in the knowledge base.

D. Weighted Product

In [6], Weighted Product Method (WP) use multiplication to connect the attribute ratings, where the ratings of each attribute must be raised first with the relevant attribute weights. This process is similar to the process of normalization. Preferences for alternative A_i is given as follows:

$$S_i = \prod_{j=1}^n x_{ij}^{w_j} \quad \text{with } i = 1, 2, \dots, n \quad \text{and } w_j = 1$$

w_j is the power of positive value to attribute profits, and is negative for the cost attribute. Relative preference of each alternative, given as:

$$V_i = \frac{\prod_{j=1}^n 1x_{ij}^{w_j}}{\prod_{j=1}^n 1(x_{j^*})^{w_j}} \quad \text{with } i = 1, 2, \dots, n$$

V : Preferences alternatives considered as a vector V

x : Value of Criteria

w : Weight of Criteria / Sub-criteria

i : Alternative

j : Criteria

n : number of criteria

E. Alliance Method

In [7], stated that Alliance method is a combination of forward chaining, backward chaining and weighted product to search the name of the disease based on symptoms or vice versa as well as to determine the percentage of disease provided by the users of the system (user) and the expert.

III. METHODOLOGY

A. Object dan Research Site

1) Research Object is Expert System of Duck Diseases With Applying Alliance Method.

2) Research Site at Bali Province Livestock Department.

B. Data Type

In this research, the authors use primary data, secondary data, quantitative data and qualitative data.

C. Data Collection Techniques

In this research, the authors use data collection techniques such as observation, interviews, and documentation.

D. Analysis Techniques

Analysis techniques used in this research is descriptive statistical.

IV. RESULT AND DISCUSSION

A. Result

1) Early Trial

At this early trial, the authors conducted a limited scale testing of the duck diseases expert system that have been made previously by involving five staff of Bali Provincial Livestock Office to perform *white box* and *black box* testing. This test can be done by giving 10 questionnaires early trials duck disease expert systems to staff of Bali Provincial Livestock Office. Diagram form of answers score percentage given by the respondents in early trial can be described as follows:

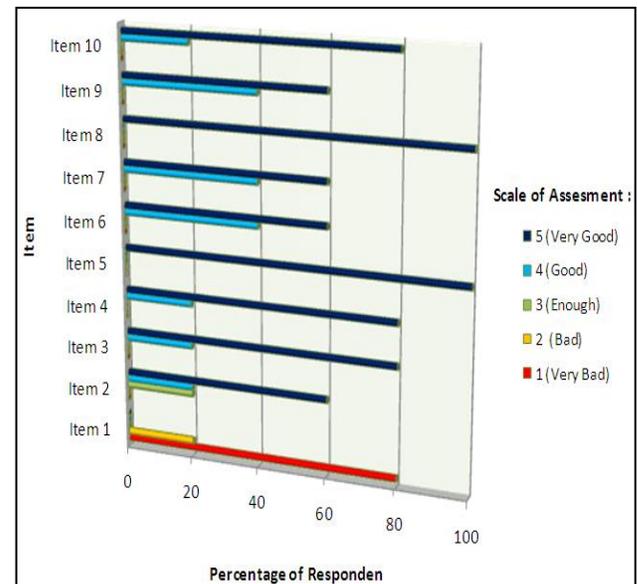


Fig. 1. Percentage Diagram of Respondents Answer Score In Early Trial

Based on the diagram above, it can be seen that the results of early trials of the duck disease expert system, find a constraint that is the answer to a very bad score by 80% of the questions on the questionnaire 1st initial trials. This is due to the unavailability of the form for the manufacture of a new username and password in the future if there is a mutation of the staff who operate the duck disease expert system. Given these constraints, then the system needs to be revised again.

2) Field Trial

At this field trial, the authors tested in a larger scale, involving an expert (vet) is understood about the duck diseases and seven staff of Bali Provincial Livestock Office. This test can be done by giving 16 questionnaires field trials duck disease expert systems to the vet and the staff of the Bali Provincial Livestock Office.

Diagram form of answers score percentage given by the respondents in field trial can be described as follows:

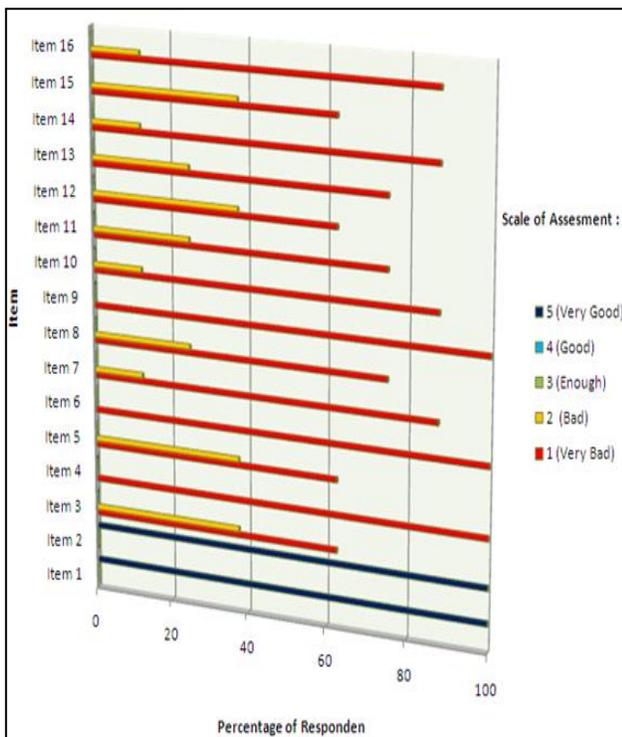


Fig. 2. Percentage Diagram of Respondents Answer Score In Field Trial

Based on the diagram above, it can be seen that the results of a field trial of the duck diseases expert system, the presence of obstacles that scores are very bad answer the score of 62.5% to the question 3rd, 5th, 12th, and 15th, 75% of the questions 8th, 11th and 13th, at 87.5% of the questions 7th, 10th, 14th, and 16th, and at 100% of the questions 4th, 6th and 9th on field trial questionnaire.

This is due to the unavailability of the form to enter or edit the physical symptoms and behavioral symptoms duck if in the future there is a new symptom on the physical and behavior of ducks, as well as the unavailability of the form to enter or edit the rule, and the weight of duck disease attack rate.

Of the constraints are found, then the system needs to be revised to obtain duck disease expert systems more interactive and dynamic.

3) Usage Test

At this usage test, the authors conducted a trial involving with the use of 20 people (breeder duck). The test is performed to test the operation of the overall form available on duck diseases expert system that has undergone revisions to field trials. This test can be done by giving the user satisfaction questionnaire to the expert system diseases duck to duck breeders who visited Bali Provincial Livestock Office.

Diagram form of answers score percentage given by the respondents in usage test can be described as follows:

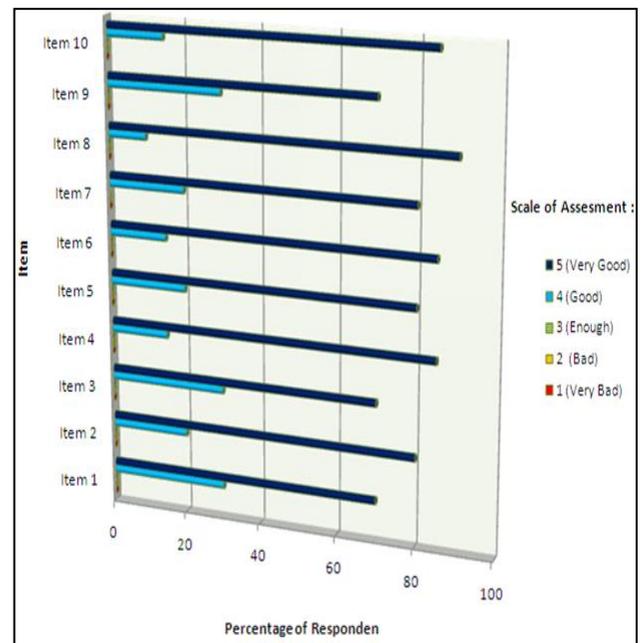


Fig. 3. Percentage Diagram of Respondents Answer Score In Usage Test

Based on the diagram above, it can be seen that the results of testing the use of the duck diseases expert system outline already looks very good and not found again the constraints in terms of technical operation (inputting and editing a new symptom on the physical and behavioral duck) as well as the principle method of expertise (alliance method). This is evidenced by the percentage scoring very good response by 70% of statements 1st, 3rd, and 9th.

Percentage scoring very good response by 80% against the statement of the 2nd, 5th, and 7th. Percentage scoring very good response by 85% of statements 4th, 6th, and 10th. As well as scoring 90% of the questions on the questionnaire 8th trial usage. And it would be even better if the duck diseases expert system added amenities help programs written in accordance with the suggestions of the respondents to the improvement of the system, so as to explain the performance of the expert system and the function of the buttons in the design of an duck diseases expert system overall with easy to understand and simple language.

B. Discussion

1) Shows alliance method has been run according to the rules

In analyzing the validity of this method alliance rule, the author will check the compatibility between the results obtained from the trial decision table rule forward chaining, backward chaining and weighted product made by respondents to the decision table rules forward chaining and backward chaining, and the results of the calculation of weighted existing product the duck diseases expert system program. As for the decision table forward chaining rules and backward chaining as well as the results of the calculation of weighted products that exist in the duck diseases expert system and table test program conducted by the respondent and can be explained as follows:

a) Forward chaining and backward chaining rules decision table of duck diseases expert system

TABLE I. FORWARD CHAINING AND BACKWARD CHAINING RULES OF DUCK DISEASES EXPERT SYSTEM

No	The Symptoms	Name of Diseases			
		Avian Influenza	Duck Cholera	Salmonellosis	Botulism
1.	Fur				
a	Moult	√	-	√	-
b	Dull	-	√	-	√
2.	Wing				
a	Hanging	√	-	√	-
b	Moult	-	√	-	√
3.	Foot				
a	Limp	√	-	-	√
b	Swelling	-	√	√	-
4.	Dirt				
a	Watery Yellow	-	-	√	√
b	Watery White	√	√	-	-
5.	Behavior				
a	Often rub the head to the ground	√	-	-	-
b	Stand with one leg	-	-	-	√
c	Often sleepy	-	√	-	-
d	Totter	-	-	√	-

b) Analyze the calculation correctness of the weighted product method in duck disease expert systems

To check the calculation correctness of this weighted product method done several things, among others:

- Determination of the weight of each physical and behavioral symptoms of ducks were observed by the user, the attack rate and weight values include: low attack rate with weight value = 0.25, enough attack with weight value = 0.50, high attack with weight value = 0.75, and very high attack with weight value = 1.
- Determination of the weight of each physical and

behavioral symptoms of ducks were observed by experts, the attack rate and weight values include: low attack rate with weight value = 0.10, enough attack with weight value = 0.20, high attack with weight value = 0.30, and very high attack with weight value = 0.40.

- Seeking percentage duck disease attack rate by multiplying the weight of all the results of powers between the physical and behavioral duck symptoms observed by user to weight the physical and duck behavioral symptoms observed by expert. And the results of these calculations multiplied by 100%.

For example:

- The weight data of every physical and behavioral symptoms of ducks were observed by the user are as follows:

TABLE II. WEIGHT DATA FROM USER OBSERVATION

Duck	Physical Symptoms	Weight	Attack Level
A	Fur	0.5	Enough
	Wing	0.75	High
	Foot	0.5	Enough
	Dirt	0.25	Low
B	Behavior	0.5	Enough
	Fur	1	Very High
	Wing	0.75	High
	Foot	0.5	Enough
C	Dirt	0.25	Low
	Behavior	0.75	High
	Fur	0.5	Enough
	Wing	1	Very High
	Foot	0.5	Enough
	Dirt	0.75	High
	Behavior	0.5	Enough

- The weight data of every physical and behavioral symptoms of ducks were observed by the expert are as follows:

TABLE III. WEIGHT DATA FROM EXPERT OBSERVATION

Duck	Physical Symptoms	Weight	Attack Level
A	Fur	0.2	Enough
	Wing	0.3	High
	Foot	0.2	Enough
	Dirt	0.1	Low
	Behavior	0.2	Enough
B	Fur	0.4	Very High
	Wing	0.3	High
	Foot	0.2	Enough
	Dirt	0.1	Low
C	Behavior	0.3	High
	Fur	0.2	Enough
	Wing	0.4	Very High
	Foot	0.2	Enough
	Dirt	0.3	High
	Behavior	0.2	Enough

- From these data it can be searched percentage of duck disease attack rate in the following way:
- S Vector to duck-A :

$$S-A = (0.5^{0.2}) * (0.75^{0.3}) * (0.5^{0.2}) * (0.25^{0.1}) * (0.5^{0.2}) = 0.526859$$

- S Vector to duck-B :

$$S-B = (1^{0.4}) * (0.75^{0.3}) * (0.5^{0.2}) * (0.25^{0.1}) * (0.75^{0.3}) = 0.637712$$

- S Vector to duck-C :

$$S-C = (0.5^{0.2}) * (1^{0.4}) * (0.5^{0.2}) * (0.75^{0.3}) * (0.5^{0.2}) = 0.605202$$

Then the S vector of the results that have been obtained above, then:

- The percentage rate of the disease in duck-A is = $0.526859 * 100\% = 52.69\%$
- The percentage rate of the disease in duck-B is = $0.637712 * 100\% = 63.77\%$
- The percentage rate of the disease in duck-C is = $0.605202 * 100\% = 60.52\%$

c) Trials alliance method performed by respondents

Respondents who did this trial was a veterinarian as experts and seven staff Bali Provincial Livestock Office conducted the field trials. The trial results are shown in the following table.

TABLE IV. TRIALS ALLIANCE METHOD

Respondent	Physical Evidence						Behavior		DS	% AL		
	Fur		Wing		Foot		Dirt					
	S	A L	S	A L	S	A L	S	A L				
RS.01	F1	E	W1	H	T1	E	D2	L	B1	E	AI	52.69
RS.02	F2	V	W2	H	T2	E	D2	L	B3	H	DC	63.77
RS.03	F2	E	W2	V	T1	L	D1	E	B2	L	BL	57.43
RS.04	F1	L	W1	L	T2	L	D1	H	B4	L	SL	52.69
RS.05	F1	V	W1	V	T1	V	D2	V	B1	H	AI	91.73
RS.06	F1	H	W1	E	T2	H	D1	E	B4	E	SL	55.52
RS.07	F2	L	W2	L	T1	H	D1	V	B2	V	BL	69.52
RS.08	F2	H	W2	E	T2	V	D2	H	B3	V	DC	73.25

Explanation :

- | | |
|----------------------------------|-----------------------------------|
| S : Symptoms | F1 : Moul |
| AL : Attack Level | F2 : Dull |
| DS : Name of Diseases | W1 : Hanging |
| %AL : Percentage of Attack Level | W2 : Moul |
| L : Low | T1 : Limp |
| E : Enough | T2 : Swelling |
| H : High | D1 : Watery Yellow |
| V : Very High | D2 : Watery White |
| AI : Avian Influenza | B1 : Often rub head to the ground |
| DC : Duck Cholera | B2 : Stand with one leg |
| BL : Botulismus | B3 : Often sleepy |
| SL : Salmo nelosis | B4 : Totter |

Based on the table results of trials alliance method performed respondents mentioned above, it can then be analyzed by comparing the results of Table IV with rule tables owned by duck diseases expert systems (Table I) were applied to the Bali Provincial Livestock Office.

The results of the matches between the two tables can be analyzed that the alliance method has been run in accordance

with the rules. This is evidenced by the correspondence between the code and the calculation of the percentage of symptom attack rate obtained by testing respondents and based on the existing rules in an expert system that generates the name of the disease which is also in accordance with the rules.

To view the alliance method has been run in accordance with the rules can be seen in the percentage diagram of response trials suitability rules.

Answer percentage diagram form of rules conformance testing given by the respondents can be described as follows:

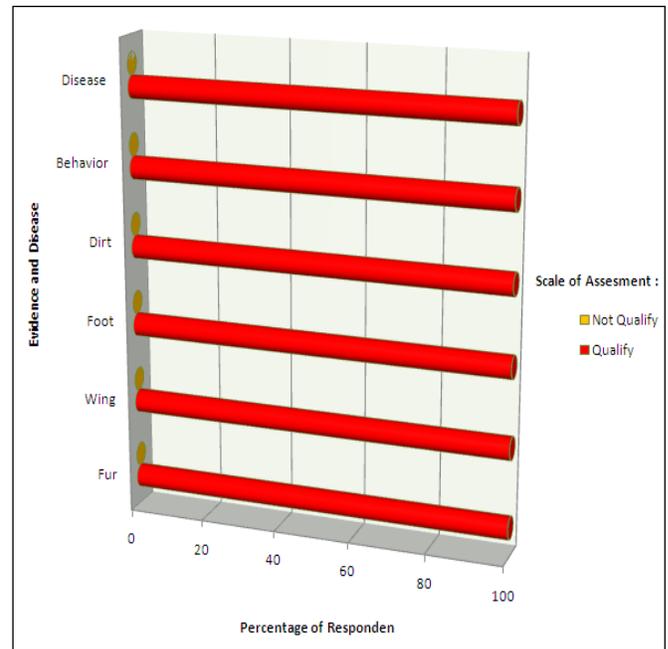


Fig. 4. Answer Percentage Diagram of Rules Conformance Testing

Based on the diagram above, it can be seen that the results of testing the suitability of duck diseases expert system rules is an outline already looks qualify. This is evidenced by the percentage of the answer symptoms fur, wing, foot, dirt, behavior and disease name according to the rules in the field of testing and each get a percentage of 100%.

2) Implementation of Duck Diseases Expert System

a) Login Form



Fig. 5. Login Form

This login form is used by staff of Bali Provincial Livestock Office to be able to come into main menu form, especially to activate of master menu, search, and report found on duck diseases expert system.

b) Main Menu Form

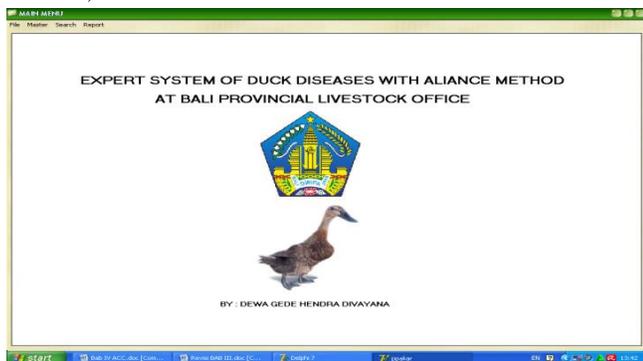


Fig. 6. Main Menu Form

This main menu form used as link to file menu, master, search, and report.

c) Membership Registration Form

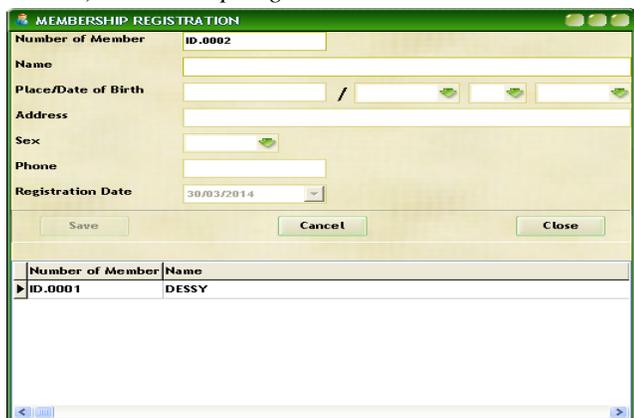


Fig. 7. Membership Registration Form

Membership registration form is used as registration facility of incoming member looking for information about duck diseases.

d) Duck Diseases Data Input Form



Fig. 8. Duck Diseases Data Input Form

Duck diseases data input form is used by staff of Bali Provincial Livestock Office to enter detail explanation about duck diseases.

e) Symptoms Data Input Form



Fig. 9. Symptoms Data Input Form

This symptoms data input form can be used for the input of new symptom for the fur, wing, foot, dirt, and behavior symptoms.

f) Rules Data Input Form

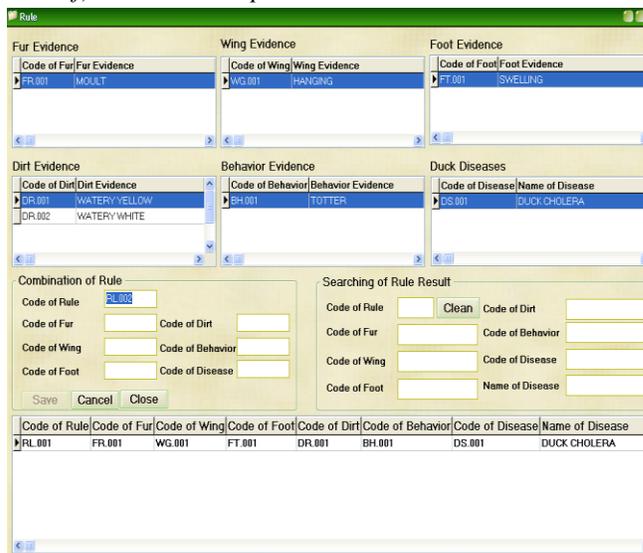


Fig. 10. Rules Data Input Form

This rules data input form is used to make symptoms combination (behaviour and physical) which is input into a order so that give an conclusion of duck disease name.

g) Weight Data Input Form



Fig. 11. Weight Data Input Form

This weight data input form is used to make attack level weight value given by user as well as by expert.

h) Form of Consultancy With Alliace Method

Fig. 12. Form of Consultancy With Alliace Method

This form of consultancy with alliace method is owning facility seeking of disease pursuant to physical and behaviour duck symptoms by applying forward and backward chaining concept. While concept of weighted product used to determine attack level to every physical and behavior duck symptoms. The attack level used to determine user and expert weight value, is so that obtained by attack level percentage of duck diseases with correct calculation.

V. CONCLUSIONS

Based on the analysis that has been made and the results of the discussion in the previous section, then some conclusions can be drawn as follows:

a) Expert systems are applied at Bali Provincial Livestock Office to facilitate duck breeders in acquiring knowledge and information about duck diseases.

b) Expert systems are applied at Bali Provincial Livestock Office has been able to provide information in accordance with the rules of alliance method. This has been proven in testing the suitability of alliance method with the calculation method of the weighted percentage of respondents at 100%.

c) With usage of this expert system, can solve problems faced at Bali Provincial Livestock Office in the case gift of service to society, specially duck breeders which searching duck diseases information and also the way of solution technique.

d) Expert system which woke up can fulfill fundamental characteristic of computerization system which concerning information quality, user interface, and technical ability compared to which is manual.

e) This duck diseases expert system can solve an problem of complicated become easier overcome.

f) With existence of this duck diseases expert system, user can find accurate solution or information about duck diseases.

g) At this expert system, every symptoms, diseases, and solution can be added, edited and deleted.

ACKNOWLEDGMENTS

The authors express their gratefulness to staff Bali Provincial Livestock Office for inspiring words and allowing them to use the examination data. They generously thank Mr. Dayung, President of Indonesia Technology University, and Mr. Semadi, Dean of Computer Faculty, Indonesia Technology University.

REFERENCES

- [1] J.C. Giarratano, and G. Riley, Expert Systems : Principles and Programming 4th Edition. USA : PWS Publishing Co, 2004.
- [2] E. Turban, and J. E. Aronson, Decision Support Systems and Intelligent System. NJ, USA: Prentice-Hall Inc, 2001.
- [3] Y. Qu, F. Tao, and H. Qui, "A Fuzzy Expert System Framework Using Object Oriented Techniques," in IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, 2008, pp. 474-477.
- [4] A. A. Hopgood, Intelligent Systems for Engineers and Scientists (2nd Edition). USA : CRC Press, 2001.
- [5] S.Kusumadewi, Artificial Intelligence (Technique and Application) 1st Edition. Yogyakarta : Graha Ilmu, 2003.
- [6] S. Kusumadewi, S. Hartati, A. Harjoko, and R. Wardoyo, Fuzzy Multi Attribute Decision Making (FUZZY MADM). Yogyakarta : Graha Ilmu, 2006.
- [7] H. Divayana, Development of Duck Diseases Expert System with Applying Alliance Method at Bali Provincial Livestock Office. USA : Corllins University, 2014.

AUTHORS BIOGRAPHY



Dewa Gede Hendra Divayana, Ph.D. was born in Denpasar, Bali, in 1984. He received the B.IT degree in Information Technology in 2008, M.IT degree in Information Technology in 2010 and in 2014, he received his Ph.D. in Information Technology from Corllins University, USA. From September 2009 until now, he worked as Lecturer of Information Technology, Chair of Information Technology Department, Faculty of Computer, Indonesia Technology University, Bali, Indonesia. His research interests include Expert Systems, Artificial Intelligence, and Object Oriented Programming.

Toward Accurate Feature Selection Based on BSS-GRF

S.M. ELseuofi

Inf. System Dept. Ras El bar High inst., Damietta, Egypt

Samy Abd El -Hafeez

Math. &Comp.Sci.Dept. Science faculty, Port Said University

Wael Awad

Math. &Comp.Sci.Dept. Science faculty, Port Said University

R. M. El-Awady

Electronic. &Communication.Dept. Faculty of engineering Mansoura University

Abstract—in recent years, Feature extraction in e-mail classification plays an important role. Many Feature extraction algorithms need more effort in term of accuracy. In order to improve the classifier accuracy and for faster classification, the hybrid algorithm is proposed. This hybrid algorithm combines the Genetics Rough set with blind source separation approach (BSS-GRF). The main aim of proposing this hybrid algorithm is to improve the classifier accuracy for classifying incoming e-mails.

Keywords—rough set; Genetic; blind source separation; E-mail Filtering; Machine Learning

I. INTRODUCTION

Due to the increasing volume of unwanted email called as spam or Junk email, the users as well as Internet Service Providers (ISPs) are facing a lot of problems. Email spam also creates a major problems to the security of networked systems. Email classification is able to control the problem in a variety of ways. Detection and protection of spam emails from the e-mail delivery system allows end-users to regain a useful means of communication. Many researches on content based email classification have been centered on the more sophisticated classifier-related issues [4]. Currently, machine learning for email classification is an important research issue. The success of machine learning techniques in text categorization has led researchers to explore learning algorithms in spam filtering [6]. However, it is amazing that despite the increasing development of anti-spam services and technologies, the number of spam messages continues to increase rapidly. Consequently, novel approaches are desired to deal with ever-increasing flood of spam and the persistent attempts by spammers to break the existing anti-spam barriers. Generations of spam filters have emerged over the years to deal with the spam issue. Most of these filters succeeded to some point in discriminating between spam and legitimate e-mails, however they require manual intervention. For example content based methods require human efforts to build lists of characteristics and their scores. Over the last five years, statistical filters have gained more attention as they are able to tweak themselves; getting better and better with less manual intervention. The most popular statistical approach is the Bayesian filter, which assigns

probability estimates to e-mails. When dealing with very large-scale datasets, it is often a practical necessity to seek to reduce the size of the dataset, acknowledging that in many cases the patterns that are in the data would still exist if a representative subset of instances were selected. Further, if the right instances are selected, the reduced dataset can often be less noisy than the original dataset, producing superior generalization performance of classifiers trained on the reduced dataset. The search for spam words in the incoming email can be viewed as a feature selection problem that can be formulated as follows: Given N data points $x_i \in R^n$, $i = 1, \dots, N$, with labels $y_i \in \{-1, 1\}$ select an L -element subset of features $\{x_{ik} | K \in S, S \subseteq \{1, \dots, N\}\}$ while preserving or possibly improving discriminative ability of a classifier. The number of relevant features L is usually chosen arbitrarily. The e-mails can be interpreted as signals in the Universe (U) that can be separated into statistically independent components. Magnitudes of those components denoted by a_i represent the original points x_i in a new feature space. Dimensionality of a_i 's is usually much smaller than dimensionality of x_i 's making classification and feature selection problem easier. The new feature set will then be reduced to attributes relevant to the given classification. Each attribute of a_i is associated with a computed component that is still in the Universe (U). Therefore, relevant features point to relevant components where difference between e-mails in Universe (U) can be observed. Optima of those components for a particular class indicate values higher or lower than usually. The following sections present a Blind Source Separation (BSS) technique used to compute components and their magnitudes in each e-mail, Rough set tools used for reduction of the new feature set and classification of the incoming e-mail based on feature selection in Universe (U).

II. BLIND SOURCE SEPARATION ALGORITHM

In blind source separation (BSS), multiple observations are carried out by an array of sensors are processed in order to recuperate the initial mixing of the source signals. The term blind refers to the fact that there is no specific information about the mixing process or about the existing source signals. Blind source separation (BSS) is the technique that anyone can separate the original signals or latent data from their mixtures

without any knowledge about the mixing process, but using some statistical properties of latent or original source signals. The perception of blind source separation is related to independent component analysis (ICA)[3]. However, ICA can be viewed as a general-purpose tool replacing principal component analysis (PCA) which means it is applicable to a wide range of problems. Some application domains of blind source separation are biomedical signal analysis, geophysical data processing, data mining, wireless communications and sensor array processing [2].

Each sequence of attributes x_i will be interpreted as signal and will be denoted by a column vector. It is assumed here that each signal is a mixture of some underlying source of activity.

It is assumed that each input signal is a linear combination of some statistically independent source.

$$X_i = Ma_i + e_i, \quad (1)$$

Where each column of $M \in R^n \times m$ is a basis function $M_j \in R^n$, $j = 1, 2, \dots, m$, $a_i \in R^m$ is a column vector of coefficients – magnitudes of each basis functions in the signal x_i and $e_i \in R^n$ represents noise or error of the model. M and a are unknown parameters that need to be estimated. Statistical independence of the basis functions can be satisfied by minimizing mutual information between the basis functions. Thus M and a are estimated by solving the following:

$$M, a = \arg \min a (\arg \min m (I(M_1, M_2, \dots, M_m) + \lambda \|x - Ma\|/2)), \quad (2)$$

Where λ is a scaling factor, and $I(M_1, M_2, \dots, M_m)$ is a mutual information between random variables M_1, M_2, \dots, M_m defined as:

$$I(M_1, M_2, \dots, M_m) = \sum_{j=1}^m H(M_j) - H(M_1, M_2, \dots, M_m), \quad (3)$$

Where $H(M_j)$ is entropy of a random variable M_j . Optimization using (2) may be performed with a gradient descent algorithm. The two quantities to be computed, M and a , make this problem complex. The minimization can be solved by estimating only M :

$$M = \arg \min m (I(M_1, M_2, \dots, M_m) + \lambda \|x - Ma\|/2), \quad (4)$$

Where the estimate A of a in each step of the algorithm is the solution of the following:

$$A = \arg \min a \|x - Ma\|/2, \quad (5)$$

Where the value of M is a partial solution of (4)

III. GENETIC ALGORITHM

Genetic algorithms (GA) are inspired by Darwin's theory about evolution. Simply said, solution to a problem solved by genetic algorithms is evolved. (GA) are a part of evolutionary computing, which is a rapidly growing area of artificial intelligence. It iteratively applies a series of genetic operators such as selection, crossover, and mutation to a group of chromosomes where each chromosome represents a solution to a problem. The initial set of chromosomes is selected randomly from solution space. Genetic operators combine the genetic information of parent chromosomes to form a new generation of the population; this process is known as reproduction. Each chromosome has an associated fitness value, which quantifies its value as a solution to the problem. A chromosome representing a better solution will have a higher fitness value.

The chromosomes computed to reproduce based on their fitness value, thus the chromosomes representing better solution have a higher chance of survival. After many generations, a chromosome, which has the maximal fitness value, is the best solution for the problem. Encoding of a Chromosome: The chromosome should in some way contain information about solution which it represents. The most used way of encoding is a binary string. The chromosome then could look like this:

Chromosome 1 1101100100110110

Chromosome 2 1101111000011110

Each chromosome has one binary string. Each bit in this string can represent some characteristic of the solution. Or the whole string can represent a number. Crossover: After we have decided what encoding we will use, we can make a step to crossover. Crossover selects genes from parent chromosomes and creates a new offspring. The simplest way how to do this is to choose randomly some crossover point and everything before this point copy from a first parent and then everything after a crossover point copy from the second parent. After a crossover is performed, mutation take place. This is to prevent falling all solutions in population into a local optimum of solved problem. Mutation changes randomly the new offspring. For binary encoding we can switch a few randomly chosen bits from 1 to 0 or from 0 to 1

IV. ROUGH SET ALGORITHM

In 1982 Rough set (RS) theory was developed by Pawlak. The most advantage of rough set is its great ability to compute the reductions of information systems. In an information system there might be some attributes that are irrelevant to the target concept (decision attribute), and some redundant attributes. Reduction is needed to generate simple useful knowledge[11] from it. A reduction is the essential part of an information system. It is a minimal subset of condition attributes with respect to decision attributes. The Rough set scheme is provided as follows. An information system is a pair $S = \langle U, A \rangle$, where $U = \{x_1, x_2, \dots, x_n\}$ is a nonempty set of objects (n is the number of objects); A is a nonempty set of attributes, $A = \{a_1, a_2, \dots, a_m\}$ (m is the number of attributes) such that $a : U \rightarrow V_a$ for every $a \in A$. The set V_a is called the value set of a . A decision system is any information system of the form $L = (U, A \cup \{d\})$, where d is the decision attribute and not belong to A . The elements of A are called conditional attributes. Let $S = \langle U, A \rangle$ be an information system, then with any $B \subseteq A$ there is associated an equivalence relation $INDS(B)$: $INDS(B) = \{(x, x') \in U^2 \mid \forall a \in B a(x) = a(x')\}$ $INDS(B)$ is called the B -indiscernibility relation. The equivalence classes of B -indiscernibility relation are denoted $[x]_B$. The objects in $\underline{B}X$ can be certainly classified as members of X on the basis of knowledge in B , while the objects in $\overline{B}X$ can be only classified as possible members of X on the basis of knowledge in B . Based on the lower and upper approximations of set $X \subseteq U$, the universe U can be divided into three disjoint regions, and we can define them as: $POS(X) = \underline{B}X$, $NEG(X) =$

$U - \bar{B}X$, $BND(X) = \bar{B}X - \underline{B}X$ The equivalence classes of B-indiscernibility relation are denoted $[x]_B$. indiscernibility is a binary equivalence relation that divides a given set of elements (objects) into a certain number of disjoint equivalence classes.

An equivalence class of an element $a_i \in U \subseteq X$ consists of all objects $a_i \in U \subseteq X$ such that $a_i R a_j$, where R indicates a binary relation. Let $IS = (R, A)$ be an information system of objects from universe Rm described by the set of attributes A, then with any $B \subseteq A$ there is an associated equivalence relation $INDS(B)$:

$$INDS(B) = \{(x, x') \in U^2 | \forall a \in B a(x) = a(x')\} \quad (6)$$

Based on the concept of indiscernibility relation, a reduction in the space of attributes is possible. The idea is to keep only those attributes that preserve the indiscernibility relation. The rejected attributes are redundant since their removal cannot affect the classification.

V. BSS-GRF PROPOSED ALGORITHM

Blind Source Separation-Genetic Rough Filter (BSS-GRF) is a proposed Hybrid algorithm that use the Blind source separation technique hybrid with Genetic algorithm to enhance the feature selection process and then the classification process done by rough set algorithm.

- 1) Discarding from X (incoming email) the column consisting of low value entered due to noise
- 2) Calculate the sub matrix.
- 3) Sort Words according to word ranks
- 4) Choose the number of generations (we'll use 10)
- 5) Read the spam and ham corpora
- 6) Randomly mix the lines of spam
- 7) Divide spam corpus into 10 slices
- 8) Loop until 10th generation:
 - a) Generate chromosomes based on the current slice of spam using the 'automatic' formula
 - b) Score chromosomes
 - c) Print results for the current generation
 - d) Keep the fittest 3rd
 - e) reproduction survivors
 - 2 survivor's reproduction via a crossover function to create a child
 - Use 'Roulette Wheel' selection top choose the 2nd parent
 - f) Mutate some of the children by randomly deleting some genes
 - g) Move to next slice of spam
- 9) Print Final results

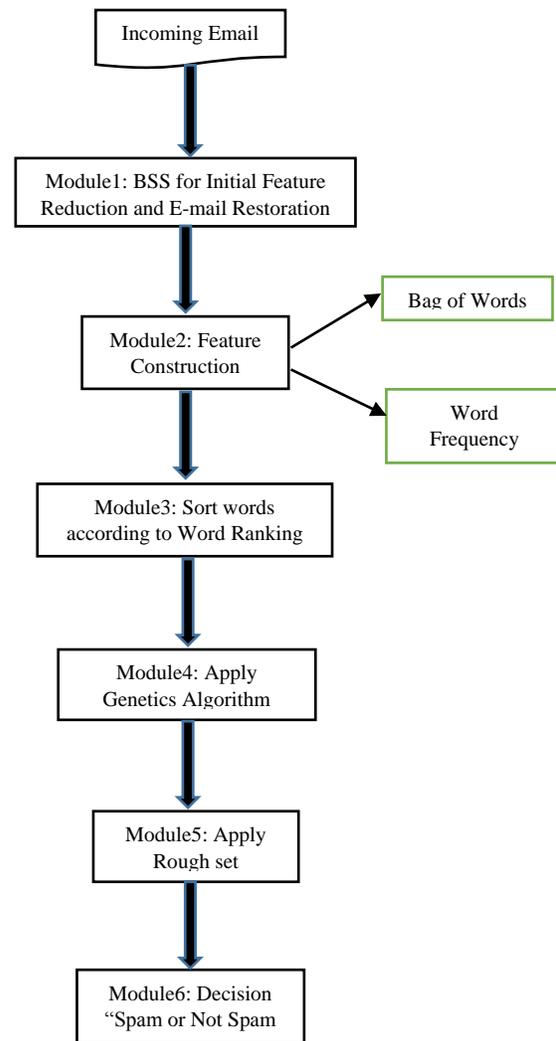


Fig. 1. BSS-GRF filtering Steps

VI. EXPERIMENT IMPLEMENTATION

In order to test the performance of above method, some corpora of spam and legitimate emails had to be compiled; there are several collections of email publicly available to be used by researchers.

SpamAssassin (<http://spamassassin.apache.org>) will be used in this experiment, which contains 6000 emails with the spam rate 37.04%. Thus we have divided the corpora into training and testing sets keeping, in each such set, the same proportions of ham (legitimate) and spam messages as in the original example set. Each training set produced contained 62.96% of the original set; while each test set contain 37.04% as Table 1.

TABLE I. CORPORA OF SPAM AND HAM MESSAGES

Message collection	Training Set	Testing Set
Ham Messages	2378	1400
Spam Messages	1398	824
Total Messages	3776	2224

The idea here is the email is usually comes with a set of words, web links, that affect the classification process accuracy, these unwanted words can be removed manually or by special module. BSS-GRF (Blind Source Separation-Genetic Rough Filter) is presented to imbed BSS algorithm with GA to first: perform an email restoration process “an email without unwanted words“. Second: GA perform the feature reduction process and word ranking that can be used for the classification process using Rough set method.

A. Performance evaluation

In order to test the performance of above mentioned methods, we used the most popular evaluation methods used by the spam filtering researchers. Spam Precision (SP), Spam Recall (SR), Accuracy (A). Spam Precision (SP) is the number of relevant documents identified as a percentage of all documents identified; this shows the noise that filter presents to the user (i.e. how many of the messages classified as spam will actually be spam)

$$SP = \frac{\text{\# of Spam Correctly Classified}}{\text{Total \# of messages classifies as spam}} = \frac{N_{\text{spam} \rightarrow \text{spam}}}{N_{\text{spam} \rightarrow \text{spam}} + N_{\text{ham} \rightarrow \text{spam}}}$$

Spam Recall (SR) is the percentage of all spam emails that are correctly classified as spam.

$$SR = \frac{\text{\# of Spam Correctly}}{\text{Total \# of}} = \frac{N_{\text{spam} \rightarrow \text{spa}}}{N_{\text{spam} \rightarrow \text{spam}} + N_{\text{spam} \rightarrow \text{ham}}}$$

Accuracy (A) is the percentage of all emails that are correctly categorized Where Nham→ham and Nspam→spam

$$A = \frac{\text{\# of e-mails correctly categorized}}{\text{Total \# of e-mails}} = \frac{N_{\text{ham} \rightarrow \text{ham}} + N_{\text{spam} \rightarrow \text{spam}}}{N_{\text{ham}} + N_{\text{spam}}}$$

are the number of messages that have been correctly classified to the legitimate email and Spam email respectively; Nham→spam and Nspam→ham are the number of legitimate and spam messages that have been misclassified; Nham and Nspam are the total number of legitimate and spam messages to be classified.

B. Performance Comparison

In order to test the proposed Hybrid system we run the same data onto four different machine learning algorithms. We summarize the performance result of the presented method in term of spam recall, precision and accuracy.

Table 2 and Fig.2 summarize the results of the classifier. Very competitive results can be seen from the BSS-GRF, in terms of spam recall, precision and accuracy the percentage here is much more than rough set . While in term of accuracy GRF[5] still has the high percent. Support Vector Machine System and the RS give us approximately the same lower percentage.

TABLE II. PERFORMANCE COMPARISON OF FOUR ALGORITHMS

Algorithm	Spam Recall (%)	Spam Precision (%)	Accuracy (%)
GRF	98.46	97.80	99.66
BSS-GRF	92.36	94.56	96.7
RS	92.00	90.12	94.90
SVM	95.00	93.12	96.90

VII. CONCLUSION AND FUTURE WORK

The prevoius results presented leads to new approach have to be taken by researchers in the future, Blind source separation show us a good result when it hybrid with Genetics in the purpose of feature separtion and reduction process.

The presented method need more improvement in case of noise types, email corpora, more effort has to be done to improve the feature selection process in terms of accuracy, more classifiers type can be used to be hybrid with the BSS instead of the rough set method.

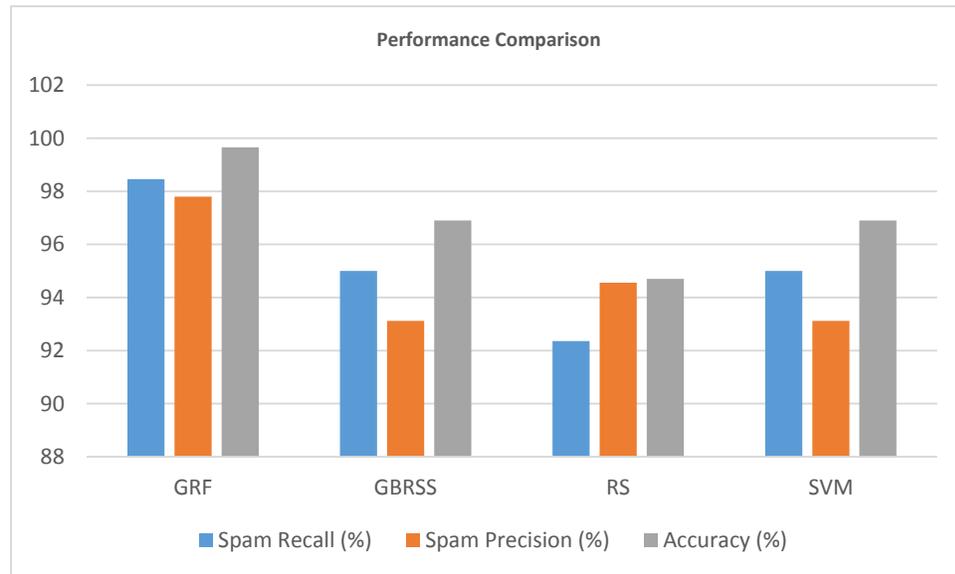


Fig. 2. Spam Recall, Spam Precision and Accuracy curves of four classifiers

REFERENCES

- [1] Boratyn G.M., Smolinski T.G., Zurada J.M., Milanova M.G., Bhattacharyya S., and Suva L.J., Hybridization of Blind Source Separation and Rough Sets for Proteomic Biomarker Identification, Proc. of the 7th International Conference on Artificial Intelligence and Soft Computing (ICAISC 2004), June 2004.
- [2] A. Budipriyanto, Blind source separation based dynamic parameter identification of a multi-story moment-resisting frame building under seismic ground motions, Procedia Engineering, vol. 54, pp. 299-307, 2013.
- [3] F. Gu, H. Zhang, and D. Zhu, "Blind separation of non-stationary sources using continuous density hidden Markov models," Digital Signal Process., vol. 23, no. 5, pp. 1549-1564, Sep. 2013.
- [4] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: a literature survey," IEEE Communications & Surveys Tutorials, vol. 15, no. 4, pp. 2091-2121, 2013
- [5] S.M. ELseofi, W.A. Awad, S. A. El Hafeez, R. M. El-Awady, Enhancing E-mail Filtering Based on GRF, International Journal of Computer Science Issues, Vol. 11, Issue 3, No 1, May 2014
- [6] M. Behdad, L. Barone, M. Bennamoun, and T. French, "Nature-inspired techniques in the context of fraud detection," IEEE Transactions on Systems, Man, and Cybernetics C: Applications and Reviews, vol. 42, no. 6, pp. 1273-1290, 2012.
- [7] A. ALmomani, T.-C. Wan, A. Altaher et al., "Evolving fuzzy neural network for phishing emails detection," Journal of Computer Science, vol. 8, no. 7, pp. 1099-1107, 2012.
- [8] B. Biggio, G. Fumera, I. Pillai, F. Roli, A survey an experimental evaluation of image spam filtering techniques, Pattern Recognition Letters 32 (10) (2011) 1436-1446
- [9] A.H. Mohammad, R.A. Zitar, Application of genetic optimized artificial immune system and neural networks in spam detection, Applied Soft Computing 11 (4) (2011) 3827-3845.
- [10] A.H. Mohammad, R.A. Zitar, Application of genetic optimized artificial immune system and neural networks in spam detection, Applied Soft Computing 11 (4) (2011) 3827-3845.
- [11] El-Sayed M. El-Alfy, Radwan E. Abdel-Aal "Using GMDH-based networks for improved spam detection and email feature analysis" Applied Soft Computing, Volume 11, Issue 1, January 2011
- [12] Almeida,tiago. Almeida, Jurandy.Yamakami, Akebo " Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers" Journal of Internet Services and Applications, Springer London , February 2011.

A New Efficient Method for Calculating Similarity Between Web Services

T. RACHAD

Architectures and systems team, LISER Laboratory
ENSEM, BP 8118 Oasis
Casablanca, Morocco

J.Boutahar, S.El ghazi

Systems, architectures and networks team
EHTP, B.P 8108, Oasis
Casablanca, Morocco

Abstract—Web services allow communication between heterogeneous systems in a distributed environment. Their enormous success and their increased use led to the fact that thousands of Web services are present on the Internet. This significant number of Web services which not cease to increase has led to problems of the difficulty in locating and classifying web services, these problems are encountered mainly during the operations of web services discovery and substitution.

Traditional ways of search based on keywords are not successful in this context, their results do not support the structure of Web services and they consider in their search only the identifiers of the web service description language (WSDL) interface elements.

The methods based on semantics (WSDLS, OWLS, SAWSDL...) which increase the WSDL description of a Web service with a semantic description allow raising partially this problem, but their complexity and difficulty delays their adoption in real cases.

Measuring the similarity between the web services interfaces is the most suitable solution for this kind of problems, it will classify available web services so as to know those that best match the searched profile and those that do not match. Thus, the main goal of this work is to study the degree of similarity between any two web services by offering a new method that is more effective than existing works.

Keywords—web service; semantic similarity; syntactic similarity; WordNet; word sense disambiguation; Hausdorff distance

I. INTRODUCTION

Web services have emerged in the last decade as an innovative technology solving several problems related to the integration of heterogeneous systems. At the beginning it was used only by some large business groups to facilitate the exchange of data between remote and heterogeneous information systems (from a technological point of view), but later and thanks to its efficiency and performance, the majority of companies have adopted it to publish the public part of their information systems in order to facilitate openness to other markets and promote communication with heterogeneous external systems.

Currently, with the democratization of the Internet, the emergence of broadband, the advent of cloud computing and

large-scale popularization of e-commerce, web service technology has found its reason for being. Its use has become a necessity, even an obligation to find a place in the electronic market and be able to exchange easily data with third parties.

The existence of a large number of web services on the Internet has led to the emergence of new problems (for discovery, selection and invocation of web services) resulting primarily from the aggravation of the problem of (semantic) heterogeneity: many web services that do the same thing, but do not have the same interfaces; web services that belong to the same business domain and do the same thing, but do not share the same vocabulary; defective web services, which must be replaced by other operational web services, etc.

A typical example where this kind of problems are encountered is the substitution of Web services (Figure 1) which consists in replacing a defective Web service by another that is similar and operational. This operation requires discovering from a Web services registry those who are similar to the defective one. Often this discovery operation is performed manually by an administrator, but given the large number of web services that exist, it will be costly in terms of time devoted to study the similarity with all available web services and it may therefore be ineffective. Automation of this process of discovery requires to have an efficient method to calculate the degree of similarity between available web services and the web service to replace.

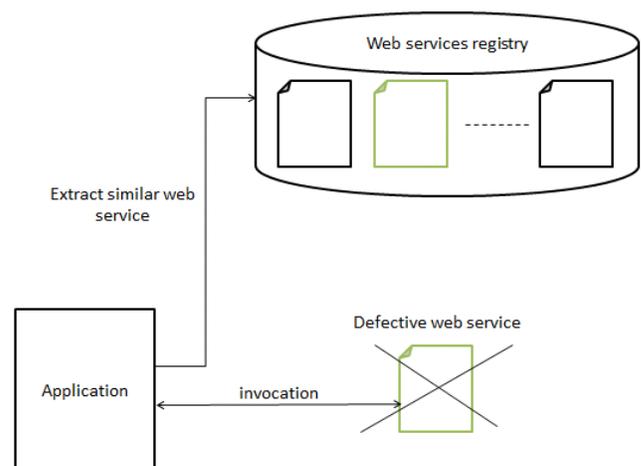


Fig. 1. web services substitution

The main goal of this work is to study the degree of similarity between any two web services by offering a new method that is more effective than existing works.

In section II of this paper we present some basic methods and tools that were used to reach the goal of our work. In Section III we present our approach in the calculation of similarity between any two web services. In section IV we evaluate our approach of similarity measurement and in section V we compare our work with results obtained by existing works. Section VI, concludes this paper and opens some perspectives of future work.

II. TOOLS AND BASIC METHODS

A. syntactic similarity

The syntactic similarity consists in assigning to a string pair S1 and S2 a real number r , which indicates the degree of syntactic similarity between S1 and S2. There are mainly two ways for measuring the degree of similarity between two concepts:

- Edit distances : in which the distance is the cost of the optimal sequence of editing operations that transform S1 to S2 or S2 to S1. Editing operations are character insertion, deletion and substitution. A small value of r indicates greater similarity. There are several algorithms based on edit distances, the most well-known are: Minkowsky (1964), manhaten, Levenstein (1965), Monger-Elkan (1996), Smith-Waterman (1981). In [1] the authors carried out a comparative study of edit distances based methods and concluded that Monger-Elkan [2] provides the best result.
- Similarity functions: are analogous to the edit distances based methods, except that higher values indicate greater similarity. The most known algorithms are: Jaro [3, 4]; Jaro-Winkler [5, 6]. Work [1] shows that jaro-winkler is the most powerful and fastest measurement.

In this work we use the Jaro-Winkler algorithm to measure the syntactic similarity between two strings.

B. Semantic Similarity

The semantic similarity consists in assigning to a pair of words w_1 and w_2 a real number r , which indicates the degree of semantic similarity between them. The similarity measure is done by comparing the senses of the two words. Thus, two words are similar (with a certain degree of similarity) semantically if they mean the same thing (synonyms), they have opposite meaning (antonyms), they are used in the same way or inherit the same type, they are used in the same context or if one is a type of the other. To measure the semantic similarity between words, we will need a lexical hierarchy such as WordNet [7].

WordNet is a lexical database which aims to identify, classify and relate the semantic and lexical content of the English language. Nouns, verbs, adjectives and adverbs are grouped as sets of cognitive synonyms (synsets) contents, each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

There are several methods and techniques to measure the semantic similarity between two concepts, the most known are: Resnik (1995); Lin (1998); Wu & Palmer (1994); Jiang and Conrath's (1997); Leacock and Chodorow (1998); Hirst & St-Onge (1998). Currently, we cannot say that there is a method that is the best or most optimal than others, because each of the studies that have examined these algorithms has been considering some evaluation criteria and neglecting others. We identified three evaluation methods; mathematics evaluations, evaluations based on human judgment and evaluations measuring performance in the context of a particular application.

In [8] authors compared experimentally five measures of semantic similarity in wordnet (ie Hirst and St-Onge, Leacock and Chodorow, Resnik, Lin and finally Jiang and Conrath) by examining their performance in spelling correction systems and by comparing their performance with human judgments. They found that Jiang and Conrath's method and Hirst-St-Onge method offer the best results, followed by measurement of Lin and of Leacock and Chodorow, Resnik measure comes in the last rank. In [9] based on human judgment, authors argue that Leacock-Chodorow measure is the best one, followed by that of Resnik , Wu-Palmer is in third place. They also argue that Lin measure and Jiang-Conrath measure are not efficient. In [10] authors evaluated the similarity measures in three different domains (transport, book and business) with reference to human judgment and experts judgment, they concluded that at recall, WordNet with Jean Conrath provide the best result at three domains, at Precision, there is no significant method can provide dominant result and At f-measure (that combine recall and precision measures), WordNet with Wu-Palmer has tendency better than the others.

The third evaluation work discussed in [10] seems the most rigorous for us, because it uses both human and experts judgments and because the tests are carried out in three different areas. In our work, to measure similarity between web services, we decided to use WordNet with Wu-Palmer because it is the measure that provides the best result.

C. Word sense disambiguation

Measurement of semantic similarity between two words refers to the measure of similarity between the senses of the two words. All algorithms for measuring semantic similarity, consider either the most common sense or senses that offer highest similarity during the comparison process. But the meaning of a word changes according to the context in which the word appears. That is why, we must extract the exact senses of different words before addressing the similarity measure. Word sense disambiguation is the scientific expression that has been attributed to the process of searching the exact meaning of a word in a specific context.

Adapted Lesk algorithm described in [11], [12] and [13] is adopted to remove the ambiguity of meaning in a given context.

Below in Table 1 the implementation that we have adopted for Adapted Lesk algorithm.

TABLE I. ADAPTED LESK ALGORITHM

Function Wsd_Simplified_Lesk(word, context)
best-sense <- most frequent sense for word max-overlap <- 0 for each sense in senses of word do signature <- set of words in of sense description overlap <- ComputeOverlap (signature,context) if overlap > max-overlap then max-overlap <- overlap best-sense <- sense return best-sense
Function ComputeOverlap (signature,context)
count=0 commonWords=("the","of","to","and","a","in","is","it","you", "that","he","was","for","on","are","with","as","i",.....) signature.removeAll(commonWords) context.removeAll(commonWords) for each word1 in signature do for each word2 in context do if SimSyntactic(word1,word2)>0.5 count++ return count
Function SimSyntactic(word1,word2)
return JaroWinkler(word1,word2)

D. distance between two sets

Throughout this work, we will need to compute the degree of similarity between two sets of concepts which elements (concepts) are connected by a similarity measure (Figure 2).

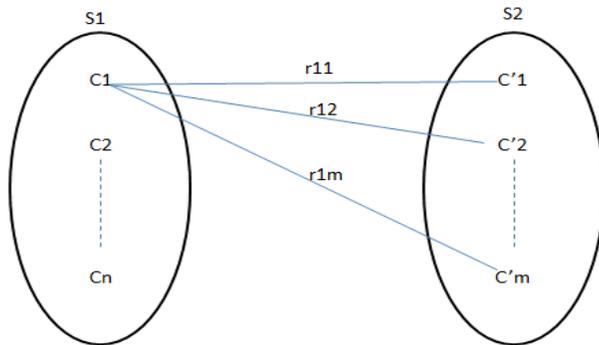


Fig. 2. relations between two sets of concepts

In this work we chose the Hausdorff algorithm [14] to calculate the degree of similarity between two sets of concepts. It is used to calculate the similarity between two objects represented by two sets of points. The problem is thus brought back to computing the distance between the two sets of points.

In [14] authors affirm that there are 24 possible ways to measure the distance between two sets of points using the Hausdorff distance and they concluded that the modified Hausdorff distance (MHD) has the highest performance to measure similarity between two objects.

The modified Hausdorff distance between two sets of points S1 and S2 is defined by the expression:

$$MHD(S1, S2) = \max \{ g_d(S1, S2), g_d(S2, S1) \} \quad (1)$$

Where d is any distance (in our case, it must be Jaro-Winkler measure or Wu-Palmer measure) and g_d is the modified Hausdorff distance. It is defined by:

$$g_d(S1, S2) = \frac{1}{|S1|} \sum_{p \in S1} \min_{q \in S2} \{d(p, q)\} \quad (2)$$

III. SIMILARITY MEASURE BETWEEN WEB SERVICES

A. Structure of a wsd file

WSDL is an XML file that follows a standard format for describing a web service. It mainly describes the operations provided by the web service and how to access them.

A WSDL file has the following structure (Table 2):

TABLE II. WEB SERVICE STRUCTURE

<definitions>
<types>
data type definitions.....
</types>
<message>
definition of the data being communicated....
</message>
<portType>
set of operations.....
</portType>
<binding>
protocol and data format specification....
</binding>
</definitions>

- Definition: is the root element of the WSDL document. It describes the Web service name and declares several namespaces.
- Types: is an XML schema that describes the data types used by wsd operations.
- Message: an abstract definition of the data exchanged with a wsd operation, it can describe the inputs and the outputs.
- Operation: an abstract definition of an action performed by the web service.
- Port type: An abstract set of operations supported by one or more endpoints.
- Binding: Describes how the operations are invoked.

In our work we consider only the operations and their inputs and outputs. We will consider that a web service is a set of operations and the operations receive and return elements that will be a parts of the WSDL schema. All other elements are ignored, because their names are often generated in an automatic way and depend on the tool used when generating web service and therefore will not intervene in the similarity measure.

B. Preliminary declarations

Let WS1 and WS2 are two web services for which we want to compute the similarity. Let S1 and S2 their schemas. Let F and G two sets of operations such as: $F = \{f / f \text{ is an operation of WS1}\}$ and $G = \{g / g \text{ is an operation of WS2}\}$. Let D and D' the departure sets, respectively, of f and g and let A and A' the arrival sets, respectively, of f and g with $D, A \in P(S1)$ and $D', A' \in P(S2)$, with P(S1) and P(S2) are respectively the sets of parts of S1 and S2.

Our goal is to measure the similarity $WsdSim(WS1, WS2)$ between two web services WS1 and WS2. This calculation depends on the similarity between the operations of the two web services. So $\forall f \in F$ and $\forall g \in G$ we must measure the similarity $OpSim(f, g)$. Calculating the similarity between the two operations f and g depend on the similarities of their sets of departures and arrivals. So $\forall f \in F$ and $\forall g \in G$, we must compute $SetSim(D, D')$ and $SetSim(A, A')$.

C. similarity between two data sets

In our work we consider that any data set E is a sub part of a web service schema S. This data set has a tree structure (Figure 3), such as the name of the set E is the root of the tree, the internal nodes correspond to the elements of complex types and leaves of the tree correspond to the elements of simple types.

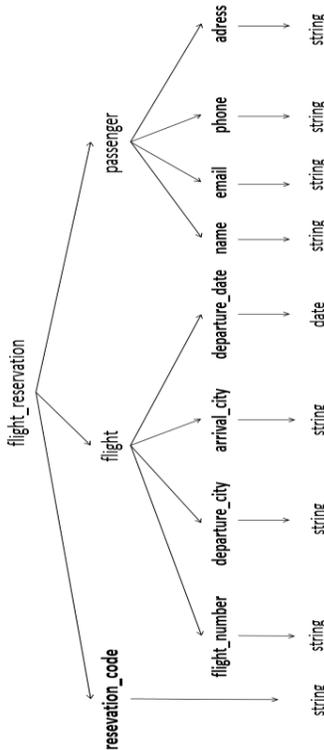


Fig. 3. example of a tree structure of a data set

For measuring the similarity between two data set, all existing works trying to compare all the node of the two sets, ie $\forall e_i \in E$ and $e'_j \in E'$, they calculate the similarity between e_i and e'_j by considering both the syntactic, semantic and structural similarity of the two representations of E and E', thing that makes calculations very complex. However, only the

elements that mainly concern us in the similarity measure are leaves of the tree structure as they are the elements involved in the transformation of data at invocation. The internal nodes do not intervene directly in the calculation of similarity.

In our work, before starting the calculation of similarity between two sets of data, we apply on them a transformation that will provide them with a structure with one level (the root directly connected to the leaves), the leaves names will be concatenated with the names of nodes that connect them with the root, in this way a leaf will represent a whole path in the tree without giving any importance to the tree structure (Figure 4).

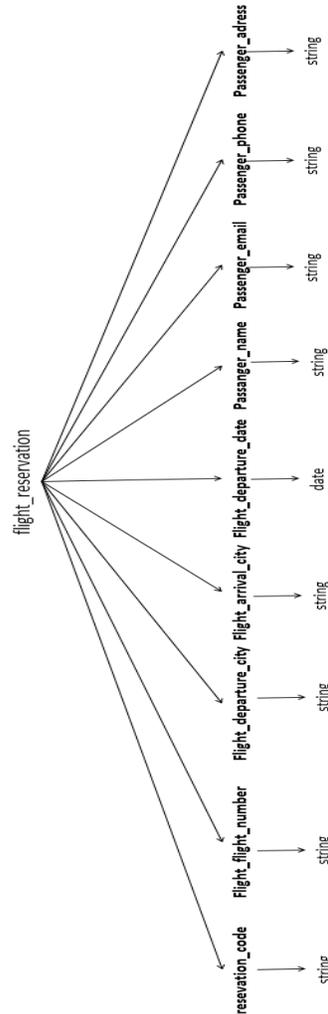


Fig. 4. tree structure with a single level

In fact, a set of data will be considered as a set of sentences, each one represents a leaf of the tree. Then, calculating the similarity $SetSim(E, E')$ between two data sets E and E' will be reduced to the calculation of the similarity between two sets whose elements are sentences. The Hausdorff distance is very suitable for this kind of calculation, it can measure the distance between two sets of points. In our case the points are sentences.

Hausdorff distance uses a similarity matrix MS such as $\forall (S_i, S'_j) \in E \times E'$ $MS(i,j) = SentenceSim(S_i, S'_j)$.

To measure the similarity *SentenceSim* (S, S') between two sentences S and S' , these two last ones shall be divided into words (tokenization), they will then represent two sets whose elements are words. The Hausdorff distance is still the case by calculating the distance between the two sets of words.

At this second level, the distance hausdorff use a similarity matrix MW such as $\forall (W_i, W_j) \in S \times S' MW(i,j) = \text{wordSim}(W_i, W_j, \text{context})$.

The *wordSim*(W, W') returns the similarity between the two words W, W' in a well determined context, it first tries to measure the semantic similarity between words using Wu-Palme algorithm, if one of the two words do not exist in WordNet then it returns a syntactic similarity using the JaroWinkler algorithm.

Table 3 describes all functions used for the calculation of similarity between any two sets of data.

TABLE III. CALCULATION ALGORITHM OF SIMILARITY BETWEEN TWO DATA SETS

Function SetSim(E, E')
return $\text{dist_hausdorff1}(E, E')$
Function $\text{dist_hausdorff1}(E, E')$
return $\min(\text{dist1}(E, E'), \text{dist1}(E', E))$
Function $\text{dist1}(E, E')$
return $\frac{1}{ E } \sum_{s1 \in E} \max_{s2 \in E'} \{ \text{SentenceSim}(s1, s2) \}$
Function $\text{SentenceSim}(S1, S2)$
return $\text{dist_hausdorff2}(S1, S2)$
Function $\text{dist_hausdorff2}(S1, S2)$
return $\min(\text{dist2}(S1, S2), \text{dist2}(S1, S2))$
Function $\text{dist2}(S1, S2)$
return $\frac{1}{ S1 } \sum_{w \in S1} \max_{w' \in S2} \{ \text{WordSim}(w, w') \}$
Function $\text{wordSim}(w1, w2, \text{context})$
If $w1$ is not in WordNet or $w2$ is not in WordNet return $\text{JaroWinkler}(W1, W2)$
else $s1 = \text{WSD_SIMPLIFIED_LESK}(W1, \text{context})$ $s2 = \text{WSD_SIMPLIFIED_LESK}(W2, \text{context})$ return $\text{WuPalmer}(s1, s2)$

D. Similarity between two operations

Let f and g be two operations such that $f \in F$ and $g \in G$ with $f : D \rightarrow A$ and $g : D' \rightarrow A'$, the similarity between the two operations f and g is the sum of the similarities between their arrival and departure sets (inputs and outputs) and the similarity between their names:

$$\text{OPSim}(f, g) = p1 * \text{SetSim}(D, D') + p2 * \text{SetSim}(A, A') + p3 * \text{SentenceSim}(f, g) / (P1 + P2 + P3) \quad (3)$$

In calculation we use a weighting to determine the order of importance of each of the similarity variables. In the measurements that we have made, it was considered that $p1 = 1, p2 = 1$ and $p3 = 2$.

E. Similarity between two web services

In our work, a Web service is considered as a set of operations. The similarity between two web services $WS1$ and $WS2$ will be the Hausdorff distance between the two sets that representing operations.

Hausdorff distance use a similarity matrix MO such as $\forall (opi, opj) \in WS1 \times WS2 MO(i,j) = \text{OpSim}(opi, opj)$.

The table below describes all the functions used to calculate similarity between two any web services.

TABLE IV. CALCULATION ALGORITHM OF SIMILARITY BETWEEN TWO WEB SERVICES

Function $\text{WSDLSim}(wsdluri1, wsdluri2)$
$F = \text{ExtractOperations}(wsdluri1)$
$G = \text{ExtractOperations}(wsdluri2)$
return $\text{distHausdorff3}(F, G)$
Function $\text{dist_hausdorff3}(F, G)$
return $\min(\text{dist2}(F, G), \text{dist2}(G, F))$
Function $\text{dist3}(F, G)$
return $\frac{1}{ F } \sum_{f \in F} \max_{g \in G} \{ \text{OPSim}(f, g) \}$

IV. EXPERIMENTAL RESULTS

To evaluate our method of calculating similarity between two web services, we chose three areas among the most visited by internet users, namely weather information, sending SMS and books research. In order to ensure the obtained results, we recuperate six web services by domain from search engines (Xmethode¹, web services search engine², webservicelist api³).

To measure the performance of our method of similarity measurement, it will be compared with the interpretations of an expert. The latter has a right to assign to a pair of web services one of the following five values: dissimilar, little similar, averagely similar, very similar and identic.

To make the comparison of expert interpretations with our measures, obtained using the method explained in section III, and considering that the obtained similarity measures belongs to the $[0 ; 1]$ interval, we split it into five parts, each one corresponds to a value of the expert interpretations. Dissimilar= $[0 ; 0.2]$, little_similar= $[0.2 ; 0.5]$, averagely_similar= $[0.5 ; 0.7]$, very_similar= $[0.7 ; 0.9]$, identic= $[0.9 ; 1]$.

Below are three tables (Table 5, Table 6 and Table 7) that correspond to the results obtained for the three domains:

¹ <http://www.xmethods.com/ve2/ViewTutorials.po>

² <http://ccnt.zju.edu.cn:8080/>

³ <http://www.webservicelist.com/webservices/>

TABLE V. MEASUREMENTS COLLECTED IN WEATHER DOMAIN

Weather domain	Pairs of services		Expert interpretation	Similarity Measurement	Error
		Service1	Service2	very similar	0.877849788899921
	Service1	Service3	Averagely similar	0.4981597637112343	0.002≈0
	Service1	Service4	Averagely similar	0.7858368347338935	0.09
	Service1	Service5	Averagely similar	0.7681897759103642	0.07
	Service1	Service6	Averagely similar	0.6828835890416772	0
	Service2	Service3	Little similar	0.48631110773757835	0
	Service2	Service4	Averagely similar	0.8395570286195286	0.14
	Service2	Service5	Averagely similar	0.8335290577478078	0.14
	Service2	Service6	Averagely similar	0.6865557185869686	0
	Service3	Service4	Averagely similar	0.6290711428413635	0
	Service3	Service5	Averagely similar	0.5790711428413635	0
	Service3	Service6	Little similar	0.49846146471204517	0
	Service4	Service5	Very similar	0.95	0.06
	Service4	Service6	Very similar	0.7724431818181817	0
	Service5	Service6	Very similar	0.7986336580086579	0
					Error≈3.4%

TABLE VI. MEASUREMENTS COLLECTED IN SMS DOMAIN

SMS domain	Pairs of services		expert interpretation	Similarity Measurement	Errors
		Service1	Service2	Averagely similar	0.6311958922550287
	Service1	Service3	Little similar	0.5608538040463417	0.07
	Service1	Service4	Little similar	0.5493516663359421	0.05
	Service1	Service5	Averagely similar	0.6948070143692332	0
	Service1	Service6	Averagely similar	0.6236555172921825	0
	Service2	Service3	Little similar	0.4187195136565853	0
	Service2	Service4	Little similar	0.4239873343390204	0
	Service2	Service5	Averagely similar	0.6332345052356138	0
	Service2	Service6	Very similar	0.7899074233058608	0
	Service3	Service4	Averagely similar	0.6754026951205351	0
	Service3	Service5	Averagely similar	0.5522594031981931	0
	Service3	Service6	Little similar	0.46428017617071077	0
	Service4	Service5	Averagely similar	0.5736106485880657	0
	Service4	Service6	Averagely similar	0.5015862333019195	0
	Service5	Service6	Averagely similar	0.6792470780206274	0
					Error≈1%

TABLE VII. MEASUREMENTS COLLECTED IN BOOKS DOMAIN

Search book domain	Pairs of services		expert interpretation	Similarity Measurement	Errors
		Service1	Service2	Identic	1.0
	Service1	Service3	Very similar	0.832998750381563	0
	Service1	Service4	Little similar	0.39732173036521107	0
	Service1	Service5	Averagely similar	0.6356054701638942	0
	Service1	Service6	Little similar	0.4287498686598919	0
	Service2	Service3	Very similar	0.832998750381563	0
	Service2	Service4	Little similar	0.39732173036521107	0
	Service2	Service5	Averagely similar	0.6356054701638942	0
	Service2	Service6	Little similar	0.4287498686598919	0
	Service3	Service4	Little similar	0.4534630160857285	0
	Service3	Service5	Little similar	0.3496053193811293	0
	Service3	Service6	Little similar	0.4604549944415463	0
	Service4	Service5	Little similar	0.33387472124846296	0
	Service4	Service6	Little similar	0.5082926323728428	0.01
	Service5	Service6	Little similar	0.1651098158022917	0.04
					Erreur≈0,4%

Using measurements stored in the tables above and to compare our results with the results of existing studies we calculated the precision and recall of our method in all three assessment areas (table 8).

TABLE VIII. RECALL AND PRECISION MEASUREMENT

	Recall	precision
weather	100%	100%
sms	100%	83.5%
book	100%	100%

The average recall of our approach is 100% and the average precision is 95.16%, which proves that our method is very effective and it is very close to human interpretation.

V. RELATED WORKS

The similarity measurement between the web services is a very discussed subject in the literature, the existing works use different techniques and therefore differ in their performance.

In [15] authors use google Normalised distance to calculate the semantic similarity between two concepts, it is a statistical method based on results returned by the Google search engine and does not take into account the context of concepts in which we want to compute the similarity. They have ignored the structure of a web service that is for them a set of terms. The similarity between two Web services will be the total similarity between the two sets of terms that represent them. By comparing their recall and precision with the mine, it is found that our method has a higher performance than their method.

In [16] authors use at the same time several metrics to calculate the semantic similarity, and use several metrics to calculate the syntactic similarity. They do not use sense disambiguation of terms for which they want to calculate the similarity. In [16] the authors did not measure the precision of their method.

In [17] authors measure the similarity between two web services by measuring the similarities between the descriptions of the different concepts included in the wsdl file. But the majority of web service we found are not documented, which shows that this method is not very convenient. They use TFDIDF algorithm to calculate the similarity between terms that for us unreliable.

In [18] authors use the same approach as the work cited in [15] using several kinds of functions to evaluate a similarity matrix except that their method does not exceed 70% in precision and recall.

In [19] authors have ignored the names of the operations in the calculation of similarity, and they considered only the inputs and outputs of simple type, while the operations of a web service have often input and output with the complex type. The precision of their method in computing similarity between two web services interfaces does not exceed 65%.

VI. CONCLUSION

In our work we have proposed a profounder approach than existing work in calculating similarity between web services combining syntactic and semantic similarity. In the semantic

part we rely on the WordNet lexical base by applying Wu-palmer algorithm and disambiguation word sense algorithm and by using the Hausdorff distance and all that with the objective of improving the precision of the similarity.

Our method has achieved very high values for the precision and recall which proves that our method is very effective and it is very close to human interpretation.

The similarity computation is not always sufficient. At invocation stage, the application using the defective web service must replace its operations with those of the similar operational web service, so it will be necessary to detect the correspondence between the operations of substituted web service and substituent web service. So our future work will be to exploit the results obtained in this paper to realize the mapping (correspondence) between two similar web services.

REFERENCES

- [1] William W. Cohen, Pradeep Ravikummar, and Stephen E. Fienberg, "A Comparison of String Distance Metrics for Name Matching Tasks", Proceedings of IJCAI-03 Workshop on Information Integration, page 73–78, August 2003.
- [2] Alvaro E. Monge, and Charles P. Elkan, "The Field Matching Problem: Algorithms and Applications", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining KDD, page 267–270, 1996.
- [3] Jaro, M. A., "Advances in record linkage methodology as applied to the 1985 census of Tampa Florida", Journal of the American Statistical Society, vol. 84, no 406, p. 414-420, 1989.
- [4] M. A., "Probabilistic linkage of large public health data file", Statistics in Medicine, vol. 14, p. 491-498, 1995.
- [5] Winkler, W. E., "The state of record linkage and current research problems", Statistics of Income Division, Internal Revenue Service Publication R99/04, 1999.
- [6] Winkler, W. E., "Overview of Record Linkage and Current Research Directions", Research Report Series, RRS, 2006
- [7] Christiane Fellbaum, "WordNet: An Electronic Lexical DatabaseReferences", Ed. Cambridge: MIT Press, 1998.
- [8] Alexander Budanitsky, Graeme Hirst, "Semantic distance inWordNet: An experimental, application-oriented evaluation of five measures", in workshop on wordnet and other lexical resources, second meeting of the north american chapter of the association for computational linguistics, 2001.
- [9] Warin, Martin, "Using WordNet and Semantic Similarity to Disambiguate an Ontology", D-Level Thesis, University of Stockholm.
- [10] Wayan Simri Wicaksana, and Bambang Wahyudi. "Comparison Latent Semantic and WordNet Approach for Semantic Similarity Calculation", CoRR, 2011.
- [11] Leacock, M Chodorow, "Combining local context and WordNet similarity for word sense identificationC" "WordNet: An Electronic Lexical DatabaseReferences", Ed. Cambridge: MIT Press, 1998.
- [12] Z Wu, M Palmer, "Verbs semantics and lexical selection", Proceedings of the 32nd annual meeting on Association for Computational Linguistics(1994).
- [13] S Banerjee, T Pedersen, "An adapted Lesk algorithm for word sense disambiguation using WordNet", Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, Pages 136-145, Springer-Verlag London, UK (2002).
- [14] Marie-Pierre Dubuisson and Anil K. Jain, "A modified Hausdorff distance for object matching", In Proceedings of 12th International Conference on Pattern Recognition, pages 566-568, Jerusalem, Israël, october 1994.
- [15] Jeberson Retna Raj, "web service discovery based on computation of semantic similarity distance and qos normalization", Indian Journal of Computer Science and Engineering, pages 566-568 Vol. 3 No. 2 Apr-May 2012.

- [16] Okba Tibermacine, Chouki Tibermacine and Foudil Cherif, "WSSim: a Tool for the Measurement of Web Service Interface Similarity", in proceedings of the french-speaking conference on Software Architectures (CAL'13), Toulouse, France. May 2013.
- [17] Natallia Kokash., "A Comparison of Web Service Interface Similarity Measures", STAIRS.,Pages 220-231, 2006.
- [18] Fangfang Liu, Yuliang Shi, Jie Yu, Tianhong Wang, Jingzhe Wu , "Measuring Similarity of Web Services Based on WSDL" , In proceeding of: IEEE International Conference on Web Services, ICWS 2010, Miami, Florida, USA, July 5-10, 2010 .
- [19] Jian Wu, Zhaohui Wu, "Similarity-based Web Service Matchmaking", In proceeding of: Services Computing" , IEEE International Conference on, Volume: 1, 2005.

Role of Knowledge Reusability in Technological Environment During Learning

O. K. Harsh

Group Director,
Amritsar College of Engineering & Technology,
Amritsar, India

Abstract—Role of technology and reusability on the knowledge management and knowledge transformation has been analyzed by considering the extended model of Nonaka and Takeuchi which includes the knowledge reuse in the three dimensional environment. Knowledge transformation has been further refined (and boosted) to get the more qualitative and quantitative knowledge by applying the concept of knowledge reification, indexing and adaption. By extending these concepts and related processes, ADRI quality model on higher education learning has been analyzed. Present work suggest that reusability along with above mentioned concepts during ADRI cycle can boost the qualitative knowledge in higher educational setting and observed that ADRI model has the similar trends as that of Nonaka model in the three dimensional environments. In addition, discussion also prevails that best practices required in a higher educational setting correspond to ADRI model.

It has been suggested that time along with reusability supports to the tacit as well as explicit knowledge management during learning. The knowledge transformation achieved this way is more qualitative. Finally it can be concluded that tacit and explicit knowledge required to reuse is an important aspect now days in managing higher educational knowledge in a fast growing contemporary environment provided knowledge is exploited appropriately.

Keywords—Knowledge; Reuse; Tacit Knowledge; Explicit Knowledge; Technology; ADRI model; Quality

I. INTRODUCTION

Managing knowledge has been becoming a competitive advantage in a global economy. Many companies have been engaged in identifying, managing and sharing experience of employees [1, 2]. Most of the companies involved in the electronic management of knowledge so that they can enhance their capability to manage vast knowledge hidden within the organization [3]. However, we now realize that managing knowledge is more difficult than previously thought. Krogh et al. [4] studied the knowledge reuse in open source software projects. They found that behaviour factors are responsible for the form of the knowledge reuse.

According to Hedlund [5], there are four different levels of carriers of knowledge which are individual, the small group, the organization and the inter-organizational domain (important customers, suppliers competitors, etc.). Hedlund [5] further suggested that each carrier is having different forms or aspects of knowledge egg, cognitive knowledge in the form of mental constructs and precepts, skills and knowledge embodied

in products or well-defined service or artifacts. Kusunoki et al. [6] applied the concept of multilayered knowledge to describe organizational capabilities. According to them there are three different layers of knowledge each of them provides different types of capability. They described that “knowledge is the layer that includes distinctive individual units of knowledge (e.g. functional knowledge embodied in a specific group of engineers, databases, patents, etc.)”.

Research of Kusunoki et al. [6] describes the features of the managerial potential possessed by every layer which may be observed in two dimensions. First one describes the modularity of the organizational potential, capturing whether organizational potential are supported on individual knowledge components and join every component of knowledge while the second dimension describes the design skill or manageability of organizational capabilities. “This dimension which symbolizes the ‘designable knowledge’ versus ‘embedded knowledge’ focuses on whether the management can directly design and control the capabilities” [6].

Work of Nonaka and Konno [9] suggested that there is a gap for knowledge construction in organizations. They mentioned that this gap “can be thought of as a shared space for emerging relationships”. According to Nonaka and Konno [9] this space can be either physical (e.g. office, dispersed business space), virtual (e.g. e-mail, teleconference), mental (e.g. shared experiences, ideas, ideals) or any combination of these”. They established four diverse forms, which fit into each phase of the SECI model (SECI stands for Socialization, Externalization, Combination and Internalization, which are the familiar four diverse forms of knowledge translation [10]).

Software reuse and software knowledge reuse is an important aspect to describe quality and productivity. Under such processes reuse is a complex procedure and during this procedure it is extremely tricky to choose or forecast suitable kind of metrics to permit an organization to get optimum advantage. It should be noted that arrangement of software component reuse should be suitable to every stage of complete life cycle of the software growth. This can allow us to understand the power and weak points in our knowledge leveraging ability. Technology used in each phase of software development or knowledge transformation can facilitate the process of reuse provided we know that how the reusability varies in the space.

Important question is that what should be the job of the technology for the learners particularly in the conversion of

one kind of knowledge into (tacit to explicit) a different kind and vice versa as a result of knowledge reusability? How the quality of data, information and knowledge can be enhanced in the Nonaka Model [7, 8] particularly in higher educational learning environments by applying the concept of knowledge reuse, reification, indexing and adaption etc. Another issue is the time obstruct in the [7, 8] Nonaka model which is the issue that requires to be suitably answered in the rapid varying technological humanity. An interesting question at this stage could be that how various processes in the refined Nonaka model are useful for higher educational environment? None of the answers of the above questions are available therefore it has decided to work on a framework to correlate all these aspects.

II. KNOWLEDGE MANAGEMENT AND TECHNOLOGY

The utilization of technology to administer and hunt compilation of explicit knowledge is well recognized. A case is to employ text categorization to allocate documents mechanically to a topic plan. A characteristic task may be to present a manuscript into a joint database.

It is a fact that technology like the World Wide Web, GPS etc. can enormously boost the allocation of knowledge both inside and outside of organizations. However knowledge management implies further than what we call databases and networks. Companies applying such means have observed that only 20 percent of their complete efforts occupy technical concerns; the residual 80 percent of their time is exhausted with their institutional problems.

Knowledge can only be managed adequately if the issue like technology, human resource practices, organizational structure and culture all are considered together. Because of this fact that we desire to make sure that the correct knowledge is brought to allow at the correct time. The majority of the companies attempt to execute modern technology but then discover that variation of knowledge with time as a result of technological developments is tricky to understand.

The objective of present work is to uncover the processes of reuse, information transfer, coordination and transformation of knowledge by applying Information Technology, internet and associated tools with the variation of time. The issue of quality of knowledge as a result of reuse, reification, indexing and adaption will also be taken into account during learning. For this, models of knowledge management in three dimensional environments as well as an extended ADRI (Approach, Deployment, Result and Improvement) quality model in higher educational environment have been considered where reuse is measured as an explicit quantities. For this we require to begin with the existing knowledge management and knowledge reuse models in the literature such as Nonaka [10] and Nonaka and Takeuchi [11] and Harsh [12-15, 26-28]. In other words the ADRI quality model [18, 22] of higher educational environments has been extended as well elaborated by applying the concepts associated with the revised Nonaka model from two dimensions to three dimensions to uncover useful knowledge transformations and reuse.

Support to the configuration and communication of tacit knowledge (including its reuse), and allow it for transforming into explicit knowledge are currently not a great deal, though

numerous optimistic modification may be observed, such as the employing of text-based chat, expertise location, and unrestricted bulletin boards.

To deal the affect of technologies, it is better to categorize the technologies by means of orientation of tacit and explicit knowledge as initiated by Polanyi [16] and employed by Nonaka [10, 11]. This creates a hypothesis of organizational learning. This hypothesis connects the revolution of knowledge amid tacit and explicit shapes. Tacit knowledge implies that that what knower recognizes, which is supported on the practice and consists of reliance and value. Reuse of tacit knowledge is simple and manageable. Tacit knowledge behaves like as an act, and therefore it is extremely significant.

One should remember that tacit knowledge is the largely significant cause for the creation of novel knowledge, because according to Nonaka [10]: “the key to knowledge creation lies in the mobilization and conversion of tacit knowledge.”

Explicit knowledge may be signified by some sort of artifact for instance text or a video, which has typically been generated with the objective of communicating with a different individual.

In Knowledge Management, there are three major components [17] which are People, Processes and Technology. People are accountable for creating, sharing, and using knowledge, and who jointly comprise the organizational civilization that takes care for and inspires for the sharing of knowledge. Processes are accountable for acquiring the techniques and creating, organizing, sharing and transferring knowledge while Technology is responsible for the tools for accumulating and supplying right to use data, information, and knowledge generated by people in a variety of settings. Reuse of all these three major components (People, Processes and Technology) is the key to success in a contemporary organization.

III. DETAILED FORMULATION

A. Nonaka and Takeuchi Knowledge Management Model

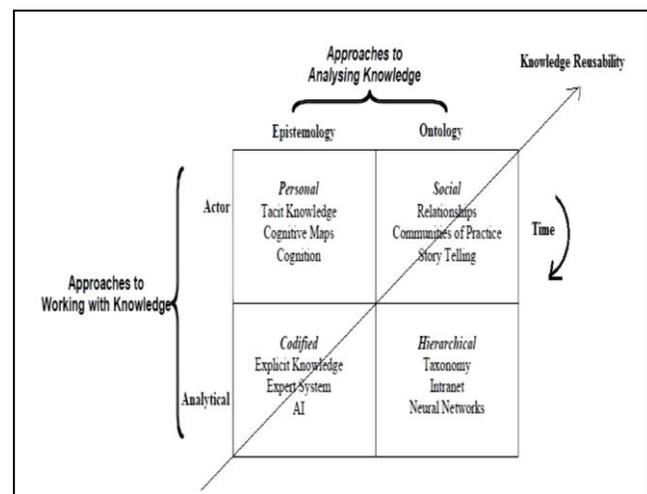


Fig. 1. Knowledge Management Matrix [references 7, 29] based on Nonaka and Takeuchi Model (1995) or SECI Model (Revised by Present Author [References 13-15]).

Nonaka and Takeuchi [10] described the conversions of tacit to explicit knowledge during the socialization, externalization, combination and internalization processes. They mentioned that tacit to explicit and explicit to tacit knowledge is constantly moveable during one or extra procedure in an institute. Moteleb and Woodman [29] submitted a Knowledge Management Model based on the Nonaka model which has been further revised by the present author by inserting reusability and time [15] (see Fig.1). This model identifies the Approaches of working with knowledge and to analyzing knowledge. Likewise, tacit to tacit and explicit to explicit knowledge transfer in a similar way. According to them there is a continual learning procedure in an organization during which knowledge amplified similar to a spiral. Therefore the knowledge is supplemented when shared like Fig. 1. Knowledge relocates amid individuals through information (externalized) “and then converted back from information back to knowledge” (Nonaka and Takeuchi Model [10]).

B. Technology and Proposed Three Dimensional Model based on Nonaka and Takeuchi Model [10]

In this article it is being suggested that relocation of knowledge definitely takes time and valuable knowledge of an institute boost up during the reuse of knowledge [12-15]. Time behaves as a critical feature for any institute to gather information, to systematize knowledge, to relocate knowledge from one shape to another in constructive shape, therefore the institute has extra knowledge. Harsh suggested that [12-15] time should be measured through an additional axis in Fig. 1 of Nonaka and Takeuchi Model [10]. Enhancement of time improves knowledge in a three dimension (see Fig. 2 below) (through translation from information) similar to a solid cone as mentioned in reference [15]. Such cone consists of all kind of knowledge of conversion processes as suggested by Nonaka and Takeuchi Model [10] (such as during socialization, externalization, combination and internalization processes).

Both explicit and tacit knowledge can be reused. Therefore we require another axis to symbolize the knowledge reusability.

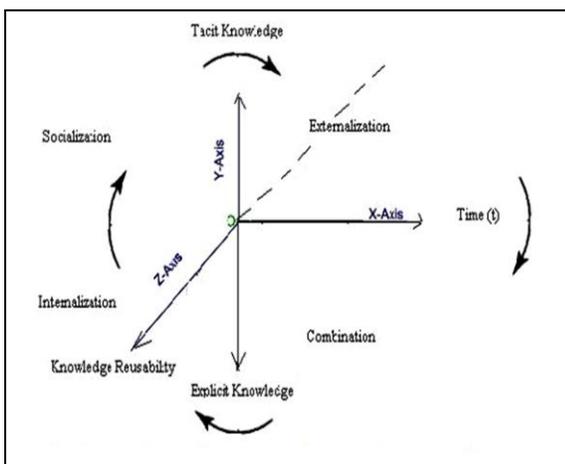


Fig. 2. Extended Nonaka and Takeuchi (1995) Model taken from reference [15].

The successful knowledge of an organization considered to be enhanced with the time since each instant we add further knowledge (it may be tacit or explicit or both) due to new ideas, new concept or new interpretation. Due to knowledge reusability we get qualitative knowledge, however, it reduces the effective quantity of entire knowledge.

Similar to Nonaka and Takeuchi Model [10], we can describe these conversion processes in three dimensional environments by the Fig. 2 above [12-14, 23]. There are some processes in Fig. 2 which are being discussed in the next paragraphs.

Socialization which is the procedure of dealings between individuals occurs during the knowledge sharing (Nonaka and Takeuchi Model [10]) and increases due to reuse of experiences (over the time), mental models and beliefs by employees. This manner the knowledge exists in public's minds augmented due to transformation of knowledge over the time. Thus here the tacit knowledge is converted into another tacit knowledge. Moreover, due to reusability the quality of knowledge also increases with the time as a result of refinement.

Face-to-face meetings are the burning example of building tacit knowledge which is shared by people. Shared experiences are the example of informal knowledge where role of information technology is nominal. However, concept of groupware is adopted during large number of on-line meetings and interpersonal exchanges. These techniques have been using either to poise usual gathering, or to some extent alternate means.

Groupware: Two significant effects are imperative to appreciate the deployment of groupware during socialization. They are shared skills and belief. Shared skills are vital for the tacit knowledge sharing. In the current three dimensional model, shared space will be more with extra choice of reusable knowledge. This reusable knowledge increases with the increase of virtual space. Thus it helps to enhance the tacit knowledge. Reusable knowledge boosts more faith because it is a provable knowledge. Lotus Notes is a type of a groupware and may be employed for sharing of documents. Groupware also assists in the sharing of explicit knowledge.

Externalization which is the method of detaining or capturing information about knowledge (Nonaka and Takeuchi Model [10]), for example communicating to someone, writing a manuscript, sketching a Figure, demonstrating a presentation, or instructions, will be faster as a result of reuse of knowledge with the time. Therefore there will be extra knowledge accessible to institute as contrast to the Nonaka and Takeuchi Model [10].

IV. RESEARCH PROBLEM

Objective of present work is to demonstrate the role of knowledge transformations and reuse in the three dimensional educational environments considering time as a factor in the Nonaka model by utilizing:

- How the quality and quantity of knowledge is affected based on the students' profile?

- How the students' can adjust the knowledge in terms of availability of reusable knowledge?
- How the Analogies between Nonaka and ADRI model exist?
- How the ontology is responsible for the learning?
- How in the theory the acquisition of reusable knowledge takes place?
- Can we consider ADRI model as equivalent to revised Nonaka model for learning?
- In what way present ADRI model is more useful for students?

V. CONTRIBUTION: REUSABILITY AND PROPOSED KNOWLEDGE REUSE PROCESSES IN AN EDUCATIONAL SETTING

Authors [24] have submitted a method for knowledge reuse in communities of practice of e-learning in which they proposed the process of reuse of knowledge within the communities of practice by the two types of sub-processes: namely the process of reification means transformation of tacit and elicited knowledge to a novel knowledge that is elicited knowledge, and secondly by the process of indexing by applying earlier used knowledge further (as a reusable knowledge).

Using their concepts and revised Nonaka model as discussed earlier including the role of technology, present author (Fig. 3 and Fig. 4) is suggesting an interesting (revised) knowledge reuse process for educational environment.

If someone wants to reuse the knowledge then he or she will have to personalize it for his or her respective educational environment.

Thus one has to adapt as per his or her own profile. If a learner wants to concentrate on his or her educational activities then knowledge is to be shared or captured accordingly. According to authors [24] we can represent a generic approach and hence a generic profile model that can be applied to any persons as well as to group in the environment of communities of practice.

Since this is the problem of tacit and explicit knowledge transformation or exploitation therefore we can apply this approach to the educational environments where knowledge transfer is possible at individual as well as at group levels.

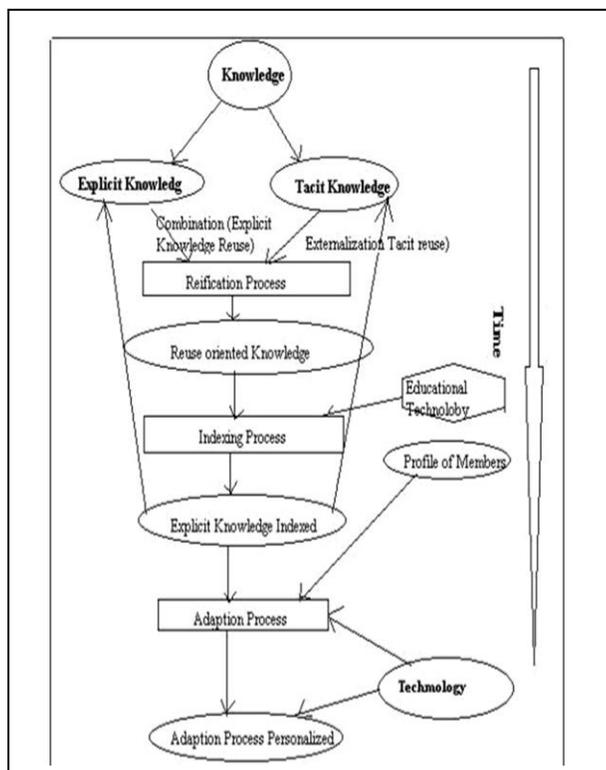


Fig. 3. Proposed Knowledge Reuse Process in an Educational Institute

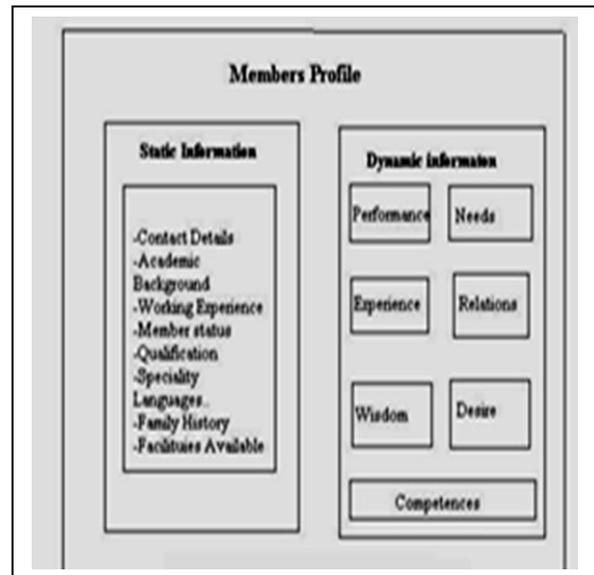


Fig. 4. Students Profile Model

Fig. 3 suggests a reification process as an approach which is responsible for deploying reusable knowledge as result we get reusable indexing process which can be further improved to achieve explicit as well as tacit knowledge. Finally the explicit or tacit knowledge can be adapted to get the personalized relevant knowledge by the students. Thus the needs of students can be personalized by the repetitive process.

Thus similar to authors [24], in the present problem one can assume two types of information in educational environments, egg static information and dynamic information (Fig. 4). Static information includes individual information like contact details, his or her academic background, qualifications, type of work experience etc while dynamic information consists of personalized behavior during individual participation in educational activities which includes knowledge sharing and knowledge capturing activities. Similar to these we propose seven dynamic elements (See Fig. 4) which can be used in educational environments, namely, needs, preferences in connection to the knowledge being related to existing resources. Other elements are relations (how to relate with a group or other learners); learning (learning scenario for the explanation), experience (learners' know-how ability),

competences (related to cognitive attribute), wisdom (ability of making judgment) and desire (aspiration for the task) (see Fig. 4).

VI. ADRI HIGHER EDUCATION QUALITY MODEL AND REUSABILITY

To appreciate the idea of knowledge reusability for learners, an application supported by ADRI model (Approach, Deployment, Result, and Improvement) [18] is being suggested which is recognized for the quality assertion and improvement properties in higher educational settings. It is being suggested that our revised ADRI model is similar to three dimensional Nonaka [8] model which not only consists tacit and explicit knowledge conversion processes with time; while all above mentioned knowledge reification and related processes can also be considered. It is noteworthy to quote that ADRI model is a general tool for appraisal and development [18] in many means. Current author have previously been engaged on ADRI model in explaining the varieties of features of knowledge management [19, 20]. To deal with the present problem work of Jantti [21] has been considered which accounts the ADRI model and has the following four approaches (Fig. 5):

- An Approach means how to relate and imagine about the mapping
- Deployment means how to correlate to execute
- Results (performance) means how to observe and assess and
- Improvement means how to relate with further rectification and adjustment.

Similar to the approach to Fig. 1 and Fig. 2, present author suggesting first time Fig. 5 (see below) which consists of tacit and explicit knowledge in opposite directions as well as knowledge reusability orthogonal to both tacit and explicit knowledge.

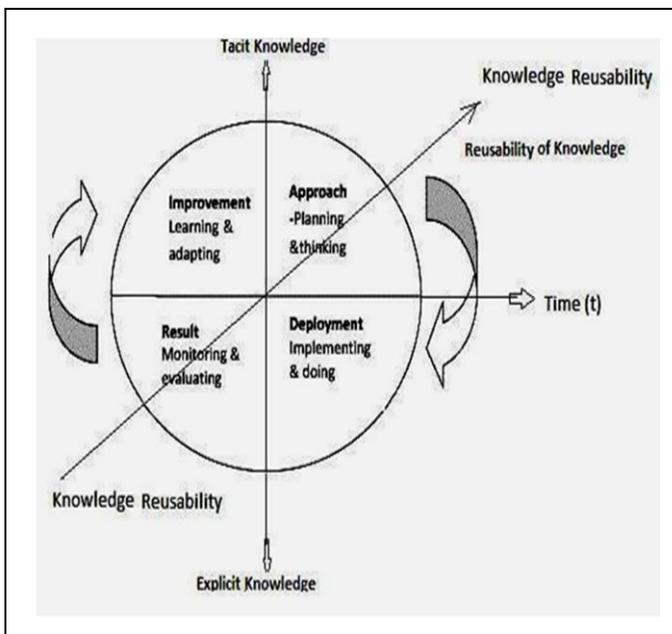


Fig. 5. Revised ADRI Model

Here we need to take a decision that what approach is to be taken up? It typically relates with expansion of goals, policies,

The first stage in the ADRI model is Approach which usually consists of imagining and scheduling jobs and therefore it corresponds to conversion from tacit to tacit principles, outcomes, map and aims. Present suggestion is that setting up of knowledge and its reuse should obviously identify qualitative and quantitative aims and can state path to achieve the aims [18].

The second stage (Fig. 5) is operation or Deployment which tenders a dais to execute or understand tasks. It is just the combination process in the Nonaka model which is the stage that joins existing knowledge with novel produced knowledge.

Here the present author would like to suggest that presence of reusability of knowledge (with time) will generate and joins more knowledge as the time enhances and hence conversion from explicit to more focused explicit knowledge takes place.

People in the organization feel more linked with knowledge in less time. Consequently it is clear that suitable arrangement can be build at the principal step to comprehend the purpose of knowledge management and reuse [18].

The third stage (Fig. 5) is the consequence or result, which signify yield or outcome as a result of the first and second stages of knowledge administration and reuse as stated above. We should remember that important thing is that output (result stage) and goal (approach stage) should be connected mutually. Result's stage presents assessment between predictable yield and achieved yield. Present statement is that result will be more qualitative due to the application of reusable knowledge. Technology also play here significant role in the deployment because of type of tool used for displaying the results as well as reusing the data, information and knowledge. This step can offer us a possibility to inspect yield (output) and demonstrate conclusions.

The fourth and final stage is enhancement or improvement which exhibits the conclusions achieved from the above results and analysis stages. This stage recommends that what is to be needed for enhancement or improvement [22]? As outlined earlier, ADRI model is a continual progression of improvements; consequently the technique of added improvement in knowledge administration and reuse depends on the subsequent achieved ADRI cycle. Since in the earlier phase we were involved in converting tacit knowledge into explicit while in this phase we have an opportunity to transform explicit to more explicit knowledge. Thus here we are getting more focused explicit and qualitative knowledge. Here results will be highly useful and interesting if we exploit the features of Fig. 3 and Fig. 4 into the ADRI model of Fig. 5. Thus as a method of reification, indexing and adaption (by classifying into static and dynamic knowledge), a process of systematic and comprehensive learning can be developed which will be much more adequate as compared to considering only ADRI model.

In the current research we have suggested to recognize the task of the technology and reusability and related issues on the three dimensional Harsh as well as ADRI model by having the

essential features of Nonaka and Takeuchi [10] model. Our models are not only enhancing the effective knowledge of the organization whereas it helps us in picking suitable issues such as:

- Categorizing inaccurate knowledge pertinent to the technology.
- Relate knowledge in an exacting technological state because we are more certain about the kind of knowledge which is convenient to technology.
- Categorizing and unraveling tacit and explicit knowledge further quicker as well as opportunity to translate tacit to explicit knowledge (and vice versa) speedily by the suitable employment of technology.
- Estimating the useful knowledge of the organization positively and recurrently as a result of straight effect of technology.
- Customers can have more choices to select the applicable technology appropriate to the wanted knowledge.
- It offers quicker knowledge relocation once we become sure about the technological strictures.
- Reification process which makes the knowledge exactly reusable and accountable.
- Indexing process to make enable knowledge reuse and sharing because students have a common vocabulary which could demonstrate the concrete concepts of students in learning environments. In order to demonstrate this vocabulary we need ontologies [25].
- Throughout adaptation process reuse purpose can be reassured along with the information of students' profile.
- Presence of indexing process at an early stage results into available reusable knowledge in the beginning.
- Fig. 6 can identify the actor and analysis methodology in conjunction with Nonaka model which further facilitates the learning matrix.

VII. KNOWLEDGE MANAGEMENT IN SOFTWARE ENGINEERING AND OUR PRESENT MODELS

Since the software process is a complex, cooperative and incessant improvable procedure, therefore People, technology, organization and connected measures are the most significant features of software making and growth. Quality is not merely a concern for the software expansion while at the same time processes involved in software creation are likewise significant. Due to the insertion of knowledge reusability and time, it becomes extra intricate to choose the appropriate technology. Computer based devices can resolve the intricacy of the software processes to some degree. Our three dimensional representation can improve the opportunity of enhanced understanding of the skills of software development since identification, indexing and adaption of reusable components can shape the quality of the organizations.

VIII. LATEST WORK ON BEST PRACTICES IN HIGHER EDUCATION

Butnariua and Milosanb [7] outlined and argued the professional abilities, skills, training and competences necessary in line to realize best practices and to thrust onward the development in knowledge management within Higher Education. They proposed a diagram through which they could represent the requirements of a methodology for change in knowledge, attitude, and or behavior (See Fig. 6). In the present work it has been shown that this Fig. works similar to ADRI model as discussed above like Approach, Deployability, Results, Improvements and arrow of Feedback.

Fig. 6 suggest that the ADRI model already included in the best practices which could be desired at a University level.

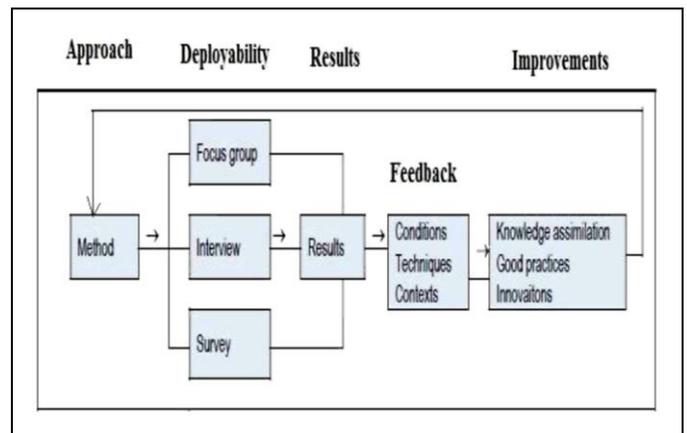


Fig. 6. Requirements of a methodology for change in knowledge, attitudes and or behavior (Reference 7)

As the authors suggested [7], a university may be the source of creating skills and capability through information and coordinate with the help of activities like Fig. 6 which fulfill are our requirements. Off course application of these activities involve the intellectual psychology. We must be aware of the fact that the teaching activities involve an intellectual analysis and imaginative components which should be as per needs of the society [7].

Thus approach of construction and humanizing the excellence of the education movement comprise a standard of continuous innovation as the time changes.

IX. CONCLUSION

In the present work author has discussed knowledge management model which involves reuse and time as an orthogonal to each other. It has been suggested logically that how tacit and explicit knowledge (including reusable) advanced and translated into non-reusable and reusable improved knowledge by the application of the suitable consideration of reification process, indexing, adaptation and off course application of technology. We appreciate that further helpful reusable knowledge may be exploited if one can co-relate amended knowledge management model (which consider the explicit role of reusability and technology) to the software expansion case studies or plans.

As a result of novel three dimensional model extra space and knowledge will not only be accessible while new knowledge (including reusable knowledge) is exchangeable from explicit to tacit and vice versa. Ultimately added choice of new kind of tacit and explicit knowledge will be obtainable. It should be noted here that the role of the technology such as visualization instruments, browsable video/audio of presentations are some of the fine examples which could be appropriately exploited for the learning in a reusable environment.

As a consequence of application of reusability of knowledge in a specified technological environment, not merely quality of knowledge while the efficiency will also be improved. Therefore by the appropriate amalgamation of metrics with the technology in a three dimensional environment, reusability may be achieved to an immense amount. In addition, best practices in higher education environment has also been briefly discussed which reflects the use of ADRI model. Thus the current research can help us in improved way in the deployability and accomplishment of the reusable components. We can think in the future to have an entire new technological structure which could play a vital part in picking the needed reusable knowledge components.

ACKNOWLEDGMENT

Author would like to thank to his institution's Chairman cum Chief Executive Officer Shri Amit Sharma (Amritsar College of Engineering & Technology, Amritsar) for supporting this research work by providing a good research environment and related facilities.

REFERENCES

- [1] G. M. Steyn, "Creating Knowledge Through Management Education: A Case Study of Human Resources" Management. Education, vol. 123(3), pp. 514-531, 2002.
- [2] M. Martensson, "A critical review of knowledge management as a management tool," Journal of Knowledge Management, vol. 4(3), pp. 204-212, 2000.
- [3] K. Eginton, "Knowledge management-law firms can do it too. AustralianLaw Librarian", vol. 6, pp. 247-255, 1998.
- [4] G. von Krogh, S. Spaeth, S. Haefliger, "Knowledge Reuse in Open Source Software: An Exploratory Study of 15 Open Source Projects," System Sciences, 2005 [HICSS '05. Proceedings of the 38th Annual Hawaii International Conference pp.198b,198b, 03-06 Jan. 2005].
- [5] G. Hedlund, "A Model of Knowledge Management and the Nform Corporation." Strategic Management Journal vol. 15, pp. 73-90, 1994.
- [6] K. Kusunoki, and I. Nonaka, Akiya Nagata "Organizational Capabilities in Product Development of Japanese Firms: A Conceptual Framework and Empirical Findings." Organizational Science vol. 9(6), pp. 699-718, 1998.
- [7] Monica Butnariu and Ioan Milosan, Best practices to increase progress in knowledge management Procedia - Social and Behavioral Sciences 62, pp. 739 - 743, 2012.
- [8] I. Nonaka, "The Knowledge Creating Company," Harvard Business Review 69, 96-104 (November-December 1991).
- [9] I. Nonaka and N. Konno, "The concept of 'Ba': building a foundation for knowledge creation", California Management Review, Vol.40, No.3, Spring, pp. 40-54, 1998.
- [10] I. Nonaka and H. Takeuchi, The knowledge-creating company. How Japanese companies create the dynamics of innovation, Oxford University Press, Oxford, 1995.
- [11] I. Nonaka, "A dynamic theory of organizational knowledge creation", Organization Science, Vol.5, No.1, February, p. 14, 1994.
- [12] O. K. Harsh, "Data, Information and Knowledge & Reuse Management Techniques", World Congress in Engineering held in London from July 2 to July 4, 2007.
- [13] O. K. Harsh, "Explicit Knowledge Management and Reuse", Presented in KMO workshop held in Lecce, Italy on Sept 10-11 2007.
- [14] O. K. Harsh, "Reusable Data, Information, Knowledge and Management Techniques", Journal of Knowledge Management Practice vol 9(3), 2008.
- [15] O. K. Harsh, "Three Dimensional Knowledge Management And Explicit Knowledge Reuse" Journal of Knowledge Management Practice, Vol.10, No. 2, June 2009
- [16] M. Polanyi, "The Tacit Dimension," Knowledge in Organizations, L. Prusak, Editor, Butterworth-Heinemann, Woburn, MA 1997.
- [17] R. Bose, "Customer Relationship Management: Key Components for IT success", Industrial Management & Data Systems, vol. 102 (2), pp. 89-97, 2001.
- [18] B. A. Abuid, ADRI – Self assessment model for teaching and learning, presented at 2nd international conference on global trends and challenge in higher education and quality assurance, 12-13 June, 2010, Oman.
- [19] A. Alani, O. K. Harsh, and S. Iqbal, Role of Information and Communication Technology on Knowledge Management in a higher educational environment, presented at International Conference on Information and Communication Technology, March 22-23, 2009, Middle East College of Technology, Muscat, Oman.
- [20] A. Alani, O. K. Harsh, and S. Iqbal, "Qualitative Knowledge Management and Knowledge Reuse in Higher Educational Setting", presented at the second international conference on quality and higher education, Muscat, Oman, June 12 to June 13, 2010.
- [21] M. H. Jantti, Minding your own business: can a business excellence Framework translate to the education sector? in Proc. Quality conversations on the Annual Higher Education Research and Development Society of Australasia Conference, presented at 25th Annual International HERDSA Conference, Perth, 7-10 July, 2002.
- [22] Quality Frameworks. (2006). Reflections from Australian Universities. edited by Jeanette Baird. Australian Universities Quality Agency. [Online]. Available At: <http://www.auqa.edu.au>. Accessed on 10th Sep 2009.
- [23] O. K. Harsh "Knowledge Reuse and Management in the Information Systems", A thesis submitted for the degree of Doctor of Philosophy of the University of New England, March, 2010.
- [24] Lamia Berkani and Chikh, Azeddine, A Process for knowledge reuse in communities of practice of e-learning. Procedia Social and Behavioral Sciences Vol.2, pp. 4436-4443, 2010.
- [25] T. R. Gruber, A Translation Approach to Portable Ontology Specifications. In Knowledge Acquisition, Vol. 5(2), pp. 199-220, (1993).
- [26] O. K. Harsh. and Sharma, Sanjiv, Software Management and Reuse: Knowledge Perspective, International Journal of Software and Web Sciences (IJSWS), Vol.1 (2), 50-53, 2013.
- [27] O. K. Harsh and Banga, Rainu, Qualitative Knowledge Management and Reuse in Software Engineering Environment, International Association of Scientific Innovation and Research (IASIR), Vol.6 (1), 14-17, 2013.
- [28] O. K. Harsh, Qualitative Knowledge Management and Reuse in Software Engineering Environment, International Journal of Software and Web Sciences (IJSWS), 14-139, Vol. 7(1), 52-56, December 2013-February 2014.
- [29] Aboubakr, A. Moteleb, and Mark Woodman, Notions of Knowledge Systems: A Gap Analysis, Electronic Journal of Knowledge Management, Vol. 5 (1), 55-62, 2007.

Neural Network Based Lna Design for Mobile Satellite Receiver

Abhijeet Upadhyaya
M.Tech. Scholar, E.C.E
Ajay Kumar Garg Engineering College,
Ghaziabad, India

Prof. P. K. Chopra
Head of Department, E.C.E.
Ajay Kumar Garg Engineering College,
Ghaziabad, India

Abstract—Paper presents a Neural Network Modelling approach to microwave LNA design. To acknowledge the specifications of the amplifier, Mobile Satellite Systems are analyzed. Scattering parameters of the LNA in the frequency range 0.5 to 18 GHz are calculated using a Multilayer Perceptron Artificial Neural Network model and corresponding smith charts and polar charts are plotted as output to the model. From these plots, the microwave scattering parameter description of the LNA are obtained. Model is efficiently trained using Agilent ATF 331M4 InGaAs/InP Low Noise pHEMT amplifier datasheet and the neural model's output seem to follow the various device characteristic curves with high regression. Next, Maximum Allowable Gain and Noise figure of the device are modelled and plotted for the same frequency range. Finally, the optimized model is utilized as an interpolator and the resolution of the amplifying capability with noise characteristics are obtained for the L Band of MSS operation.

Keywords—Satellite; Mobile; Artificial Neural Networks; Scattering Parameters; Noise Figure

I. INTRODUCTION

Historically, Mobile Satellite Systems (MSSs) have been used in Marine and Military Communications for providing the seamless connectivity to mobile users in regions where no other means of network can be established. MSS have also been successfully utilized in disaster management situations when all other forms of communication systems fail to respond. Recently, Hybrid Mobile Satellite Systems (HMSS) [1], [2], [3] have been opted in several countries due to the possibility of retrieving the combined advantages of terrestrial Cellular networks and Satellite Systems. Cellular Systems provide high bandwidth, low latency data and voice services. On the other hand, use of Satellite Systems extends the geographical coverage of the overall network. MSS systems like Thuraya, ACeS, and INMARSAT etc. have contributed a lot in the development of hybrid network by coordinating with cellular vendors across the globe.

Choosing the right equipment with right service is the key to growth in mobile industry. Dual Mode Sat Phones combine the functional benefits of the cellular-based equipment with the advantage of reliable and uniform coverage provided by extended foot print of the Satellite. Mobile Terminal must facilitate a seamless handover between the two types of communication systems as and when required.

The main aim of the present work is to obtain insights about Satellite links and their availability issues. Henceforth, a

Low Noise Amplifier's model is created and optimized using Artificial Neural Networks (ANN) based learning in the frequency range starting from 500 MHz to 18 GHz. The advantages of Neural Networks (NN) when compared to conventional methods lies in the ease with which they may be modelled to generalize any nonlinear relationship between variables. NN models have existed over six decades after McCulloch and Pitts [4] presented the model of the basic human nervous system and its basic unit termed a neuron. But implementation of models based on ANN approach in design of a microwave devices have been reported only in the last decade and a half [5], [6], [7].

Modelling issues encountered using other methods such as excessive time consumption, required level of designer's expertise, etc. disappear due to Universal Approximation capability of the of the NN [8]. One of the best works ever presented in literature in [9], the authors have demonstrated a step by step method of how NN may be obtained for understanding the behavior of a Radio Frequency (RF) device. This research work we have used Agilent ATF 331M4 pHEMT made of InGaAs/InP material.

The remaining of the paper is segmented in the following manner. Section II carries a review of the various degradations that the Satellite signal suffers. Section III and IV provide RF receiver and LNA design issues while section V discusses ANN modelling in brief. Finally, simulated results are presented in section VI and VII.

II. SATELLITE LINK IMPAIRMENTS

To incorporate a low cost, high spectrally efficient wireless communication system, a complete description of all the propagation impairments is necessary. Satellite signal consists of a direct component, called Line of Sight (LoS), and multiple copies of this direct component known as multipath components. In [10], the authors have modelled the received Satellite signal as

$$C(t) = A_0 e^{j(\omega_0 t + \varphi_0)} + \sum_{m=1}^M A_m e^{j(\omega_m t + \varphi_m)} + n(t) \quad (1)$$

Where A_0 , ω_0 , φ_0 are amplitude, Doppler shift and phase of the LoS component, while summation part quantifies the ' M ' multipath components and $n(t)$ is the White Gaussian Noise.

According to equation (1), the signal received by the mobile handset receiver comprises of a dominating LoS and several other components shifted in frequency and relative phase. Following are some link degrading factors:

A. Free Space Path Loss

FSL (Free Space Loss) is determined by the amount of distance that the signal has to travel and on the system's operating frequency. High operating frequency and larger distance results in higher losses. Even in the LoS component of received signal, FSL impacts the signal strength and mobile receiver front end must provide the legitimate amplification in order to perceive the information.

B. Ionospheric Effects

The ionized section of the atmosphere extending from a height of 30 km to 1000 km has adverse effects on earth-Satellite radio propagation. Two of the more important signal effecting mechanisms encountered by the signal propagating in ionosphere are Scintillation and Polarization depending upon impact of individual layer over the signal. Due to high inhomogeneity of Ionosphere, signal varies in amplitude, phase and angle of arrival as it propagates causing what is called Ionospheric Scintillations, the effects of which decreases with increase in frequency [11]. Rotation of plane of polarization of the EM (Electro-magnetic) wave as it passes through the irregular motion of free electrons in Ionosphere under the influence of earth's magnetic field degrades the spectral efficiency of the Satellite radio channels.

C. Tropospheric Meteorites

Lower portion of earth's atmosphere extending upto a height of 15 km above sea level introduces many hydrometeor effects such as those of clouds, rain, snow, fog etc. to the propagating radio signal and becomes significant above frequency of 1 GHz. Ippolito [11] suggests many models for quantifying atmospheric attenuation. At operating frequency above 10 GHz (e.g. Ku Band), rain attenuation effects Satellite communication links adversely as the radio wave energy is absorbed and scattered by rain drops.

D. Sun Outage

Radiations from the Sun effects the Satellite signal which ranges from partial degradation to total destruction. The nonionizing component of the radiation consists of electromagnetic waves providing a constant source of heating to the part of the spacecraft facing the Sun with a rate of 1400 W. With respect to MSS systems, Sun Transits/ Outages/ Fades are of more concern as they interfere with the Geo Stationary Satellite Signals and in some cases during the equinoxes tend to completely interrupt them.

E. Multipath Fading

Received radio signal in urban dense environment is the resultant of diffused scattered waves produced from LoS component. Combining constructively or destructively, these multipath components lead to certain signal level at mobile receiver front end and must be considered during decision over RF component's specifications is carried.

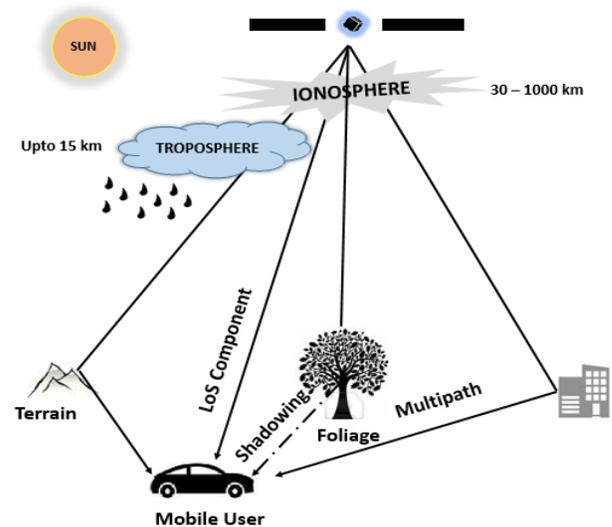


Fig. 1. Mobile Satellite Scenario

F. Shadowing

MSS face complete demolition of the LoS radio signal strength due to presence of large obstacles in the LoS between the Satellite and the mobile user. Vegetation and foliage cause extensive signal absorption. Many mathematical models have been demonstrated in literature to quantify the effect of shadowing [12]. Factors discussed in this segment are illustrated in Fig. 1.

III. MOBILE RECEIVER FRONT END

MSS reception at the mobile terminal is affected by several key factors as discussed in the previous section. Schemes must be opted to overcome the same at the mobile receiver handset. In order to integrate the best of the functionalities of the Satellite and Cellular receiver, generally the architecture shown in Fig. 2 is opted. In this cost effective architecture, the user level selection of the preferred network is possible with Common Control unit in the uplink. While on the downlink, generally this issue is controlled by the network level management entities. Some carriers also implement dual LNA in the separate Satellite links but authors believe that this is not an economically viable solution.

Terminals supporting Dual Mode are designed to cater seamless handover between Cellular and Satellite networks. Two types of implementations are noted i.e. a Hybrid solution and an Integrated solution. Hybrid implementation is more complex than the later due to the fact that although the cellular-satellite switching intelligence is housed in a single enclosure, they work as stand-alone terminals. Integrated solution has recently gained momentum because of their simplicity due to integration of cellular and satellite modems over a single wafer resulting in better inter-network-switching characteristics with reduced cost.

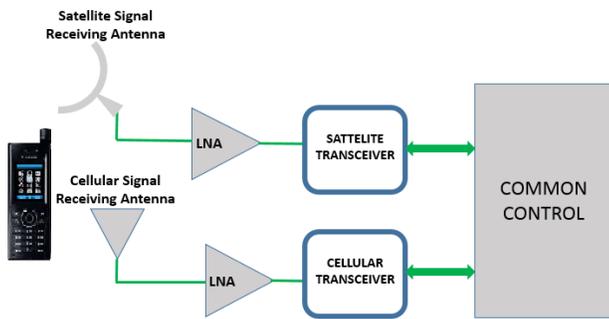


Fig. 2. Mobile RF Front End

IV. LNA DESIGN

In order to relax the Satellite link margin, the LNA must assist with a minimum amount of gain and keep the noise figure below a predefined level. The characterization of the microwave device is generally performed using the Scattering (S) parameters as obtaining the short circuit or an open circuit for devices operating at microwave frequencies is not practically feasible. For a two port network, four complex S parameters are defined and when analysed, provide correct insight to device performance. S_{11} and S_{22} provide insights about the reflection from the device while S_{12} and S_{21} are reverse and forward transmission gains respectively. Reflection coefficient quantifies the mismatch at concerned port and is represented on Smith chart. Remaining two (S_{12} and S_{21}) are plotted on Polar plots for simplicity. Next, some important aspects of LNA design are considered and described in brief.

A. Key Parameters

Input to the LNA is the high frequency noise effected received signal at the load end of the RF antenna. According to the famous Friis Formula, the noise performance of an RF receiver is determined by the Noise Figure (NF) and Gain of the LNA. Minimization of NF and maximization of the Maximum Available Gain (MAG) generally have opposite requirements as discussed in [13]. Minimum NF is obtained when the LNA input impedance is made equal to the optimum impedance Z_{opt} calculated at operating frequency. On contrary to this, MAG is obtained under perfectly matched input and output terminations in characteristic impedance Z_0 . These two complex impedances are never equal and an optimization scheme needs to be addressed.

B. Selection of Semiconductor Material

With the evolution of nanometre technology, the Monolithic Microwave Integrated Circuits (MMIC), a scaled down solution to integration several RF components on silicon substrate is possible. Microwave amplifier device technology brings the High Electron Mobility Transistor (HEMT) to achieve very fast switching speeds using compound semiconductors for crystal growth. Gallium Arsenide (GaAs) was the first to be opted for its high carrier mobility and remains to lead the lot. Recently, wide band semiconductor materials such as Silicon Carbide (SiC) and Gallium Nitride (GaN) have emerged as robust options at high power withstanding capabilities. Present research work has opted InGaAs based material which is an alloy of GaAs and Indium

Arsenide (InAs) which has superior performance for space applications. InAs/GaAs alloy is described as $In_xGa_{1-x}As$ where 'x' is the proportion of InAs and '1-x' is the proportion of GaAs. Literature survey confirms the fact that most convenient substrate for InGaAs is InP (Indium Phosphide) and the combination is termed InGaAs/InP. Due to its advantages in terms of low noise and high gain, InGaAs/InP has recently gained recognition in space applications [14].

C. Design Methodology

As discussed in segment A of the present section, the different parameters important for designing the LNA are of opposing in nature. Different optimization schemes are demonstrated in [15], where the authors advocated Power Constrained optimization scheme for the design of active amplifier. Since the Voltage Standing Wave Ratio (VSWR) quantifies the amount of mismatch of the device ports taking the characteristic impedance as reference value, in [16], a practical trade-off between NF and VSWR is claimed using a graphical approach. Designer must also take into consideration various other aspects like bias conditions, device geometry, operating bandwidth, RF power generated etc. all at a time Due to the complexity involved, generally an unacceptable amount of round off error results during the optimization process. Also, the models are restricted to specific case study and may not be generalized. Keeping the above stated under consideration, Artificial Neural Network presents an accurate, flexible approach to design active device models. The Universal Approximation property given in [8] forms the basis of employing a Neural Network (NN) for a generalized model. Next section investigates ANN models.

V. ANN LNA MODEL

Basic processing unit of the human nervous system is a biological neuron. In 1942, the first man made neuron model was presented [4] and till date the human investigation to imitate the complex parallel processing nature of the human nervous system remains a topic for research.

In order to generalize a massive interconnection between input and output variables, a Back Propagation algorithm [9] is utilized. An interconnection of various layers of artificial neurons is formed with the first layer accepting the input variables and is termed the Input layer while the layer at which output parameters are obtained is termed the Output layer. All the layers existing between these layers are called the Hidden layers and optimization of the model depends upon the number of Hidden layers and the value of the weights and biases of individual neurons in this layer. The ANN model must be trained with accurately measured values of desired output variables corresponding to input parameters. When optimized, the model has the capability to generalize the input/output nonlinear relationship [17], [18].

An innovative approach to an LNA design was presented in [19], where the model based on input variables frequency and temperature of device operation and biasing voltage and current was implemented. Training of this model performed in present paper utilizes this concept to optimize the model in order to retrieve Scattering parameters, NF and MAG of the LNA.

Fig 3 shows a schematic representation of modelling strategy opted in present work. The NN training is based on experimental data as a function of frequency of operation, operating temperature, drain to source voltage and drain current of the pHEMT amplifier. Parameters are varied between some predefined ranges. It is only due to such modelling that a neural model helps in simultaneous optimization of several variables involved.

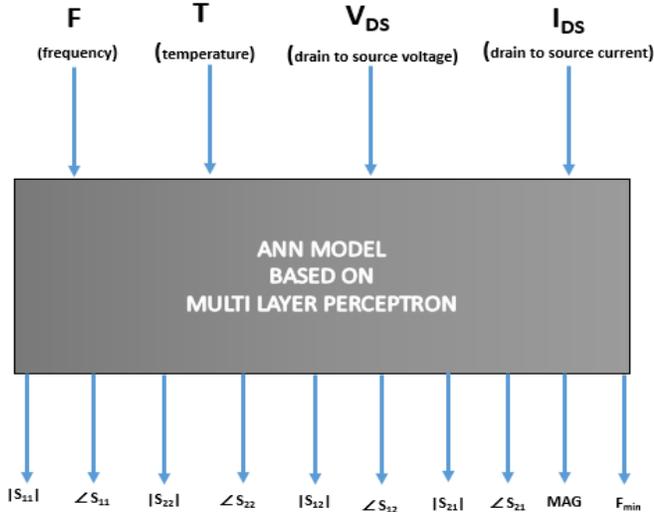


Fig. 3. ANN Modelling Strategy

VI. RESULTS AND DISCUSSIONS

In order to model the LNA, the training of the neural model is done using the Agilent ATF 331M4 pHEMT FET family datasheet values [20]. MATLAB ANN Tool Box is used as software platform for generating the feedforward network with Levenberg Marquardt learning algorithm. The mean square error (MSE) for the present model's learning gradient is set to 10^{-6} and network is optimized to plot the Smith charts and the Polar charts of the scattering parameters. Inputs to the model are frequency, biasing conditions and operating temperature.

Accuracy of the model increases with increase in training data set values and as may be easily followed from Fig. 4(a), 4(b), 4(c) and 4(d), the Neural Network is trained efficiently.

These models are optimized for biasing condition of 2 Volts V_{DS} and 40 mA I_D for the pHEMT for the frequency range starting from 0.5 GHz to 18 GHz. MAG and NF are also calculated and plotted in Figures 5(a) and 5(b) respectively for the same frequency range. The values of the microwave LNA MAG and NF, drop and rise with increase in operating frequency, respectively, which seems intuitively correct.

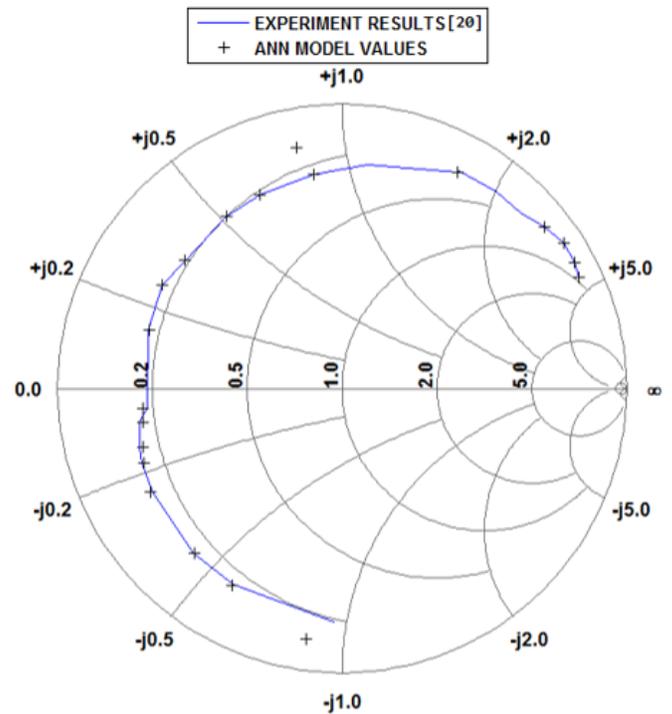


Fig. 4. (a). S₁₁ Smith Chart

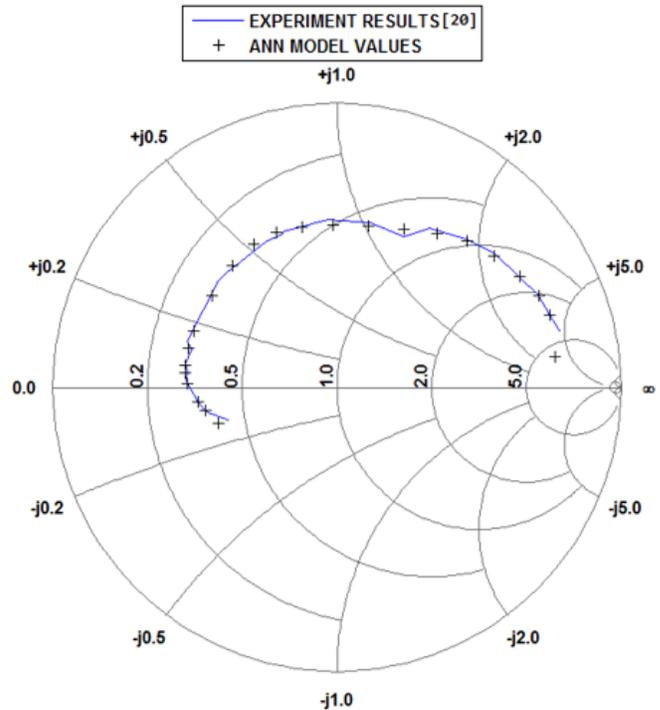


Fig. 4. (b) S₂₂ Smith Chart

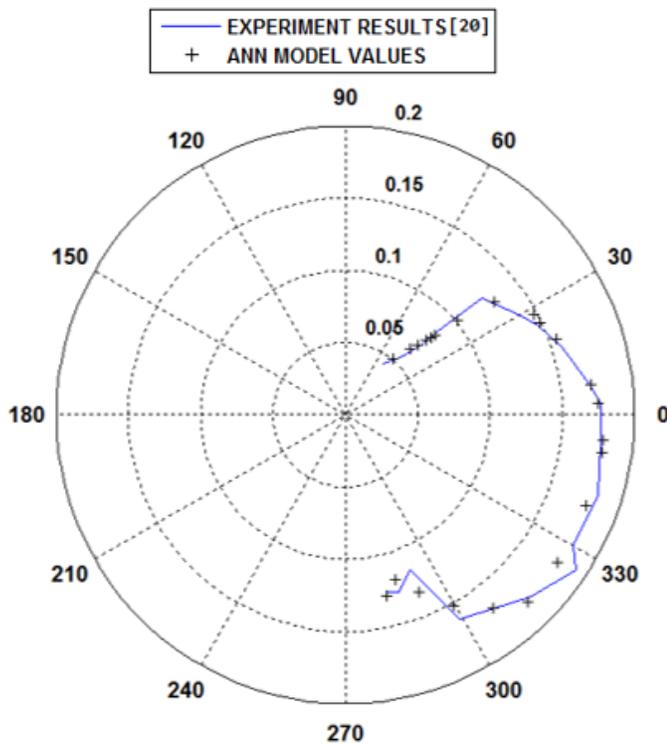


Fig. 4. (c). S_{12} Polar Plot

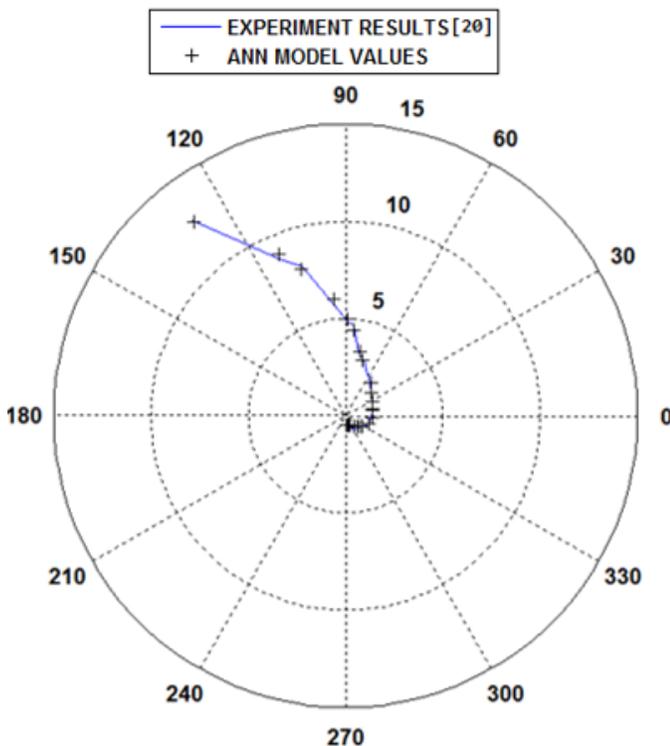


Fig. 4. (d). S_{21} Polar Plot

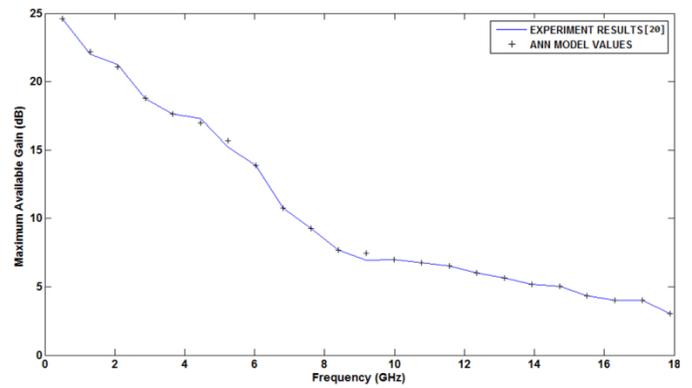


Fig. 5. (a). MAG (dB) vs Frequency (GHz)

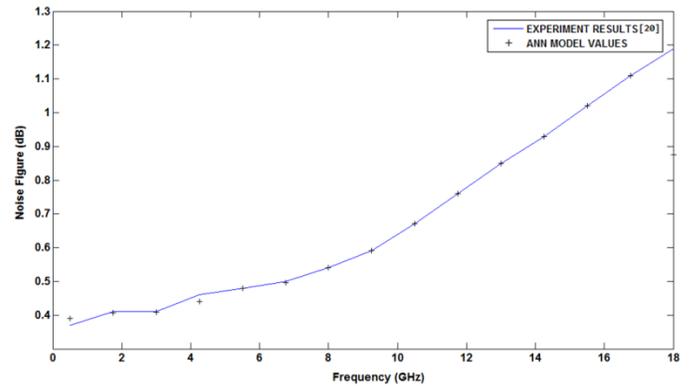


Fig. 5. (b). Noise Figure (dB) vs Frequency (GHz)

VII. PAPER CONTRIBUTION

As discussed initially, the designed LNA is to be implemented in the RF front of the mobile receiver of a Dual Mode SatPhone. If the HMSS system is operated in the L Band designated by the ITU-R, then resolution of the results must be improved in the 2 GHz spectrum. In order to achieve the same for the two most important parameters for LNA in MAG and NF, the model developed in section VI is used to interpolate the values in the concerned frequency range. The interpolated values are plotted in Fig. 6(a) and 6(b) for MAG and NF respectively with increased number of data set for 0 to 2 GHz frequency range. By comparing Fig. 5(a) and 6(a) it may be clearly noticed that the latter is more informative than the former in the L Band when the LNA is to be analysed for amplification. Same is true for the resolution of NF in Fig. 5(b) and 6(b). The interpolation capabilities of the Neural Network may easily be noted. It is very important to note that before using the ANN model for interpolation, it was trained with all the datasheet values and the MSE (Mean Square Error), regression coefficients and error estimation on point to point basis is computed and optimized.

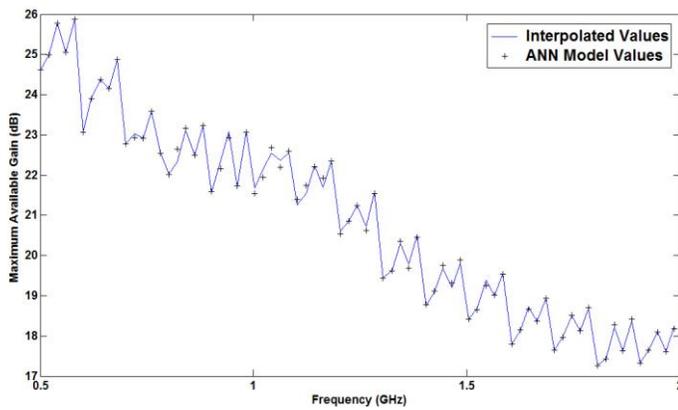


Fig. 6. (a). MAG (dB) vs Frequency (GHz)

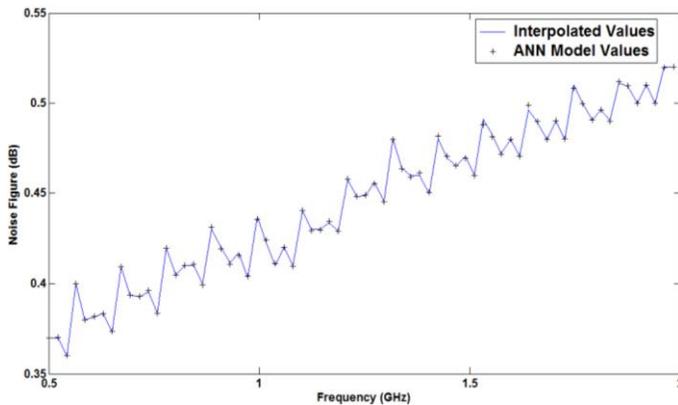


Fig. 6. (b). Noise Figure (dB) vs Frequency (GHz)

VIII. CONCLUSION

Signal deteriorating factors in Mobile Satellite System are discussed. A Hybrid MSS mobile receiver schemes are presented and an LNA is designed keeping in view its compatibility with the mobile unit's RF front end. A brief summary of Artificial Neural Network is provided to develop a quick insight into why designers are interested in such modelling strategy.

Optimization of the LNA model is carried out using the ANN approach. All the scattering parameters are calculated and their respective Smith charts and Polar plots are plotted and the results show that the ANN model has been trained efficiently. The trained model is utilized for interpolation to obtain performance analysis in the L Band range of MSS operation for Gain and Noise Figure of the pHEMT MMIC LNA. The performance of an LNA depends on the input and output matching networks, which is dictated by the scattering parameter description of the device. As a future scope to the present work, the authors recommend the design of matching networks for the LNA designed in the present work.

ACKNOWLEDGMENT

I sincerely thank Ajay Kumar Garg Engineering College, Ghaziabad for providing the opportunity and guidance for research work.

REFERENCES

- [1] Sastri L. Kota, Giovanni Giambene, Paolo Chini, "A Mobile Satellite Systems Frame Work For Network CENTRIC Applications", Conference on Military Communications (MILCOM), IEEE, Volume., No., pp.1-8, Nov. 2008.
- [2] Zhang Tao; Zhang Jun; Liu Zhong Kan, "A Delay Constraint Minimum Cost Routing Algorithm for Mobile Satellite Networks," Information, Communications and Signal Processing, IEEE 2005.
- [3] Umehira, M.; Fujita, S.; Zhen Gao; Jing Wang, "Dynamic channel assignment based on interference measurement with threshold for multi-beam mobile satellite networks," Asia-Pacific Conference on Communications (APCC), August, 2013.
- [4] Warren S. McCulloch, Walter H. Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity", Bulletin of Mathematical Biophysics, 1943.
- [5] Zlatica Marinković, Aleksandar Stošić, Vera Marković, Olivera Pronić., "ANNs in Bias-Dependent Modeling of S-parameters of Microwave FETs and HBTs", Microwave Review, June 2006.
- [6] Anil Ahlawat, Manoj Pandey, Sujata Pandey "A Fast and Robust Approach for Modeling of Nanoscale Compound Semiconductors for High Speed Digital Applications, Journal Of Semiconductor Technology And Science, 2006.
- [7] Ronan G. Brady, Christopher H. Oxley, Thomas J. Brazil 'An Improved Small-Signal Parameter-Extraction Algorithm for GaN HEMT Devices', IEEE Transactions on Microwave Theory and Techniques, Vol. 56, No. 7, July 2008.
- [8] Kur Hornik, Maxwell Stinchcombe and Halber White, "Multilayer Feedforward Networks are Universal Approximators", Pergamon Press, 1989.
- [9] Qi-Jun Zhang et. al., "Artificial Neural Networks for RF and Microwave Design—From Theory to Practice", IEEE Transaction on Microwave Theory and Techniques, 2003.
- [10] Lu Lu, Daoxing Guo, Aijun Liu, Maoqiang Yang, "Analysis of Channel Model for GEO Satellite Mobile Communication System", National Conference on Information Technology and Computer Science, Atlantic Press, 2012.
- [11] Dr. Louis J. Ippolito, "Propagation Effects Handbook For Satellite Systems Designs", ACTS Conference 2000, Sixth Ka-Band Utilization Conference, Cleveland, Ohio, Issue No. 2, 2008.
- [12] Chun Loo. "A Statistical Model for a Land Mobile Satellite Link". Vehicular Technology, IEEE Transactions Volume. 34, Issue No. 3, 1985.
- [13] Guillermo González, "Microwave Transistors Amplifiers: Analysis and Design", 2E, Prentice Hall, 1997.
- [14] Onur Esame, Yasar Gurbuz, Ibrahim Tekin, Ayhan Bozkurt, 'Performance Comparison of the state of art heterojunction bipolar devices (HBT) based on AlGaAs /GaAs, Si/SiGe and InGaAs/InP', Microelectronics Journal, Elsevier, Nov., 2004.
- [15] Trung-Kien Nguyen et. al., "CMOS Low - Noise Amplifier Design optimization Techniques" IEEE Transactions on Microwave Theory and Techniques, 2004.
- [16] Alan Victor, "An Analytic and Graphical Method for LNA Design with Feedback" Summit Technical Media, High Frequency Electronics, 2010.
- [17] Vikas Chaudhary, R.S. Bhatia, Anil Ahlawat, 'An efficient self-organizing map learning algorithm using the set of nearest neurons', International Conference on Contemporary Computing (IC3), Aug. 2013.
- [18] Anil Ahlawat, Manoj Pandey, Sujata Pandey, 'A Fast and Robust Approach for Modeling of Nanoscale Compound Semiconductors for High Speed Digital Applications', Journal of Semiconductor Technology and Science, IEEE, Sept., 2006.
- [19] P. K. Chopra et. al. , "ANN Modeling Approach for Designing Low Noise pHEMT Amplifier in Wireless Communication Systems", Optical Memory and Neural Networks (Information Optics), Springer, 2011
- [20] Agilent ATF 331M4 Low Noise Pseudomorphic HEMT in a Miniature Leadless Package available at <http://www.agilent.com>.

Multi-Agent Architecture for Implementation of ITIL Processes: Case of Incident Management Process

Youssef SEKHARA, Hicham MEDROMI, Adil SAYOUTI

Equipe Architectures des Systèmes (EAS), Laboratoire d'Informatique, Systèmes et Energies Renouvelables (LISER)
Hassan II University–Ain Chock, Ecole Nationale Supérieure d'Electricité et de Mécanique (ENSEM)
Casablanca, Morocco

Abstract—ITIL (Information Technology Infrastructure Library) is the most widely accepted approach to IT service management in the world. Upon the adoption of ITIL processes, organizations face many challenges that can lead to increased complexity. In this paper we use the advantages of agent technology to make implementation and use of ITIL processes more efficient, starting by the incident management process.

Keywords—ITIL; Process; Multi-agent system; Incident Management Process

I. INTRODUCTION

The use of computers has become essential for business organizations. Having the best technology will not guarantee the expected needed service reliability. It is necessary for them to have a full service around these technologies. There is a large number of repositories that reflect the best practices developed over the years, ITIL presents a guide of best applicable practices suitable to all types of organizations providing services to a business organization [1]. ITIL shows the framework of the organization, objectives, processes related to major activities of information services and their interactions. It is a kind of canvas that can be used by the directions of information systems to design their own organization.

The successful adoption of ITIL processes is a major challenge for many managers of information systems. Agent technology has shown great potential for solving problems especially for complex applications involving interaction between several entities. Our job is to use the advantages of agent technology to make more efficient implementation and use of ITIL processes.

In this paper, characteristics and the importance of ITIL in information systems are presented in the second section. The third section gives a view on problems encountered in the implementation of ITIL and the contribution of multi-agent systems to make it easier and more efficient. The fourth section shows the multi agent architecture proposed to implement incident management process chosen to be the starting point for the implementation. In the fifth section, we describe the implementation.

II. ITIL & MULTI-AGENT SYSTEM

A. Presentation of ITIL

The IT Service Management ITIL is based on five groups of activities (each containing its multiple processes) to manage

the service throughout its life cycle [2]; ITIL recommends taking into account management of services, from the phases of study and defining the needs of IT projects. This, provides several benefits to all teams of the production or development, but serves primarily to provide a more reliable service to the customer, it also validates that we have many resources and skills necessary for the operation of this new application, which involves taking into account the impact on physical infrastructure in areas such as capacity, performance, availability, reliability and maintenance.

The five phases are[3]:

- **Service Strategy** : The phase of strategic planning of service management capabilities, and the alignment of service and business strategies.
- **Service Design** : The phase of designing and developing appropriate IT services, including architecture, processes, policy and documents; the design goal is to meet the current and future business requirements
- **Service Transition** : The phase of realizing the requirements from previous stages, and improving the capabilities for the transition of new and modified services to production
- **Service Operation** : The phase of achieving effectiveness and efficiency in providing and supporting services in order to ensure value for the customer and the service provider.
- **Continual Service Improvement** : The phase of creating and maintaining the value for the customer by design improvement, and service introduction and operation

In the ITIL framework, the quality of service is based on a structure of measurable possible repeated activities in interrelated processes. This approach to service management by process is now recognized as the most effective by a large number of companies. The implementation of processes providing efficient operation of these systems is therefore an important element to enable businesses to take full advantage of their entire IT infrastructure.

A process consists of several activities generating results to clients. It is called after a trigger event. A well-studied process must achieve the objectives, using optimal time, money and

resources. A process is provided by roles and not directly by the people, what makes this notion generic and independent of the organization.

B. Challenges of ITIL adoption & Proposition of multi-agent solution

Despite the flexibility of ITIL and its processes, organizations are faced with many challenges when adopting ITIL best practices that can lead to increased complexity. The inability to control this complexity leads to increased costs and longer time lines for implementation. Such as:

- **Lack of management commitment:** it presents the main cause of the failure of the implementation of ITIL because without it nothing tends to occur.
- **Choosing an inefficient point of departure:** for the adoption of ITIL there is not a standard starting point, every business must start with a gap analysis between existing and ITIL benefits to be able to find areas where organization will receive more value by adopting ITIL.
- **Resistance to ITIL:** ITIL is a major organizational change which is always resisted. Departments, teams and individuals tend to defend the current status.
- **The close relationship** between activities of different processes.

Faced with these problems, the leaders of IT departments have recourse to incorporate ITIL's compliant software permitting to accompany and make the implementation of ITIL processes easier[4].

There are several free and commercial service software for the automation of repetitive tasks and reducing human error and cost, but these are just tools that help to meet the objectives as explained by Claude Durant, founder of itSMF France (The IT Service Management Forum France): they are only supports, prior to use, you must define and implement objectives of ITIL processes such as incident management and problem management for example. The tool will then help to meet the set objectives of performance and quality [5].

We propose to design and implement a multi-agent distributed platform that implements the various ITIL process through autonomous entities (agents) that interact with each other and share the knowledge, reasoning, and decision, that leads to profit from the benefits of multi-agent systems to facilitate the adoption of ITIL and fully enjoy the process.

Distribution using multi-agent systems is very useful given the nature of the problem addressed and the complexity of the system permitting the implementation of the various tasks of ITIL processes; we may explain the need for distributing by multi-agent systems for the following reasons:

- **Complexity of the problem:** ITIL processes involve many subsystems very diverse in nature, with many features, interacting with several human experts (operators, experts, technicians, etc..), who are often distributed in the physical space.
- **Need for local vision:** the solutions based on local

approaches such as agent approach often allow to fastly solve the problems especially those are too large to be analyzed globally.

- **Adaptation to changes in the structure or environment:** the multi-agent systems, thanks to their distributed nature -because they always involve a local reasoning- allow the integration and the appearance or disappearance of agents during operation, which enables to build suitable architectures to consider scalability and adaptation needed to operate the system

C. Definition & advantages of multi-agent systems

Multi-agent systems (MAS) are systems where multiple agents can interact with each other in various modes including cooperation, competition or mutual existence[6].

An agent is a hardware or (more usually) a software-based computer system that enjoys the following properties[7]:

- **Autonomy :** agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state;
- **Social ability :** agents interact with other agents (and possibly humans) via some kind of agent-communication language;
- **Reactivity:** agents perceive their environment and respond in a timely fashion to changes that occur in it;
- **Pro-activeness:** agents do not simply act in response to their environment; they are able to exhibit goal-directed behaviour by taking initiative.

The need for autonomy makes the agent support activities in an intelligent and flexible way and can adapt to the environment without requiring human intervention [8].

The main characteristics of a multi-agent system are:

- Group of agents acting and working independently of each other,
- Each agent is a part of the system
- Each agent works to accomplish its tasks, each agent in the proposed architecture allows to perform the tasks related to the concerned process
- Each agent is able to communicate and interact with other agents, which allows a rich exchange between different processes
- An agent is able to coordinate its activities with other agents to access shared resources
- Agents' common goal in our architecture is the proper functioning of the process
- Each agent has a partial view of the MAS.

Some advantages of using multi-agent systems [9] compared to other technologies:

- **Scalability and flexibility;** it is easy to add new agents to the system.

- **The development and reusability**, since it is easier to develop and maintain a modular software
- **Robustness and reliability**

III. INCIDENT MANAGEMENT PROCESS

A. Approach of implementation

ITIL has many processes. Each ITIL process has a great impact on the organization. It is preferable that the full implementation of ITIL is spread over several phases. A big bang ITIL has the potential to disrupt the SI as well as commercial operations [10]. This leads us to choose the bottom-up approach for implementing process relying on the scalability and flexibility of multi-agent systems; we therefore start by the implementation of each process in order to integrate all in one platform.

We chose to start with the implementation of the incident management process. Whatever the quality of the information system set up in the company or the skills of technicians who operate, incidents occur. These incidents always have an important effect on the trust that users place in the team who manages this information system. How to handle these "crises" and their rapidity of resolution is an indicator of the maturity of the IT team. Incident Management has strong links with the management problem for the identification of causes, and change management for the implementation of changes after identification of causes. There are also links with configuration management [11]. That is why starting with the implementation of the incident management process is particularly important.

B. Incident Management process

Incident Management Process is responsible for all steps from the detection and recording of an incident until it is resolved and closed. The aim is to restore service as quick as possible with minimal disruption to the business.

In ITIL terminology [12], an 'incident' is defined as: An unplanned interruption to an IT service or reduction in the quality of an IT service. Failure of a configuration item that has not impacted service yet is also an incident, for example failure of one disk from a mirror set.

Incidents can be triggered in several ways. The most common way is when a user calls the service center or completes an online form of the incident in a tool or via the Internet. However, many incidents are recorded by event management tools.

The incident management process consists of the following steps [13]:

- 1) **Identification** : The incident is detected or reported.
- 2) **Registration** : An incident record is created.
- 3) **Categorization** : The incident is coded by type, status, impact, urgency, SLA, et cetera.
- 4) **Prioritization** : Every incident gets an appropriate prioritization code to determine how the incident is handled by support tools and support staff.
- 5) **Diagnosis** : A diagnose is carried out to try to discover the full symptoms of the incident.
- 6) **Escalation** : When the service desk cannot resolve the incident itself, the incident is escalated for further support (functional escalation). If incidents are more serious, the appropriate IT managers must be notified (hierarchical escalation).
- 7) **Investigation and diagnosis** : If there is no known solution, THE INCIDENT IS INVESTIGATED.
- 8) **Resolution and recovery** : Once the solution has been found, the issue can be resolved.
- 9) **Incident closure** : The service desk should check that the incident is fully resolved and that the user is satisfied with the solution and the incident can be closed

IV. PROPOSED IMPROV

A. Proposed architecture

Our architecture (Figure 1) consists of two layers: a reactive layer and a deliberative layer. The reactive layer consists of reactive agents (simple processing units that perceive and react to changes in their environment [14]), in our architecture the User Agent is reactive agent, The deliberative layer is composed of other agents that are cognitive agents (in which decision making depends upon the manipulation of data structures representing the beliefs, desires, and intentions of the agent [15]). The communication between agents is done using messages.

B. Description of the architecture

Figure 2 presents the sequence diagram which gives an overview on the different interactions between agents in our architecture in order to better understand its behaviour [16]. After reporting the incident to User Agent linked to an interface, it sends the incident's information to the Diagnosis agent, which records it in the incident database and determines whether the incident has already occurred.

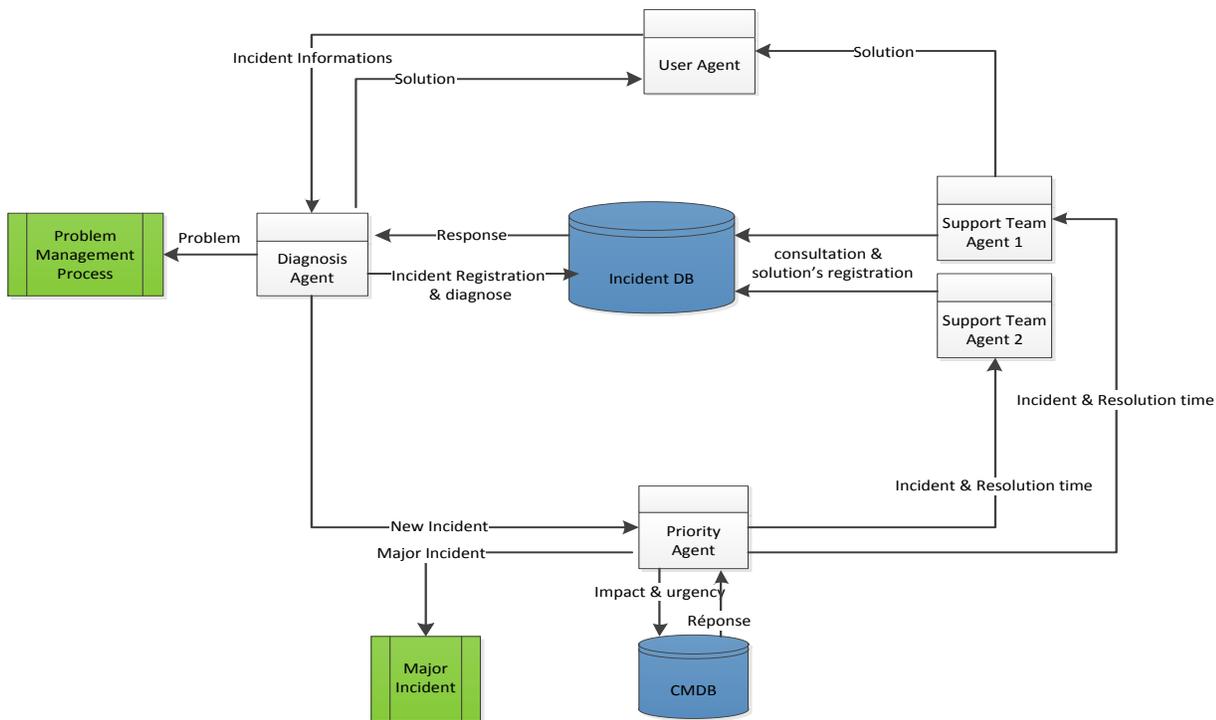


Fig. 1. Architecture of incident management process implementation

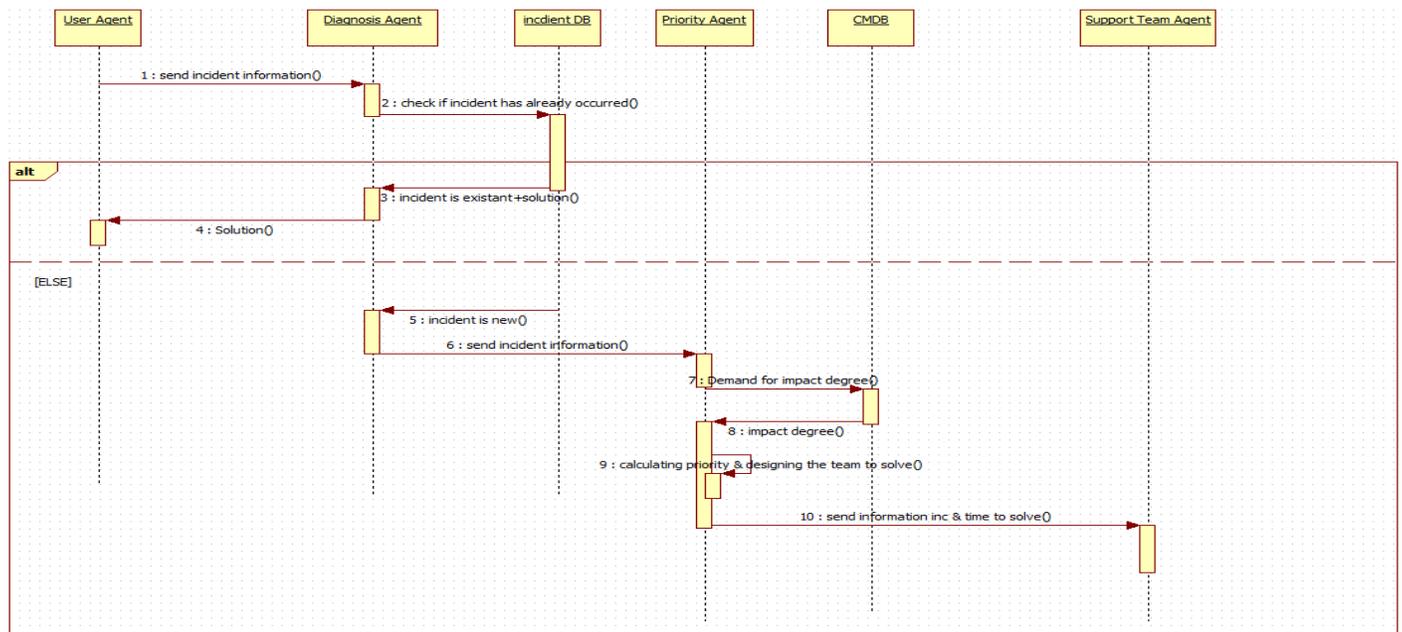


Fig. 2. Sequence diagram of the architecture

In this case, it returns to the user the solution present in the database, this operation saves considerable time. The Diagnosis Agent calculates the rate of production of this incident and in case it exceeds a predetermined value in the service level agreement (SLA) the Diagnosis Agent shall forward the incident to the responsible of problem management process to find the source of the problem.

In the case where the incident occurs for the first time, the Diagnosis Agent sends incident's information to the Priority

Agent that calculates the priority of the incident according to the degree of impact and urgency presents in the configuration management database (CMDB).

This calculation is used to define the time allocated to the resolution and this is communicated to the appropriate Support Team Agent according to the type of the incident (for example if the incident concerns the software, the incident and resolution time presented to the team that deals with incidents softwares). If the problem has a very high degree of priority,

the Priority Agent shall communicate it to the team that deals with major incidents. Once the solution is found, it's recorded by the Support Team Agent in the incident database and sent to the user.

V. APPLICATION

We implemented this architecture using Java language and Java Message Service (JMS) using Apache ActiveMQ language to develop the different agents and also their interactions. Figure 3 presents the user interface where he puts information about the incident who wanted to be solved.

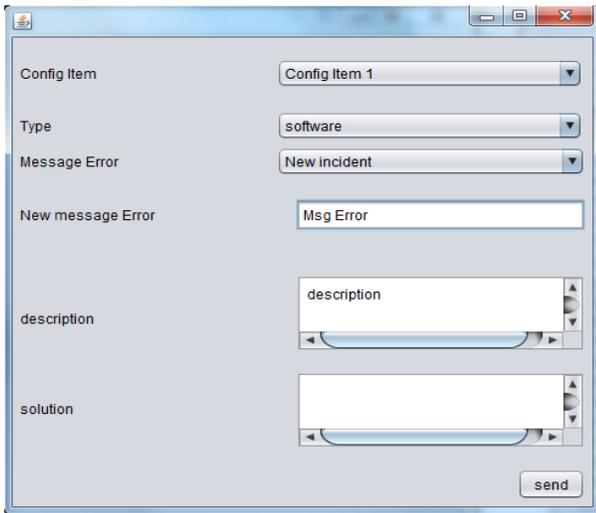


Fig. 3. User interface

VI. CONCLUSIONS

The first part of this paper presented characteristics and the importance of ITIL in information systems. The second part showed problems encountered in the implementation of ITIL and the contribution of multi-agent systems to make it easier and more efficient. In the third part we talked about the choice of incident management process as a starting point for the implementation by the agents. In the last part, we proposed multi agent architecture to implement incident management process.

We started with the process of incident management, chosen to be the core of this architecture because it has strong links with other processes such as process problem management, availability management and continuity management that we aim to develop in the following step and merge them into a complete system

REFERENCES

- [1] Y. Sekhara, H. Medromi and A. Sayouti, "Système multi-agent pour l'implémentation du processus de gestion d'incidents ITIL", in Proc. WOTIC'13, 2013.
- [2] itSMF France "ITIL : Information Technology Infrastructure Library". [Online]. Available: <http://www.itilfrance.com>
- [3] A. de Jong, A. Kolthof, M. Pieper, R. Tjassing, A. van der Veen, T. Verheijen, "ITIL® V3 Foundation Exam - The Study Guide" pp. 23-26.
- [4] Y. Sekhara, H. Medromi and A. Sayouti, "Etude des logiciels conformes à ITIL et proposition d'une Solution Distribuée à base multi-agent pour la gestion de la CMDB au Coeur des Processus ITIL", in Proc. JD TIC'12, 2012.
- [5] INDEXEL. "Les logiciels Itil, indispensables mais pas autosuffisants". [Online]. Available: <http://www.indexel.net/materiels/les-logiciels-til-indispensables-mais-pas-autosuffisants.html>
- [6] J. Ferber, "Les Systèmes Multi Agents: vers une intelligence". InterEditionsBradshaw 1995. pp. 28.
- [7] M. Wooldridge, and N. R. Jennings, "Agent theories, architectures, and languages". In Wooldridge and Jennings, eds. Intelligent Agents, Springer Verlag, 1995. pp. 1-22.
- [8] N. R. Jennings, M. Wooldridge, "Agent technology: foundation, application, and markets". Springer, New York 1998.
- [9] Katia P. Sycara, "Multiagent Systems" the American Association for Artificial Intelligence 1998.
- [10] BMC software, "ITIL pour les PME/PMI". [Online]. Available: www.bmc.com/fr-CA/documentation
- [11] C. Dumont. "ITIL Pour un service informatique optimal" Eyrolles. pp. 87.
- [12] Office of Government Commerce, "ITIL V3 – Service Operation". pp. 86.
- [13] A. de Jong, A. Kolthof, M. Pieper, R. Tjassing, A. van der Veen, T. Verheijen, "ITIL® V3 Foundation Exam - The Study Guide" pp. 90.
- [14] Adina Magda Florea, "Introduction to Multi-Agent Systems". International Summer School on Multi-Agent Systems, Bucharest, 1998.
- [15] M. Wooldridge, "Intelligent Agents", Multiagent Systems: A Modern Approach to Distributed Modern Approach to Artificial Intelligence, edited by Gerhard Weiss. Massachusetts Institute of Technology 1999. pp. 27-79.
- [16] X. Blanc, I. Mounier. "UML2 pour les développeurs". pp. 64-77.

Path Planning in a Dynamic Environment

by the approach of the sliding on edge

Mohamed EL KHAILI

Laboratory SSDIA, ENSET, Mohammedia
HASSAN II University, Casablanca, Morocco

Abstract—Path planning is an important area in the control of autonomous mobile robots. Recent work has focused on aspects reductions in processing time than the memory requirements. A dynamic environment uses a lot of memory and hence the processing time increases too. Our approach is to reduce the processing time by the use of a pictorial approach to reduce the number of data used. In this paper, we present a path planning approach that operates in three steps. First, a construction of the visibility tree is performed. The following treatments are not performed on the original image but on the result tree whose elements are specific points of the environment linked by the relationship of visibility. We construct thereafter a visibility graph which one seeks the shortest path. This approach has a great interest because of its fast execution speed. The path search is extended also for the case where obstacles can move. The moving obstacles may be other mobile robots whose trajectories and speeds are known initially. At the end, some applications are provided on solving similar problem such civil aviation in order to guide plane avoiding collisions.

Keywords—component; path planning; navigation; robotics; visibility graph; obstacles contours; moving obstacles; space-time representation

I. INTRODUCTION

Much research in image processing has been made in order to determine the shortest path between two points in a given environment: sea or air navigation, a runs between two cities, installation of distribution: Water, electricity, telephone.... And that is the case for the motion planning for a mobile robot. The problem of planning is determining the path of the mobile robot from its current position to a final position within an environment with static or moving obstacles.

When obstacles are static, a necessary condition is that the path is constructed by a sequence of line segments connecting a subset of points of edges visible between them obstacles. Several algorithms based on this condition have been developed; the best known is called Algorithm VGraph using a visibility graph constructed from the image taken on the environment. The shortest path is determined by the method of Dijkstra. Alexpoulos et al. presented two algorithms for path search using a visibility graph constructed from a tessellation of contours and removal concave highs [11]. These two algorithms are merely variations of the A* algorithm. In the first algorithm called V*Graph, it is assumed that the obstacles are static while the second called E*Graph, the obstacles can move along linear paths at a constant speed.

In this paper, we present two algorithms for path search using a visibility graph constructed by sliding on the edges of

obstacles. In both algorithms, the path is also determined by the A* Algorithm applied to a reduced graph by applying the concept of a visibility recursive procedure and sliding on the edges of the obstacles. The final path consists of edge points of obstacles and line segments connecting the points visible. When obstacles are mobile, a technical of 3-D modeling is used. Our approach can also be applied to cases where the value and the direction of speed of obstacles can change.

The following assumptions are considered below:

- Obstacles are represented by their boundaries;
- Obstacles are dilated by the size of the robot. Therefore, the mobile robot can be considered as a moving point between obstacles;
- The mobile robot has a constant speed, acceleration and deceleration instant.

This paper is composed of three parts. The first presents the problem of path planning with a comparison with the navigation problems. The second part explains our approach and shows some results. The third part presents the extension of our works to moving obstacles environment. Comments and conclusions are provided in the last part.

II. PRESENTATION OF THE PATH PLANNING PROBLEM

In this section, we define the planning problem while noting the difference from the navigation problem. The different approaches to solve this problem will be cited.

A. Path Planning Versus Navigation

Path planning and navigation are two important areas in the control of autonomous mobile robots. In both cases, solving the problem is to move the mobile robot while taking into account the internal and external constraints (eg, limits of the actuators and obstacles) [1] [4] [10].

The navigation problem is to move the mobile robot in an unknown environment, but also said partially visible path planning problem with incomplete information. Navigation systems are divided into two systems according to Payton, Crowley, Faverjon and Slack [8] [9] [12].

The first is the system that manages local navigation sensorimotor interactions of the robot with the environment. The second is the global navigation system that reasoning at a level of abstraction rentals and their interconnections to guide the activities of the local navigation system. In this context, the global navigation system receives commands to perform tasks from an expert system and produces a sequence of primitive

tasks locally defined. These local tasks will be translated later in sensorimotor activities thanks to the local navigation system [7].

However, the path planning is to determine the minimum length path between two points in a known environment. The vision sensor is not on board the mobile robot must overhang but all the site where the robot should move.

B. Different approaches

There are several research focused on solving the problem of path planning between a starting point and a destination without collision with obstacles supposed point in our case fixed. A comprehensive and detailed view of the different approaches developed to solve this problem can be found in [10]. In the following, the most approaches encountered in this field is mentioned.

1) Mathematical approaches

The most common mathematical approach is based on the notion of graph. This approach has data structures for easy carrying. The second approach is based on the differential geometry, requires highly advanced scanning including finite difference techniques and finite element.

a) Graphs

In image processing, an image is associated with an undirected graph whose vertices are individual pixels of the image and the arc length between two vertices i and j is the Euclidean distance $d(i,j)$. The result graph is a graph with lengths of positive arcs.

The shortest path between two vertices i and j traverses corresponds to $u(i, j)$ from i to j , the total length $l(u) = \sum d(k,m)$ is a minimum.

Moore and Dijkstra proposed algorithm with their names for the search of the shortest path from one vertex to another in a graph whose arc lengths are positive. This algorithm is widely used for separation exploration and evaluation (Branch and Bound). Berge proposes associating to each vertex i of the graph a quantity $h(i)$ said potential i and implements an algorithm that calculates h potential until the potential difference between two vertices i and j is less than or equal to the length of arc (i, j) [3] [13].

There is a widely used for determining the shortest path algorithm which carries the A*algorithm name. This algorithm is a strategy guided by an eligible research evaluation function.

b) Differential Geometry

Kimmel et al. [17] presented a model based on a new approach to determine the minimum length path between two points on a surface of 3-D algorithm. The numerical resolution used is based on finding equal geodesic distance contours on a given surface. The relationship between critical path, surveying and equal distance contours were drawn from the literature of differential geometry.

Indeed, local geodesic are critical paths in the sense that each disturbance of the geodesic curve increases their lengths. This becomes a problem and digital topological level as encountered during the implementation of such an algorithm, which is added complexity assessments curves and parametric

representations. The core of the algorithm consists of a Kimmel digital resolution based on the finite difference method.

c) Distance fields construction

Distance fields are frequently used in computer graphics, geometric modeling, robotics and scientific visualization. Their applications include shape representation, model simplification, remeshing, morphing, Constructive Solid Geometry (CSG) operations, sculpting, swept volume computation, path planning and navigation, collision and proximity computations, etc. These applications use a signed or unsigned distance field along a discrete grid [26][27].

Different algorithms have been proposed to compute the distance fields in 2D or 3D for geometric and volumetric models. The computation of a distance field along a uniform grid can be accelerated by using graphics rasterization hardware. The most used algorithm computes 2D slices of the 3D distance field by rendering the three dimensional distance function for each primitive. However, rendering the distance meshes of all the primitives for each slice may become expensive in terms of transformation and rasterization cost. These algorithms for 3D distance field computation may be slow and not work well for deformable models or dynamic environments.

The mathematical approach is always present and the approaches mentioned in the following are only a means of reducing the execution time or reduce the memory requirements of search algorithms of the shortest path.

2) Symbolic approaches

In the 80s, the design of database images had attracted much attention. Applications that use databases of image data, interest in robotics, the acquisition of the Earth's resources, archiving of medical images, weather ... These systems require the need of a model and a relative path to the image data. Such a model should be independent of the orientation of the image (eg, invariant to rotation).

This approach is to represent an image by checking an expression syntax generated from a picture alphabet (object) linked by operators. This syntax is used to generate a set of rules capable of providing the information necessary for an expert system to assimilate unambiguously perceived environment. The expert system receives the points of departure and arrival. This approach is very suitable in the case of an environment densely populated obstacle. It gives good result also for the case of moving obstacles.

3) Pictorial approaches

Pictorial approaches include work on the image as a data base of the problem. These approaches are implementing several tools for image processing.

Crowley divides the space into convex regions. A convex region has the property that two points can be connected to it by a line segment included in this area entirely. The convex regions of the free space are eroded by the size of the robot in such a way as to represent the robot by a pixel.

Kuo-Chin Fan et al. provide a decomposition of environmental problems in homogeneous regions using the algorithm of McClusky. An adjacency graph is constructed in

which a path search is calculated with a variant of the known A* algorithm modified A* algorithm which takes into account both the translation and rotation of the mobile robot [13].

Rao represents the environment populated by obstacles in the form of generalized polygons disjoint ground. A generalized polygon is constituted by a connected sequence of circle arcs and straight segments. Solving the problem occurs in three different ways: Graph visibility generalized Voronoi Diagram and generalized trapezoidal decomposition. The algorithm proposed by Rao, gives good results for solving the problems of navigation [18].

Brooks offers generalized cones to represent the free space whose complement is the space robot prohibited. It subdivides the space in the form of generalized and mobile robot moves along the axes of the cones [7].

Conte et al. have developed a path planning algorithm for a mobile robot based on a construction of distance fields. Each pixel of the image, a label that matches the length of the shortest path between the starting points is assigned. The advantage of such an approach is the obtaining of all the shortest paths connecting all points of the image at a starting point [14][15][16]. Elmesbahi et al. proposes to calculate the distance field on a broken end environment to reduce the execution time and the need for main memory [19][20].

Wesly et al. described the environment by a graph whose nodes correspond to particular points of the image. Arcs connecting two nodes with visible Euclidean distances as the arc lengths [1]. The graph produced is called the visibility graph. Asano et al. and Welzt worked in the same direction by developing a fast algorithm for constructing the visibility graph [5][10]. They made a pre-treatment to reduce the number of nodes considered in the visibility graph and uses the A* algorithm for finding the shortest path. Reducing the number of nodes is due to the approximation of obstacles by closed polygons and the fact of considering only the peaks corresponding to convex angles.

4) Genetic approaches

Genetic approaches include work on the image as a data base of the problem by implementing several tools of Genetic Algorithm GA for image processing [23][24][25][28].

5) Our approach

Our approach is pictorial where a construction of the visibility tree is performed. Obstacles are dilated by the radius of the enclosing circle of the mobile robot. The mobile robot occupies a considerable area of the free space. But after expansion of the obstacles, the mobile robot is regarded as a point moving in the free space (Figure 1). The following processing is not performed on the original image but the result tree whose elements are specific points of the environment linked by the relationship of visibility. We built later a visibility graph on which we apply any scheduling algorithm. The path found is optimal. This approach is of great interest given its fast execution speed and minimizing memory requirements to implement such an algorithm.

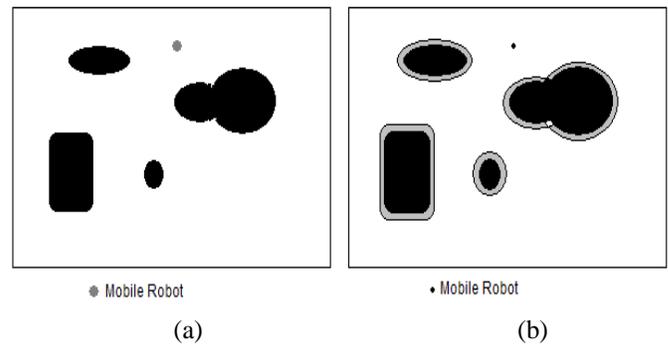


Fig. 1. Obstacles are dilated by the radius of the enclosing circle of the mobile robot (a) The blackened area represents the forbidden space so be it moving. The mobile robot occupies a considerable area of the free space SI. (b) After expansion of the obstacles, the mobile robot is regarded as a point moving in the free space SI.

III. PATH PLANNING WITH STATIC OBSTACLES

A. Principle

The algorithm presented in this section, comprises three essential steps. The first step is to detect recursively individual points belonging to the assumed stationary obstacles.

The second step reduces the number of points detected in the first step and therefore gives a visibility graph which will be based on which to calculate the shortest path in the final stage.

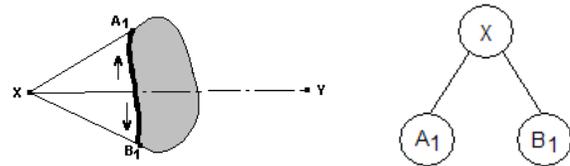


Fig. 2. Principle of the process used to build the visibility tree

The construction of the tree is a recursive process visibility enabling a single data structure (binary tree), one element is linked with his son's relationship with visibility. Figure 2 illustrates the technique used for the construction of such a tree.

X is an element of a visibility tree. It is the arrival point Y. If X is visible in the construction steps because the branch is terminal. If not, is slipped over the contour of the obstacle to the achievement of the two last visible point X (A1 and B1 in the case of Figure 1); then just do the same thing for point A1 and B1 to the end point Y. We repeat the process until the Y terminal as part of all branches of the tree visibility.

Figure 2 provides an example of visibility tree element constructed from a given environment. This operation is repeated until connecting all pair of visible points of the global environment.

Figure 3 illustrates an example of environment with obstacles and its tree of visibility. The visibility tree can be considered a multi-level graph.

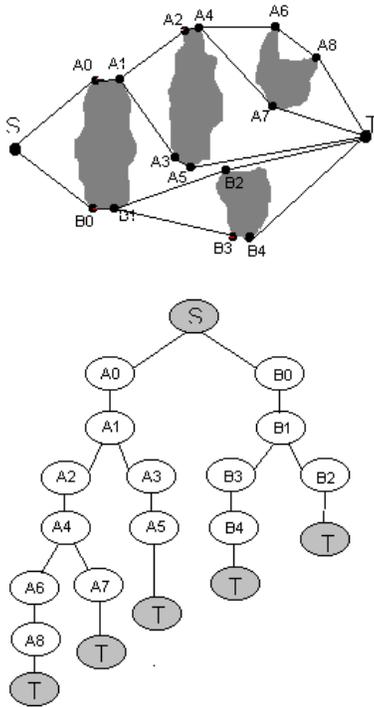


Fig. 3. A populated obstacle environment is given with the starting point S and the point of arrival T. The tree of visibility is computed to represent this environment.

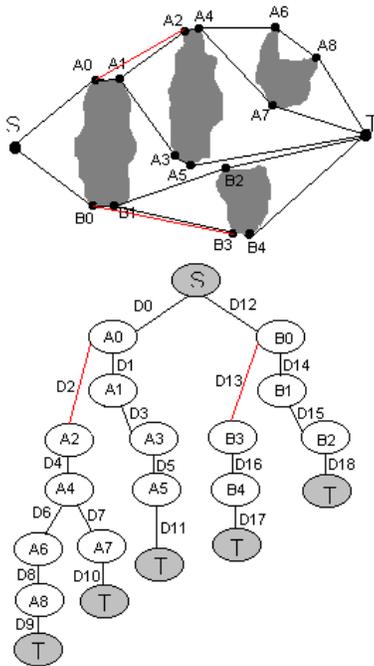


Fig. 4. Reducing the visibility tree with evaluation of different distances

B. Reducing visibility tree

As shown in Fig. 4, two visible points can be connected in order to reduce the tree because the segment connecting two visible points is the shortest path. Each element of the tree is affected by an amount corresponding to the length of the shortest path from the starting point S.

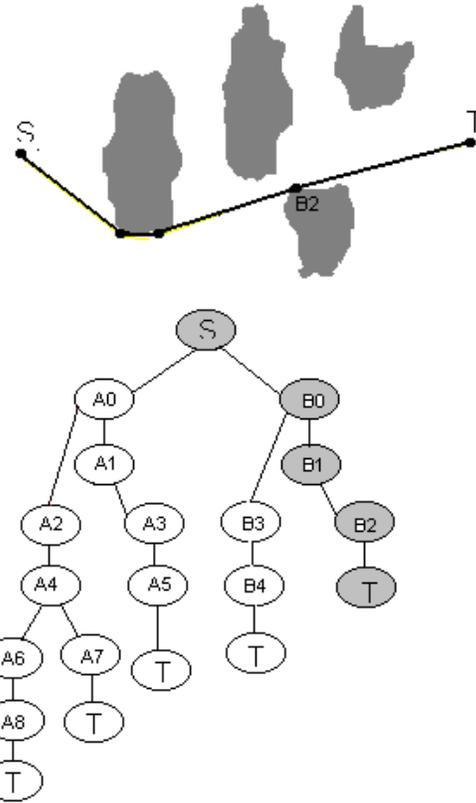


Fig. 5. Construction of final path

C. Shortest path construction

The procedure for construction of the road is very simple. Each node of the tree takes a value corresponding to the distance from the start point S. Terminal node with minimal distance is selected. Other endpoints of the segments constituting the path correspond to take the fathers of nodes starting from selected T. The Figure 5 illustrates an example of the final path construction.

D. Results

To test the proposed algorithm, two different environments are chosen. The first is simple. You can see lots of free space between obstacles when there is no concave shape (See Figure 6). In the second, a concave obstacle includes the starting point (See Figure 7). The results are correct in both cases.



Fig. 6. The first example illustrates a path planning of a boat. Such an environment is populated by a group of islands seen as obstacles to what you can add boats supposedly immobile. The problem then is to determine the shortest path without collision from a starting point S to a point of arrival T. The obstacles are dilated by the size of the moving object (the boat in our case): the shaded area.

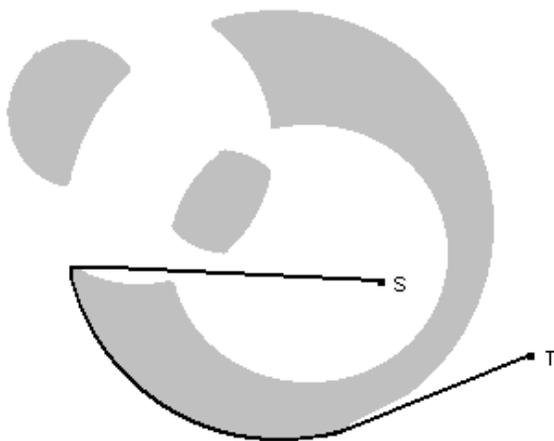


Fig. 7. The second example illustrates a path planning of a mobile robot in a world populated by a set of geometric constraints of any environment. The algorithm solves the problem without any influence of the shape of the obstacles on the execution time. The barriers also here are expanded by the size of the moving object.

IV. ENVIRONMENT WITH MOVING OBSTACLES

In this section, it is assumed that the obstacles are moving along a linear trajectory with a constant speed. Those moving obstacles can be others mobile robots. As already mentioned, the problem of path planning for a mobile object at a constant speed in the presence of moving obstacles, is very tricky[21][22]. We show that our algorithm, presented in the previous section can be extended to find the path without collision with moving obstacles, the path is optimal; it is brought to the readers that the prismatic obstacles in the figures is chosen in order to facilitate the graphical representations execution. The barriers also here are expanded by the size of the moving object.

A. Extension of the Slip Algorithm

Given the mobile obstacles in the plane (X, Y), the dimension "time" is used to define the position of obstacles in the space. This three-dimensional space is called Space-Time Representation. Each obstacle defined in Space-Time Representation is shown as prism form (See Figures 8 and 9). The static obstacles are orthogonal to the plane (X, Y) while the form of moving obstacles is oblique.

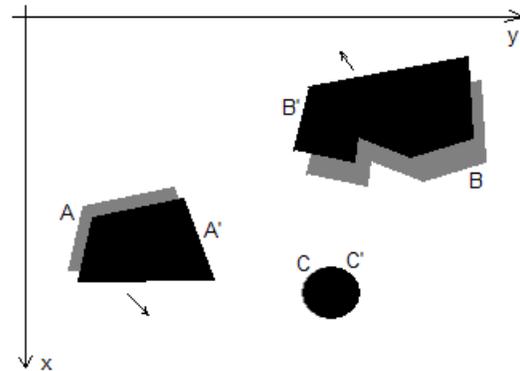


Fig. 8. Environment 2-D captured at two different instants.

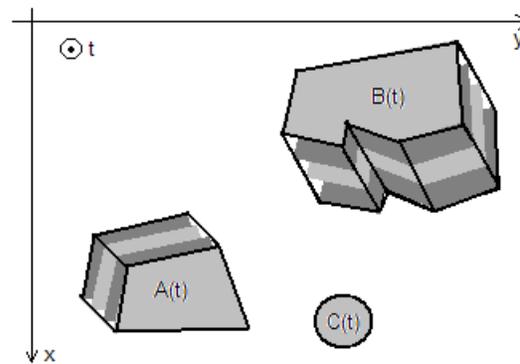


Fig. 9. In space-time environment representation, A and B have moved upon their speed directions therefor C stands at the same place

To find the path between S and T, the three-dimensional space is used. In this case, providing that the shortest path is comprised of straight line segments connecting the vertices visible of the obstacles is always true. Therefore, the algorithm uses the visibility graph and with the assumptions already made, a simple modification of our algorithm can find the path without collision over time.

As the moving object moves at constant speed v and occupies a position W of the Time-Space Representation, it can jump to the possible positions defined by the surface of the cone (C) from the vertex W . This surface can be generated as follows (See Figure 10):

- Constructing the vector which origin is W and making an angle θ with the plane (P) containing W with $\theta = \arctan(v^{-1})$.
- Sweep around this vector while keeping constant the angle θ .

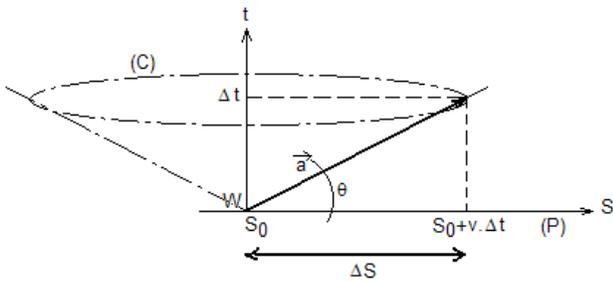


Fig. 10. The cone points representing the moving object can reach

Since there are obstacles (volume in space-time), the intersections of the edges of surfaces with cone (C) are interesting. A set called W-Visible containing all the points of the visible space under angle θ , is used for the construction of the graph searched for our algorithm. These points define the W-Visible sequences under the angle θ and correspond to the vertices W-Visible in the plane (X, Y) (See figure 11).

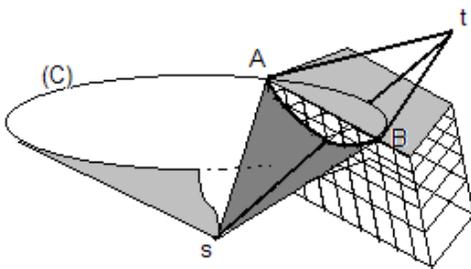


Fig. 11. Determination of visible points on an obstacle with a 3-D slide

B. Complexity of the algorithm

The complexity of the algorithm is determined in the worst case as follows:

Each iteration of the algorithm consists of four steps. The first is to look for a node in the tree a set of points on the obstacle in $O(n)$ operations. The second gives the set of pair of points visible under the angle θ in $O(n)$ operations. The third step determines the extreme points in $O(n)$ operations. The last step removes the points corresponding to the vertices obtuse $O(n)$ operations. Then each iteration requires $O(n)$ operations. The algorithm performs at most n iterations resulting in a worst case complexity of $O(n^2)$.

This complexity increases if the moving object is supposed to have different speeds or obstacles change direction of movement repeatedly.

V. APPLICATION : AIRCRAFT TRAJECTORY PLANNING WITH COLLISIONS AVOIDANCE

Two specific models of aircraft trajectory path planning and optimization, centralized and distributed, have been examined and analyzed. Richards has models which solve this problem, each with their own respective benefits and drawbacks. Simulation results for each model and method have been presented in [29][30]. In general, the centralized model provides the best solution for multiple aircraft and collision avoidance,

whereas the distributed model offers a worse solution in shorter computational time.

Our approach can be considered as a method for finding optimal trajectories for multiple aircraft avoiding collisions. Developments in spacecraft path-planning have shown that trajectory optimization including collision avoidance can be drawn as a cubic form. An approximate model of aircraft dynamics using only cubic construction enables our approach to be applied to aircraft collision avoidance (See figures 12 and 13).

The formulation of our approach can also be extended to include multiple waypoint paths planning, in which each vehicle is required to visit a set of points in an order chosen within the optimization.

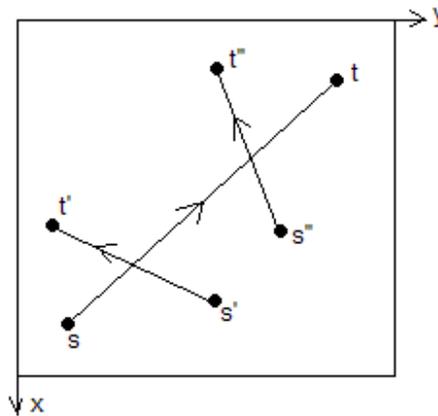


Fig. 12. Example of three trajectories that intersect on the same altitude

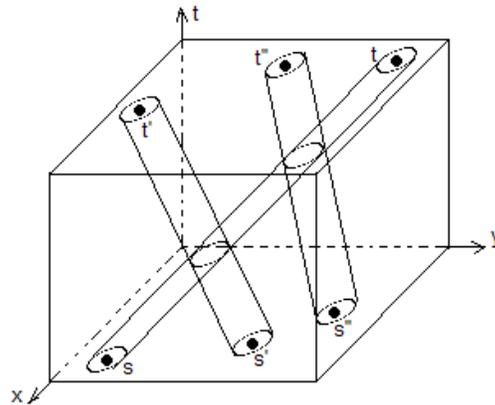


Fig. 13. The representation of trajectories in space-time

VI. CONCLUSIONS

The algorithm presented in this paper requires no pre-treatment or constraints on geometric shapes of obstacles. In addition, the binary tree is a data structure accessible and easily feasible with any language.

The first algorithm allows exploiting the latest work by using a small number of nodes to find the shortest path in a more improved while keeping the same time optimization criteria.

The extended algorithm determines a collision-free path is not necessarily optimal in the presence of moving obstacles. The environment is modeled in a space-time three-dimensional. As we found it difficult to simulate, a 3D viewing environment is being designed.

In the future work, this approach can be applied to resolve several problems such as multi robot environment and multiple waypoint paths planning, in which each vehicle or robot is required to visit a set of points in an order chosen.

REFERENCES

- [1] T. Lozano-Perez, M. Wesley, « An algorithm for planning collision free path among polygonal obstacles », Communication of the ACM, Vol. 22, no. 10, pp. 560-570,1979.
- [2] T. Lozano-Perez, « Spatial planning : a configuration space approach », IEEE Trans. Comput. , vol. 32, pp. 108-120, 1983.
- [3] R.A Brooks, « Solving the find-path problem by good representation of free space », IEEE Trans. on SMC, vol. 13, pp. 190-197, 1983.
- [4] J.L. Crowley, « Navigation for an intelligent mobile robot », IEEE J. Robotics automation, Vol. 1, pp. 31-41, 1985.
- [5] G. Borgefors, « Distance transformations in digital images », Computer vision graphic image processing, Vol. 34, pp. 344-371, 1986.
- [6] J. L. Crowley, « Dynamic world modeling for an intelligent mobile robot using a rotating ultrasonic ranging device », Proceeding IEEE Int. Conf. Robotics Automations, pp. 128-135, March 1985.
- [7] R.A. Brooks, « A robust layered control system for a mobile robot », IEEE J. Robotics Automation, Vol. 2, pp.14-23, March 1986.
- [8] M.G. Slack, « Planning path through aspatial hierarchy: Eliminating stair-stepping effects », Proceeding SPIE conf. Sensor Fusion, Nov. 1988.
- [9] J.S.B. Mitchell, « An algorithmic approach to some problem in terrain navigation », Artificial Intelligent, Vol. 37, no. 1, pp. 171-201, 1988.
- [10] J. Latombe, « Robot motion planning », Kluwer Academic Press, 1991.
- [11] C. Alexopoulos, P.M. Griffin « Path planning for a mobile robot », IEEE Trans. SMC, Vol 22, no. 2, pp. 318-322, March/April 1992.
- [12] M.G. Slack, « Navigation templates: Mediating qualitative guidance and qualitative control in mobile robots », IEEE Trans. SMC, Vol. 23, pp.452-466, 1993.
- [13] Kuo-Chin Fan, Po-Chang Lui, « Solving find path problem in mapped environment using modified A* algorithm », IEEE Trans. on SMC, vol. 24, pp. 1390-1396, 1994.
- [14] G. Conte, R. Zulli, « Motion planning and collision avoidance robots using distance field », Proceeding of the 3rd international workshop on robotics in Alpe-Adria Region, Bled, juil. 1994.
- [15] G. Conte, S. Longhi, R. Zulli, « An algorithm for non-holonomic motion planning », Proceeding ISMCR, Slovakia, juin 1995.
- [16] G. Conte, R. Zulli, « Hierarchical path planning in multi-robot environment with a simple navigation function », IEEE Trans. on SMC, vol. 25, pp. 651-654, 1995.
- [17] R. Kimmel, A. Amir, A. Bruckstein, « Finding shortest path on surfaces using level sets propagation », IEEE Trans. PAMI, Vol. 17, no. 6, pp. 635-640,1995.
- [18] N. S. V. Rao, « Robot navigation in unknown generalized polygonal terrains using vision sensors », IEEE Trans. On SMC, Vol. 25, pp.947-962, 1995.
- [19] J. Elmesbahi, A. Raihani, M. Elkhaili, O. Boattane, « A region decomposition methode for path planning using a distance field », Proceeding IFIP-WG7, Noisy-le-grand, France, avril 1996.
- [20] J. Elmesbahi, A. Raihani, M. Elkhaili, « A fast algorithm for path planning problem using region decomposition », Proceeding ITHURS-96, Leon, Espagne, juillet 1996.
- [21] I.Ulrich, J. Borenstein, VFH* : Local Obstacle Avoidance with Lookahead Verification In IEEE int. conf. on Robotics and Automation, San Francisco, USA, 2000
- [22] J. Thomas, A. Blair, N.Barnes, “Towards an efficient optimal trajectory planner for multiple mobile robots, Proceedings of the 2003 International Conference on Intelligent Robots and Systems, 2291-2296.
- [23] K.H. Sedighi, K. Ashenayi, T.W. Manikas, R. L. Wainwright, H.M. Tai, “Autonomous Local Path Planning for a Mobile Robot Using a Genetic Algorithm”, Proc. 2004 IEEE Congress on Evolutionary Computation (CEC2004), p. 1338-1345.
- [24] Kramer, J.,Scheutz, M., “Development environments for autonomous mobile robots: A survey”, C _Springer Science + Business Media, LLC 2006.
- [25] L. Montesano, J. Minguez, L. Montano. “Modeling Dynamic Scenarios for Local Sensor-Based Motion Planning”, Autonomous Robots, Vol. 12, No3, 231-251, 2008
- [26] Meng, Rui, Su, Wei-Jun, Lian, Xiao-Feng. ”Mobile robot path planning based on dynamic fuzzy artificial potential field method” Computer Engineering and Design, Vol. 31, no. 7, pp. 1558-1561. 16 Apr 2010
- [27] H.Adeli, M.H.N. Tabrizi, A. Mazloomian, E. Hajipour,M.Jahed, “Path Planning for Mobile Robots using Iterative Artificial Potential Field Method”, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 2, July 2011
- [28] Tamilselvi, Mercy shalinie, Hariharasudan, “Optimal Path Selection for Mobile Robot Navigation Using Genetic Algorithm”, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011
- [29] A. Richards, “Aircraft trajectory planning with collision avoidance using mixed integer linear programming”, American Control Conference, 2002. Proceedings of the 2002
- [30] Matthew W. Zinn, “Analysis of Aircraft Trajectory Models with Collision Avoidance as a Mixed Integer Linear Program”, College of William & Mary, July 10, 2013

Online Monitoring System Design of Intelligent Circuit Breaker Based on DSP and ARM

Meng Song¹

College of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Songjiang District, Shanghai 201620, China

Liping Zhang^{2*}

College of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Songjiang District, Shanghai 201620, China

Yuchen Chen³

College of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Songjiang District, Shanghai 201620, China

Weijin Zheng⁴

College of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Songjiang District, Shanghai 201620, China

Abstract—In order to accurately analyze the dynamic characteristics of the vacuum circuit breaker, a dual-core master-slave processor structure for online monitoring system based on DSP and ARM is proposed. This structure consists of host computer, lower computer and signal processing modules. The lower computer uses DSP as the core, which completes acquisition and data preprocessing of circuit breaker's dynamic characteristics through sensors and signal conditioning circuits. The host computer uses ARM as the core which is responsible for task management, analysis, processing and displaying collected data via Ethernet. The communication between DSP and ARM is conducted by HPI. This design improves the reliability of intelligent control unit for the circuit breaker. The experiment showed that this system works steadily and accuracy.

Keywords—circuit breaker; online monitoring; ARM; DSP

I. INTRODUCTION

The vacuum circuit breakers is not only the most important control and protection equipment in power system, but also the equipment with most frequent mechanical and electrical operation in substations, which affects the stability and reliability of the power grid directly[1]. How to implement online monitoring of vacuum circuit breakers more efficiently has become an important technical topic for domestic and foreign engineers. With the popularization of smart grid technology, the power system is developing to intelligent control system. As an important part of the grid, the breaker's intelligent level has a significant impact on intelligent power grids. The embedded monitoring system in circuit breakers for online intelligent monitoring has become a major research topic for intelligent circuit breaker [2~4].

Online condition monitoring is the basis for the condition maintenance. Existing online monitoring system has the following problems: less monitoring parameter and single function. There are still some deficiencies in management and exchange of the data, scalability and remote maintenance in the system. Moreover, it is difficult to make a systematic and comprehensive evaluation in the system. In order to overcome these deficiencies mentioned above, a dual-core master-slave processor structure for online monitoring system based on

ARM and DSP is proposed. ARM is typical of low power consumption, rich interfaces and strong ability to control, and DSP process data fast with high accuracy. The combination with ARM and DSP will improve the real-time and versatility of the circuit breaker online monitoring system.

II. SYSTEM STRUCTURE

The intelligent vacuum circuit breaker online monitoring system is composed by the host computer, the lower computer and signal processing modules. The lower computer hardware platform consists of a TMS320F2812 DSP and peripheral hardware circuits. It collects mechanical parameters, divide-shut brake circuit current signal and vibration signal of vacuum circuit breaker. The host computer uses ARM as a core, mainly working as remote communication with the host computer, and getting the results of data processing and eigenvalues from the DSP at the same time. The preprocessed data is transferred from DSP to ARM via a HPI interface, and transmitted via the Ethernet interface to the host computer for data analysis and processing, so as to determine the current status of the circuit breaker, and analyze its operation situation and diagnose malfunctions. The system structure is shown as Fig 1.

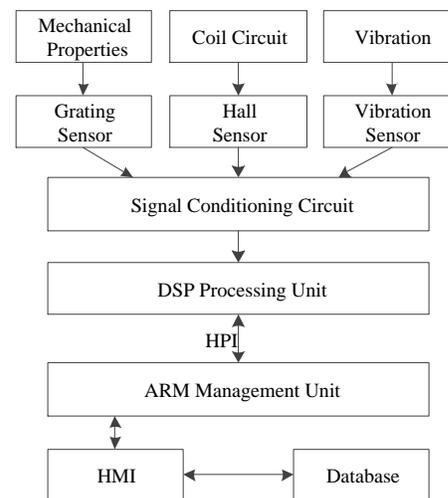


Fig. 1. The Online Monitoring System Diagram

III. THE EXTRACTION OF ONLINE MONITORING CHARACTERISTIC QUANTITY

Main monitoring items for vacuum circuit breaker include: mechanical characteristics, divide-shut brake circuit signal and vibration signal.

Main monitoring items of the mechanical properties of the vacuum breaker include the time, speed, distance and overtravel of the divide-shut brake [5]. The connection between contacts and actuator of circuit breaker is shown as Fig2. Among them, 1 is the moving contact, 2 is the static contact, the length of the link rod 3, link rod 4, link rod 5 are L_1, L_2, L_3 respectively.

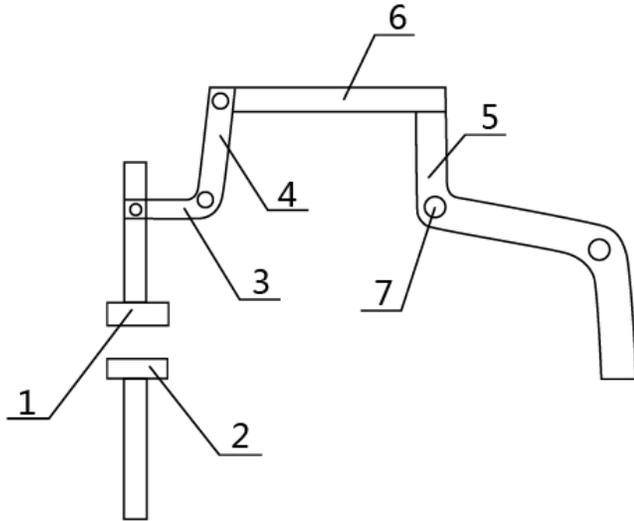


Fig. 2. The Structure for Connection of Contact and Actuator

Due to the structure is very compact and with high potential isolation problem, sensor can't be installed in the insulated rod of moving contact. Thus, the system is installed the angular displacement sensor on the actuator spindle of circuit breaker. According to the relationship between the measured voltage change amount ΔU and the reference voltage U_0 , angular displacement sensor turned as the angle α and measured spindle turned as the angle β with the following relationship:

$$\beta = \alpha \cdot \Delta U / U_0 \quad (1)$$

The actuator spindle turned the angle as β , the displacement of insulated pillar linear is X , the link rod 3, 4 turned angle as γ and the displacement of dynamic contact linear Y has the following relationship:

$$\begin{aligned} L_3^2 - X^2 &= 2 \cos \gamma \cdot L_3^2 \\ L_2^2 - X^2 &= 2 \cos \gamma \cdot L_2^2 \\ L_1^2 - Y^2 &= 2 \cos \gamma \cdot L_1^2 \\ Y &= (L_1 \cdot L_3 / L_2) * \sqrt{1 - 2 \cos \beta} \end{aligned} \quad (2)$$

Meanwhile, the test system reads the pulse signal with a sampling frequency and obtains the trip - time characteristic curve of the circuit breaker during the operation. Using the disposed principal axis angular displacement - time curve of the divide-shut brake we can obtain the linear displacement -

time curve for the moving contact, and then obtain the time, velocity, distance and overtravel of the divide-shut brake [6].

The main monitoring items of the divide-shut brake circuit include divide-shut brake coil current, the current and voltage of charging motor. Since the judgment is based on whether the current pass through the coil of divide-shut brake or not, the measure of coil currents has direct impact on the measurement accuracy and stability of system. Therefore, hall sensor with high precision, good linearity, small size and good dynamic performance is used to complete data acquisition.

The vibration signal, which consists of a series of transient waveforms, contains a large number of status information of device. Each transient waveform is reflected by the "incident" signal during breaker operation. Vibration is a response to the internal excitation source of the device, identifying the vibration excitation source through appropriate means of detection and signal processing methods so as to find out the source of the fault.

IV. HARDWARE DESIGN

The hardware platform of the dual-core master-slave processor online monitoring system for vacuum circuit breaker consists of ARM, DSP and signal conditioning circuit. The main processor consists of ARM and peripheral circuits, works as embedded operating system between the user and the co-processor. ARM manages and maintains the system configuration and communicates with the host computer via Ethernet. At the same time, ARM accesses the processing data and calculated eigenvalues from DSP in real-time.

As the coprocessor of lower computer, DSP is divided into three parts in accordance with the function: ARM interface circuit, DSP coprocessor circuits and power systems. ARM interface circuitry is used to expand peripheral interface circuit of ARM core board, including UART, USB, Ethernet. The lower computer uses TMS320F2812 DSP as core, and uses host interface HPI port as master-slave processor synchronization and communications hardware interface. ARM completes the management task for the lower computer, and power system supply power for these two chip modules.

A. Interface Design

Communication between ARM and DSP is performed by HPI (Host Port Interface). HPI is the chip peripheral to communicate with the host computer. The host computer can easily access all the address space of DSP and control it via HPI. Compared to the dual-port RAM, serial interface, HPI does not need to add peripheral logic circuits. There is not extra hardware and software overhead in communication with the host computer via HPI. DSP doesn't interrupt normal program running because it can coordinate with hardware conflict itself. HPI is suitable for the applications of high real-time requirements, a large quantity of data rate [7, 8].

TMS320F2812 is connected with ARM by the separate 16-bit data lines HD and nine control lines as shown in Fig 3. In the process of communication between ARM and DSP via HPI, DSP only needs active participation in interrupting state. DSP is in a passive state in other state. DSP is equivalent to an external memory for ARM.

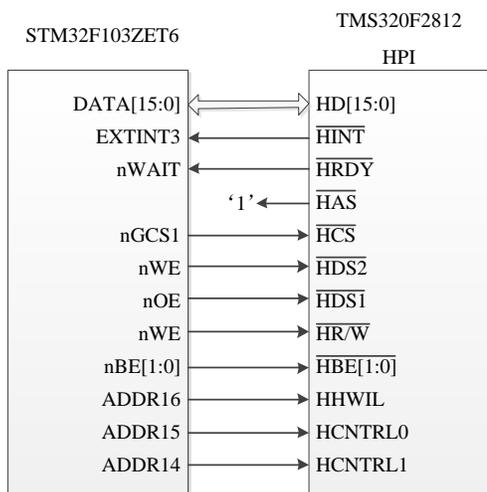


Fig. 3. Hardware Interface

B. Program Development

Firstly, the DSP firmware is divided by the according functions and required to complete initialization, data acquisition, data processing, data storage, which correspond to three task threads, defined as TskSplit, TskProcess and TskSave. Data acquisition is to obtain real-time data from the sensor, and then the data is divided by channel. The data processing mainly is to process real-time data by channel. There may have one or more arithmetic unit within the channel depending on the application needs, by which to analysis data and obtain the status information of the object. Data storage is to storage processed data to the specified physical address space according to the agreed format.

The processes for the entire program are:

Firstly, the device driver notifies TskSplit to access data from the object data via SIO after obtaining data from the external device. TskSplit will be used after data is read and divided into channel data zones.

Secondly, TskProcess executes the processing program by channel sequentially, and stores the processed data to the processing data buffer.

Finally, TskSave gets data through the pointer of data buffer to organize and store, then returns the empty data buffer after be read to TskProcess.

V. SOFTWARE DESIGN

The system software coordinates with hardware to achieve the monitoring of the circuit breaker's dynamic characteristics. The program flow is shown as Fig 4.

The software system's task is to complete initialization of the system, extraction of the signal, data processing, information monitoring and communications with the host computer. Firstly, the program initializes each module, and opens the interrupt event after the system is powered.

When vacuum circuit breaker is opened and closed, an A/D channel starts to collect the opening and closing current signal, the vibration signal parameters and mechanical properties by the way of interruption. Then the data processing program performs. Finally, the data is transferred to the liquid crystal display module. If the lower computer has action, the host computer receives data recovery, then the host computer call data from the lower computer.

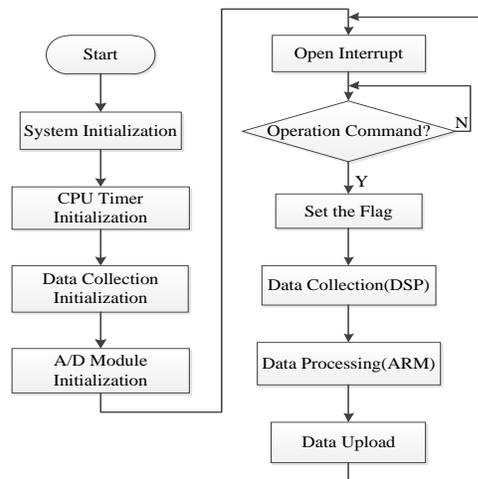


Fig. 4. Software Program Flow Chart

VI. TEST RESULT

The online monitoring system device is installed on ZN63A-12 vacuum circuit breaker. The parameters of the coil circuit signal and mechanical properties are tested. The results of the coil circuit shown in Table I is the comparison between oscilloscope and experimental data.

TABLE I. CIRCUIT ELECTRICAL PARAMETERS RESULTS

Parameter	Control Voltage(V)	Closing Current(A)	Tripping Current(A)	Storage Current(A)
Oscilloscope	113.25	1.97	1.93	2.01
System	114.03	2.01	1.96	1.98
Error	0.69%	2%	1.6%	1.5%



Fig. 5. The Current Curve of Coil Circuit

The current curve of coil circuit is shown as Fig 5. The result shows that the error between the system values and standard values is small.

In order to test the mechanical properties of the circuit breaker, the monitoring system is compared with mechanical properties tester.

The results are shown in Table II and the tripping displacement - time characteristic curve and closing displacement - time characteristic curve are shown as Fig 6 and Fig 7.

The result shows that the error between the system values and standard values is small.

TABLE II. THE TEST RESULTS OF MECHANICAL PROPERTIES

Parameter	Distance (mm)	Closing Time(ms)	Closing Speed(m/s)	Tripping Time(ms)
Tester	11.879	32.246	0.79	22.03
System	11.537	33.014	0.81	21.74
Error	2.9%	2.4%	2.5%	1.3%

Parameter	Tripping Speed(m/s)	Ultra-trip (mm)
Tester	0.98	2.97
System	0.965	3.08
Error	1.5%	3.7%

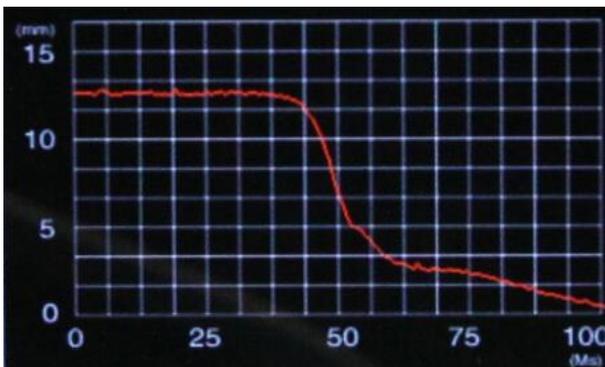


Fig. 6. Tripping Displacement - Time Curve

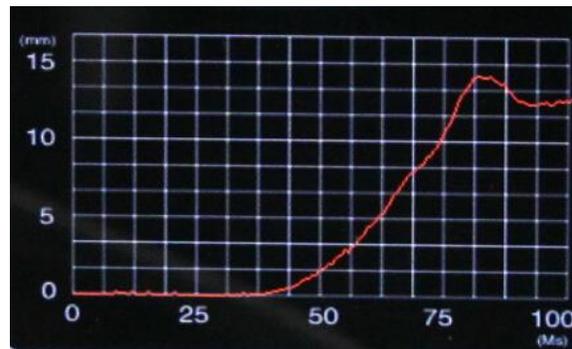


Fig. 7. Closing Displacement - Time Curve

As seen from the above tables and figures, the mechanical characteristics, the tripping, closing switching circuit of the circuit breaker is consistent with the measurement requirements. Moreover, it also truly reflects the actual work situation and effectively reaches its monitoring goals.

ACKNOWLEDGMENT

The project has been supported by the Innovation Program of Shanghai Municipal Education Commission "Vehicle Collision Avoidance System based on Vehicle Wireless Communication" (No.12YZ151).

REFERENCES

- [1] XuGuo-zheng, et al. Principles and applications of high voltage circuit breakers [M]. Beijing Tsinghua University Press, 2000.
- [2] Lü Yi-hang, Li Jing, Dai Huai-zhi, et al. Research and Development of the On-line Synthetic Monitoring System of High Voltage Circuit Breakers[J]. Electric Power, 2004, 37(3) : 68-71.
- [3] Ma Qiang, Li Kun, Rong Ming-zhe. Design of System on Chip for Switchgear On-line Monitoring [J]. Automation of Electric Power Systems, 2005, 29(3):73-76
- [4] Dai Huai-zhi, Lü Yi-hang, Jia Shen-li, et al. Design of On-line Monitoring System for Circuit Breaker [J]. High Voltage Apparatus, 2004, 40(2): 104-106
- [5] H. H. Zeineldin, et al. High Voltage Circuit Breaker Modeling for Online Model-Based Monitoring and Diagnosis[J]. IEEE Transactions on Power Delivery, 2008, 3(4): 317-322.
- [6] Xiong Xiao-fu, et al. Distributed monitoring system based on DSP and ARM for mechanical characteristic of high voltage [J]. Power System Protection and Control, 2009, 3(16)
- [7] Liu Qiao, Miao Si-en, Implementation of the Communication between ARM and DSP Based on HPI[J]. Communications Electronics, 2010, 47(3):52-54.
- [8] Liu Chang, Tao Ran, Host-Port Interface (HPI) Application Based on DSP and ARM[J]. Dual-use technologies and products, 2006, 41-42

Evaluating Usability of E-Learning Systems in Universities

Nicholas Kipkurui Kiget
Department of Computer Science
Masinde Muliro University of
Science and Technology
Kakamega, Kenya

Professor G. Wanyembi,
Department of Information
Technology
Mt. Kenya University
Thika, Kenya

Anselemo Ikoha Peters
School of Computing and
Informatics
Kibabii University College
Bungoma, Kenya

Abstract—The use of e-learning systems has increased significantly in the recent times. E-learning systems are supplementing teaching and learning in universities globally. Kenyan universities have adopted e-learning technologies as means for delivering course content. However despite adoption of these systems, there are considerable challenges facing the usability of the systems. Lecturers and students have different perceptions in regard to the usability of e-learning systems. The aim of this study was to evaluate usability attributes that affect e-learning systems in Kenyan universities. The study had two fold objectives; determining status of e-learning platforms and evaluating usability issues affecting e-learning adoption in Kenyan universities. The research took a case study of one of the public universities which has implemented Moodle e-learning system. The usability attributes evaluated were user-friendliness, learnability, technological infrastructure and policy. The research made recommendations which could help universities accelerate the adoption of e-learning systems.

Keywords—e-learning; moodle; usability; learnability

I. INTRODUCTION

E-learning systems are becoming accepted tools for teaching and learning. The e-learning systems provide a platform for using computers to improve education. According to [1], computers have become useful not only in corporate world but also in Education.

Popularity of e-learning systems is attributed to their key benefits. When applied correctly, e-learning systems can have the following benefits; reduced teaching and learning costs, reduced teaching and learning time, more effective learning / better lecturer productivity, more consistent learning, flexible delivery / distance delivery, measurable learning, recognition of prior learning and multi-cultural learning. The value of E-Learning is to fully enable “learning anywhere at any time” by providing an array of resources, opportunities for active participation, mastering content and self learning [2].

As noted by Nielsen [3], inadequately equipped e-learning systems can result in frustration, anxiety, confusion, and reduce learners’ interest. Miller [4], states that poor usability is a major contributor to lack of adoption of most e-learning systems. Usability of e-learning systems influence the way learners evaluate their learning experience, if usability of e-learning system is bad, learners fail in their attempt to use the

system. These factors hinder the usability of e-learning systems adopted in an institution of higher learning.

II. CONCEPTUAL FRAMEWORK

The conceptual framework in fig 1 shows the relationship linking the usability issues and the e-learning platforms; the usability issues that affect an e-learning platform are user friendliness, user satisfaction, learnability and errors [5]. The contravening factors are institutional strategies and policies, cultures and legal issues, demographic factors and technological infrastructure.

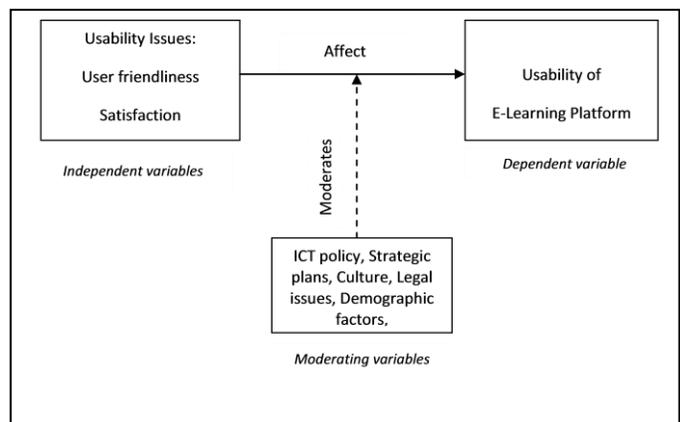


Fig.1. Conceptual framework

Source: Author (2014)

III. HYPOTHESIS OF THE STUDY

The following hypotheses were tested:

- H₀₁: There is statistically no significant difference between the learnability and the usability of an e-learning system
- H₀₂: There is no statistically significant difference between the user-friendliness and the usability of an e-learning system
- H₀₃: There is no statistically significant difference between the usability policy and the usability of an e-learning system
- H₀₄: There is no statistically significant difference between the infrastructure and the usability of an e-learning system

IV. REVIEW OF RELATED LITERATURE

E-learning systems, also called virtual learning environments (VLE's), are systems that use modern

information and communication technology to aid education and training efforts [6]. A purpose of an e-learning system is to distribute the learning materials to the users. It enables instructors and learners to post content, participate in discussions, maintain a grade book, keep a roster, track participation, and generally engage in and manage learning activities in an online environment [7].

Learning Management Systems (LMS) fall in the category of technology delivered e-learning. It is used to support teaching and learning. LMS is the backbone of e-learning, which is a software system integrating web-based training, classroom delivered courses, online courses and human resources system as stated by [8].

The role of Learning Management System (LMS) is to deliver e-learning courses in a self paced approach. Through LMS lecturers are able to publish courses in an online catalog and also be able to assign online courses to the learners who then log in to the LMS using an internet browser and start the courses. LMS will then track the learners' activities and provide upto date reports for each course and each learner. Common LMS are Moodle, WebCT, Blackboard, eLeap and Desire2Learn. This study targeted the users of Moodle e-learning system.

A. The need to evaluate e-learning system

Evaluation is a course of action for determining the value and effectiveness of a learning system with benefits such as error correction, establishing the users' point of view and reducing unsupportable design issues in a system [9]. Shepherd [10] in his article, states that there are four reasons for evaluating e-learning system; validating training as business tool, justifying costs incurred in training, help improve design of training, and to help in selecting training methods. In the context of this study, the evaluation of the e-learning system is to help enhance the design of the training and the design of the e-learning system.

B. Usability of E-learning Systems

According to International Organization for Standardization (ISO) 9241 [11], usability is defined as the degree to which a particular product is used by particular users to accomplish specific goal with efficiency, effectiveness and satisfaction in a precise standpoint used. Majority of the past studies on the usability of E-learning systems have been on exploring the usability of interface of E-learning systems and the links between usability features and the E-learning success.

Usability has been defined differently as specified in components that are more specific i.e. learnability, memorability, errors and efficiency [12]. Nielsen [3] gives attention to expert users when talking about efficiency though learnability is directly related to efficiency. Memorability mostly relates to casual users and errors deal with those errors not covered by efficiency, which have more disastrous results. A comparable definition is given by Shneiderman[13]; Shneiderman while defining usability of e-learning system looks at it as five measurable human factors central to evaluation of human factors goals; speed of performance, time to learn, retention over time, rate of errors by users and

subjective satisfaction. Dix [14] defines concepts entailing system usability; learnability, flexibility and robustness signifying that those concepts are on the similar abstraction level.

C. User-Based Evaluation of usability of e-learning systems

User-based evaluation presently provides complete form of evaluation, since it assesses usability by picking samples of real users. A suitable technique used in the research was the system inquiry in which users are asked to give their opinions or views on the way they perceive the system after using for some time.

V. METHODOLOGY

The research used case study approach which is one of the most widely used qualitative research method in information systems research [15], owing to its ability to understand the interaction between information technology and organizational contexts in a thorough manner. The population was drawn from students and lecturers using learning management system (Moodle). A sample of 20 lecturers and 30 students was used for the study. The study used questionnaires and interviews to collect the information from the respondents. Questionnaires from twenty five (25) students and fifteen (15) lecturers were dully filled, returned and used for analysis. This represented a response rate of 83% and 75% respectively.

VI. FINDINGS AND DISCUSSION

A. Gender Distribution of the Respondents

Table I shows the gender disparity of the respondents. From the results, it is evident that 70.0% were male while 30% were female respondents. This shows that there are few female participants using the e-learning system. Among the lecturers, 73.3% were male while 26.7% were female. As for the students 68.0% were male while 32% were female.

TABLE I. GENDER DISTRIBUTION

Category		Gender		Total	
		Male	Female		
1	Lecturers	Frequency	11	4	15
		%	73.3%	26.7%	100.0%
2	Students	Frequency	17	8	25
		%	68.0%	32.0%	100.0%
Total		Frequency	28	12	40
		%	70.0%	30.0%	100.0%

B. Level of Education and Category of Lecturers

Table II shows staff category and level of education of the lecturer respondents. The most common educational qualification of the respondents was the Masters, with a total of 10 representing 66.67% of the lecturers; followed by degree holders representing 13.3% while the least was the Doctorate representing 13.33%. On the staff category most of the respondents were lecturers (7) representing 46.67% followed by Senior Lecturers and Assistant Lecturers each 3 representing 20%, the least was Graduate Assistant / Tutorial Fellow representing 13.33%

TABLE II. LEVEL OF EDUCATION AND STAFF CATEGORY

		Staff Category				Total
		Graduate Assistant / Tutorial Fellow	Assistant Lecturer	Lecturer	Senior Lecturer	
Degree	Count	2	1	0	0	3
	%	13.33	6.67	0	0	20
Masters	Count	0	2	6	2	10
	%	0.00	13.33	40.00	13.33	66.7
PHD	Count	0	0	1	1	2
	%	0	0.00	6.67	6.67	13.3
	Count	2	3	7	3	15
	%	13.33	20	46.67	20	100

C. Experience of the Respondents with E-learning System

On the use of e-learning systems, all the respondents answered on affirmative as shown in table III. This means that all the respondents have an experience with e-learning systems. From the results, uploading, assignments, quiz and forum are the most frequently used modules recording 97.5%, 100%, 65% and 95% respectively. The least used modules are journal, chats, workshop and choice at 32.5%, 15%, 32.5% and 42.5% of the respondents respectively.

TABLE III. E-LEARNING SYSTEMS USED

E-learning System	Yes	No
Wiki	57%	43%
Moodle	100%	0%
WebCT	5%	95%
Blackboard	5%	95%
Sakai	0%	100%

D. Factors hindering implementation of e-learning systems

As shown in table IV, 97.5% of the respondents agreed that lack of equipment (computers) hinders e-learning implementation, 85% of them didn't agree with the course quality concerns as a factor. 95% of the respondents agreed on the factor of access speeds while 57.5% did not think that lack of skills was a major contributing factor in hindering e-learning systems. All respondents (100%) did not agree to the following factors; lack of interest, not aware of its availability, legal concerns, plagiarism and course not suited to be implemented on the e-learning platform. 32.5% of the respondents agreed that institutional traditional culture and lack of motivation are hindering implementation of e-learning while 37.5% of them agree that policy is also affecting its implementation. A considerable number 42.5% also thought that having no usability policy in place was also affecting use of e-learning system.

TABLE IV. FACTORS HINDERING E-LEARNING

Factor	Yes	No
Lack of Equipment (Computers)	97.5%	2.5%
Course quality concerns	15%	85%
Access Speeds	95%	5%
Lack of Skills	42.5%	57.5%
Lack of Interest	0%	100%
Not aware of its availability	0%	100%
University Traditional Culture	32.5%	67.5%
Lack of Motivation	32.5%	67.5%
Lack of policy	37.5%	62.5%
Legal Concerns	0%	100%
Course not suited for E-learning	0%	100%
Plagiarism concerns	0%	100%

Factor	Yes	No
Lack of Equipment (Computers)	97.5%	2.5%
Course quality concerns	15%	85%
Access Speeds	95%	5%
Lack of Skills	42.5%	57.5%
Lack of Interest	0%	100%
Not aware of its availability	0%	100%
University Traditional Culture	32.5%	67.5%
Lack of Motivation	32.5%	67.5%
Lack of policy	37.5%	62.5%
Legal Concerns	0%	100%
Course not suited for E-learning	0%	100%
Plagiarism concerns	0%	100%

E. Moodle modules used by respondents

Table V shows that uploading, assignments, quiz and forum are the most frequently used modules recording 97.5%, 100%, 65% and 95% respectively. The least used modules are journal, chats, workshop and choice at 32.5%, 15%, 32.5% and 42.5% of the respondents respectively.

TABLE V. MOODLE MODULES USED

Modules used	Yes	No
Uploading	97.5%	2.5%
Assignment	100%	0%
Quiz	65%	35%
Journal	32.5%	67.5%
Chats	15%	85%
Workshop	32.5%	67.5%
Forum	95%	5%
Choice	42.5%	57.5%
Any other	-	-

F. Hypotheses tested

a) Hypothesis 1

There is no statistically significant difference between the learnability and the usability of an e-learning system.

Null Hypothesis

Learnability factors do not affect the usability of e-learning system.

Alternative Hypothesis

Learnability factors affect the usability of e-learning system.

To determine whether there was or no statistical significant difference between learnability and the usability of an e-learning system, a linear regression was used. Usability factor was tested with the following learnability factors: "Learning to use LMS system is easy for me", "Exploring new modules by trial and error is easy", "Easy to be skilful with the LMS" and "Easy to upload and download using LMS". Table VI shows

linear regression results on learnability issues affecting the usability of e-learning system. From the results, the learnability factors significant are “Learning to use e-learning system is easy” with $p=0.044$, “exploring new modules by trial and error is easy”, with $p=0.701$ and “Easy to be skillful” being strongly significant with $p=0.009$ while “Easy to Upload / Download” has significance of $p=0.346$

TABLE VI. SIGNIFICANCE OF LEARNABILITY FACTORS

Model		Sig.
	Learning to use is easy	0.044
	Exploring new modules by trial and error easy	0.701
	Easy to be skillful	0.009
	Easy to Upload / Download	0.346

Thus the Null hypothesis is disproved and the alternative accepted that there is a significant difference between the learnability factors and usability of the e-learning system. This means that the learnability factors affect the usability of e-learning system considerably. If a system is not easy to learn then it affects its usability. This is in agreement with Ghaoui [5] who while surveying usability issues affecting e-learning systems stated that learnability was one of them. Higher learnability therefore relates to greater usability.

b) Hypothesis 2

There is no statistically significant difference between the user-friendliness and the usability of an e-learning system.

Null Hypothesis

User-friendliness factors do not affect the usability of e-learning system.

Alternative Hypothesis

User-friendliness factors affect the usability of e-learning system.

To determine the significant difference between the user-friendliness and usability of e-learning system, a linear regression was done on the following variables: usability on one hand and user-friendliness factors on the other hand. The user-friendliness factors identified were “Accessing menus and commands is easy for me” and “My interaction with LMS is clear and understandable”. Table VII shows the results of the correlations.

TABLE VII. ANALYSIS OF USER-FRIENDLINESS FACTORS

Model		Sig.
	Accessing Menus and Commands easy	0.007
	Interaction with LMS clear	0.002

From the results, “Accessing Menus and Commands easy” is strongly significant with $p=0.007$, while “Interaction with LMS clear” is also strongly significant with $p=0.002$. According to the Pearson Correlation the user-friendliness factors affects the usability of e-learning system. Therefore this negates the hypothesis hence there is a significant difference between the user-friendliness and usability of e-

learning system. This therefore means that user-friendliness affects the usability of e-learning systems.

The results conforms with Yildrin [8]; according to Yildrin [8], there are four key issues to successful LMS; general features and functionality / user-friendliness, content/ID, Support tools and management and technical; infrastructure

c) Hypothesis 3

There is no statistically significant difference between the usability policy and the usability of an e-learning system.

Null Hypothesis

Usability policies do not affect the usability of e-learning system.

Alternative Hypothesis

Usability factors affect the usability of e-learning system.

To determine whether there was significant difference between the usability policy and usability of e-learning system, a linear regression analysis was carried out between the usability factor and need for a policy factor. Table VIII shows the results of the analysis.

TABLE VIII. USABILITY POLICY IN RELATION TO USABILITY FACTOR

Model		Sig.
1	(Constant)	0.000
	Usability Policy	0.739

With a significance of 0.739 as indicated in the linear regression analysis, the statistical significance difference between usability policy and usability of e-learning system is weak. This means that though the significance is weak, usability policy affects the usability of e-learning system. This seems to agree with Al Rawi [7] who indicated that lack of a policy framework on e-learning has hampered development of technology in institutions of education.

d) Hypothesis 4

There is no statistically significant difference between the infrastructure and the usability of an e-learning system.

Null Hypothesis

Technological infrastructural factors do not affect the usability of e-learning system.

Alternative Hypothesis

Technological infrastructural factors affect the usability of e-learning system.

To test the significance difference between the infrastructure and the usability of e-learning system, a linear regression was performed on usability factor and the infrastructural factors. The infrastructural factors identified were “There is need for more computers for e-learning use” and “I can access e-learning system while in LAN and while on a WAN (Through the web on the Internet)”. Table IX shows the results of the analysis.

TABLE IX. INFRASTRUCTURAL FACTORS IN RELATION TO USABILITY FACTOR

Model		Sig.
1	(Constant)	0.001
	More Computers	0.007
	Can be accessed on LAN and WAN	0.006

VII. CONCLUSION

The aim of the research was to study the usability of e-learning system being used in one of the public universities in Kenya. The factors of usability evaluated were learnability, user-friendliness, technological infrastructure, usability policy, culture and gender. The results from this study showed that a significant number of users agreed that learnability of e-learning system was affecting its usability. The learnability factors tested were the ability of e-learning system to be learnt, exploring new modules by trial and error, ability to be skilful with an e-learning system and ability of users to upload and download using e-learning system. Generally most of the respondents agreed that learning e-learning and using e-learning system was not easy.

This is in tandem with Dix [14] who noted that learnability affects usability of e-learning system. To enhance the adoption of e-learning systems, universities have to enhance the learnability of e-learning systems. Smulders [18], described usability of e-learning as a precursor of learnability.

The research also identified user-friendliness as a factor that affects usability of e-learning system. Majority of the responses agreed that e-learning system has to be user friendly for it to be usable. User-friendliness factors investigated were the ease to access menus and commands and clarity of interaction between the user and the e-learning system.

According to the findings universities need more computers and more training for both lecturers and students to enhance adoption of e-learning system. The e-learning system should also be accessible on a local area network and on a wide area network over internet.

VIII. RECOMMENDATIONS

Though 90% of the respondents agree that they have had training organised by the university, still 90% of them agree to the fact they still need more trainings / workshops on e-learning system. The research therefore recommends more training / workshops are done to enhance learnability of e-learning system.

Three modules of Moodle were frequently used according to the respondents. These were uploading / downloading, assignment and forum. It is recommended that the lecturers and students be trained and encouraged to use other modules which could enhance learning; these include chat, workshop, assignment and quiz. According to Moodle website www.moodle.com (2010), all the seven modules; uploading / downloading, forums, chats, quizzes, Assignments, grades and wikis makes e-learning process complete.

The research recommends enough computers be purchased by the universities for successful implementation of e-learning system. The e-learning system should be accessible both on Local Area Network (LAN) and on internet.

Lack of e-learning policy has affected the usability of e-learning system. Newhouse [19], states that it's through e-learning policy that students can know what the instructor expects from them. It is recommended that universities come up with e-learning policies such as usability policy to guide the learners, lecturers and management staff as they implement the systems. The policy will encourage professionalism in creating, uploading and sharing of digital content by the lecturers and learners.

IX. SUGGESTIONS FOR FURTHER RESEARCH

This research focused on Moodle as a Learning Management System that had been implemented by the university under the study. Comparative study is suggested to look at usability issues of other e-learning systems not covered by this research.

The study took a sample of students and lecturers from computer science department, a further study could be done to understand the perception of other lecturers and students in other departments and other universities on e-learning systems. Additionally, a comparative study on open source and proprietary e-learning systems is also suggested.

REFERENCES

- [1] Sambrook, S., (2003), 'E-learning in Small Organizations', Education + Training, vol.45, no.8/9, pp. 506-516.
- [2] Norman, V. (2007). Perspectives on blended learning in higher education. *Journal of ELearning*, January.
- [3] Nielsen, J. (2005). *Heuristic Evaluation*. Retrieved 22/06/2014, from <http://www.useit.com/papers/heuristic/>
- [4] Miller, M.J. (2005). *Usability in E-learning*. Retrieved 11/07/2014, from www.learningcircuits.org/2005/miller.htm
- [5] Ghaoui, C. (2003). *Usability Evaluation of Online Learning Programs*. Information Sciences Publishing.
- [6] Dowming (2008). What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction, *Computers & Education*, Vol. 50 (4), pp 1183-1202.
- [7] Al-Rawi, A. (2010). Using a learning management system to foster independent learning in an outcome-based university: A gulf perspective. *Proceedings of Issues in Informing Science and Information Technology* (pp. 73-87). Retrieved from <http://iisit.org/Vol7/IISITv7p073-087Lansari733.pdf>
- [8] Yildirim and Temur, H., (2004), Handbook on LMS: *Success factors in e-learning implementation*. John Wiley, Inc.
- [9] A.M. Lund (2004): Measuring usability with the use questionnaire, http://www.stcsig.org/usability/newsletter/0110_measuring_with_use.htm, last accessed: 05/06/14
- [10] Shepherd, C. (2002). *In search of the perfect e-tutor*. Accessed on 15/07/2014 from: http://www.fastrakconsulting.co.uk/tactix/Features/perfect_etutor.htm
- [11] ISO (1998) ISO 9241: *Guidance on Usability Standards*. [On-line]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=16883&ICS1=13&ICS2=180&ICS3
- [12] Nielsen, J. (1993). *Usability Engineering*. New Jersey: Academic Press.
- [13] Smulders, D. (2003). Designing for Learners, Designing for Users, *eLearn Magazine*, last accessed 08/07/2014, from [Http://www.elearnmag.org/](http://www.elearnmag.org/).
- [14] Shneiderman, B. (2004). *Designing the User Interface: Strategies for effective Human-Computer Interaction* (3 ed.). Addison-Wesley.
- [15] Dix, K.L. (2006). A longitudinal study examining the Impact of ICT adoption on students and teachers. Unpublished Doctoral Thesis. Flinders University: Adelaide, South Australia.

- [15] Yin R.K. (2003), Case study research – Design and Methods, 3rd edition, thousand Oaks, California-Sage Publication.
- [16] Yildirim, S., Temur, N., Kocaman, A. and Goktas, Y. What makes a good LMS: An analytical approach to assessment of LMSs. In *Proceedings of the Fifth International Conference on Information Technology Based Higher Education and Training*. (Istanbul, Turkey, May 31–June 2, 2004), 125–130.
- Gachenge, B. (2008). E-learning taking root in Educational Institutions. Date accessed 12/06/2014 from <http://allafrica.com/stories/200807011188.html>
- [17] Smulders D. (2001). Web Course Usability. <http://www.learningcircuits.org/2001/aug2001/elearn.html>.
- [18] Newhouse (2004), The Impact Of ICT On Learning And Teaching, A Literature Review, Western Australia Department Of Education.

Agent Based Personalized Semantic Web Information Retrieval System

Dr.M.Thangaraj, Associate Prof.

Department of Computer Science
Madurai Kamaraj University, Madurai, Tamilnadu

Mrs.Mchamundeeswari, Associate prof.

Department of Computer science
V.V.V College for Women (Affiliated to Madurai
Kamaraj university)
Virudhunagar, Tamil Nadu

Abstract—Every user has an individual background and a precise goal in search of information. The goal of personalized search is to search results to a particular user based on the user's interests and preferences. Effective personalization of information access involves two important challenges: accurately identifying the user context and organizing the information to match with the particular context. In this paper, the system uses ontology as a knowledge base for the information retrieval process. It is one layer above any one of search engines retrieve by analyzing just the keywords. Here, the query is analyzed both syntactically and semantically. The developed system retrieves the web results more relevant to the users query. The level of accuracy will be enhanced since the query is analyzed semantically. The results are re-ranked and optimized for providing the relevant links. Based on the user's information access behavior, an ontological profile is created, which is also used for personalization. If the system is deployed for web information gathering, search performance can be improved and accurate results can be retrieved.

Keywords—Agent; Personalization; Semantic web; information retrieval; ranking algorithm

I. INTRODUCTION

The main purpose of this section is to justify the need for an integrating approach that combines both intelligent agents and personalized semantic web service technologies. The study concentrates on personalized semantic web services and then intelligent agents and multiagent systems which are enumerated and the most pressing problems of agent technology pointed out.

A. Personalization using Semantic web:

Semantic technologies promise a next generation of semantic search engines. General search engines don't take into consideration the semantic relationships between query terms and other concepts that might be significant to the user. Thus, semantic web vision and its core ontology's are used to overcome this defect. The order in which these results are ranked is also substantial. Moreover, user preferences and interests must be taken into consideration so as to provide the user a set of personalized results.

B. Query Expansion using ontology:

Ontology is to create a shareable and agreeable semantic resource over a wide range of agents. The important goal of building ontology is it may serve as an index into a repository

of information to facilitate information search and retrieval and also used to identify the user context accurately, so that the search results can be personalized by reorganizing the results returned from a search engine for a given query. In this research, context is extracted from Domain Ontology in terms of concepts and used to extract the semantic patterns in queries which can represent actual users' requirement.

Through personalization, one can improve the navigation on a web site by, for example, highlighting content and links of interest, hiding those that are irrelevant, and even providing new links in the site to the users likely web destinations. While personalization can help to identify relevant new information, new information can create problems in re-finding when presented in a way that does not account for previous information and interactions. This study presents a model of what people remember about search results, and shows that it is possible to merge new information invisible into previously viewed search result lists where information has been forgotten. Personalizing repeat search results in this way enables people to effectively find both new and old information effectively using the same search result list.

C. Agent based personalization:

The main characteristic of agent-based technology is that the structure of the software is represented by a group of agents who collaborate in achieving the goal of the task in hand. The combination of information retrieval and Multi-agent technology has the following features: Adaptability, initiative and collaborative. Among different types of agents, the personal assistant agents are particularly interesting to this research. This type of agents operates at the user interface level and actively assists users by offering information and advice to the users (Wasson *et al.*, 2001). These agents usually apply a kind of intelligent learning algorithm so that they can intercept the users input, examine it and take actions that are more specific to those particular users' needs at that moment. These agents are also called learning or adaptive agents. Agent can initiatively retrieve the corresponding information based on users' demand, and even can monitor the changes of information sources and agents also share the information with other Agents.

This paper introduces a personalized information retrieval system based on multi-agent, which can accomplish information retrieval according to user interest knowledge via multi-agent collaboration for providing personal service to the

user. In the process of personal information retrieval, the precision and quality depend on the veracious degree that the system master user interest. Therefore, the paper solves problems how to construct user interest model based on vector space, and how to update user interest model in time when user's interest changes:

II. LITERATURE REVIEW

Web personalization is understood in various dimensions. One way of doing this is categorization of users based on demographic information provided by the users at the time of selecting the style for personalization. An example of this is Google Personal search through *igoogle*. This approach requires that the user must exactly know what information is needed prior to searching. The research is also going on to modify the structure of the web documents and make it semantic so that the documents are then retrieved on the basis of the meaning of the query and not the terms present in the query [2]. This approach seems very promising but is a long term project, the acceptability and usability of which depends on the user community. Another way to personalize the search is to classify users on the basis of pre-calculated classes. The classes may be pre-calculated through users browsing history. A classification of the on-line users to one of the predefined classes is typically based on similarity calculation between each predefined pattern and the current session. The current session is assigned to the most similar cluster [6, 8]. Further this approach is modified to accommodate fuzzy classification so as to prevent some users to become outliers [7].

Some authors have constructed user profiles on the basis of modified collaborative filtering with detailed analysis of user's browsing history in one day [5]. User profiles are also constructed on the basis of ontology [1]. Some efforts have also been done to refine the search process by re-defining the queries and then submit it to the search engine. Refined queries are then clustered to form user's profiles [6]. In this approach also only visiting a page makes it interesting enough to update user profile. Another method to personalize is to discover association among various links accessed by the user through its sessions [3].

Another interesting effort has been done in actual personalization of users' interest in which they have considered that every user's behavior is different on same search results obtained through same search query [3]. They have used two properties of a document for modeling users i.e. attractiveness and perseverance. They have assumed that these properties depend on the popularity of the document among the similar user community and distance of that document from last selection. Normal user behavior suggests that after a certain no of unattractive documents the user stops navigating the search results. Efforts have also been done to construct user profiles using relevancy between the terms of the queries presented in current session and in earlier sessions [4].

Due to the intelligent agent technologies shortcomings, the inherent need for autonomous software entities in SWS environments, and the promising benefits of having both intelligent agents and (semantic) web services working cooperatively, numerous research projects have been carried out that try to put these two technologies together into

integrated frameworks. The author Hendler (2001) proposes a method for describing the way the invocation of services should be done by agents by means of an ontology language. The Semantic Web FRED project (SWF) combines agent technology, ontologies, and SWS in order to develop a system for automated cooperation. The GODO (Goal-Oriented Discovery for SWS) system (Go'mez et al., 2006), which is based on a software agent that is located between different SWS execution environments (e.g. WSMX, METEOR-S, OWL-S Virtual Machine, etc.) and final users.

The authors Buhler and Vidal [10] highlight the passive behavior of web services and propose to wrap them in proactive agents. The problem of this approach is that semantically described web services are not considered at all. Another related solution is the one provided by the "Agents and Web Services Interoperability Working Group (AWSI WG)"³, which is part of the IEEE FIPA Standards Committee which can handle the fundamental differences between agent technology and web services, that is, the use of different communication protocols (ACL vs. SOAP), service description languages (DF-Agent-Description vs. WSDL) and service registration mechanisms (DF vs. UDDI). With this approach, the so called Agent Web Gateway middleware (Shafiq et al., 2006) facilitates the required integration without changing existing specifications and implementations of both technologies. This category focuses on the overlapping features of the technologies under question. However, we believe that most of the functionality provided by Intelligent Agents and Web Services is complementary, so that each of these technologies must be situated at a different abstraction level.

The model proposed here is a frame work for building a user model in addition to explicit & implicit feedback from user and find the relevancy between the terms presented for query and the document using past sessions by user and the contents of the documents. Then the documents with the higher relevance ratio are presented to the user. The current user session data is used to update the user's profile for future reference. Two types of parameters are considered for constructing user model: static parameters and dynamic parameters. Static parameters are relevancy of documents with the specific category measured by the popularity of the document. Static parameter used in the model is: Term-document relevancy which is maintained in a 2-D matrix T. The relevancy calculation done here is based on the occurrence of terms in the document. We have tried to improve upon the modality of updating the matrix. The matrix is updated every with every user session with the browsing patterns of a user and for first 'n' sessions it keeps on constructing new columns with respect to the terms that relates to a document. Our system proposes a different ranking based on URL. The ranking is *query-dependent*. The *proposed* algorithm assigns a score that measures the quality and relevance of a selected set of pages depending on their URL to a given user query. The basic idea is to build a query-specific two dimensional vector table, called a related vector table, and perform URL analysis. The present paper proposes a slightly different ranking based on URL. In our research we use hybrid approach to find ranked webpage.

III. PROPOSED ARCHITECTURE:

This paper we proposes, an architecture for an Agent Based Personalized Semantic Web Information Retrieval (APSIR), which can help users to get the relevant web pages based on their selection from the domain list , so that users can obtain a set of related web pages from the system .

APSIR is a crawler-based search engine that makes use of crawler to collect resources from both semantic as well as traditional web resources

This section explains the basic architecture of our system. In section III the working mechanism of the proposed system is describe. Section IV shows the performance evaluation results. Finally, section V sums up all the above said points.

The system APSIR (in Figure 1) consists of different components like User Agent, Semantic *Extraction Agent*, and Semantic *Searching Agent*, *Filtering Agent*, *Personalized Ranking Agent* and Knowledge Base. All agents are monitored entirely to fulfill proprietary system functions, including information retrieval and Knowledge Base update.

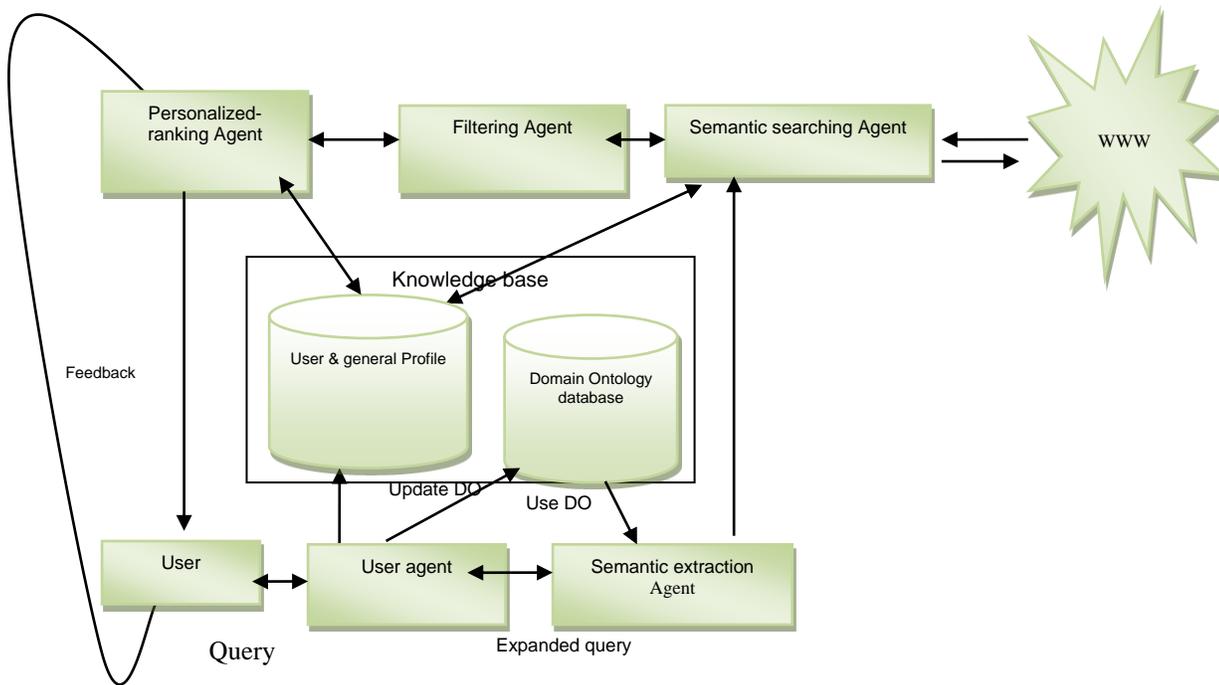


Fig. 1. Structure of Agent –based personalized Semantic web information retrieval system

A. *User Agent*: User Agent is the mutual interface between user and system, and provides a friendly platform to users. It can construct user interest model according to User's browsing history record and registration data. (System comes with an easy to use Google like search interface. After submitting his query, results are displayed.) User Agent incepts user's retrieval request (needed search) which is transformed to prescriptive format, and transmits the formatted user request to Semantic Extraction Agent to expand the query based on the respective domain and related terms based on ontology. User Agent also takes over all result from Personalized Ranking Agent, and presents personally these results to user. In addition, User Agent presides over creating a profile user for new user. User's browsing or evaluating behavior can also be stored as profile and it is learned by User Agent, so user interest model may be updated and improved in time. User Agent includes Environment view, Memory Base, Knowledge Base, Learning mechanism and Inference Engine.

- The view of the environment module in the User Agent is the user's input and output interface.
- Memory Base records the original information entered by the user.
- Knowledge Base defines the user's personal knowledge, classified information and the user model.
- Learning mechanism is used to summarize the behavior of users and formats the information.

B. *Semantic Extraction Agent*: Semantic Extraction Agent aims to find the semantic features in the users' queries. It will make use of agent technologies and ontology technologies to analyze the association relation in the users' queries and document to extract semantic features. This module contains the following components:

- Query preprocessing: Meaningless words like neuter pronouns, articles, and symbols in the content will be removed from query.
- Semantic Analyzing: This component identifies semantics elements like Subject, Property, and Object in the Query content and analyzes their semantic relations.
- Semantic matching : In the personalized information retrieval system, Semantic matching agent takes charge of receiving formatted user request from User Agent, and the user request is expanded (based on ontology)according to user interest. Afterwards, the perfected user request is transmitted to Semantic Searching Agent. It analyses the returned data from Searching Agent, filtrating useless information, and the processed results are send to user.

Alg. For QE using domain ontology

Input: Original query terms set (Q_{or}) where $Q_{or} = \{ t_1, t_2, \dots, t_n \}$

Output: Query terms set (Q_{set}) where $Q_{or} \cup Q_{ex} = Q_{set}$ is the expanded query terms

```
queryexpand( $Q_{or}$ )
{
 $Q_{set} = \{ \text{empty} \}$ 
// expand query based on ontology
Get  $Q_{or}$  and add it to  $Q_{set}$ . // split each word in the query
and stored as  $Q_{set}$ 
for all term  $t_i$  in  $Q_{set}$ 
{
If (  $t_i$  in  $D_{ontology}$  )
{
If(  $t_i$  in  $P_{ontology}$  ) and (  $t_i$  in  $R_{ontology}$  ) // find its
possible and related terms
 $Q_{ex} = P_{ontology} + R_{ontology}$ 
}
}
ElseIf (  $t_i$  in  $p_{ontology}$  )
{
If (  $t_i$  in  $D_{ontology}$  ) and (  $t_i$  in  $R_{ontology}$  )
 $Q_{ex} = D_{ontology} + R_{ontology}$ 
}
}
ElseIf (  $t_i$  in  $R_{ontology}$  )
{
If (  $t_i$  in  $D_{ontology}$  ) and (  $t_i$  in  $P_{ontology}$  )
 $Q_{ex} = D_{ontology} + P_{ontology}$ 
}
} // for loop
} // end
```

C. *Semantic Searching Agent*:

This component is responsible for searching and retrieving relevant results. Semantic Searching Agent mainly includes Search Strategy, Search Optimization and Crawler.

- Search Strategy includes depth-first search strategy
- Search Optimization includes the way accessing to the page that should be subjected to management of websites and the frequency of visiting, collecting important web pages which have high page weight and have changed, ensuring pages that will not be repeated crawled.
- Crawler is a program that crawl pages based on the hyperlinks between webs to collect information

Semantic search(Q_{or})

```
{
```

```
Queryexpand(Qor) // expand user query based on
ontolgy
If the term is found
{
Retrieve the relevant web page from in knowledge
base/web
Store the query and web pages to db
}
}
Else // new ontology term
{ Update ontolgy with new terms
}
}d
```

D. Filtering Agent:

In this process, matching algorithms are presented to enable fast matching and searching for content. Components included in this module are: Knowledge Base and Semantic Matching.

Knowledge base: It includes the user's personalized information transmitted by the User Agent. When matching, Semantic Matching will make use of users personalize information (users' interest behavior and search history) to match and search more accurate and useful information for users.

Semantic Matching: According to the Users behavior, this component will match semantics in users' queries and semantics in the documents, and in accordance with the relevant, the results will be submitted to the User Agent.

E. Personalized Ranking Agent:

In the personalized information retrieval system, Personalized Ranking Agent is the decision-making center of personalized information retrieval system based on multi-agent. Using Re-ranking alg. find the new score based on users interest.

Algorithm of calculation of relevance score(re-ranking) for the Web pages.

Input : web page (P) Query term {t₁,t₁,t₃...t_n} from the expanded query

Output relevance score (ranking) for all the web page (WP)

Urlset={url₁,url₂...url_n} for the given query

Re-rankign(urlset ,Q_{ex})

{

For all url in the urlset

{

read webpage(p);

Query set {} ←--- Separate each term in Q_{ex}

N=total number of term

I=1

If t_i in WP

{

TF =TF+∑D(T_i)

}

For all terms t_i in Query set

{

Wcount(t_i)=wcount(t_i)+∑_{i=1..n}∑_iWP_{i=1}(D(t_i))

}

termcount=wcount/N

Count=termcount+TF

}

F_s=Get feedback score{0,1,-1}

t_s=viewtime(WP)

Save count,f,c, t_s in DB table

}

F. Knowledge Base:

Knowledge Base is used for storing every user interest model, user- record, and rules or parameters that serve for ensuring system well-balanced circulation.

G. User interest profile:

Two general methods are used to discover user interest (i) apparent feedback and(ii) connotative feedback.

In apparent feedback, user can input the data of personal interest or evaluation to current work.

1) *Apparent feedback*: When information retrieval, user gives a weigh value W_v, which represents user's satisfactory degree to the provided document D, formalized expression is described as follows.

$$\text{Satis_De}(D) = f(W_v) \quad 0 \leq \text{Satis_De}(D) \leq 1 \quad (1)$$

In system implement, user can select whether evaluation page appear or evaluation page may appear constrainedly. The satisfactory degree setting may be an option bar , so user can adjust to set W_v.

2) *Connotative feedback*: The system may obtain user interest information via tracking user behavior and operation.

The under-mentioned factors may be used to discover impliedly user's interest. a) History record - User is interested in the pages, which are browsed before time, the more accessing times the higher interest degree. b) User behavior. Some operations (e.g. saving, printing or copying) indicate user interest when user browsing page. In addition, browse time are also related with user interest. So the mine above-mentioned data is to discover user's interest.

H. Construction of user interest Profile:

First of all, the following process may be used to classify browsing history record documents.

Step1: if QUERY match exactly (BH(browsing history)) which is standardized vector set of browsing history record.

Step2: For $i=1; j=2, 3, \dots, |BH(D)|$, calculate the relativity of document Di with document Dj in $BH(D)$ set.

$$Sim(D_i, D_j) = W_i * W_j / |W_i| * |W_j| \quad (2)$$

All documents D_j with $Sim(D_i, D_j) \geq \text{thersold value}$ (thersold value is no. of web page to be displayed). So the classified document vector set $S1(D), S2(D), \dots, Sn(D)$ are gained. All specific terms in document vector set $Si(D)$ is sorted according to the weight descending. In this way, we can get a user interest vector $UserIni = ((ti1, wi1), (ti2, wi2), \dots, (tik, wik), \dots, (ti, DT_Limit, wi, DT_Limit))$ wik is standardized weight of specific term tik. Then, the user interest model is constructed as:

$$\text{User Interest Model (UIM)} = UserIn_1 \cup UserIn_2 \cup UserIn_3 \dots \cup UserIn_n \quad (3)$$

There, n is the classified set number of history record documents.

I. User behavior factor:

When collecting user interest information, should be paid attention to other factors. For example, user attitude to browsing page is very important factor of user interest information. Some pages are saved, some pages are copied, or some pages are printed. By all appearances, user is interested much more in those copied, saved, or printed pages related to merely browsed pages. So the users domain interest degree is introduced whereas before-mentioned reason. $User_Interest(UserIni)$ denotes the degree of interest of interest domain $UserIni$ that document belongs to. Therefore, Knowledge Base also stores user behavior data besides user interest model.

1) $FreqInDi$, which is the citing frequency of user interest domain $UserIni$ that user browsed document belongs to, and

2) $SaveInDi$, which is the saving frequency of user interest domain $UserIni$ that user saved document belongs to, and

3) $SpeedInDi$, which is the viewing timing of $UserIni$ documents, and

So we can construct a numerical function of domain interest degree, which may reflect interest information of user behavior.

$Fi(FreqInDi, SaveInDi, SpeedInDi)$ All interest domains may be resorted at any moment according to the function Fi of domain interest degree. In this way, changes of user interest along with time can be reflected in user interest model

J. Result storing and viewing :

Step1: Convert retrieval request query string Q to vector.

$$V(Q) = ((qt1, qw1), (qt2, qw2), \dots, (qtr, qwr))$$

Step2: Construct document vector for any returned document FDi . (from browsing history)

$$V(FDi) = ((fwi1, fwi1), (fwi2, fwi2), \dots, (fwi, NT_Limit, fwi, NT_Limit))$$

Step3: Calculate the comparability between document vector $V(FDi)$ and query vector $V(Q)$.

$$Wsim(FD_i, Q) = nfW_i * nqW_j / |nfW_i| * |nqW_j| \quad (4)$$

There, n is total number of specific terms in query Q or in document.

Step4: Calculate the comparability between document vector $V(FDi)$ and user interest vector $V(UserInj)$.

$$Psim(FD_i, Userin_j) = nfW_i * nW_j / |nfW_i| * |nW_j| \quad (5)$$

n is total number of specific terms in document FDi or user interest model

Then, the comparability of document FDi , query Q and user interest model. $User_Model$ is represented as follow.

$$Sim(FDi, Q, User_Model) = Wsim(FDi, Q) + Psim(FDi, Userin) \quad (6)$$

If all returned documents are processed then go to Step5, otherwise go to Step2.

Step5: Output sorted searching results according to $Sim(FD, Q, User_Model)$ value descending for the returned documents.

K. User interest model update

When user browses output documents, the system memorizes user's behavior (browsing, saving etc.) to Knowledge Base in real time. The system may give an evaluation page for asking user to do satisfactory degree. Evaluating all documents, which satisfactory degree $Satis_De(D)$ is more than a default minimal value, (threshold value) is extracted for constructing new user interest domain vector.

New user interest domain vector is used to replace old user interest domain vector, which is cited seldom. The storage capacity of user interest model is commonly limited to finite space capability, for example: $N_Class_MAXTIME$. When the number of user interest domain vector exceeds the capability limit (\geq particular time interval); some user interest domain vectors, which are cited seldom (scaling by domain interest degree function $Fi(FreqInDoi, CopyInDoi, PrintInDoi, SaveInDoi, SpeedInDoi)$), may be deleted from user interest model and moved to dump table. So the number of user interest domain vectors is limited to definite scope, and the system can track user interest in time.

IV. IMPLEMENTATION AND EXPERIMENTATION

In this section experiments carried out to evaluate the performance of proposed system will be discussed from a quantitative point of view by running some experiments to evaluate the precision of the results. The basic idea of the experiment is to compare the search result from keyword based search engine with proposed one on the same category and the same keywords. The criteria of our experiments include suitability (the ratio of the amount of useful information to the total amount of information) age (the period of the document

post) and semantic matching (the accuracy of matching). After several time of similar information search, our system will get better results than the current search engine expectedly by updating user profile based on the user feedback autonomously. A test set collection is which consists of set of documents, queries and a list of relevance documents are used to evaluate the proposed system. These are used to compare the results of proposed system by performing relevance based evaluation method.

The proposed system is implemented in C#.Net as Web-based system using Visual Studio 2008, .NET Framework 3.5, and SQL Server 2005. The number of stored documents is more than 3 lakhs documents. These Web documents are about computer science domain. The improvement is measured by performing different experiments using the relevance based evaluation method. It uses the metrics: precision, recall, f-measure, average precession (AP) and mean average precision (MAP), to measure the performance of proposed system.

A set of queries has been manually for comparative performance measurement. The set of sample queries is given in Table 1. It Show the different levels of performance for different queries, the proposed semantic information retrieval method that improves the document ranking.

TABLE I. AP AND F-MEASURE USING PERSONALIZED AND UNPERSONALIZED RANKING FOR SINGLE USER AND MULTIPLE USER

TABLE 1A: SINGLE USER

Keyword	AP		F-measure	
	existing Using keyword query	Current Using semantic query	existing Using keyword query	Current Using semantic query
Java	.76	0.56	.61	0.41
Constructor	1.00	0.81	.69	0.52
Polymorphism	1.17	1.00	.89	0.67
memory compaction	1.1	0.79	.58	0.53
Encapsulation	1.3	0.79	.69	0.53
disk space management	1.2	0.79	.76	0.53
abstract classes	.98	0.84	.89	0.56

TABLE 1B: MULTIPLE USER

User	AP		F-measure	
	existing Using keyword query	Current Using semantic query	existing Using keyword query	Current Using semantic query
User 1	.65	0.55	.55	0.41
User 2	.688	0.55	.99	1.08
User 3	.459	0.375	1.258	1.125
User 4	.65	0.375	1.356	1.125

Fig 2 and Fig. 3 shows comparative study of the results of the both systems that retrieves the documents based on similarity between the query and the collected documents. This experiment shows the average precession that is based on retrieving results for different query of single user and single query of multiple users. Graph shows that the system gives high precision during retrieving documents.

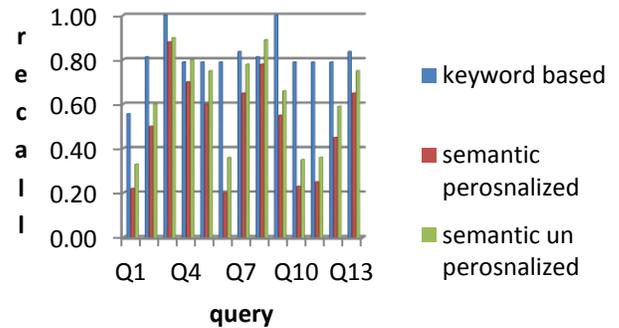


Fig. 2. AP measure precision of personalized vs. unpersonalized (for single user)

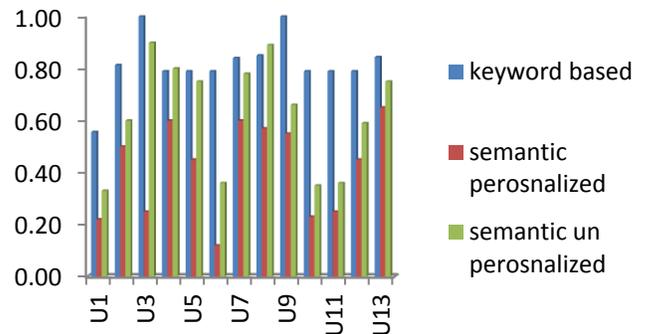


Fig. 3. AP measure precision of personalized vs. un-personalized (for multiple users with multiple keywords)

The retrieval efficiency is a major challenge when the size of the database increases. This shows the importance of semantic similarity during determining the documents that are relevant to the user query. The second sets of experiments, which are user centered, are focused on the overall performance of the search engine and the evaluation of real interactions with users. Fig.4 discusses the performance efficiency of both systems when the system uses to retrieve the result. This graph shows that agent based personalized search is better than other method because it highlight user profile and study user behavior to determine ranking for each time. It can also be observed that the contextualization technique consistently results in better performance with respect to simple personalization, as can be seen in the average precision and recall depicted by Fig.5, which shows the average PR results over the different user cases

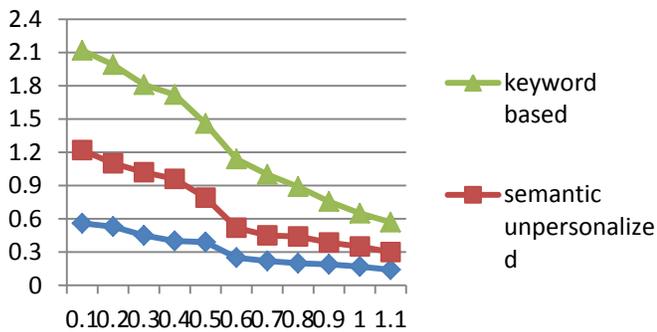


Fig. 4. Performance evaluation

The next experiment aims at determining the importance of personalization by using generated dynamic user model during using the system. The user model is used to re-rank the retrieved documents to match the user interest.

Personalization time:

Time to retrieve any information depends on the type of search engine, size of data set, relevancy between query and doc. user history & re-ranking algorithm used. The personalization performance can be expressed:

Personalization performance = $\sum_{i=1}^n Grank + UserRank$ and

For each page find $UseRank = \sum_{i=1}^n UR + VT + FC$ Where UR – user rating VT page view time and FC-frequency count and n represents threshold value

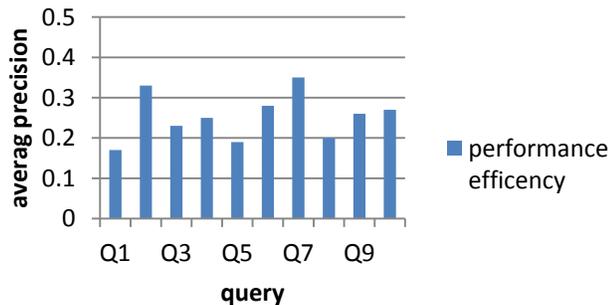


Fig. 5. Performance evaluation

This Fig.5 focuses the usage efficiency of the systems when the system uses to retrieve the result.

It is observed that 80% users, have found improved precision with the proposed approach in comparison to the standard search engine (Google) results, while 20% users have achieved equal precision with both the approaches. It has been observed that users who posed Queries in unpopular context than well liked context got better performance In addition, when the system can extract the exact context of user’s need,

the Precision and recall is found better than other search engine results.

The experimental result indicates that the efficiency of information retrieval, by the use of the above-mentioned personalized information retrieval system based on multi-agent, precedes evidently current information retrieval tools in common use to sum up the precision is improved 15% - 35%.

The system realizes individuation and intelligence of information retrieval for providing personal service to user via multi-agent collaboration according to user interest characteristics and different information needs. The construction algorithm and update algorithm of user interest model, which are based on user browsing history record and user browsing behavior, can discover user interest in time, control safely the scale of user interest model, and increase effectively document filtration efficiency.

V. CONCLUSION

In this study, a new information retrieval system based on Semantic Web and Multi-Agent has been presented to effectively the offset existing defects and constraints of the traditional keyword-based search, and help users to obtain required information.

The proposed system experimentation shows that, it can improve the accuracy and effectiveness for retrieving the web documents. It aims at providing the relevant web-document in certain domain that is matched to user’s request. It can be used in other domain by editing the domain ontology using export option of APSIR and building the domain concepts weight table. A user model is proposed to improve the ranking of the relevant documents retrieved to user based on its interest.

REFERENCE

- [1] Ahu Sieg, Bamshad Mobasher, Robin Burke Ontological User’s Profiles for Personalized Web Search In Proceedings of Conference of AAAI 2007
- [2] Bettina Berendt, Andreas Hotho, and Gerd Stumme, Towards Semantic Web Mining ISWC 2002, LNCS 2342, pp. 264-278
- [3] Filip Randlinski, Thorsten Joachims Query Chains : Learning from Implicit Feedback. In the Proceedings of KDD’05 August 21-24 2005
- [4] Jingqiu XU, Zhengyu ZHU, Xiang REN, Yunyan TIAN, Ying LUO Personalized Web search using User Profile, Journal of Computational Information System 2007
- [5] Kazuneri Sugiyama, Kenji Hatano, Masatoshi Yoshikawa Adaptive Web Search based on user Profile Constructed without Any Effort from Users ACM Press 2004
- [6] Kazienko P, Kiewra M (2003) Link Recommendation Method Based on Web Content and Usage Mining. In: the Personalized Web Search: User Modeling Using Implicit Feed Back From Click Through Data
- [7] Maciej Kiewra Iterative Discovering of User;s Preferences Using Web Mining, International journal of Computer Science and Applications Vol. II No. II pp. 57-66
- [8] Mobasher B, Dai H, Luo T, Sun Y, Zhu J (2000) Integrating Web Usage and Content Mining for More Effective Personalization. LNCS 1875 Springer Verlag, 156-176
- [9] Bernes-Lee T., Hemdler Agents and the Semantic Web IEEE Intelligent Systems Journal (march/April 2001)
- [10] Paul Buhler and José M. Vidal (2005) Towards Adaptive Workflow Enactment Using Multiagent Systems ,Information Technology and Management Journal , volume 6 pages:61-87

Social Learners' Profiles in a Distance Learning System Powered by a Social Network

HROR Naoual

Laboratory Systems and Telecommunications Engineering
Of Decision, University IbnTofail,
Kenitra, MOROCCO

OUMAIRA Ilham/MESSOUSSI Rochdi

Laboratory Systems and Telecommunications Engineering
Of Decision, University IbnTofail,
Kenitra, MOROCCO

Abstract—This work is integrated into the general problem of the research systems of distance learning; and more particularly in the monitoring and the positioning of social profile learners in a distance learning system powered by a social network. In this article, we propose a model multi-entry to determine the learners' profile and positioning his social type. This template allows you to exploit the different traces generated by a learner in the platform and produce indicators on his own profile.

Our approach leverages the tools of the fuzzy set theory, the modal and temporal logic. The motivation of this research is to create a tool which helps the tutor to better observe and follow the actions of the learners at the level of the learning platform, and to anticipate potential discouragements or abandonment of the learner.

Keywords—Distance learning system; Social profile; Traces; fuzzy; social network

I. INTRODUCTION

Our research focuses on the IT environments for human learning, and particularly on the assistance of learners in their learning journey online. This mode of learning has profoundly benefited from the important technological innovation of the 21st century. Despite this great innovation, the dropout rate of learners remains high. According to Bernatchez and Gendreau [1], this rate can vary from 37 to 50% depending on the context of learning.

The purpose of our research work firstly consists in defining the needs referred by internet in social networks, and determine the various inputs offered by these networks, and secondly in integrating the functionality provided by these systems in the available platforms of learning online.

We will then be offering the learner a system of comprehensive learning which responds to both the cognitive and social needs, and then we are aiming to exploit the traces produced on these new workspaces, and generating indicators to help the tutor to better follow the social behavior of learners within these complex systems.

In the first paragraph, we are presenting the social needs of internet users, and then we are charting a state of the art on the platforms of distance learning that have used social networks. In the second paragraph, we present the most important features of social networks that we are trying to exploit in our target system.

The third paragraph describes the architecture of our Model System of the social profiles of learners SMPSA. The last paragraph is dedicated to the modeling of traces collected by the method of fuzzy logic, to produce the learners' profiles.

II. SOCIAL NETWORKS (SN)

A. The Features of Social Networks

All the social networks attract surfers by their features, below you will find a list of the typical features of a social networking site like Facebook, Tiwtter, Google, FriendFeed... Each platform provides these features. Next we propose a brief description of some features.

- Sharing Documents

Most of the social networks offer the feather of sharing documents; a member of the SN can upload a text file, PDF, or even an image ... The other members will be able to download it as well, send back a new version or display it directly.

- Sharing Comments

Comment modules allow your users to interact with the content and other members of your social network. Our flexible infrastructure enables administrators to attach comments to virtually any kind of content: wall notes, blog posts, images, etc[5].

- Sharing Tags

Similar to comments, tags can be attached to different types of content, allowing users to build an independant form of navigation and/or categorization. [5]. The surfer can use the tags "I like" "like" "seen" etc, if he likes a shared document or a comment, he can mark it with a tag; it is a simple way to express his point of view.

Thanks to the previously mentioned features, Social networks like Facebook, Twitter and YouTube have rapidly become a part of many people's everyday lives, especially for those younger generations, like students, who have grown up with so much technology at their fingertips.

There are lots of possible reasons for student's social media usage – to stay in touch with friends, share a funny video, keep up with news, and build professional contacts. Yet, why do students use social media? This was one of the questions we attempted to answer in the next paragraph.

B. The Sociol Needs of Surfers

HyderKabani, Shaamah, author of The Zen of Social Media Marketing [6], affirms the existence of two reasons that prompt users to integrate a social network:

- The expression of the identity: internet users need to put their identity in value, by trying to be better than other.
- The need to keep in touch with their friends and create new bonds of friendship.

Julie Schlack, Michel Jennings and Manila Austin [7] have also cited six essential social needs which are pushing the internet users to join a social network:

- Express their identities using the profiles.
- Help the others and ask them for help.
- Find people sharing the same center of interest.
- Create networks and a relationship with the each other.
- Develop a true sense of belonging to a community.
- Be reassured on their values, by having the feeling being useful and having influence on the world that surrounds them.

Similar results have been deducted by a study made by DacharyCarey[8] :

- Help connected people.
- Search for the popularity.
- Build a community.
- Marketing
- ...

The social needs of a learner vary from an online system to another (learning platform, games application etc.) and from a context to another (obtain an academic degree, training of language etc.), also these needs vary between the systems of learning and the sites of social networks. For this purpose, we have proposed to classify them according to the diagram below (Figure 1): the degrees of the color of the arrows depict the importance of the need

III. DISTANCE LEARNING PLATFORMS SUPPLIED BY A SOCIAL NETWORK

The integration of social networks in the platforms of distance learning is a recent approach; several works [9] [10] [11] have been established on this new method of learning. Among these works, we cite the experience of the University of Leicester, one of these teachers has opened the social network "FriendFeed" with the students to provide them with information relating to the course - links, folders...

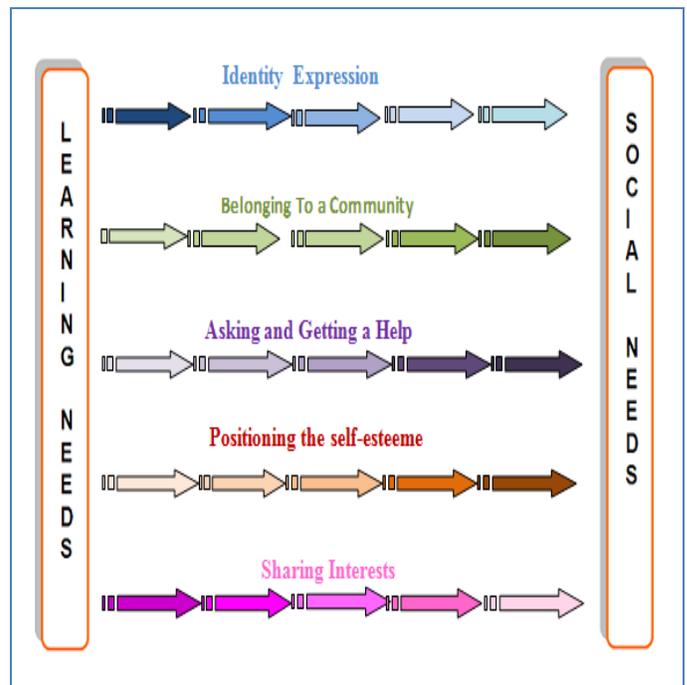


Fig.1. The social Needs

Each participant has a personal page, in which he can indicate the status of their work and the difficulties encountered [10].

Another experiment launched by a teacher from the University of Vienna, who asks his students to make feedbacks on his course through the site of micro-blogging in order to assess the success of his lessons [12].

These experiences have highlighted the general results, like the importance of the attendance rate of network, the regularity of connected pupils, the usefulness of the comments to improve the job ... [13].

However, this information is still insufficient for the tutor to have a visibility on the social participation of the learner at the level of the platform that will help him to better intervene during the training.

In our university, we work with the platform "Moodle" , a free, online Learning Management system[14] ,Technically, to satisfy our need to feed our distance learning platform "Moodle" with a social network, we have made studies about open sources social networks; In the next paragraph, we present the summary of our research.

IV. THE SOCIAL NETWORKS OPEN SOURCE

Currently, several platforms of social networks are used. Below we show a few examples [15]:

TABLE.I. THE PLATFORMS OF SOCIAL NETWORKS OPEN SOURCES

Social Network	Langue	Technology	Compatibility Moodle	Documentation
Elgg [16]	English	Apache,MySQL, PHP	Yes	Low
Oxwall [17]	English/ German	Apache,MySQL, PHP	Yes	Low
LiveStreet [18]	English/ Russian	Apache,MySQL, PHP	Yes	Low
Mahara [19]	French /English /Arab	Apache,MySQL, PHP	Yes	Strong

- **Mahara** is fully featured electronic portfolio, weblog; resume builder, and social networking system for connecting users and creating online communities.
- **Elgg** is an open source social networking platform developed for LAMP (Linux, Apache, MySQL, PHP) which encompasses weblogging, file storage, RSS aggregation, personal profiles, FOAF functionality and more.
- **Oxwall** is a free and open source community software distributed under the Common Public Attribution License. It is written in PHP and is used as a platform for social networking and community sites.
- **LiveStreet** is open team collaboration software that uses social networking to unify team workspaces, written in PHP, Javascript and MySQL.

To make our choice, we define some important technical variables (Table 1) for the platforms previously presented.

The decisive criterion of our choice was the documentation available on the integration of the social network chosen by the Moodle platform. For this, our choice has focused on the Mahara platform; the integration documentation is very available [19] ,and its features meet perfectly our needs.

V. PROPOSAL

Many works in social psychology have highlighted the impact of forms of interactions or social memberships on the cognitive mechanisms [20], as well as the importance of social factors in cognitive development, including the acquisition of cultural knowledge [21].

In this perspective, we propose to integrate the social network "Mahara" to our learning platform "Moodle", which renamed "SocioMoodle ", and we exploit the traces produced on this workspace to generate the new social indicators by our system SYMPA (System of modeling profile learner),and in order to help tutor to make a best decision in the best time.

For the establishment of our proposal, we have adopted the following approach:

- Make a thorough study of the features of social networks the most used and identify internet users within these environments
- Make a study on the social networks open source available, and make the choice of the social network that best suits our needs.
- Integrate the social network to the platform used by the university (Moodle).
- Experiment with the complete system Moodle+ social network(SocioMoodle) in a real context of training
- Analyze the collected data and calculate the social indicators using our system SYMPA (system of modeling profile learner).

VI. EXPERIMENTATIONS

A. Target Audience

The duration of our first experimentation was spread over thirty weeks with 140 learners. The first batch of students who have made the experiment are the students of MQL (Master Quality Software) graduation 2011/2012, they have worked on varied subjects of application development. The second batch concerned students from ENSA (National School of Applied Science) graduation 2011/2012 who have all worked on some subject of application development.

The second experiment was conducted with the learners of MQL (Master Quality Software) of Kenitra, graduation 2012/2013, and the students of the ENSA (National School of Applied Science) graduation 2012/2013, during a period of 14 weeks. The tutor of MQL asked students to develop different application by group, and the tutor of ENSA gave the same subject, of application development, to all learners. To succeed our experiment, we asked learners to:

- To be connected daily to the platform.
- To created their public page on Mahara platform.
- To integrated their documents (Files,script,PDF...) on the platform.
- To create a group for a work.
- To create a page of group on Mahara

VII. GENERAL PRESENTATION OF SYMPA

A. Overview of SYMPA

Our System of Modeling Profile Learner SYMPA, exploits the traces generated by the learners to produce social indicators, which will help the tutor to identify a social profile for each learner.

The social profile of the learner depends on his social production; it means by the number of Tags, comments, or sharing produced by the learner.

Our SYMPA (Figure2) is composed of three components: a component which is engaged in the collection of the traces generated by the learners, the second which is the treatment of

traces and the calculations, and the last component for the generation of profiles.

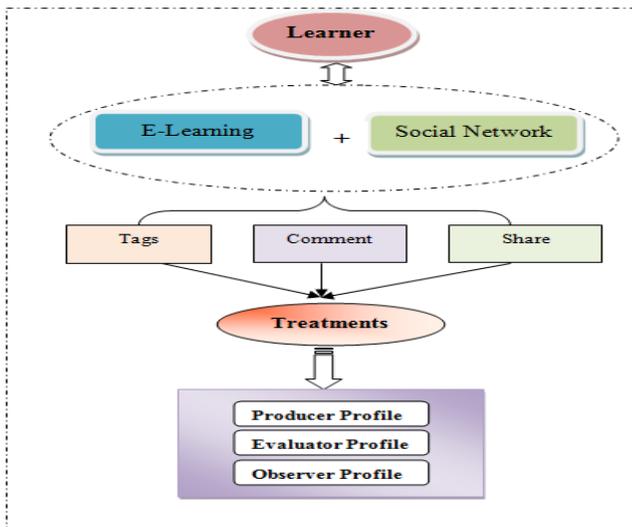


Fig.2. Modeling of SYMPA

In our case, the variable of the time is a parameter taken by the component "treatment" in order to assess the profile of the learner during the period of the training.

B. The Socials Profiles Learners

With the advent of Information and communication technology (ICT), we talk about "Social Learning" which allows the collaborative learning exchange between individuals. Students learn how to communicate, eventually to reformulate for a better understanding. During the collaborative learning, the exchanges resulting from it are substantial; they allow students to get feedbacks and get back to their learning. Through social networks, we will allow the learners to be better at sharing and comparing their knowledge, offering a new style of dialogue.

Olivier Le Deuff[22] in an article in his blog "Guidedesegares.info" investigated on the concept of social networks made for leisure, as new ways of learning. It shows that people have developed skills beyond the mere techniques of creative leisure. It shows that people can progress in that field and take part of it.

Therefore, we can deduce that the innovation of distance learning, using social networking encourages inter-student interactions between teachers and students; it also highlights the educational resources brought by digital platforms, which prompt users to learn collaboratively.

In a distance learning situation, the characteristics of different social behaviors are among learners. However, distance and media exchanges do not make as immediate perception of these social behaviors. To overcome this limitation, we believe it is useful to have a computer system that can automatically analyze these behaviors.

In particular, in the context of online education, it is interesting to try to determine automatically profiles of social behavior among learners, as they can be useful both for the tutor and learners themselves.

So each learner connected to the platform produces a set of acts that we have previously presented (sharing tags, sharing documents, sharing comment). These acts are represented by the traces; they are stored at the database level. Using our system SYMPA, will allow us to classify them according to their types of actions and define profiles learners.

For our case and based on the work of Plety [23], we have defined three sets of profiles:

TABLE.II. THE PROFILES OF LEARNERS FOLLOWING A PERSONAL SYNTHESIS OF THE WORK OF PLETY .

Profile	Main Acts on the platform	Volume of participation of other acts
Producer	Share	Important/Medium/Low
Evaluator	Comment	Medium/Low
Observer	Tags	Low

After the definition of the social profiles, with which we have chosen to classify our learners, we present in the next paragraph the steps of their modeling.

VIII. MODELING OF COLLECTING DATA

A. Presentation of the Modeling

Our job is to analyze the traces generated at the level of our experimentation {Tags (Tag), Comment (Cmt), Sahre (Prtg)}, and modeling the profile of the learner through the use of fuzzy set. For each element of the set {Tag,Cmt,Prtg}, we will get a value that characterized the profile.

The choice of fuzzy logic for modeling the data collected is explained by the relative nature of this data, which designate knowledge that is not perceived or defined clearly.

The positioning of the learner profile according to the fuzzy logic is due to the combination of entered elements {Tag, CMT, Prtg} , with their degree of importance {Low, Medium, High}.

B. Steps of Fuzzy Logic

- Step of Fuzzification

This step has for aim to transform the variables of digital inputs in linguistic variables [24]. For our system we have as variables of entered the values below:

- Percentage of share Pctg(Prtg): It is the percentage of documents that share a learner on the platform (Figure 3).
- Percentage of comments Pctg(CMT) : it is the percentage of comments filed by a learner on the shared documents on the platform (Figure 4).

- Percentage of tags Pctg(Tags): this is the percentage of tags that a learner has marked on documents or comments shared on the platform (Figure 5).

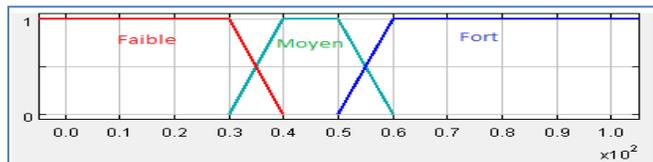


Fig.3. Function belonging to Tags

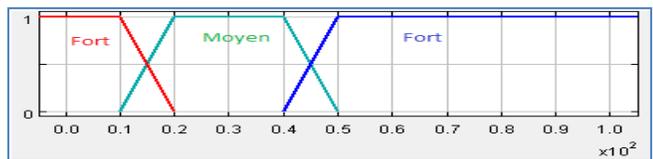


Fig.4. Function belonging to Share

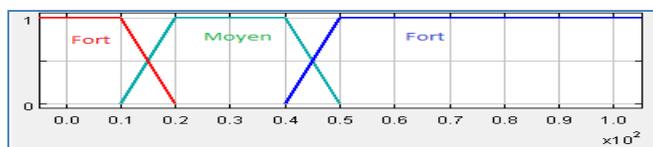


Fig.5. Function belonging to Comments

As output variable, we get the profile of the learner {Observer profile, Creator Profile, Producer Profile}, defined as follows (Figure 6).

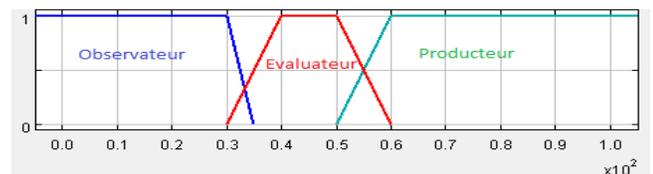


Fig.6. Function belonging to the output

• Step of Rules Inferences

The rules of fuzzy logic are of the form "If (X is A) then (Y is B) "with the fuzzy variable X which belongs to the Class A with a certain degree of membership and in the same way, the variable Y belongs to the class B with a degree of membership [25]. Below is sample of rules of our system:

If (Percentage share is Low) and (Percentage Tag is Low) and (Percentage Comments is Fort)
Then (Evaluator Profile)

If (Percentage share is Strong) and (Percentage Tag is Low) and (Percentage Comments is Fort)
Then (Producer Profile)

• Step of Defuzzification

The last step to have a blur system operational is called the defuzzification. During the second step, it has generated a lot of commands in the form of linguistic variables (one command per rule). The purpose of the defuzzification is to merge these commands and to transform the resultant settings into digital data [26].

Under the fuzzy set associated to the output variable are {Observer, Evaluator, Producer}.

The generation of the output variable is done by the system by using the method of the center of gravity, depending on the result, we determine the profile learner (Figure 6).

C. Results

The beaches of the results obtained according to the statistics of our experimentation, are as follows (Figure 6):

- Between 0% and 35% the result is Observer Profile.
- Between 30% and 70% the result is Evaluator Profile.
- Between 65% and 100% the result is Producer Profile.

To give a tutor interface to observe the social profile of each learner, we develop a new model in the platform "Moodle", in the Figure 7 an example of profile learner according to time, the blue color present "Observator Profile", when the learner change the kind of contribution {Comment , Share}, the learner have a new Profile {Evaluator, Producer}.

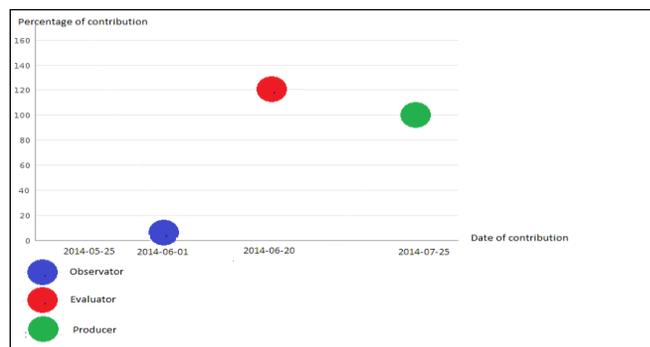


Fig.7. Viewing social profiles on Moodle

In addition, to validate our model, we have asked our tutor to make an assessment of their students and to assign a profile to each learner from their perceptions and what they observed during experimentation.

We have compared the experimental results obtained with the perception of the tutor, and we have noticed a great similarity in the results (table III).

TABLE.III. COMPARISON BETWEEN THE VALUES OF THE GUARDIAN AND THE MODEL OF POSITIONING

Learner	1	2	3	4	5	6	7	8	9	1	1	1	1	1	1	1
										0	1	2	3	4	5	6
Tutor	O	O	P	P	O	P	P	E	E	P	O	O	P	E	P	E
SYMPA	O	P	P	P	E	O	P	E	E	P	O	O	P	P	P	E

The difference noticed between a few profiles of learners defined by the tutor, can be explained either by the timidity of the learners during the sessions of courses or directed tasks (case of learners 2, 5), the preferences for the use of the

platform (case of learners 14), or technical problems on the platform (case of learners 6).

IX. CONCLUSION

In this paper we presented a method for modeling social profile learner, in the "Moodle" system powered by the social network "Mahara". Calculating the learner profile is influenced by among others inaccuracy, due to errors and approximations involved in gathering information, for this we used the method of fuzzy logic.

In this article, we have also presented the results of our experiments to validate our model for positioning the social profiles of learners in distance education; our job is to guarantee a better analysis of the different tracks that were produced by students in our experiment, to generate social indicators. With the help of our system SYMPA, the tutor can optimize his interventions with the learners: motivate, encourage, warn etc.

The results of our experiments were encouraging, and the profiles determined by the expert tutor were very close to the values given by our system SYMPA.

Our next goal is to analyze the traces obtained at the platform for more social indicator of learners.

REFERENCES

- [1] Bernatchez, J. et Gendreau, L. (2005). L'opération « contrat de performance » des universités québécoises : dans la perspective de l'objectif de réussite étudiante.
- [2] Rob Reinhardt ,2011 , About social community network features
- [3] Encyclopedia Britannica, 2008. « Encyclopedia Article for blog »
- [4] BoLeuf and WardCunningham. ,2002 , The WikiWay, by your hosts
- [5] Monox,2010,Social Networking features
- [6] Shama Kabani's book The Zen of Social Media: An Easier Way to Build Credibility 2013.
- [7] Julie Wittes Schlack , Michael Jennings, Manila Austin « Meeting Business Needs by Meeting Social Needs in Small Communities,2007 ».
- [8] Dachary, *Use the Right Tools for the Job*, March 9, 2010.
- [9] Cédric Lemery, « Du web 2.0 à l'éducation 2.0 ? », 2007.
- [10] Alan Cann, les réseaux sociaux améliorent la compréhension de la formation ,2010
- [11] Jean-Paul Pinte, Les réseaux sociaux sont des outils d'apprentissage en devenir ,2009.
- [12] Terry Anderson, Le Web 2.0, les réseaux sociaux et l'éducation. Les environnements technologiques, les standards et les pratiques innovantes.2009
- [13] Timothée Sicot, SMS et réseaux sociaux améliorent l'écriture des étudiants, 2013
- [14] <http://docs.moodle.org/27/en/Features>
- [15] Viva Logo Design Resources, 2011 Top 40 free downloadable open source social networking software
- [16] Elgg - Logiciel de réseau social Open Source,2004
- [17] Oxwall,Start a social network,a fan site,an education project with Oxwall-free open source community software,2010.
- [18] LiveStreet CMS, LiveStreet CMS is a great tool to create social networks and team blogs, 2010.
- [19] Mahara, « Mahoodle :Integrating Mahara with Moodle »
- [20] Sherif, M. et alii, 1988. The Robbers Cave Experiment: Intergroup Conflict and Cooperation. Middletown (Conn.), Wesleyan University Press.
- [21] Doise, W. et Mugny, G., 1997. Psychologie sociale et développement cognitif. Paris, Armand Colin.
- [22] Olivier Le Deuff ,2010 « Réseaux de loisirs créatifs et nouveaux mode d'apprentissage», Distances et savoirs. Vol.8, n°4, p.601-621.
- [23] Robert Pléty ,l'apprentissage coopérant, Ethologie et psychologie des communications. [24]William DERIGENT, Vincent BOM BARDIER « Extraction de connaissances et reconnaissance Automatique d'entités minces en aéronautique par Règles linguistiques floues » ; 11ème Colloque National AIP PRIMECA La Plagne - 22-24 avril 2009.
- [24] Franck Deroncourt, Le raisonnement en logique floue,2013
- [25] Gerald huguenin, Processeurs Floushe-arc, baptiste savoye 26, ch-2610 st-imie

Regression Testing Cost Reduction Suite

Mohamed Alaa El-Din

Arab Academy for Science,
Technology and Maritime Transport
(AASTMT)
Cairo, Egypt

Ismail Abd El-Hamid Taha

Arab Academy for Science,
Technology and Maritime Transport
(AASTMT)
Cairo, Egypt

Hesham El-Deeb

Modern University for Technology
and Information (M.T.I) Cairo, Egypt

Abstract—The estimated cost of software maintenance exceeds 70 percent of total software costs [1], and large portion of this maintenance expenses is devoted to regression testing. Regression testing is an expensive and frequently executed maintenance activity used to revalidate the modified software. Any reduction in the cost of regression testing would help to reduce the software maintenance cost. Test suites once developed are reused and updated frequently as the software evolves. As a result, some test cases in the test suite may become redundant when the software is modified over time since the requirements covered by them are also covered by other test cases.

Due to the resource and time constraints for re-executing large test suites, it is important to develop techniques to minimize available test suites by removing redundant test cases. In general, the test suite minimization problem is NP complete. This paper focuses on proposing an effective approach for reducing the cost of regression testing process. The proposed approach is applied on real-time case study. It was found that the reduction in cost of regression testing for each regression testing cycle is ranging highly improved in the case of programs containing high number of selected statements which in turn maximize the benefits of using it in regression testing of complex software systems. The reduction in the regression test suite size will reduce the effort and time required by the testing teams to execute the regression test suite. Since regression testing is done more frequently in software maintenance phase, the overall software maintenance cost can be reduced considerably by applying the proposed approach.

Keywords—Software maintenance cost; reduced test suite; reduced regression test suite; regression testing cost reduction.

I. INTRODUCTION

In regression testing as integration testing proceeds, number of regression tests increases and it is impractical and inefficient to re-execute every test for every program if one change occurs.

Test suite reduction techniques decrease the cost of software testing by removing the redundant test cases from the test suite while still producing a reduced set of tests that covers the same level of code coverage as the original suite.

Optimizing the cost of the regression testing without compromising the fault exposing capability is always challenging for the testing team. Testing team always face constraints like lack of resources, squeezed testing schedule, changing and ambiguous requirement, which in terms impacts and reduces the effectiveness of regression testing. The Test

automation tool will help testing team speed-up the test execution.

Due to the differences in the execution costs between the test cases, the representative set with the smallest number of tests may not be the one with the minimum execution cost. As such, the cost of a test should be a more important consideration for achieving cost-effective testing than the size of the test suite. Thus, it is necessary to consider individual execution costs when choosing the test cases.

The traditional HGS algorithm is one of the most common algorithms aiming to reduce the cost of regression testing. It is proposed by Harrold, Gupta and Soffa to test suite reduction “Selecting a representative set of test cases from a test suite, providing the same coverage as the entire test suite” that has received considerable attention. This algorithm assumes that we could have

T_i (for $i = 1, 2, 3, \dots, m$) represent the subsets of T , with each subset T_i containing all of the test cases that satisfy the i -th test requirement. The HGS algorithm could determine the representative test cases for each subset and include them in the representative set. The HGS algorithm follows the following four steps:

- 1) Initially, all requirements are unmarked.
- 2) for each requirement that is exercised by only one test case each, add each of these test cases to the minimized suite and mark it.
- 3) Consider the unmarked requirements in increasing order of the cardinality of the set of test cases exercising a requirement. If several requirements are tied since the sets of test cases exercising them have the same cardinality, select the test case that would mark the highest number of unmarked requirements tied for this cardinality. If multiple such test cases are tied, break the tie in favor of the test case that would mark the highest number of requirements with testing sets of successively higher cardinalities; if the highest cardinality is reached and some test cases are still tied, arbitrarily select a test case among those tied. Mark the requirements exercised by the selected test. Remove test cases that become redundant as they no longer cover any of the unmarked requirements.
- 4) Repeat the above steps until all testing requirements are marked.

The traditional HGS algorithm suffers from some disadvantages since no clear reason is shown for the initial

choice of the test cases as starting point. Also, it did not assure the cover all tests with all possible cases of all the selection statements.

II. PROBLEM STATEMENT

Given a set T of test cases $\{t_1, t_2, t_3, \dots, t_n\}$, a set of testing requirements $\{r_1, r_2, \dots, r_m\}$ that must be covered to provide the desired coverage of the program, and the information about the testing requirements exercised by each test case in T , the test suite minimization problem focus on finding a minimal cardinality subset of T that exercises the same set of requirements as those exercised by the un-minimized test suite T .

Most of the existing approaches to reduction aim to decrease the size of the test suite disregarding the time/cost. Yet, the difference in the execution time/cost of the tests is often significant and it may be costly to use a test suite consisting of a few long-running test cases. [2]

The reduction in the original test suite could be computed according to the following formula:

$$C_{red} [\%] = ((CR - C_{min}) / CR) * 100 \quad (1)$$

Where:

CR Original regression test suite
C min Reduced regression test suite

III. ALTERNATIVE APPROACHES

Many techniques have been proposed to obtain the near-optimal solution for the test suite reduction problem. Even though the representative sets produced by these techniques are not guaranteed to be optimal, they can significantly decrease both the size of the test suite and the cost associated with its execution.

These approaches could include the usage of Greedy algorithm, selective redundancy approach and irreplaceability algorithm.

A. Greedy Algorithm

The Greedy algorithm is a commonly-used method for finding the near-optimal solution to the test suite reduction problem. This algorithm repeatedly removes the test which covers the most unsatisfied test requirements from the test suite set T to the requirements set until all of the requirements are covered. Many existing test suite reduction methods are based on the concept of the Greedy algorithm. In other words, many algorithms repetitively choose the "best" test case to obtain the near-optimal solution from the locally optimal solutions. [3]

B. Test Suite Reduction with Selective Redundancy

Test suite reduction that attempts to selectively keep redundant tests in the reduced suites. Experiments show that this approach can significantly improve the fault detection effectiveness of reduced suites without severely affecting the extent of test suite size reduction. This assures the achievement of high suite size reduction while simultaneously allowing for low fault detection effectiveness loss. The

intuition driving is that when a non-reduced suite contains lots of redundancy with respect to a coverage criterion, it may be helpful to selectively keep some of that redundancy in the reduced test suite so as to retain more fault detection effectiveness in the reduced suite, hopefully without significantly affecting the amount of suite size reduction. [4]

C. Irreplaceability Algorithm

This algorithm is based on the concept of test irreplaceability which creates a reduced test suite with a decreased execution cost. Leveraging widely used benchmark programs, the empirical study shows that, in comparison to existing techniques, the presented algorithm is the most effective at reducing the cost of running a test suite. [5]

IV. RELATED WORK

Researchers, practitioners and academicians proposed various techniques on test suite reduction, test case prioritization, and regression test selection for improving the cost effectiveness of the regression testing.

Rothermel and Harrold presented a technique for regression test selection. Their algorithms construct control flow graphs for a procedure or program and its modified version and use these graphs to select tests that execute changed code from the original test suite [6].

James A. Jones and Mary Jean Harrold proposed new algorithms for test suite reduction and prioritization [5]. Saifur-Rehman Khan, Aamer Nadeem proposed a novel test case reduction technique called Test Filter that uses the statement-coverage criterion for reduction of test cases [8]. T. Y. Chen and M. F. Lau presented dividing strategies for the optimization of a test suite [4]. M. J. Harrold et al presented a technique to select a representative set of test cases from a test suite that provides the same coverage as the entire test suite [8]. This selection is performed by identifying, and then eliminating, the redundant and obsolete test cases in the test suite. This technique is illustrated using data flow testing methodology.

A recent study by Wong, Horgan, London, and Mathur [3], examines the costs and benefits of test suite minimization. Rothermel et al [2] described several techniques for using test execution information to prioritize test cases for regression testing, including: techniques that order test cases based on their total coverage of code components, techniques that order test cases based on their coverage of code components not previously covered, and techniques that order test cases based on their estimated ability to reveal faults in the code components that they cover. Most of the techniques described in the above papers assume that source code of the software is available to the testing engineer at the time of testing. But in most of the organizations the testing is done in black box environment and the source code of the software is not available to the testing engineers. A simple greedy algorithm for the set-cover problem (and therefore for the test suite minimization problem) is described in [4]. The work presented in [9] uses a greedy technique for suite reduction in the context of model-based testing. This work showed that while suite sizes could be greatly reduced, the fault detection capability of the reduced suites was adversely affected. This

situation increases the degree of complexity of the proposal solutions for the test suite minimization problem.

Existing test suite minimization techniques are defined in terms of test case cover-age as they attempt to minimize the size of a suite while keeping some coverage requirement constant. A related topic is that of test case prioritization.

In contrast to test suite minimization techniques which attempt to remove test cases from the suite, the test case prioritization techniques [8, 10, and 11] only re-order the execution of test cases within a suite with the goal of early detection of faults. In [11], the ATACMIN tool [6] was used to find optimal solutions for minimizations of all test suites examined. This work showed that reducing the size of test suites while keeping all uses coverage constant could result in little to no loss in fault detection effectiveness. In contrast, the empirical study conducted in [12] suggests that reducing test suites can severely compromise the fault detection capabilities of the suites.

A new model for test suite minimization [7] has been developed that explicitly considers two objectives: minimizing a test suite with respect to a particular level of coverage, while simultaneously trying to maximize error detection rates with respect to one particular fault. A limitation of this model is that fault detection information is considered with respect to a single fault (rather than a collection of faults), and therefore there may be a limited confidence that the reduced suite will be useful in detecting a variety of other faults.

From the previous demonstration of the above related work, it could be concluded that suite size and fault detection effectiveness are opposing forces in the sense that more suite size reduction would intuitively imply more fault detection and effectiveness loss, since throwing away more test cases, in effect, throws away more opportunities for detecting faults. Thus, there seems to be an inherent tradeoff involved in test suite reduction: one may choose to sacrifice some suite size reduction in order to increase the chances of retaining more fault detection effectiveness.

V. ENHANCED HGS ALGORITHM (EHGSA)

The research approach target is to get the original regression testing and the reduced regression test suite reduction with selective redundancy by modifying the HGS algorithm. This approach is general and can be applied to any test suite minimization technique. EHGSA finds the minimum regression test with minimum machine time of the test suite covering all possible paths primary variables values of the all selection branch cases (IF) statements of both cases True/False (T/F) of the program tested.

The EHGSA algorithm have several advantages since it take into consideration all the possible braches cases of selection statements included in the program being tested. Also, it computes the real machine time for each branch case and the total time for each test of the test suite. The pseudo code of the EHGSA algorithm is illustrated in Fig. 1.

```
Begin:  
Stage I: Create the Test Suite Text File  
Input: n; // number of selection statements  
m: = 2n; // m is all possible tests ti  
Open Test Suite Text File;  
i:=0;  
While (i < m)  
j:=0;  
While (j < n)  
convert i to binary number b;  
//ti: set primary values pj to binary i vales b  
set primary values pj to b bit j vale;  
j:=j+1;  
End While  
write Text Test Suite Line i ti;  
i:=i+1;  
End While  
Close Test Suite Text File;  
// Test Suite Text File created  
  
Stage II:  
Step 1: Establish Test Link List Class  
Test Class Node Structure {  
Test_id;  
Array Test_Coverage_marked_Selection_Cases;  
Counter_Marked_Selection_Cases h;  
Test_Machine_Time TT;  
Pointer next_Node;  
Pointer previous_Node; }  
  
Step2: Apply Test Suit on Selection Statements  
Open Test_Suite_Text File T as Input;  
// Array Primary Values PV  
Array PV[n];  
i :=0;  
While ( ! T.eof( ) )  
Read ( T , ti);  
j := 0;  
While ( j < n )  
Set PV[ j ] := ti(j,j) ;  
j:= j +1;  
End While  
// Apply ti on Selection Statements Cases  
If ( PV[k] )  
// if statements staff  
// Coverage Cases  
// Calculate total machine test time of ti;  
End If  
Add test_Node;  
i =i+1;  
End While
```

```

// Regression Testing Reduction Proposal
Algorithm Step4:
// Find Test ti Max Coverage with Min
Machine Time
Coverage = {};
Uncoverage={ all possible Coverage};
Min_Subset_Tests = {};
// Read T Test Link List Nodes ti;
i:=0;
Max_Coverage := 0;
// Looking for Test ti with Max Coverage &
Min Machine Time
MT := Max_no;
// At Head Test_Link_List T
While ( ! T.eof() )
Read T.Node ti;
If ( h >= Max_Coverage and TT <= MT)
Max_Coverage := h;
Test := ti;
End If
i = i + 1;
End While
Min_Subset_Tests = Min_Subset_Tests + Test;
Coverage ::= Coverage + Test.Coverage;
Uncoverage := Uncoverage – Coverage;

Step5:
// Find Test ti cover Max Uncoverage with
Min Machine Time

While ( Uncoverage != Null)
i:=0;
Max_Coverage := 0;
// Looking for Test ti with Max Coverage &
Min Machine Time
MT := Max_No;
// At Head Test_Link_List T
While ( ! T.eof() )
Read T.Node ti;
If (Test_Coverage_marked_Selection_Cases
<=
Uncoverage Max_Coverage and TT <=
MT)
MT = TT;
Test := ti;
End If
i = i + 1;
End While
Min_Subset_Tests = Min_Subset_Tests + Test;
Coverage ::= Coverage + Test.Coverage;
Uncoverage := Uncoverage – Coverage;
End While
// Proposal Algorithm Output
Write Min_Subset_Tests;
End
    
```

Fig.1. EHGSA Pseudo Code

VI. IMPLEMENTATION

Pointing out the test suites with minimum machine time where the test suite covers all possible paths of the selection statements by applying algorithm in the following sample case study (Fig. 2) with four if statements n=4, each test ti has n primary variable values, p[i], (i = 0, 1, 2, 3).

```

1: read text test suite file line ti
(p[0], p[1], p[2], p[3]);
B1: if (p[0] > 0)
B1T: // Branch 1 True Statements
B1F: else

// Branch 1 False Statements
End If
B2: if (p[1] > 0)
B2T: // Branch 2 True Statements
B2F: else
// Branch 2 False Statements
End If
B3: if (p[2] > 0)
B3T: // Branch 3 True Statements
B4: if (p[3] > 0)
B4T: // Branch 4 True Statements
B4F: else
// Branch 4 False Statements
End If
B3F: else
// Branch 3 False Statements
End If
    
```

Fig.2. Sample Case Study [4]

TABLE I. THE TEST SUITE FILE FOR ALL POSSIBLE PRIMARY VARIABLES VALUES M X N. WHERE: N: NUMBER OF SELECTION STATEMENTS, M = 2N

Test	p[3]	p[2]	p[1]	p[0]
t0	0	0	0	0
t1	0	0	0	1
t2	0	0	1	0
t3	0	0	1	1
t4	0	1	0	0
t5	0	1	0	1
t6	0	1	1	0
t7	0	1	1	1
t8	1	0	0	0
t9	1	0	0	1
t10	1	0	1	0
t11	1	0	1	1
t12	1	1	0	0
t13	1	1	0	1
t14	1	1	1	0
t15	1	1	1	1

TABLE II. EHGSA ALGORITHM OUTPUT ALL POSSIBLE REGRESSION TESTING WITH MACHINE TIME.M X ((2 * N) + 1).

Test/Case	B1T	B1F	B2T	B2F	B3T	B3F	B4T	B4F	Time
t0		X		X		X			0.03
t1		X		X		X			0.03
t2		X		X	X			X	0.043
t3		X		X	X		X		0.042
t4		X	X			X			0.029
t5		X	X			X			0.029
t6		X		X	X		X		0.042
t7		X	X		X		X		0.041
t8	X			X		X			0.029
t9	X			X		X			0.029
t10	X			X	X			X	0.042
t11	X			X	X		X		0.41
t12	X		X			X			0.028
t13	X		X			X			0.028
t14	X		X		X			X	0.041
t15	X		X		X		X		0.040

The EHGSA Algorithm Final Result for the Reduction Subset Tests is: t15, t0, & t14.

VII. TRADITIONAL HGS ALGORITHM RESULTS

Apply the HGS algorithm over the same selection statements case study Fig2. The HGS is used the test suite consists of only five test {t1, t2, t3, t4, t5} [4]. The HG algorithm used the following test suite.

TABLE III. THE HGS TEST SUITE INITIAL SUITE

Test	p[0]	p[1]	p[2]	p[3]
t0	1	1	0	0
t1	0	0	1	0
t2	0	1	0	0
t3	0	1	1	1
t4	0	0	1	1

TABLE IV. HGS ALGORITHM OUTPUT REGRESSION TESTS

Test/Case	Bt1	Bf1	Bt2	Bf2	Bt3	Bf3	Bt4	Bf4
T1	X		X			X		
T2		X		X	X			X
T3		X	X			X		
T4		X	X		X		X	
T5		X		X	X		X	

The HGS Algorithm Final Result for the Reduction Subset Tests is: t1, t2, & t4.

VIII. EXPERIMENTAL RESULTS

The EHGSA algorithm stage one has generates the text test suite file for all possible variables values PV.

The EHGSA algorithm stage two its input is the text test suite file then generate the original regression testing: CR.

The EHGSA algorithm stage two has criteria to find the Reduced Regression Test Suite CMIN of the original regression testing: CR that coverage all possible selection statement branch test cases with minimum cost (machine time) "Regression Testing Cost Reduction Suite".

Apply the EHGSA algorithm over different programs contains different number of selection statements SS has following parameters:

- Number of Selection Statements: SS
- Number of Primary Variables: PR = SS
- Possible Primary variables Values of for both branch cases T/F : $PV = 2^{SS}$
- Possible selection Branch Test Cases : $BTC = 2 * SS$
- Original Regression Testing: $CR = 2^{SS}$
- Reduced Regression Test Suite C_{MIN}

The reduction in the original test suite could be computed according to the formula (1)

TABLE V. EHGSA ALGORITHM EXPERIMENT RESULTS

SS	PR	PV	BTC	CR	C _{MIN}	C _{RED} [%]
4	4	16	8	16	3	81.25%
5	5	32	10	32	4	87.50%
6	6	64	12	64	5	92.19%
7	7	128	14	128	6	95.32%
8	8	256	16	256	7	97.27%
9	9	512	18	512	7	98.63%
10	10	1024	20	1024	8	99.22%
11	11	2048	22	2048	8	99.60%

The following figure illustrate the results in a bar chart which clarify that the reduction in cost of regression testing for each regression testing cycle is ranging highly improved in the case of programs containing high number of selected statements

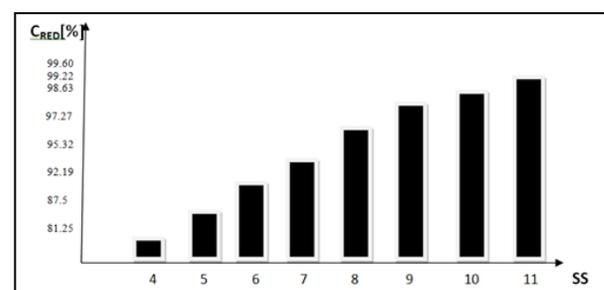


Fig.3. EHGSA Reduction Cost Results

IX. CONCLUSION

Selecting the reduced testing cases, in appropriate accurate approach; needs browsing all the possible paths of cases of the selection statements included in the cod.

The paper proposed algorithm automatically generates the test suite that cover all possible test primary variables values of all cases true/false for all selection statement of the tested program code. This algorithm computes the machine time of each test case on a dynamic base using the linked list with test node. The EHGSA finds the subset tests covering all possible test paths of all selection statements with minimum machine time which in turn reduced the regression testing cost.

REFERENCES

- [1] Sriraman Tallam, Neelam Gupta, "A Concept Analysis Inspired Greedy Algorithm for Test Suite Minimization", Proceeding PASTE '05 Proceeding of the 6th ACM SIGPLAN-SIGSOFT workshop on Program analysis for software and engineering 2005.
- [2] Prof. A. Ananda Rao and Kiran Kumar J "An Approach to Cost Effective Regression Testing in Black-Box Testing Environment", May 2011
- [3] S. Yoo and M. Harman, "Regression Testing Minimization, Selection and Prioritization: a Survey," Software Testing, Verification and Reliability, Vol. 22, No. 2, March 2012.
- [4] Dennis Jeffrey and Neelam Gupta, "Test Suite Reduction with Selective Redundancy", Dept. of Computer the University of Arizona Tucson, AZ 85721. IEEE Computer Society Washington, DC, USA
- [5] Chu-Ti Lin, Kai-Wei Tang, Cheng-Ding Chen, Gregory M. Kapfhammer, "Reducing the Cost of Regression Testing by Identifying Irreplaceable Test Cases", Aug 28, 2012
- [6] H. Zhong, L. Zhang, and H. Mei, "An Experimental Study of Four Typical Test Suite Reduction Techniques," Information and Software Technology, Vol. 50, No. 6, pp. 534-546, May 2008.
- [7] Prashant Malangave and Dr. Dinesh B. Kulkarni, "Efficient Test Case Prioritization in Regression Testing", Walchand Collage of Engineering Dept. of Computer Science & Eng, 2008
- [8] M. J. Harrold, R. Gupta, and M. L. Soffa, "A Methodology for Controlling the Size of a Test Suite," ACM Trans. on Software Engineering and Methodology, Vol. 2, No. 3, pp. 270-285, July 1993.
- [9] A. M. Smith and G. M. Kapfhammer, "An Empirical Study of Incorporating Cost into Test Suite Reduction and Prioritization," Proceedings of the 24th ACM SIGAPP Symposium on Applied Computing, Software Engineering Track, March 2009.
- [10] Luciano S. de Souza1; Ricardo B. C. Prudencio2, Flavia de A. Barros," Multi-Objective Test Case Selection: A study of the influence of the Catfish effect on PSO based strategies". Anais do XV Workshop de Testes e Tolerância a Falhas - WTF 2014.
- [11] Dennis Jeffrey, Neelam Gupta, "Test Suite Reduction with Selective Redundancy" Proceeding of ICSM '05 Proceedings of the 21st IEEE International Conference on Software Maintenance, Pages 549-558, 2005.
- [12] Anannado Rao and Kirzn Kumar J, "An Approach to Cost Effective Regression Testing in Black-Box Testing Environment" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011.

©2005

Discovering a Secure Path in MANET by Avoiding Black Hole Attack

Hicham Zougagh, Ahmed Toumanari, Rachid Latif, Y.
Elmourabit
Laboratory ESSI,
National School of Applied sciences
Agadir, Morocco

Noureddine.Idboufker
Laboratory TIM
National School of Applied sciences
Marrakech, Morocco

Abstract—In a mobile ad hoc network (MANET), a source node must rely on intermediate nodes to forward its packets along multi-hop routes to the destination node. Due to the lack of infrastructure in such networks, secure and reliable packet delivery is challenging. The performance of a Mobile Ad hoc Network (MANET) is closely related to the capability of the implemented routing protocol to adapt itself to unpredictable changes of topology network and link status. One of this routing protocol is OLSR [1] (Optimized Link State Routing Protocol) which assumes that all nodes are trusted. However, in hostile environment, the OLSR is known to be vulnerable to various kinds of malicious attacks. This paper proposes a cooperative black hole attack against MANETs exploiting vulnerabilities of OLSR. In this attack, two attacking nodes cooperate in order to disrupt the topology discovery and prevent routes to a target node from being established in the network.

Keywords—MANET; OLSR; Security; Routing Protocol Cooperative black hole attack

I. INTRODUCTION

A Mobile Ad-hoc NETWORK (MANET) is a collection of nodes which are able to connect to a wireless medium forming an arbitrary and dynamic network. Implicitly herein is the ability for the network topology to change over time as links in the network appear and disappear. In order to enable communication between pair of nodes in such a MANET, a routing protocol is employed. The abstract task of the routing protocol is to discover the topology to ensure that each node is able to acquire a recent map of the network topology to construct routes.

One way of securing a mobile ad hoc network at the network layer is to secure the routing protocols, so all possible attacks are prevented. The abstract task of the routing protocol is to discover the topology to ensure that each node is able to acquire a recent map of network topology to construct routes.

The Optimized Link Stat Routing Protocol (OLSR) is a proactive routing protocol for MANET, i.e. All nodes need to maintain a consistent view of the network topology. They are also vulnerable to a number of disruptive attacks in the presence of malicious nodes (identity spoofing, link withholding, link spoofing, miserly attack, wormhole attack and collusion attack..).

In this paper, we focus on the cooperative black hole attack [2] where two nodes cooperate to prevent routes to a target node from being established; the first attacker forces the target to choose it as its MPR node. It simply sends HELLO messages

with willingness equal to Will_always, after this it will choose the second attacker as its only multi-point relay that can drop, alter or look at any packet it forwards. The result is that the routes to target node cannot be established by nodes more than two hops away from it.

In our approach, we present, we present an improved MPR selection algorithm that can reduce the number of malicious nodes trying to be selected as Multipoint Relay by maintaining its Willingness fields equal to Will_always.

The rest of the paper is organized as follows. The next section provides a short overview on OLSR, followed by the description of cooperative black hole attack. Section IV summarizes the literature. In section V, we present our approach to secure OLSR protocol. In section VI we give an Illustration and an example. Section VII concludes the paper.

II. THE OLSR PROTOCOL

The Optimized Link State Routing Protocol (OLSR)[1], is a proactive link routing protocol, designed specifically for mobile ad hoc networks. OLSR employs an optimized flooding mechanism to diffuse link state information to all nodes in the network. In this section, we will describe the element of OLSR, required for the purpose of investigation security issues.

A. OLSR Control Traffic.

Control traffic in OLSR is exchanged through two different types of messages.

1) HELLO messages

To detect its neighbors with which it has a direct link, each node, periodically and at regular intervals (*HELLO Interval seconds*) broadcasts hello messages, containing the list of neighbors known to the node and their link status (symmetric, asymmetric, Multi-Point Relay or Lost). These messages are broadcast by all nodes and heard only by immediate neighbors; they are never relayed any further, i.e. these packets have a *Time-To-Live (TTL)* value of 1.

In addition to information about neighbor nodes, the periodic exchange of HELLO messages allows each node to maintain information describing the link between neighbor nodes and nodes which are two hops away. Based on this information, each node independently selects its own set of Multi-Point Relay (MPR) among its one-hop neighbors so that the MPR covers all two-hop neighbors.

2) Topology Control (TC) messages

TC (Topology Control) messages are also broadcast by MPR-nodes in the network at regular intervals ($TC_Interval\ second$). Thus, a TC message contains the list of neighbors that have selected the sender node as a MPR (MPR Selector Set), and an *Advertized Neighbor Sequence Number (ANSN)* is used by a receiving node to verify if the information advertized in the TC messages is more recent. The TC messages are flooded to all nodes in the network and take advantage of Multi-Point Relay to reduce the number of retransmissions.

Using information of a TC message, a node generates topology tuples ($T_des_adr, T_last_adr, T_seq, T_time$), the set of these tuples is denoted the "Topology Set". Here T_des_adr is the destination address, T_last_adr is the address of the node that generated the TC message, T_seq is a sequence number of the TC message and the T_time is the time duration after which the topology tuple expires [1].

Based on the information in the topology set, the node calculates its routing table; each entry in the table consists of $R_des_adr, R_next_adr, R_dist,$ and R_iface_adr .

Such entry specifies that the node identified by R_dest_adr is estimated to be R_dist hops away from the local node, that the symmetric neighbor node with interface address R_next_adr is the next hop node in the route to R_des_adr , and that this symmetric neighbor node is reached through the local interface with the address R_iface_adr . All entries are recorded in the routing table for each destination in the network for which a route is known [10].

B. Multi-Point Relays Selection.

Multi-Point Relays Selection is done in such a way that all the two-hop neighbors are reachable from the MPR in terms of radio range.

The two-hop neighbor set found by the exchange of HELLO messages is used to calculate the MPR set and the nodes signal their MPRs selections through the same mechanism.

The aim of Multi-Point Relays is to minimize the flooding of the network with broadcast packets by reducing duplicate retransmission in the same region Fig 1. Each node of the network selects the smallest set (MPRs) of neighbor nodes that can reach all of its symmetric two hop neighbors which may forward its messages. Each node in the network maintains an MPR selector set, which has selected this node as an MPR.

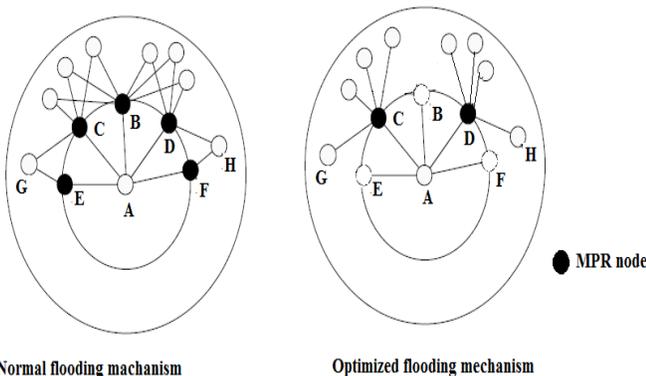


Fig. 1. Reduction of duplicate retransmission by MPR selection

III. THE MODEL OF COOPERATIVE ATTACK AGAINST OLSR PROTOCOL.

In this section, we describe how malicious node can launch a cooperative black hole attack in MANET. The first step to launch the cooperative black hole attack is that a malicious node S1 can force its election as MPR by maintaining constantly its willingness field to Will_always in its HELLO messages. According to the protocol, its neighbors will always select it as MPR. Using this mechanism, a malicious node can easily earn, as an MPR, a privileged position within the network, it can then exploit its rank to carry out deny of service attacks and alike. The second step S1 select its adjacent node S2 as MPR, after this, S2 can drops all TC messages forwarded by node S1. The attacked node, in the set of MPR selectors of S1, cannot detect this misbehavior because node S2 is out of its radio range.

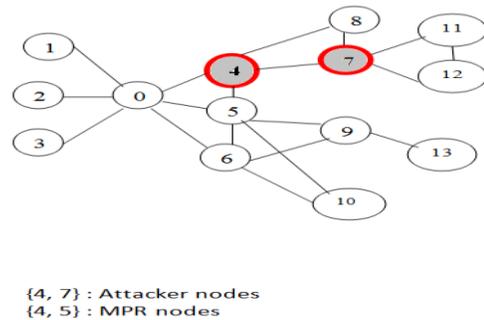


Fig. 2. A cooperative black hole attack model

Fig 2 shows an illustrative description of this cooperative black hole attack. Let {1,2,3} a set of nodes to be attacked and 4, 7 the attacker nodes, {4,5} the set of 0's MPR set nodes, {7,9} is the subset of 0's tow hop neighbors which constitutes the MPR set of nodes in 0's MPR set and {11,12,13} the set of 0's 3hop neighbors. The attack is launched as follows: node 4 sends its HELLO message with the value of willingness field as will_always, according to the protocol; all its one hop neighbors will choose it as an MPR. Then it chooses the node 7 as the only MPR node to relay its TC messages. By doing this if node 4 broadcast a TC message, then node 7 might be responsible to retransmit the message but may decide not to do so. In consequence, nodes {11,12} will never learn that the last hop to reach nodes {1,2,3} is node 0. The consequence of this attack is illustrated in Fig 3, where node C5, C6 and C7 can not build a route toward T's MPR selector because the 0's TC messages are never received (i.e. the topology information held by these nodes is incomplete).

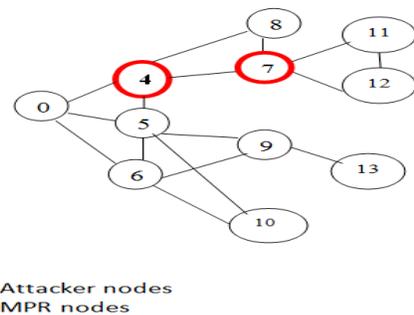


Fig. 3. Topology perceived by nodes 11, 12 after attack

IV. RELATED WORK

In [2], to detect a collusion attack the authors propose to extend the HELLO messages by including the two-hop neighbours list. Based on this extension, a node can learn its tree-hop neighbours without the need of TC message. The aim of this method is that a target node can detect the contradiction due to the attack. Though the proposed method detects an attack, it cannot differentiate between an actual attack and topology changing.

In [4] the authors propose the theoretical information framework for trust modeling. The method uses special packets to request neighbouring nodes for calculating the trust value of other nodes in the network. After a certain threshold the nodes will be blacklisted. This method involves observation of the suspected attackers and requires cooperatives of neighbouring nodes to arrive at correct results.

In [8] the authors address the problem of collusion attack in OLSR using an acknowledgement (ACK) based mechanism to detect attackers, so this scheme has a considerable overhead induced by the extra control messages.

In [7] the author proposes a method to avoid a virtual link attack by using SNVP protocol based on the Principle of checking the symmetry of the link advertised by the neighbour before confirming it. The problem of the proposed solution is that it might not detect the misbehaving nodes that launch the proper attack.

A SU-OLSR[6] is a solution to detecting malicious attack that can use either HELLO messages claiming illegitimate neighbours or TC messages claiming falsely that is has been selected as MPR. In this method the authors extend the HELLO messages by listing the selected trusted MPR set and the discovered non trusted suspicious set. The MPR selection of SU-OLSR has a different goal. Its objective is to reduce the impact of malicious nodes trying to be selected as MPR nodes. Thus, the MPR selection algorithm has to find the non trusted nodes according to the selected criterion and the trusted MPR covering a maximum subset of two-hop neighbours.

In [3] the authors address another problem called Node Isolation Attack. In this attack, an MPR node does not generate its TC message. To defend against this attack the authors propose a countermeasure that consists of two phases: detection phase and avoidance phase. In the first phase the target observes its MPR node to check whether the MPR is generating TC message or not. In the second phase, to avoid the impact of this attack, the authors include in the HELLO message a new field named Requested-value.

In the suggested technique [9], when the node detects a symptom of collusion attack, it adds the lone MPR to an AvoidanceSet after waiting for AvoidanceDelay. All entries in the AvoidanceSet of X are not included in its MPRs computation process. These entries are removed from AvoidanceSet after duration AvoidanceOld. In addition the authors discuss two possible convergences of the attack. This method is simple but it affects a network performance by repeating the processes selection of MPR set in case of legitimate node.

In method [5], the authors present a scruple when a symptom is checked right. The node waits for a fixed duration and sends scruple packet. The inconvenience of this method is that it increases the overhead.

Sanjay Ramaswamy et al. exploit data routing information (DRI) table and cross checking method to identify the cooperative black hole nodes, and utilize modified AODV routing protocol to achieve this methodology [11].

Chang Wu Yu et al. propose a distributed and cooperative mechanism viz. DCM to solve the collaborative black hole attacks. Because the nodes works cooperatively, they can analyze, detect, mitigate multiple black hole attacks. The DCM is composed of four sub-modules [12].

Weichao Wang et al. design a hash based defending method to generate node behavioral which involve the data traffic information within the routing path. The developing mechanism is based on auditing technique for preventing collaborative packet drop attacks, such as collaborative black hole and grey hole problems [13].

Zhao Min and Zhou Jiliu propose two hash-based authentication mechanisms, the message authentication code (MAC) and the pseudo random function (PRF). These two proposals are submitted to provide fast message verification and group identification, find the collaborative suspicious hole nodes and discover the secure routing path to prevent cooperative black hole attacks [14].

Vishnu K. and Amos J. Paul address a mechanism to detect and remove the black and gray hole attack. This solution is able to find the collaborative malicious nodes which introduce massive packet drop percentage. Authors, refer this method to penetrate their system model, and also add a novel scheme videlicet restricted IP (RIP) to avoid collaborative black and gray attacks [15].

Po-Chun Tsou et al. design a novel solution named Bait DSR (BDSR) scheme to prevent the collaborative black hole attacks. The proposed mechanism is composed of proactive and reactive method to form a hybrid routing protocol, and the major essence is the DSR on-demand routing [16].

The main goal in this paper is to detect successfully and isolate the data packet dropping attackers from routing path in OLSR routing protocol for MANETs [17].

V. THE PROPOSED SOLUTION

To deal with cooperative black hole attack, we present an improved MPR selection algorithm which has a different goal; its objectif is to reduce the impact of malicious nodes trying to be selected as MPR nodes by maintaining constantly its willingness fields equal to will_always in the HELLO message. In order to limit the impact of this attack the following concept of trustworthiness is used: a node S should not trust any neighbor X showing strong characteristics which can maintain its willingness to will_always and $|MPR_set(X)|=1$.

In [1] the standard way of selecting MPR set, start with an MPR set made of all members of node with willingness equal to

will_always, then it select as a MPR the node with highest willingness among the nodes in its one hop neighbor with non zero reachability (the number of nodes in two hop neighbor which are not yet covered by at least one node in the MPR set, and which are reachable through this one hop neighbor). In our algorithm we give priority to a node that covers maximum nodes in two hop neighbors without giving priority to node with highest willingness.

Before introducing this algorithm, some notations should be described first:

- $1HN_set(X)$: the set of node X's one hop symmetric neighbors. It is created by the way of changing HELLO messages between nodes.
- $2HN_set(X)$: the set of node X's two hop symmetric neighbors excluding any node in $1HN_set(X)$. It is also created by the way of changing HELLO messages.
- $Degree(X, Y)$: the degree of node X's one hop neighbor; returns the number of nodes in $2HN_set(X)$ such that $\{2HN_set(X) \cap 1HN_set(Y) \neq \emptyset\}$ assuming that $Y \in 1HN_set(X)$.
- $Reachability(X, Y)$: the number of nodes in $2HN_set(X)$ which are not yet covered by at least one node in the $MPR_set(X)$, and which are reachable through node Y
- $MPR_set(X)$: the set of nodes selected as MPR by the node E. ($MPR_set(X) \subseteq 1HN_set(X)$).
- $MPRS_set(X)$: the set of symmetric neighbours which have selected the node X as MPR. ($MPRS_set(X) \subseteq 1HN_set(X)$).
- $Isolate_set$: A subset of $2HN_set(X)$ which are covered by only node in $1HN_set(X)$.

Our proposed algorithm for selection of MPRs, constructs an MPR_set that enable a node to reach any node in the symmetrical strict 2_hop neighborhood through relaying by one MPR node without giving opportunity to node with willingness equal to will_always.

The proposed heuristic for selecting MPRs is then as follows:

- 1) Calculate degree of each node in one hop neighbor of X
- 2) Select as MPRs those nodes in one hop neighbor which cover the isolate nodes in two hop neighbor.
- 3) We remove the isolate nodes from two hop neighbor set for the rest of the computation.

While there exist nodes in two hop neighbor which are not covered by at least k nodes in the MPR set.

- Calculate the reachability of each node in $1HN_set(X)$ node in $MPR_set(X)$.
- For each node in $1HN_set(X)$, calculate the reachability, i.e., the number of nodes in $2HN_set(X)$ which are not yet covered by at least one node in the MPR set, and which are reachable through this 1-hop neighbor.
- Select as a MPR the node with lower willingness among the nodes in $1HN_set(X)$ with non-zero reachability. In case of multiple choice select the node which provides

reachability to the maximum number of nodes in $2HN_set(X)$. In case of multiple nodes providing the same amount of reachability, select the node as MPR whose $D(y)$ is greater.

- Eliminate all the nodes in $2HN_set(X)$ now covered by at least one node in the MPR_set .

Algorithm 1: MPR Selection

```
1HN*_set(X) ← 1HN_set(X)
2HN*_set(X) ← 2HN_set(X)
MPR_set(x) ← ∅
S1 ← ∅
S2 ← ∅
For all node Y ∈ 1HN_set(X) do
Degree(X,Y) ← | 1HN_set(Y) \ 1HN_set(X) \ {X,Y} |
End.
While (∃ Z: Z ∈ 2HN*_set(X) ∩ ∃! Y ∈ 1HN*_set(X): Z ∈
1HN_set(Y)) do
MPR_set(X) ← MPR_set(X) ← {Y}
1HN*_set(X) ← 1HN*_set(X) \ {Y}
2HN*_set(X) ← 2HN*_set(X) \ 1HN_set(Y)
End.
While (2HN*_set(X) ≠ ∅) do
For each Y ∈ 1HN*_set(X) do
Reachability(X, Y) ← | {F / F ∈ 2HN*_set(X) ∩ 1HN_set(Y) and
MPR_set(X) ∩ 1HN_set(F) = ∅} |
End.
For each Y ∈ 1HN*_set(X) with reachability(X,Y) ≠ 0 do
S1 ← {Y / Willingness = min (willingness(Y))}
End.
If |S1| = 1 then
MPR_set(X) ← MPR_set(X) ← {Y}
1HN*_set(X) ← 1HN*_set(X) \ {Y}
2HN*_set(X) ← 2HN*_set(X) \ 1HN_set(Y)
Else
S2 ← {Y / Reachability (X,Y) = max (Reachability (X,Y), Y ∈
1HN*_set(X) )}
If |S2| = 1 then
MPR_set(X) ← MPR_set(X) ← {Y}
1HN*_set(X) ← 1HN*_set(X) \ {Y}
2HN*_set(X) ← 2HN*_set(X) \ 1HN_set(Y)
Else
MPR_set(X) ← MPR_set(X) ← {Y / Degree(X,Y) = max {
Degree (X,Y), Y ∈ 1HN*_set(X)}
1HN*_set(X) ← 1HN*_set(X) \ {Y}
2HN*_set(X) ← 2HN*_set(X) \ 1HN_set(Y)
End if
End if
END.
```

Algorithm 1, start with an empty Multipoint Relay Set, select those one-hop neighbor nodes in $1HN_set(X)$ as MPR which are the only neighbor of some nodes in $2HN_set(X)$ with willingness different to `will_never` which covers a nodes in `isolate_set`, and add these one-hop neighbor nodes to the multipoint relay set of X. Then if there are still some node in two-hop neighbors set which is not covered by the multipoint relay set, select the one-hop neighbors with lower willingness and who could cover the most uncovered two hop neighbor as MPRs and which has de maximum degree. Repeat this step until all the two-hop neighbors are covered by MPRs.

As soon as node X receives a HELLO message from its MPR node Y which showing the same characteristics of attacker node ($Y_willingness = will_always$ and $|MPR(Z)| = 1$), it recalculates its MPR set without it. Otherwise, if Y has more than one MPR neighbor node, X will process HELLO message normally (Algorithm 2).

Algorithm 2 :HELLO reception

```

If orig_adr_willigness = Will_always and orig_adr ∈ MPR_set
(receiver_adr) then
  If |MPR(orig_adr)| = 1 then
    If |1HN_set(orig_adr)| ≠ 1 then
      Recalculate MPR_set (receiver_adr) without orig_adr
      Drops Hello message
    Else Process HELLO message
  Else Process HELLO message
Endif
Else Process HELLO message
Endif
END.

```

Based on the information in the topology set, the node calculates its routing table by application of this algorithm which discard the node with high Willingness to reach the two hop neighbor:

- 1) All the entries from the routing table are removed.
- 2) The new routing entries are added starting with the symmetric neighbors ($h=1$) as the destination nodes.
- 3) For each node in $N2$ create a new entry in the routing table:

$N2$ is the set of 2-hop neighbors reachable from this node, excluding:

- The nodes only reachable by members of $1HN_set$ with willingness equal to `WILL_Always`.
- The node performing the computation.
- All the symmetric neighbors: the nodes for which there exists a symmetric link to this node on some interface.

4) For each topology entry in the topology table, if its T_dest_addr does not correspond to R_dest_addr of any route entry in the routing table AND its T_last_addr corresponds to R_dest_addr of a route entry whose R_dist is equal to h , then a new route entry MUST be recorded in the routing table:

- $R_dest_addr = T_dest_addr$
- $R_next_addr = R_next_addr$ of the entry with ($R_dest_addr = T_last_addr$)
- $R_dist = h+1$

VI. ILLUSTRATIVE EXAMPLE.

To understand the mechanism of our solution, we present a schema which shows an example of MANET (Fig. 4). Table 1 represents the nodes in one hop neighbors of E and their Willingnesses.

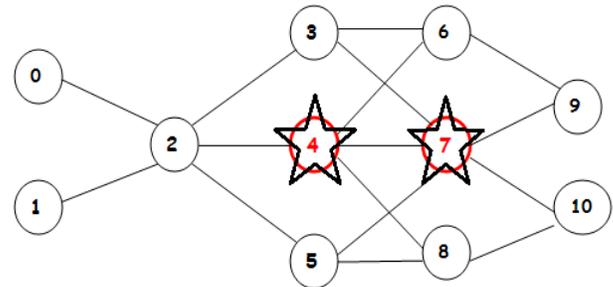


Fig. 4. Example of cooperative black hole attack model: {0,1} Target nodes and {4,7} a cooperative black hole attackers nodes.

TABLE I. WILLINGNESSES OF NODES IN $1NH_SET(2)$

Nodes	Willingnesse
3	3
4	7
5	3

The statement of our algorithm is as following:

- Calculating the degree of each node in $1HN_set(2)$: degree = {3 (2), 4 (3), 5 (2)}.
- Adds to the $MPR_set(2)$ those nodes in $1HN_set(2)$, which are the only nodes to provide reachability to a node in $2HN_set(2)$; $isolate_nodes = \{\emptyset\}$ then $MPR_set(2) = \{\emptyset\}$ and $2HN_set(2) = 2HN_set(2) \setminus \{\emptyset\} = \{6,7,8\}$.
- Since, as $2HN_set(2) = \{6,7,8\} \neq \emptyset$, the algorithm proceeds by calculating the reachability of nodes in $1HN_set(2)$: reachability (3) = 2, reachability (4) = 3, reachability (5) = 2. Then it adds nodes 3 and 5 to the $MPR_set(2)$ because willingness of 4 is equal to `will_always` and removes $1HN_set(3,8)$ from $2HN_set(2)$.
- Finally, we have $2HN_set(2) = \emptyset$ then the algorithm return $MPR_set(2) = \{3,5\}$ (Fig 5).

Suppose now, that (4,7) a cooperative black hole attacks. By the application of our approach, 4 will never be selected as MPR, because it has a high willingness and there exist other nodes with lower willingness which covers all nodes in to hop neighbors. After this when the first attacker 4 lunch the attack by selecting node 7 as its MPR node, it sends a HELLO message to a node 2. This last detects that 4 shows strong characteristics of malicious node, then it will recalculate the MPR_set (2) without 4, this operations will have result as 2 will choose {3,5} as its MPR to cover {9,10}.

In general our approach not favors nodes that have a Willingness equal to Will_always to the other nodes (Fig. 5). Otherwise, if we use the standard way of selecting MPRs [1], node 4 will be selected as multipoint relays (Fig. 6), which means the convergence of cooperatives attacks. The consequently of the attacks is that node 9,1,0 can not build a route toward 2's MPR selectors because the 2's TC messages are never received.

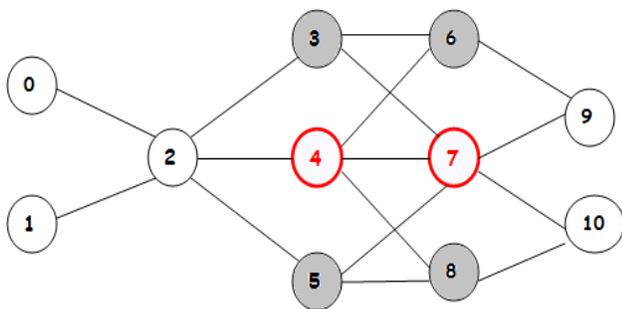


Fig. 5. An Example of selecting MPRs using Algorithm 1. MPR_set(2) = {3,5}.

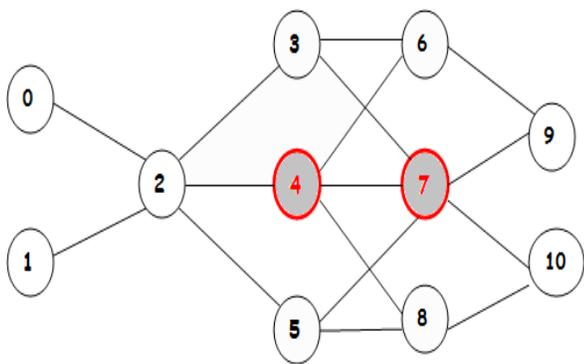


Fig. 6. An Example of selecting MPRs using standard OLSR. MPR_set(2) = {4}.

VII. SIMULATION AND RESULTS

To test the effectiveness of our solution, simulations were implemented using network simulator NS-2.35 with modified version of the UM-OLSR implementation. We embedded our scheme in implemented OLSR protocol for the detection of the cooperative black hole attack. All the default values for the OLSR protocol from [1] were used. The simulations were performed for 20 to 100 nodes with a transmission range of 250 meters, in an area of size 1000*1000 meters during 150 seconds. Random waypoint model is used as the mobility model of each node. Nodes speed is varied from 0m/s to 10 m/s. A single source generate UDP packets to the target (that has a distance further

than two hops away) from 10th second. To launch the attack, the first attacker chooses a victim node from its MPR selector set that has to be an MPR of the other neighbors at the 20th second (Table 2).

TABLE II. SIMULATION PARAMETER

Parameter		Values
Connection type		CBR/UDP
Simulation area		1000*1000
Transmission Range		250 m
Packet size		512 bytes
Number of Nodes	20-40-60-80-100	
Duration	150 s	
Pause time	0 s	
CBR_Start	10s	
Attack_start	20s	

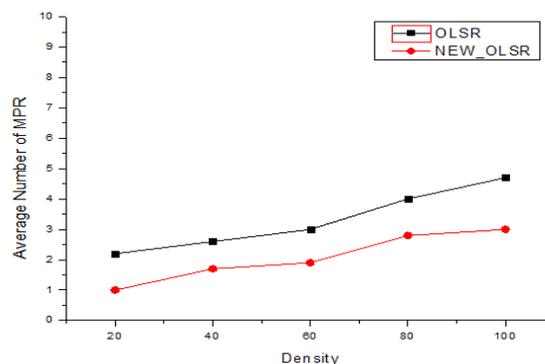


Fig. 7. Average number of MPR versus Density

Fig 7 gives the average number of MPR nodes selected by OLSR and New_OLSR for different densities (50% of nodes are willingness equal to 7). We can see that density clearly affects the number of MPR node selected by both protocols. It increases when density is increased and the number of MPR nodes selected by New_OLSR is low than the number selected by OLSR. The reason is that our algorithm of selection don't gives priority to a node with Willingness equal to Will_always but select as MPR the nodes that covers maximum nodes in its two hop neighbors with lower willingness.

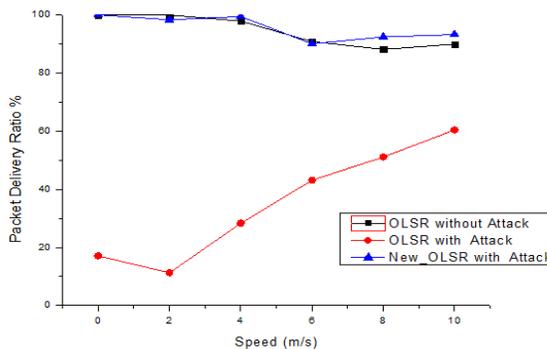


Fig. 8. PDR versus Speed under different scenarios

We also define the packet delivery ratio (PDR) as a value of the number of received data packets to that of packets being sent by the source node. Fig 8 compares OLSR and our approach New-OLSR. We observe that in presence of the attack, the PDR in OLSR is very low, the only packets received by the node are before launching the attack and we see that the PDR increase when the speed of the node increases. On the other hand when the New-OLSR is under attack we see that, generally, PDR is stable (minimum value equal to 90%) and better than the OLSR performance without attack. This is due to our approach route calculation, eliminating nodes with symptoms of malicious nodes routes to the destination node.

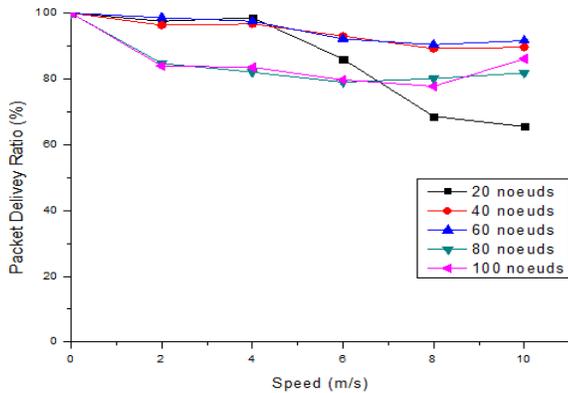


Fig. 9. Packet Delivery ratio under different number of nodes.

Fig 9, shows the relationship between Packet Delivery Ratio and speed. Generally the PDR decreases slightly with increasing velocity. Firstly with increasing speed in the case of 20 nodes the PDR does not exceed 65.5%. This is because the target has no choice in its one hop neighbor to select its MPR nodes. Secondly in case (80,100) we notice a slight decrease which exceeds 80%. Finally, for the case (40 and 60) a similar behavior can be seen with a reduction not exceeding 90%.

Fig 10 shows how our strategy offers a higher prevention to mitigate the effect of cooperative black hole attack. The percentage of detection rate is 100 % in static network, we observe an increase of detection rate in the case of large density.

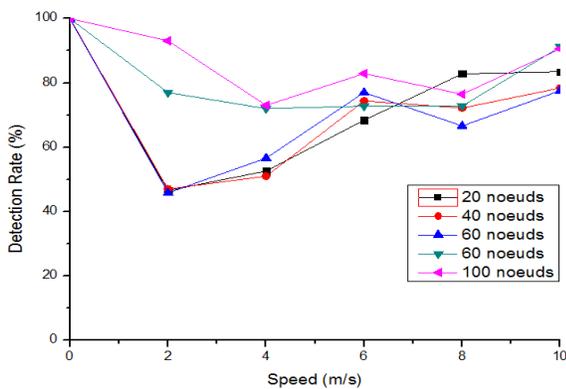


Fig. 10. Detection rate when changing mobility of nodes and a number of nodes.

VIII. CONCLUSION

The cooperative black hole attack exploits the routing protocol's vulnerabilities by forcing its election as Multipoint relay by maintaining constantly its willingness field to will_always in its HELLO message.

In order to deal with this sophisticated attack, we have proposed a novel approach to select MPR nodes. This gives priority to a node that covers maximum nodes in two hop neighbors with lower willingness which not showing strong characteristics to influence the MPR selection to be selected as MPR. We modified the procedure of calculating routes through the elimination the node with high Willingness to reach the two hop neighbor.

Simulation results demonstrate that the proposed method is effective in struggling cooperative black hole attack. It shows high packet delivery ratio and high detection rate of malicious nodes.

REFERENCES

- [1] T.Clausen, P. Jaquet, IETF Request for Comments: 3626 Optimized Link State Routing Protocol OLSR, october 2003.
- [2] A. Jamalipour B. Kannhavong, H. Nakayama.A collusion attack against OLSR-based mobile ad hoc networks. In Global Telecommunications Conference, GLOBECOM '06. IEEE, pages 1--5, November 2006.
- [3] Bounpadith Kannhavong , Hidehisa Nakayama , Nei Kato , Abbas Jamalipour , Yoshiaki Nemoto, A study of a routing attack in OLSR-based mobile ad hoc networks, International Journal of Communication Systems, v.20 n.11, p.1245-1261, November 2007.
- [4] Kishore Babu Madasu, A. Antony Franklin, and C. Siva Ram Murthy. On the Prevention of Collusion Attack in OLSR-based Mobile Ad hoc Networks. In IEEE International Conference on Networks (ICON 2008), New Delhi, India, December 2008.
- [5] Lalith Suresh P, Rajbir kaur, Manoj Singh Gaur, Vijay Laxmi. A collusion attack detection method for OLSR-based MANETS employing scruple packets. the 3rd international conference on Security of information and networks. 2010.
- [6] Rachid abdellaoui and Jean Marc Robert. SU-OLSR : A new solution to thwart attacks against the olsr protocol. Mster thesis.Height school of technology (ETS) Canada. 2009.
- [7] Soufian Djahel, Farid Nait Abslam, Avoiding virtual link attack in wireless ad hoc networks, Proceeding of the 2008 IEEE/ACS International conference of computer systems and application, p 355-360. March 31 avril 04, 2008.
- [8] Soufiene Djahel, Farid Naft-Abdesselam, Zonghua Zhang, and Ashfaq Khokhar. Defending against packet dropping attack in vehicular ad hoc networks. Security and Communication Networks, 1(3):245--258, 2008.
- [9] Suresh, P.L.; Kaur, R.; Gaur, M.S.; Laxmi, V.Collusion attack resistance through forced MPR switching in OLSR. Wireless Day IFIP 2010. Venice. Italy.
- [10] C.Adjih, A.Laouiti, P.Minet, P.Muhlethan, A. Quayyum, L.Viennot. The Optimized Routing Protocol for Mobile ad hoc Networks: Protocol Specification. Projet HIPERCOM.INRIA research report N° 5145, March 2004.
- [11] Weerasinghe H, Fu H (2007) Preventing Cooperative Black Hole Attacks in Mobile Ad Hoc Networks: Simulation Implementation and Evaluation. Paper presented at the Future Generation Communication and Networking, Jeju-Island, Korea, 6-8 December 2007.
- [12] Yu CW, Wu T-K, Cheng RH, Chang SC (2007) A Distributed and Cooperative Black Hole Node Detection and Elimination Mechanism for Ad Hoc Network. Paper presented at the PAKDD workshops, Nanjing, China, 22-25 May 2007.

- [13] Wang W, Bhargava B, Linderman M (2009) Defending against Collaborative Packet Drop Attacks on MANETs. Paper presented at the 2nd International Workshop on Dependable Network Computing and Mobile Systems (DNCMS 2009) (in Conjunction with IEEE SRDS 2009), New York, USA, 27 September 2009.
- [14] Min Z, Jiliu Z (2009) Cooperative Black Hole Attack Prevention for Mobile Ad Hoc Networks. Paper presented at the International Symposium on Information Engineering and Electronic Commerce, Ternopil, Ukraine, 16-17 May 2009
- [15] Vishnu KA, Paul J (2010) Detection and Removal of Cooperative Black/Gray hole attack in Mobile Ad Hoc Networks. *International Journal of Computer Applications* 1(22):38–42. doi: 10.5120/445-679.
- [16] Tsou P-C, Chang J-M, Lin Y-H, Chao H-C, Chen J-L (2011) Developing a BDSR Scheme to Avoid Black Hole Attack Based on Proactive and Reactive Architecture in MANETs. Paper presented at the 13th International Conference on Advanced Communication Technology, Phoenix Park, Korea, 13-16 Feb. 2011.
- [17] Ahmed Mohamed Abdalla, Ahmad H. Almazeed, Imane Aly Saroit, Amira Kotb, Detection and Isolation of Packet Dropping Attacker in MANETs. *International Journal of Advanced Computer Science and Applications*, Vol. 4, No.4, 2013.

Secure Deletion of Data from SSD

Akli Fundo¹, Aitenka Hysi² Igli Tafa³

Polytechnic University of Tirana,
Computer Engineering Department

Abstract—The deletion of data from storage is an important component on data security. The deletion of entire disc or special files is well-known on hard drives, but this is quite different on SSDs, because they have a different architecture inside, and the main problem is if they serve the same methods like hard drives for data deletion or erasing. The *built-in* operations are used to do this on SSDs. The purpose of this review is to analyse some methods which are proposed to erase data from SSDs and their results too, to see which of them offers the best choice. In general we will see that the techniques of erasing data from entire disc from hard drives can be used also on SSDs, but there's a problem with bugs. On the other hand, we cannot use the same techniques of erasing a file from hard drives and SSDs. To make this possible, there are required changes in FTL layer, which is responsible for mapping between logic addresses and physical addresses.

Keywords—*deletion-data; SSD; FTL layer; logic address; physical address*

I. INTRODUCTION

Nowadays, corporations and agency's store their data's in digital media, managing them is becoming important. In this article we will focus on challenges of SSDs for erasing the information, and some suggestions from different researches and experimental results. Data erasing is an important process and different techniques are include like built-in in ATA or SCSI commands. This techniques are effective on HDDs, where we can store the entire disc or specific files, but it's not the same thing with SSDs, because SSDs and HDDs have a different technology and algorithms of managing the information. SSDs have an indirect layer between the logic address that computer systems use to access data and the address that identify the physical storage [8]. The differences between SSDs and hard drives make it unclear of which techniques or commands are worthy on both of them. An experiment of [8] make this clear: they have write a structured data model to the drive, apply the deletion techniques, dismantle the drive and extract the data directly from the flash chips using a flash testing system [8]. To delete one file they made changes on FTL layer [1]. From the articles I've read, I have seen that there are solutions about this problem, but the performance is decreasing. On the other section I would like to show some of the deletion problems and what different engineers have done to solve them. But firstly I would like to show from [9] four levels of clearing(sanitizing) that a storage media could have:

The first level is *logic clearing*. The deleted data on logic way cannot be salvage through standard interface hard-ware like ATA or SCSI commands. The user can delete logically

one file or the entire disc by overwriting respectively the whole disc or a part of it.

The second level is *digital clearing*. In this case it's impossible to recuperate the data in a digital way.

The third level is *analog clearing*. This level can damage the signal which encodes the data, and it's impossible to rebuilt the signal even with very sensitive equipments. An alternative way to "hide" the bits is *cryptographic clearing* [4]. In this level the media uses a key for encryption and decryption of the data which enter and exit. But this is not a secure way, because someone can extract the key from physical media and can avoid the encryption.

II. RELATED WORKS

SSDs have specific characteristics. The traditional magnetic discs are composed by sectors 4 KB. The sector is the smallest unit of data which can be read or written on the disc. SSDs operate in another way, after the minimal number of bytes which can be read vary from those which can be written or delete. The flash banks have a typical block size, from 16 KB to 512 KB and the minimal number of bytes written simultaneously is 4 KB [12]. Otherwise from magnetic discs, the flash media firstly must delete a whole block before writing new data. SSDs don't have mechanical part so they don't have rotation or searching latency [11]. That's why they have a good performance. But there's a problem with the "wear" phenomenon, which is present over times. Every block of SSD can be deleted in finite times and after that it can not be written anymore. To fix this problem engineers have applied "wear leveling" in I/O controller. This is a group of algorithms used by controller to make a same allocation (diffusion) of deleting cycles on every block of the memory flash. In this way, no one of the blocks cannot be deleted not normally, to avoid SSDS fail [12]. Another way to fix "wear" problem and the deleting before writing engineers have proposed a hybrid architecture SSD-HDD called HPDA [13], which is more reliable and increases the performance. From [3] SSD uses a flash memory to store the data. This memory is divide into pages and blocks. The program operations interfere in this pages and change "1" to "0". The clearing operations are applied in the blocks and set all the bits of the blocks in "1". Usually there are 64-256 pages/blocks. FTL manages the mapping between logic blocks addresses (LBA), we can see them from ATA or SCSI interfaces and physic pages of flash memory. There is a mismatch of this two operations, so it's not possible a directly update of the LBA sector. Instead of modifying the sector, FTL will write the new content of sector in another place and will update the map.

The flash memory will keep the old version of data in a digital form[8]. Since directly updates are not possible as I said upper, the overwriting techniques which work on hard drives may not work on SSDs. This techniques suppose that overwriting of a part in a LBA area will result in overwriting on the same physic place. Except from FTL, the engineers have proposed **CAFTL** [14] (Content-Aware Flash Translation Layer) to reduce in an effective way the traffic writing on memory flash, removing the unnecessary duplicated writing. They have realized this by join together the parity data, which increase the efficiency of garbage collection and “wear leveling”.

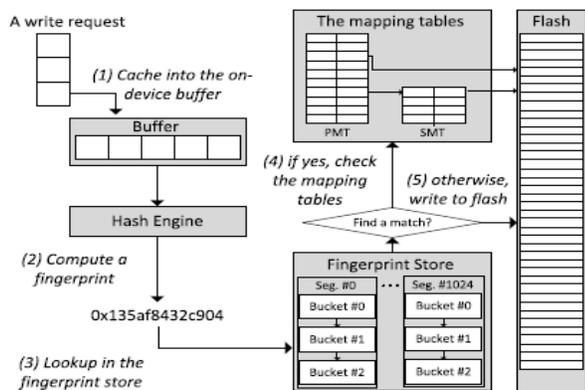


Fig. 1. CAFTL architecture [14]

The figure 1 shows the architecture of a CAFTL. It eliminates the duplications of writing and redundant data through a combination of the non duplications: in-line and out-of-line.

It examines the entry data and delete the redundant data before the writing process in flash. Anyway, it doesn't guarantee that all duplicated writes will delete immediately, that's why CAFTL scans flash memory and reduce the redundant data out-of-line.

Figures 2 and 3 shows exactly the improvement on eliminating duplicate writes with 24,2 % and the extended flash memory with 31,2 %. Here offline means that duplications are eliminated offline, no-sampling refers to the purely CAFTL with a sampling unit of 128 KB [14].

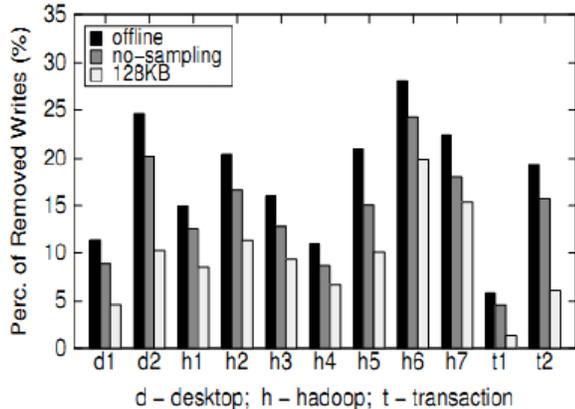


Fig. 2. Percentage of removed duplicate writes.[14]

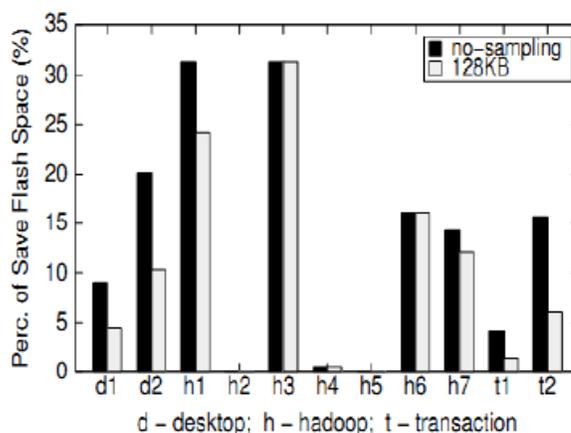


Fig. 3. Percentage of extended flash space.[14]

After these experiments engineers have seen that offline is the optimal case, but purely CAFTL is able to eliminate the duplicate writes in a significant way. Also the analog clearing is more complex on SSDs. [4] examines the garbage problem of data in flash, DRAM, SRAM and EEPROM and the “cold boot” attacks.

The simplest method is that the voltage level in the floating gate of an erase flash cell may vary, depending on the value it has before the erase command. The quantity of digital garbage may be very big. The SSDs which are tested contain 6-25 % more physical storage than they show on their logic capacity[8]. Figure 4 shows the garbage on the SSDs. There are created 1000 small files on the SSDs, the driver is dismantled and analyses. For some of this files SSD contains up to 16 old copies. This copies were created during garbage collection and un directly updates.

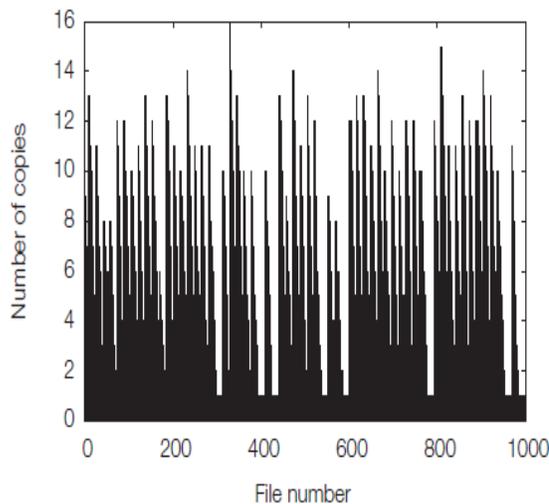


Fig. 4. Multiple copies creating from FTL, duplicating files up to 16 times.[8]

The differences between SSDs and hard drives described in upper sections will create a disconnect between what a user expects and how drive behaves. Someone who has an SSD may use a clearing technique which is characteristic for hard

drive, in way to make the data inaccessible, but the data will stay on the drive and they can be extracted with sophisticated methods.

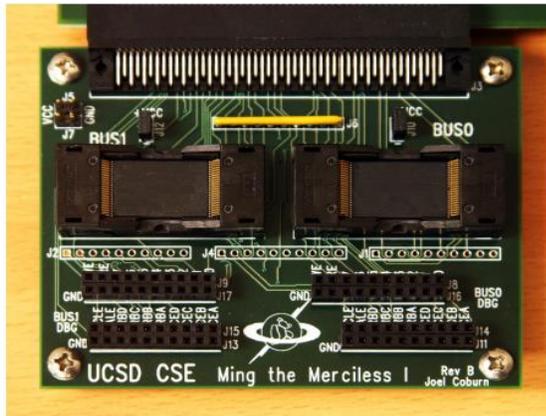


Fig. 5. FPGA based on flash testing hardware.[7][8]

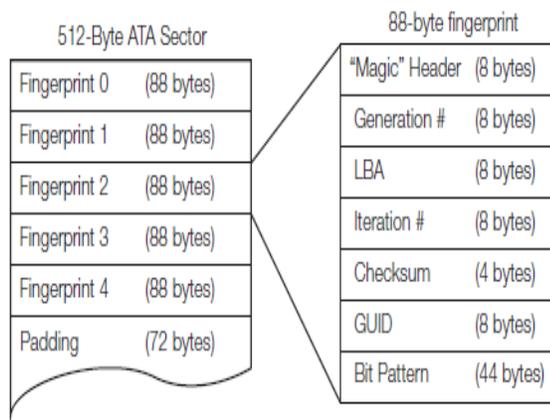


Fig. 6. Fingerprint structure. The easily-identified fingerprint simplifies the task of identifying and reconstructing remnant data.[8]

The engineers of [8] had verified the second level which is digital clearing by using the lowest level of digital interface: the pins of individual flash chips. To verify this operation, they wrote an identifiable data model called *fingerprints* and then they applied the clearing techniques under test. The fingerprint makes possible the identification of digital garbage on the chips. It also includes a sequence number that is unique at entire fingerprints. The figure 6 shows the fingerprint structure. According to the fig. 6 every fingerprint is 88 byte long repeats five times in a 512 byte ATA sector. Another method described in the article is overwriting every logic block address on the drive. This is the main method for many disc deletions. The different bits aim to change a lot of physical bits as possible on the drive, make it harder to recover the data with analog ways.

TABLE I. THE SOFTWARE OF OVERWRITING WHOLE DISC. THE NUMBER OF EACH COLUMN SHOWS THE NUMBER OF THE STEPS FOR DELETING THE DATA FROM THE DRIVE.[8]

SSD	Seq. 20 Pass		Rand. 20 Pass	
Init:	Seq.	Rand.	Seq.	Rand.
A	>20	N/A*	N/A*	N/A*
B	1	N/A*	N/A*	N/A*
C	2	2	2	2
D	2	2	N/A*	N/A*
F	2	121 hr.*	121 hr.*	121 hr.*
J	2	70 hr.*	70 hr.*	70 hr.*
K	2	140 hr.*	140 hr.*	140 hr.*
L	2	58 hr.*	58 hr.*	58 hr.*

*Insufficient drives to perform test

* Test took too long to perform, time for single pass indicated.

As it looked, these bits are important for SSDs. According to the experiments that the engineers have made, they have seen that some SSDs compress their data before storing them, they will write fewer data on flash. They suggest that for maximum effectiveness, SSD overwriting procedures should use random data. One of the drivers tested from them showed that it used compression and encrypted the data, but they can't verify the erasing process.[8]

They have tested eight drivers which don't use encryption and table 1 shows the results of these tests. The numbers show how many generations of data were needed to delete the drive. For some drives, they realized that random writes were too slow so they didn't do the test to these drives. In some cases they overwrite the drive twice which was enough to erase the disc, no matter how was its previous state. But they found exceptions on drive A, because some of the data weren't erased totally.

They compared this technique with the first one and find out that this is more reliable, but it's not totally secure, because of the type of the drivers they use.

Another method that [8] evaluate for erasing data was degaussing. Degaussing is the eliminating process or reducing the undesired magnetic fields [10]. Degaussing is fast and effective because it removes the low level formatting and damages the drive motor. They don't expect this method will work including the architecture of SSDs. And the experiment made in [6] shows that after degaussing, nothing happened to the data, so this method fails on SSDs.

III. CONCLUSION

This article was focused on the proposed techniques from different engineers to sanitize data from SSDs. These techniques were based on hard drives, because they were successfully in there. From the results, the first method had a half success (50:50), the second method was more successful except some fails it had and the third method was worthless, because it doesn't guarantee anything. So the problem of sanitizing data from SSDs actually is very present nowadays and engineers are trying to find ways to do this more reliable. From [6] an efficient way is to combine the techniques to each other for a safe and verifiable sanitizing data from SSD.

REFERENCES

- [1] J. Lee, J. Heo, Y. Cho, J. Hong, and S. Y. Shin. Secure deletion for nand flash file system. In SAC '08: Proceedings of the 2008 ACM symposium on Applied computing, pages 1710–1714, New York, NY, USA, 2008. ACM.
- [2] A. Birrell, M. Isard, C. Thacker, and T. Wobber. A design for high-performance flash disks. Technical Report MSR-TR-2005-176, Microsoft Research, December 2005.
- [3] E. Gal, S. Toledo. Algorithms and data structures for flash memories. ACM Comput. Surv., 37(2):138–163, 2005.
- [4] P. Gutmann. Secure deletion of data from magnetic and solid-state memory. In SSYM'96: Proceedings of the 6th conference on USENIX Security Symposium, Focusing on Applications of Cryptography, pages 8–8, Berkeley, CA, USA, 1996. USENIX Association. https://www.cs.auckland.ac.nz/~pgut001/pubs/secure_del.html
- [5] USB Implementers Forum. Universal Serial Bus Mass Storage Class Specification Overview, September 2008.
- [6] S. Swanson and M. Wei. Safe: Fast, verifiable, sanitization for SSDs. <http://nvsl.ucsd.edu/sanitize/>, October 2010.
- [7] L. M. Grupp, A. M. Caulfield, J. Coburn, S. Swanson, E. Yaakobi, P. H. Siegel, and J. K. Wolf. Characterizing flash memory: Anomalies, observations and applications. In MICRO'09: Proceedings of ..., New York, NY, USA, 2009. ACM, IEEE.
- [8] M Wei, L. Grupp, F. Spada, and S. Swanson. Reliably Erasing Data From Flash-Based Solid State Drives. In FAST' 2011.
- [9] Royal Canadian Mounted Police. G2-003, Hard Drive Secure Information Removal and Destruction Guidelines.
- [10] <http://en.wikipedia.org/wiki/Degaussing>
- [11] "Solid-state drive", 2012 http://en.wikipedia.org/wiki/Solid-state_drive.
- [12] David Bartizal, Thomas Northfield, "Solid State Drive Performance", www.csee.umbc.edu/~squire/images/ssd2.pdf
- [13] Bo Mao, Hong Jiang, "HPDA: A hybrid parity-based disk array for enhanced performance and reliability", Parallel & Distributed Processing, 2010 IEEE International Symposium
- [14] F. Chen, T. Luo, "CAFTL: A content-aware flash translation layer enhancing the lifespan of flash based solid state drives", 9th USENIX Conference on file and storage technologies, 2011, www.cse.ohio-state.edu/~fchen/paper/papers/fast11.pdf

Cost-Effective Smart Metering System for the Power Consumption Analysis of Household

Michal Kovalčík, Peter Fecíľak, František Jakab
Department of Computers and Informatics
Technical University of Kosice
Kosice, Slovakia

Jozef Dudiak, Michal Kolcun
Department of Electric Power Engineering
Technical University of Kosice
Kosice, Slovakia

Abstract—This paper deals with design, calibration, experimental implementation and validation of cost-effective smart metering system. Goal was to analyse power consumption of the household with the immediate availability of the measured information utilizing modern networked and mobile technologies. Research paper outlines essential principles of the measurement process, document theoretical and practical aspects which are important for the construction of such smart meter and finally, results from the experimental implementation has been evaluated and validated.

Keywords— *electrical voltage; electrical current; active power; measurement principles; intelligent metering system; smart meter; Atmel*

I. INTRODUCTION

Microcontroller based embedded control systems and intelligent sensors are nowadays commonly used in smart houses supporting the way we live, work and play. In recent years there was a lot of effort invested in making the systems smart and thus helping to reduce the amount of energy, keeping an eye on power consumption, saving money and sparing the environment. Smart technologies, mobile platforms and e-commerce systems are also changing the way how the energy consumption information is delivered. Modern technologies allow us to provide on-demand value added service like the consumption statistics or abnormality notifications for the behaviour of the powered devices and remote control.

One of the goals of smart home technologies with the immediate availability of power consumption information is to motivate consumers to adjust their behaviour in order to lower their energy consumption and energy costs. Effective usage of the power resources contributes to environmental protection, being one ingredient for the success of the energy transmission. The benefits provided by smart meters can even be maximized, for example, by combining the energy meters with home control systems such as smart homes, with which lighting and household appliances, among other things, can be automatically controlled.

The electronic energy meter which is based on digital micro technology that doesn't use moving parts is also known as static energy meter. In electronic energy meter the accurate functioning is controlled by a specially designed integration circuit. Application specified integration circuit (also known as ASIC) is constructed only for specific applications using embedded system technology. Similar application specified

integration circuits are now used in washing machines, air conditioners, automobiles or digital cameras.

The first part of the paper is about theoretical aspects of measuring the electricity by means of the newest intelligent meters. The second part of the article is about real application of constructed device that is installed in household for testing and measuring the behavior of electricity consumption.

II. MEASUREMENT OF ELECTRICAL PARAMETERS

Measurement is the process of getting the knowledge of outside world, where we try to get the desired values. Success is defined mainly with measuring accessibility, environment and the precision of measuring devices. The role of electrical measurement is to find out the values of defined electrical quantities as precisely as possible. During measurement of these values is not possible to interfere with measurement object and the measurement faults would be in the context of equipment quality. Measurement faults should be caused by many factors but not all of them we can eliminate or mitigate their effect. The aim of this chapter is to point out the measurement basis of the electrical quantities such as voltage, current and power in metering system.

A. Measurement and evaluation of electric voltage

Voltage is a physical quantity expressing the difference of electric potential between two points and represents the energy that is required for transferring the electrical charge between these two points within a certain electric field. The measurement unit of voltage is the Volt [V], which belongs to the derived units of SI.

Voltage measuring in smart meters is primarily provided by the integration circuits which consist of signal sensors and analogue-digital converter. Voltage measurement accuracy is affected by many factors such as temperature and also by the fluctuations of input power supply.

Electronic meter records the value of the voltage profile in the 10 minutes power quality. The overall 10 minutes profile consists of a 600-second values that are recorded every second into the computer memory system of electric meter. Subsequently the arithmetic average of these values is submitted according to defined formula (1).

$$U_{1/10min} = \frac{1}{600} \sum_{i=1}^{600} U_i \quad (1)$$

Each value of voltage that is entered into the computer

memory is calculated by equation (3). Electric meter reads 32 values every 20 millisecond, from which is calculated the mean voltage per second.

$$\Delta U_{1s}^2 = \frac{1}{n} * \sum_{i=1}^n U_i^2 \quad (2)$$

After modification:

$$\Delta U_{1s} = \sqrt{\frac{1}{n} * \sum_{i=1}^n U_i^2} \quad (3)$$

The recording of voltage in the meter works on a similar principle as an electronic voltmeter. The signal of sensor is placed across the voltage divider, which records the change of the output voltage signal on the resistance R.

B. Measurement and evaluation of electric current

Electric current is a physical quantity that determines the size of the charge that flows around a given point in an electrical circuit. The basic unit of current is Ampere [A], which is the basic unit of the SI (1A=1C/s). The resistance of direct current circuits to the current flow is constant so that the current in the circuit is related to the ratio of the voltage and the resistance according to Ohm's law.

Measurement of current in the smart meters is carried out by means of a current transformer or with hall sensors. On the output of it is connected a simple constant resistor, which is used for measuring the voltage of the analogue signal. A recording hysteresis of voltage is typically ± 5 V, depending on the resistance. The output voltage signal is fed to the integrator, where the A/D converter converts these voltage pulses to the digital value of the current.

Electronic meter records the value of the current in the 10 minutes profile of power quality. The overall 10-minutes profile consists of a 600-second values that are recorded every second into the computer memory system of the electric meter. Subsequently the arithmetic average of these values is submitted according to defined formula (4)(1).

$$I_{1/10min} = \frac{1}{600} \sum_{i=1}^{600} I_i \quad (4)$$

Each value of current that is entered into the computer memory is calculated by equation (5). Electric meter reads 32 values every 20 millisecond, from which is calculated the mean current per second.

$$\Delta I_{1s}^2 = \frac{1}{n} * \sum_{i=1}^n \left(\frac{U_1}{R_1}\right)^2 \quad (5)$$

After modification:

$$\Delta I_{1s} = \frac{1}{n} * \sum_{i=1}^n I_i^2 \quad (6)$$

After modification:

$$\Delta I_{1s} = \sqrt{\frac{1}{n} * \sum_{i=1}^n I_i^2} \quad (7)$$

Recording the current in smart meter provides current transducer, which consist of current transformer that ensures high linearity over a wide range of currents with ability to measure the direct currents. Meter evaluates the measured data and stores them in to the special registers according to OBIS codes.

C. Measurement and evaluation of active power

Active power can be defined as electric power, which converts on the load on work or other form of energy.

The power is defined as work over the time so:

$$p = \frac{dW}{dt} = u * i \quad (8)$$

The main unit of power is the watt (W), this unit is the SI derived unit (1W = 1J/s). Building on the equation (1) and (4), the mean value of two multiple signals is an essential part of the active power measurements. The mean value of active power in electricity meter we can calculate according (9).

$$P = \frac{1}{T} * \int_0^T p(t)dt \quad (9)$$

Substituting the equation (8) we get equation:

$$P = \frac{1}{T} * \int_0^T u(t) * i(t)dt \quad (10)$$

The electric voltage and current are recorded in one second intervals during every 15 minutes. The 15 minute value of the average power is calculated as the multiple of instantaneous values of voltages and currents at one-second intervals. On Fig. 1 is a load profile consisting of a multiple one second values of active power.

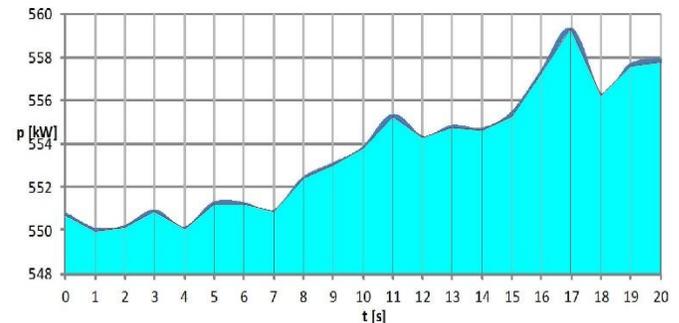


Fig. 1. Recorded second values of active power

The amount of consumed active power is recorded in electricity meter in 15 minute intervals. That active power can be calculated as area under the load curve, thus:

$$P = \frac{1}{900} * \int_{i=1}^{900} u_i * i_i \quad (11)$$

D. Sensors – General Overview

The sensor converts the information from the physical area of measured quantity into another physical area, usually an electrical signal or an electrical parameter. There are of course number of measuring devices with other than electric outputs (needle indicator, level of liquid). Currently sensors are practically used in all types of industrial products and systems. World market for sensors is still significantly growing. In Europe there are markets for more than hundred thousand different types of sensors. This number does not illustrate only the widespread use of sensors, but the fact that the choice of sensor for specific use is not a simple task. The main reason for rising up the interest for sensors is miniaturization of these devices. For this trend is responsible the continuing progress in the development of technologies compatible with information technologies. Currently there are available sensors based on silicon or similar technology for almost every value and there is also further scope for development in this area. Devices that convert information from one (physical) area to another (usually into electric field) are named converters (transducers), sensors or gauges.

The imperfections of sensors are usually written in the datasheet of the device producer. These specifications inform the users about the deviations of the sensor from the ideal conditions. Every sensor must be fully specified with respect to his further operation.

The main characteristics of sensors we can define as: sensitivity, linearity and hysteresis, resolution, maximum permissible error, zero adjustment, noise, response time, rating of frequency response.

From the energetic point of view the sensor can be divided into two groups: direct and modulating (active and passive). The distinguishing feature is the need for auxiliary energy source. Direct sensors do not require an additional power source. Consequently, the sensor draws power from the measured object, what can cause the loss of information about the original state of the measured object. Indirect or modulating sensors use an additional power source that modulates the measured object. Output energy of sensor comes mostly from auxiliary power and only a part of energy is taken from the measured object. Modulating sensors are due to not burdening the object of measurement more accurate than direct sensors.

E. Sensors of electrical voltage and current

These sensors are suitable for measuring alternating and direct currents. The most common are clamp probes consisting of drop down jaws that encircle the measured wire. And so we can measure currents without galvanic coupling. Measurement of voltage must be performed by means of test leads, respectively spikes. On the output of measurement probe is an analogue value of current, which is usually transformed to digital with usage of digital multimeter, respectively with another A/D converter.

The certain restrictions of measuring probes can be low sensitivity for small currents so the non-contact measurement of current is not appropriate for it. For small currents it is possible to use measurement direct on the contacts. [8]

1) Voltage sensors

The most common probes used for voltage measurement are passive voltage probes with damping (divider 1:10). They reduce measured voltage and increase its total output resistance during the sensing of it. When using passive probes we need to take into account the divider ration (measured value of 0.5 with converting 1:10 is 5 volts). With the use of oscilloscope mostly is set the automatic recalculation regarding the type of probe. For high-frequency measurements the oscilloscope input capacitance and parasitic capacitance of the cable together with divider resistors form a low-pass filter suppressing high-frequency parts of the signal. The impact of undesirable capacity can be compensated by adding a parallel capacity.

Transmission of measured voltage to the oscilloscope or analyser through a probe should be frequency-independent. Practical compensation is performed by using calibrator – rectangular course with different levels and frequencies. In addition to the classical low frequency compensation also is used additional high – frequency compensation using a broadband current probes.

2) Differential voltage sensors

Differential voltage probes are designed for explaining the basic principles of electric energy. By means of these probes for measuring voltages in low voltage alternating and direct currents with range of ± 6 Volts is system ideal for use in battery and lamp circuits. It is also used in combination with a current probe to explore Ohm laws and phase relationship between reactive power and much more. The main difference with voltage probes is fact that neither terminal is connected to the ground.

Differential voltage probe measures the difference of potential between terminals V+ (usually red) and V- (usually black). Voltage probes have differential inputs it means that voltage is measured with respect to the black terminal and not to a ground circuit. It allows the measurement direct on the circuit elements without limitations of general ground. Voltage probes can be used for measurement negative potential the same as for positive potential. This is considered as a major advantage with the usage of interface from 0 to 5 Volts.

Voltage probes are intended to use as the voltmeter. They should be placed across the circuit elements. Differential input range is from - 6 to +6 volts. Surge protection is provided so that the slightly higher voltage does not damage the voltage sensor. These probes are not designed for measurement of high voltage.

3) Current sensors

Current probes allow the measurement of current curves by means of oscilloscope with a bandwidth of approximately 100MHz. Mostly using the principle of Hall probe at low frequencies and current transformer at high frequency, that are designed for wider range of frequency and smaller values of currents. Hall probe offers a wide frequency response, thus measure with high precision of direct and alternate currents and complex courses too. Hall probes are usually made as compact clip-on probe with a nominal current from 5 mA to

30 A in a peak with an accuracy of $\pm 1\%$ excluding the effect of interference. These current probes provide a voltage output directly proportional to the measured current.

4) Other types of sensors

Passive probe with low input impedance (50Ω) is designed especially for measurements in high frequency circuits, where it is required to adjust the impedance and high limiting frequency. Active probes with unbalanced inputs contain a preamplifier and are designed for measuring very small signals, for which the oscilloscope sensitivity or high impedance of source signal is insufficient. Input capacitance is usually around 1pF . They are very sensitive to overvoltage, including the static electricity. Bandwidth is usually up to the GHz.

Active differential probes are designed for measuring ungrounded sources of signal. They contain a preamp and conversion from the differential input to unbalanced output suitable for connection to an oscilloscope. They are usually build as the broadband high-frequency probes that are able to process signals up to about 10GHz .

High voltage differential probe with galvanic isolation of input provide secure, precise and broadband measurements of high-voltage devices and systems up to the voltage about kV.

III. MEASURING THE ELECTRICAL PARAMETERS USING THE ATMEL MICROCONTROLLER

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Abbreviations and Acronyms

The previous chapter described essential principles of measurement of the electrical parameters such as voltage, current, active power. Next chapter will explain and illustrate the whole process of measuring and storing of these values in internal memory of microprocessor. The implementation of such smart metering system will consist of the following basic elements:

- Sensors/ probes,
- electronics,
- software.

Probes are an inseparable part of electronics that ensure a proper processing of analogue signals for inputs of analogue-digital converter. Goal of the program code (software resource) is to process the input data, calculate arithmetic average and effective values and store them into non-volatile memory so the data could be accessed for the analytical purposes.

B. Block diagram of the smart metering system

Fig. 2 shows internal structure within the block diagram. System is consisted of sensors used for voltage and current measurement, electronics for processing of measured signals, analogue-digital converter, microprocessor and its additional modules.

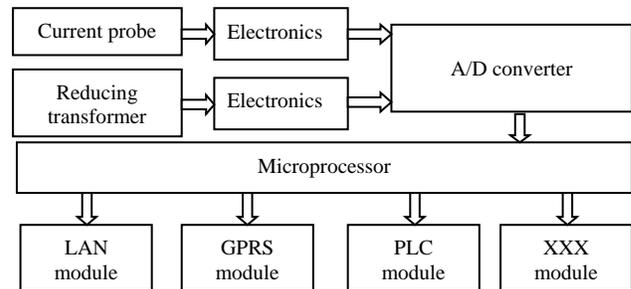


Fig. 2. Block diagram of one phase measuring system

In order to ensure the accuracy of measured parameters, the implementation of the whole system, as depicted in Fig. 2, is multiplied by the number of measured phases. Based on the specific requirements of the environment where it is important to measure power parameters, the system could be built as one phase system or with the equipment to measure more phases. Three phase meter will have the same components for every phase and one microprocessor with modules. On Fig. 2 is displayed the block diagram just for one phase metering system.

C. Schematic diagram of electronics for IMS

Schematic diagram of electronics consists of circuits for electric voltage metering and circuits for electric currents metering. Some of the intelligent electronic metering systems have a disconnection circuits which are used in order to disconnect consumers remotely without direct intervention of technician. The other functionality is limitation of current in electricity meter. This is used for limitation of the consumer load. On the Fig. 3 there is the scheme of electronics for electric voltage metering on one phase. On the left part of the figure is the voltage transformer. Its role is to transform the voltage from 230 V to the level that is acceptable for processing at the microcontroller. It is appropriate to use the transformer with 230V AC voltage on the inputs and 12V AC output.

The ratio of this transformer is expressed as:

$$k = \frac{U_{out}}{U_{in}} \quad (12)$$

In this equation k is transformer ratio, U_{out} is output voltage of the voltage transformer, U_{in} is input voltage of voltage transformer. Transformer ratio k is any value from interval between $(0;\infty)$, but it is still open interval. When the ratio is 1 it means that it is separating transformer and when the ratio is more than 1 it means that transformer so increasing. In this case we will have decreasing transformer with transformer ratio $k=0,0522$.

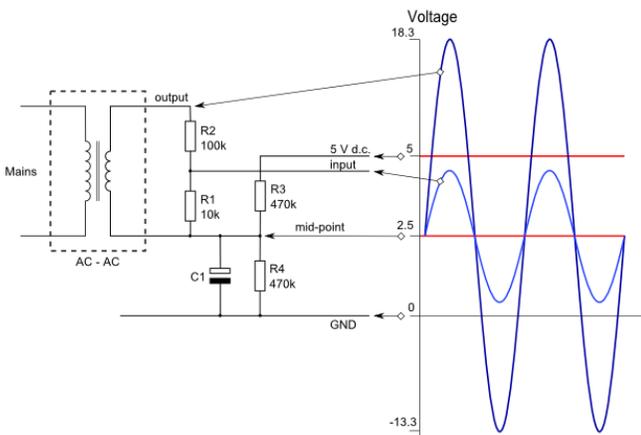


Fig. 3. Scheme of electronics for voltage metering on phase 1 [6]

The output alternating current from transformer is placed to voltage divider which consists of resistor R_1 and R_2 . The role of these transformers is to decrease the value of voltage according to their ratio from the output of transformer to the input of analogue-digital converter. In most cases the AD converter operates with values up to the 5V direct current as it was described in previous chapter. On the transformer, there is voltage around $\pm 12V_{ef}$ what means that amplitude of voltage can achieve values according to equation:

$$U_{max} = U_{ef} \times \sqrt{2} = 12 \times 1,41 = 16,97V \quad (13)$$

When we calculate the absolute value of both amplitudes, we will have the value of voltage 33,94V. Measured voltage on the output of divider we can calculate from equation:

$$U_{R1+R2} = U_{R1} + U_{R2} \quad (14)$$

After modification:

$$U_{R1} = U_{R1+R2} - U_{R2} \quad (15)$$

According to Ohm's law it is possible to express current I by voltage U and resistance R , which is known in our circuit. Subsequently from that we can calculate U_{R1} . The maximum value of U_{R1} is 3,04V. The most important is determination of decreasing ratio of divider (labelled as p):

$$p = \frac{R_{in}}{R_{out}} = \frac{R_1 + R_2}{R_1} \quad (16)$$

Input resistance of voltage divider consists of the sum of two resistors that are connected in series. As the output resistor of voltage divider is considered that one which will be connected to the input of AD converter. The ratio of voltage divider after substitution of resistors values is 11:1.

The next important part is the offset of mean value of AC voltage to the middle of the operating voltage in AD converter. It is ensured by R_3 and R_4 according to scheme on Fig. 3. These two resistors are of the same value and they are connected in series where the end of first resistor is connected to the GND of AD converter. The end of second resistor is connected to the power source of AD converter. The point where are connected both resistors will have the same

potential to the ground as to the power source terminal.

The value between them will be exactly 2,5V. This point will be connected to the output of transformer and resistor of voltage divider with lower value of resistance to ensure the offset of alternate current which will fluctuate around the value of 2,5V. The last part of circuit is capacitor C_1 which ensures low impedance path between AC current and ground of DC power source. Fig. 3 shows the real course of voltage on the resistance divider inputs and the real course of voltage on the output of voltage divider.

The scheme for measurement of currents will be slightly different from the voltage measurement. Every phase is to be measured separately, the same way as was measured voltage.

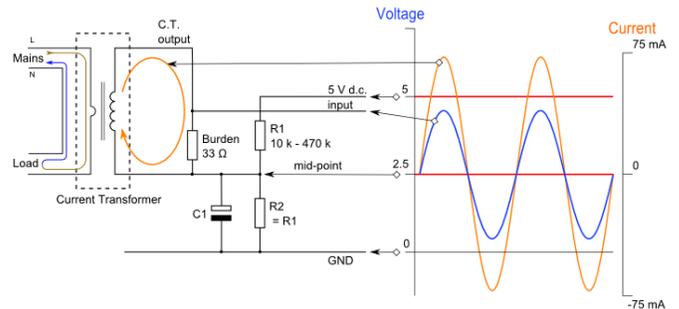


Fig. 4. Scheme of electronics for current metering on phase 1 [7]

Fig. 4 shows the initial scheme of circuit for measurement of current with components for evaluation of the current that flows via one phase of metering system. On the left side of the figure there is current transformer. It is probe for metering currents according to chapter II. This probe operates on the principle of transformation of electric voltage from primary winding. Primary winding consist of one wire which cross through interior of ferrite core of that transformer. Around the wire there is induced electromagnetic field which is transmitted over the ferrite core of the secondary winding of this transformer. Every current transformer do have its own ratio that is expressed in amperes. For this case the current transformer with factory mark: SCT 013-030 was used. Its ratio is 100A to 50mA where 100A current which flows through the primary winding is on the secondary winding just 50mAs. This physical law is true only when the load is constant.

This is the base for deriving the constant for calculating the real value on the proposed intelligent metering system. The recommended load is on the scheme shown under the label Burden.

Next part of the scheme is once again the pair of resistors that are connected in the series. The role of them is to offset the middle of oscillation amplitude to the value of 2,5V.

Basically, we are dealing with voltage measurement on the load Burden, where it is possible to calculate the current flowing over the resistor following Ohm's law and yet determine the current flowing through the measured wire thanks to ratio of current transformer.

Capacitor C_1 have the same function as it was written in description of scheme of voltage measurement. On the right side is shown the real course of voltage (blue) on the resistor –

load (Burden) and the real course of current (red).

Microprocessor that is used for this implementation do have integrated 10 bits analogue – digital converter on the outputs. In this case we don't need additional support circuits because the output of prepared circuits will be connected straight to the analogue inputs of microprocessor. In one phase metering system we will use just two analogue inputs on the microprocessor. In three phase metering system we will use six analogue inputs on the microprocessor. For every phase two inputs. One input is just for metering current and the second one is for voltage metering.

D. Software of intelligent metering system

Intelligent metering system which is build on the base of microprocessor shouldn't be functional without the software. Software is an inseparable and very important part of the system. It is used for reading of values from inputs, calculate the effective values of voltage and current and power as well. For experimental testing we make the prototype solution on the prototyping Arduino board. It is standard Atmel 8-bit family microprocessor. Arduino IDE has been used as the code development environment.

Structure of program consist of these main parts:

- importing the header files,
- declaration and initialization of variables,
- setting the initial parameters,
- the main loop with sampling signals,
- calculation of the one second effective value,
- storing of calculated values in internal registers.

1) Importing the header files

Core libraries that are used to operate the Atmel microprocessor are automatically imported in ARDUINO IDE environment. This depends on the fact which type of ARDUINO prototyping board we will use for the code compilation. Excluding this compilation we need to import the following libraries:

```
#include <SPI.h>,  
#include <Ethernet.h>,
```

SPI library is for communication via SPI bus, where are connected support modules for communication over the Ethernet. Serial Peripheral Interface (SPI) is a synchronous serial data protocol used by microcontrollers for communicating with one or more peripheral devices quickly over short distances. It can also be used for communication between two microcontrollers[11].

Ethernet library allows an Arduino board with the Arduino Ethernet Shield to connect to the internet. It can serve as either a server accepting incoming connections or a client making outgoing ones. The library supports up to four concurrent connection (incoming or outgoing or a combination).

Arduino communicates with the shield using the SPI bus. This is on digital pins 11, 12, and 13 on the Uno and pins 50,

51, and 52 on the Mega. On both boards, pin 10 is used as SS. On the Mega, the hardware SS pin, 53, is not used to select the W5100 but it must be kept as an output or the SPI interface won't work [12].

2) Declaration and initialization of variables

It is important to initialize and declare the following variables:

set the MAC address on Ethernet Shield

```
byte mac[] = {0xDE, 0xAD, 0xBE, 0xEF, 0xFE, 0xED};
```

set the IP address, that will be configured on experimental IMS.

```
IPAddress ip(192,168,1,201);
```

setting the server that will be using for receiving measured data via Ethernet

```
char server[] = "michal.kovalcik.s.cnl.sk"
```

variable of Ethernet client type

```
EthernetClient client;
```

variable for resetting the memory of last connection

```
boolean lastConnected = false
```

the variable for counting the number of samples per unit of time

```
int i = 0;
```

variables that will store data from A/D converter

```
int ii1 = 0, u = 0;
```

variable for partial calculation of measured data

```
float uuu=0;
```

```
float iii1=0;
```

variable for final calculation of measured data

```
float napatie = 0
```

```
float prud1 = 0;
```

calibration constant for conversion measured values from microprocessor to the real effective values

```
float ki = 0,053167;
```

```
float ku = 1,085731;
```

definition the time interval per one second

```
long interval = 1000000;
```

variable for counting the time

```
long previousMillis = 0;
```

3) Setting the initial parameters and creation the necessary instance in the loop setup()

Initialization the Ethernet library and the network settings

```
Ethernet.begin(mac);
```

Reading the time stamp

```
previousMillis = micros();
```

```
4) The main loop with signal sampling loop()
```

Reading the time stamp before reading the first value of the sample.

```
unsigned long currentMillis = micros();
```

Reading the first sample from A/D converter and saving it to the served samples.

```
iii1 = analogRead(i1Pin);
```

```
uuu = analogRead(uPin);
```

Summarization of reading values according the equation (3) and (7).

```
iii1 = iii1 + sq((iii1-511.5)*ki);
```

```
uuu = uuu + sq((uuu-511.5)*ku);
```

Increment of variable for calculation number of samples per time unit.

```
i++;
```

The beginning of term after that time set by variable was achieved.

```
if(currentMillis - previousMillis > interval){
```

Calculation of mean value and the square root according to the equation (3) and (7).

```
napatie = sqrt(uuu/i);
```

```
prud1 = sqrt(iii1/i);
```

Resetting and set the variables for further use in the loop () and finishing the term.

```
uuu=0;
```

```
iii1=0;
```

```
i=0;
```

```
previousMillis = currentMillis; }
```

Now we have values of measured electrical parameters per one second that are in registers (variables voltage and current).

Subsequently we can use these values for calculating the active power or exporting the parameters of voltage and current.

5) Results from designed IMS

Electrical parameters that were read, calculated and described in previous chapters are exported to the remote places by means of existing networks. We can export values secondly to the storage system or we can use 15 minutes profile for exporting. For transferring of collected data we can choose any well-known technology such as PLC, GPRS, LAN or any other. It depends on the connected module to the designed IMS system.

If we do not convert the read values from A/D converter according to equations (3) and (7) we will have on the output of the microprocessor just sampling values in range from 0 to 1023.

We put these values of measurement of the load in to the graph On Fig. 5 where we have graph of real course of the measured current.

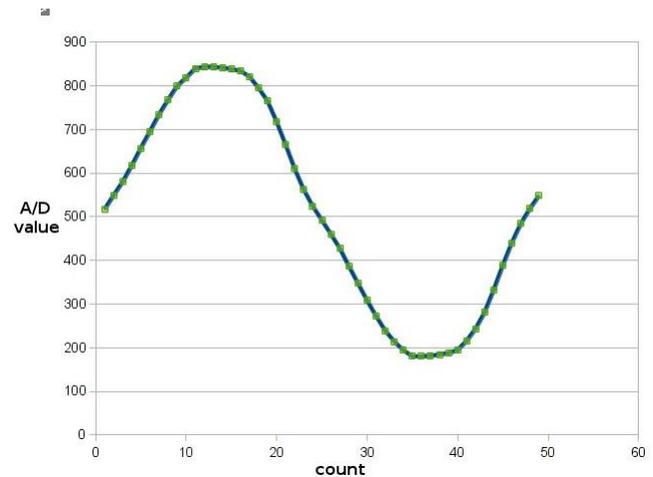


Fig. 5. The real course of the current

The minimal term is that electricity meter must read at least 32 samples per one period. From graph it is possible to see that designed IMS device measure about 50 samples per one period of time and satisfy this term in full range. After applying the equation we have results with export of currents and voltage every second. These values we put into the table and make graph as on the Fig. 6.

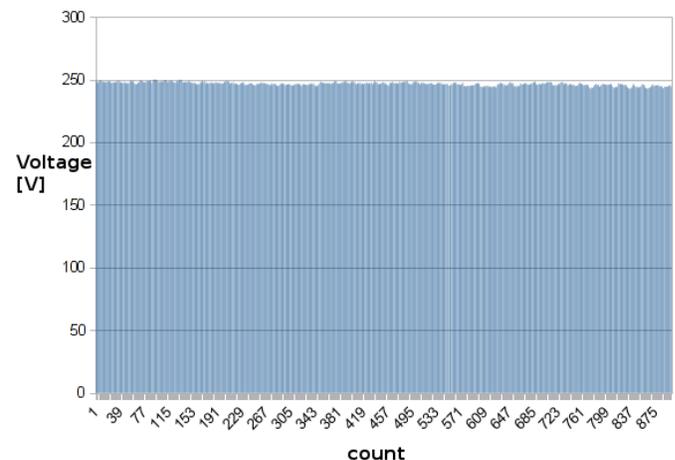


Fig. 6. The graph of measured voltage on one phase

On the x axis of this graph there is sequential number of measurement from 1 up to 900 what is number of samples per 15 minutes. y axis is the effective value of measured voltage in volts. After analyse we decided that the voltage isn't constant but fluctuates under the value 250 V. This is a little higher voltage as usual what is caused by closeness of distributional transformer which make the voltage level higher. It is still considered as normal value in the range of 230V ± 10%, so the quality of electricity is ensured. The fluctuation of voltage is caused by the variable character of load in the power system.

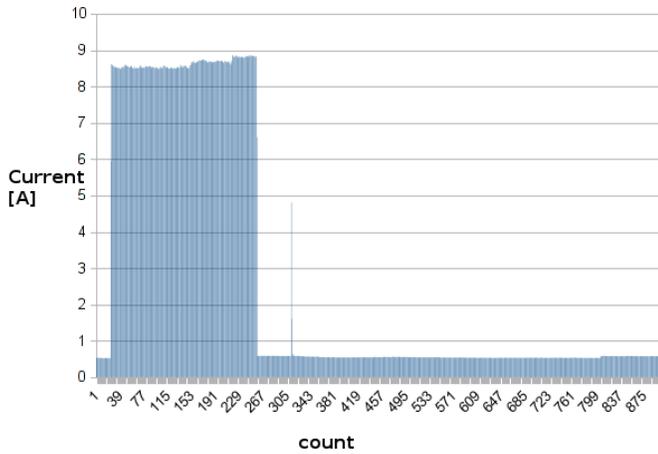


Fig. 7. The graph of measured current on one phase

The similar graph we can see on Fig. 7 which is constructed from values of current that were consumed by consumer equipment. The x-axis is the sequential number of measurement. We have 900 measurements and the y-axis is the effective value of current measured in amperes. It is possible to see that the fluctuation of current is more significant than voltage. While the voltage should be constant the current will be changing in dependence of what appliance is connected and consume the energy. There we can see that from the moment of measure 24 up to the 257 is the increase about 8 amperes. We expect that one appliance was connected to the power system. The current of 8 amperes equals to the consumption of electric kettle.

The next figure shows the graph of active power calculated on one phase. On Fig. 8 we can see that graph which is similar than currents on one phase. The axis of x is sequential number of measurement from 1 up to 900 what is number of samples per 15 minutes. The axis y shows values of active power. The main measure unit of power is Watt.

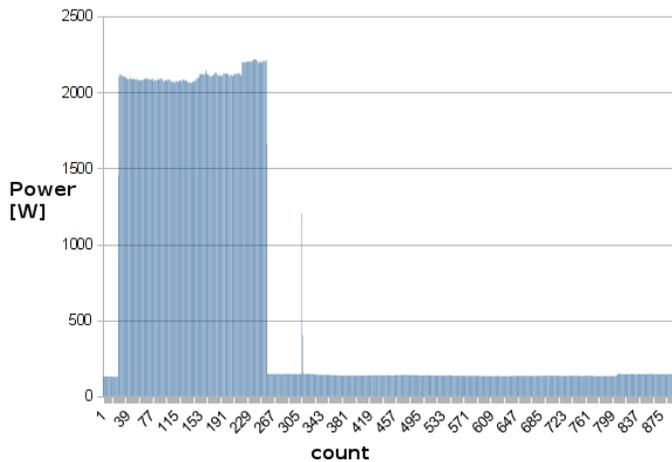


Fig. 8. The graph of measured active power on one phase

The last graph showed on Fig. 9 is the summary graph of exported 15 minutes values of power. The x-axis is sequential number of measurement and y-axis shows values of active power in Watt. There are only 10 values but every value characterize the duration of 15 minutes. The whole time of measurement showed there is two and half an hour.

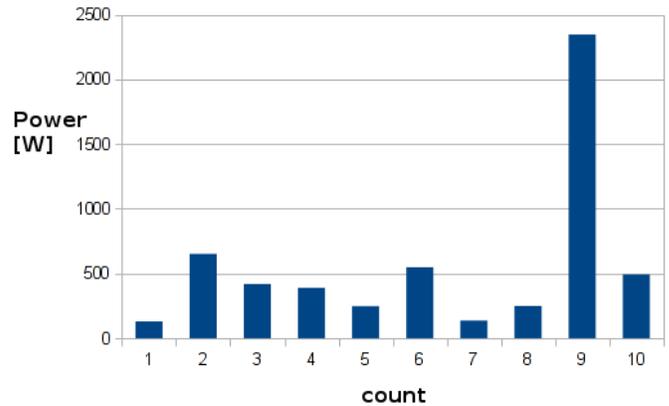


Fig. 9. The profile of one phase power in 15 minutes

IV. CALIBRATION AND TESTING OF THE IMS SYSTEM

In previous chapter was mentioned the calibration constant k_i and k_u . It is necessary to calculate these constants, after that to verify the calculation and store the result in the memory of microprocessor. This process is important when we want to have effective values calculated by IMS the same as real measured values. Constant k_i depend on the values according the equation (17) where we have ratio from the A/D converter, part for calculation of the current flowing via the resistor according the Ohms law. Next one is transformation ratio of current transformer and the number of threads on the primary winding. In this experimental implementation there was used the current transformer with maximal measured current of 100 Amps. To decrease the range and increase the accuracy we put six threads on the core of the probe.

$$k_i = \frac{5}{1023} \times \frac{1}{R_{burden}} \times \frac{100}{0,05} \times \frac{1}{6} \quad (17)$$

We can measure the value of R_{burden} with millimetre. In this case the nominal value of it is 33 Ohms but the real value is 33,6 Ohms. For further calculations we will use this measured value of resistor. The next unknown value is the ratio of current transformer or current probe. The verification of nominal values on current sensors we can make by means of measuring the currents on any circuit where we need to use two different types of measuring devices of current and our designed current probe. Our probe must be connected to the constant load. With measuring voltage on this load we can calculate the current that is flowing through load according to Ohms law. So we can determine the ratio of primary winding of current transformer to the value that we had calculated.

TABLE I. MEASURING THE RATIO OF CURRENT TRANSFORMER

Current oscilloscope	Current analogue	Average current	U_R	I_R	Ratio of transformer
0.96	1	0.98	0.112	0.00333	294
2	2.01	2.005	0.233	0.00693	289.1330472
3	3	3	0.348	0.01036	289.6551724
4.1	4.05	4.075	0.472	0.01405	290.0847457
5.03	5	5.015	0.584	0.01738	288.5342465
5.94	5.9	5.92	0.684	0.02036	290.8070175
8	8	8	0.937	0.02789	286.8729989
10	10	10	1.167	0.03473	287.9177377
12.1	12	12.05	1.4	0.04167	289.2

From the following table 1, where we have written values according to previous process we can calculate the ratio of current probe. Subsequently from the equation (18) we can determine the average ratio of transformers.

$$p_{tr} = \frac{1}{n} \times \sum_1^n \text{transformer ratio} \quad (18)$$

After calculation we have ratio of 289.578 where the nominal value is $(100/0,05) \times (1/6) = 333.33$. So we can decide that nominal value is different from our calculation for 15,1%. Thus the constant $k_i = 0,042123$.

Constant k_u was derived from the equation (19) where we have the ratio of A/D converter, the ratio of voltage divider according the equation (16) and the last part is proportion of nominal input and output values of voltage.

$$k_u = \frac{5}{1023} \times \frac{11}{1} \times \frac{230}{12} \quad (19)$$

The ratio of A/D converter is given and precise defined by producer. The ratio of voltage divider was measured and confirmed that it isn't different from the nominal value. We need to verify whether the ratio of input and output winding of transformer is correct. We connect two resistors (10 and 10 kOhms) that are connected to the series on the output of the transformer. We will measure by means of two voltmeters. One will be connected to the input and the second one will be connected to the output. The output voltage will be increasing continuously. We will make 10 measurements with the range from 0 to 250 Volts and the results we will write to the table. From the table we calculate the arithmetic average of the proportion – input and output voltage. The nominal value is 19,166. Experimentally verified value is 19,41. The nominal value is different from the verified about 1,25%. K_u we can quantify as the value 1,0424.

Due to different nominal values from measured and verified we need to make additional calibration. Calibration may be carried out on the basis of the connection of these elements:

- More measuring devices
 - multimeter,
 - oscilloscope with current probe,
 - lab pointer ampere meter.

- continuously adjustable source 0-250V AC, 0 – 20A,
- load,
- calibrated IMS.

Scheme of calibration circuit:

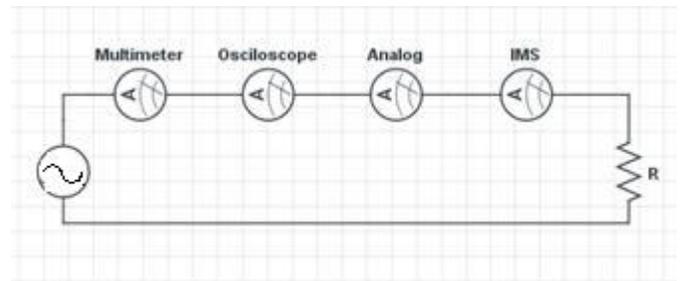


Fig. 10. Scheme of calibration circuit

The power source was designed to be continuously adjustable for the range 0 - 250V 0-20A, in this way both values can be adjusted. From the power source is in series directly connected multimeter and analog pointer ampere meter. Subsequently, from these devices, is directly connected only the load. Oscilloscope and IMS have their own probes connected contactless but for better understanding on the Fig. 10 they are shown as connection to the series.

Measurement was processed according the table2:

TABLE II. MEASUREMENT OF CALIBRATION CONSTANT K_i

S.N.	Current oscilloscope	Current multimeter	Current analogue	Current average	Current IMS
1	1.12	1.09	1.1	1.103	0.868
2	2.12	2.06	2.06	2.080	1.643
3	3.27	3.18	3.16	3.203	2.546
4	4.14	4.06	4.09	4.097	3.241
5	4.92	4.96	4.92	4.933	3.945
6	6.1	6.08	6.08	6.087	4.853
7	8	n/a	8	8.000	6.377
8	10	n/a	10.01	10.005	7.978
9	12	n/a	12.01	12.005	9.515
10	14.1	n/a	14.02	14.060	11.06
11	16	n/a	16	16.000	12.515

From measured values we can calculate coefficient k_{ix} for increasing or decreasing constant k_i .

According the equation (17) we can calculate the arithmetic average of the current ratio and measured on the IMS system.

$$k_{ix} = \frac{1}{n} \times \sum_1^n \frac{\text{current average}}{\text{current IMS}} \quad (20)$$

In this case is the value of coefficient k_{ix} 1.2621833. When we look back for the constant k_i which has been calculated from nominal value of the element, it has value of 0,042123.

From the table2 we can see that our IMS system measure a little smaller values as the lab equipment. We need to multiply the calibration constant k_i with coefficient k_{ix} to ensure the

new more precise constant $k_{in} = 0,0531669471459$.

Specification of the coefficient k_{ux} was realized by measuring input voltage on the transformer by means of two different lab metering devices and the calibrated IMS system. All measuring devices were connected parallel to the measured load and the results were written to the table 3:

TABLE III. MEASUREMENT OF CALIBRATION CONSTANT K_u

S.N.	Voltage oscilloscope	Voltage multimeter	Voltage average	Voltage IMS
1	10.5	10.047	10.2735	9.85
2	52	52.25	52.125	49.87
3	103	103.5	103.25	99.2
4	136	137	136.5	131
5	169.2	169.2	169.2	162.75
6	200.6	200.6	200.6	193
7	231.6	231.6	231.6	222.7
8	245	245	245	235

These values we put into the equation (21) and calculate the coefficient k_{ux} .

$$k_{ix} = \frac{1}{n} \times \sum_1^n \frac{\text{voltage average}}{\text{voltage on IMS}} \quad (21)$$

After calculation, the equation with measured values we have coefficient $k_{ix} = 1,041569$. Again we can see that measured value with IMS is lower than measured by lab metering systems so we need to increase value in coefficient k_{ux} . Subsequently we multiply the constant k_u with coefficient k_{ux} and we get constant $k_{un} = 1,0857315256$ which is more precise than the value before.

V. CONCLUSION

The goal of this paper was to share the information about construction of the smart metering system. We are at the beginning of difficult role with intelligent devices of the power system. The invented IMS system that is described in this paper is cheap and very simple equipment that can be used in every home for checking and analysing the power consumption. The social trend is saving money and guarding the consumption is one of the progressive tools to do it effectively.

The expectation of this project was to make a calibrated tool for metering of the power consumption in household. The results state that our aim was reached and the device is now installed in one household where it is monitoring the behaviour of family consumption. The highest advantage is that all measured data are stored in internal memory of the

computer what gives us the unique chance to make reports.

We can make reports based on hour, day or month base. Subsequently we can compare our consumption of electricity during the time periods. This detailed analysis is good for analysing the appliances that we use at home. Every appliance has different input power so we can make a diagram of appliances used at home and see which appliance is currently connected. This is very good tool for monitoring homes when we aren't at home yet.

The distribution system operators are currently at the beginning of the era of smart metering systems. It can take up to 10 years till everyone will have smart meter at home. This is very important challenge which is financially very difficult. The distribution system operators will send reports about electricity consumption to every consumer using the portal, but it is the sound of future. So this experimental tool can bypass the time from now up to the installing smart meter by distribution system operator.

ACKNOWLEDGMENT

Paper is the result of the Project implementation: University Science Park TECHNICOM for Innovation Applications Supported by Knowledge Technology, ITMS: 26220220182, supported by the Research & Development Operational Programme funded by the ERDF. We support research activities in Slovakia/This project is being co-financed by the European Union.

REFERENCES

- [1] J. Dudiak, M. Kolcun, Analýza komunikačných technológií pre inteligentné meracie systémy, Magazine EE, ed 2: Bratislava, 2014, pp. 42-44.
- [2] E. Moulin, Measuring reactive power in energy meters, Metering International, Issue 2: Tokai, 2002, pp. 52-54.
- [3] A. Brandolini, Power and energy measurements. Online at: <<http://www.eolss.net/sample-chapters/c05/e6-39a-04-05.pdf>>.
- [4] D. Mohankumar, Electronic Energy Meter or Electricity Meter. Online at: <<http://www.engineersgarage.com/contribution/electronic-energy-meter>>.
- [5] Hyper physics, Average Power. Online at: <<http://hyperphysics.phy-astr.gsu.edu/hbase/electric/powerac.html#c3>>.
- [6] Open energy monitoring voltage measurement. Online at: <<http://openenergymonitor.org/emon/buildingblocks/measuring-voltage-with-an-acac-power-adapter>>.
- [7] Open energy monitoring current measurement. Online at: <<http://openenergymonitor.org/emon/buildingblocks/ct-sensors-interface>>.
- [8] Arduino documentation, SPI library. Online at: <<http://arduino.cc/en/Reference/SPI>>.
- [9] Arduino documentation, Ethernet library. Online at: <<http://arduino.cc/en/Reference/Ethernet>>

Information Communication Technology Adoption in Higher Education Sector of Botswana: a Case of Botho University

Clifford Matsoga Lekopanye
Faculty of Computing
Botho University
Gaborone, Botswana

Alpheus Mogwe
Faculty of Computing
Botho University
Gaborone, Botswana

Abstract—Opportunities, benefits and achievements are emerging factors for institutions, lecturers, and learners from the increasing availability of Information Communication Technologies (ICT). These factors are relevant especially for new and growing higher educational institutions (HEI) whose survival depends on, among other factors, the use of ICT to develop new organizational models to enhance their internal and external communication relationship and produce quality graduates.

Given the relevance of the topic, the researchers studied positive impact of the adoption of ICT by higher educational institutions in an attempt to justify the use of ICT. In this paper, adoption refers to institutions migrating from traditional modes of paper based school management and student engagement to a computerized environment. This shift is hoped to enhance academic development and flexibility, increase level of student engagement, enhance cost-effectiveness, and create a sustainable environment through interactive learning resources.

Although the study was conducted at a single institution (i.e. Botho University), it restricts its focus exclusively to the educational motivations for institutions to adopt ICT. In order to ascertain the current state of knowledge, an extensive review, analysis, and synthesis of the collected data and literature have been undertaken. The authors conclude the paper by identifying and examining potential benefits and achievements of institutions in adopting ICT.

Keywords—ICT; Adoption; e-learning; Education

I. INTRODUCTION

The acronym ICT is an umbrella acronym that includes any communication device, including but not limited to: computer, telephone, television, radio and others. There are also services and software that are associated with the aforementioned devices such as Distance Learning Software Tools and Video Conferencing Applications. According to Unwin [1], “ICT can be a catalyst by providing tools which teachers use to improve teaching and by giving learners access to electronic media that make concepts clearer and more accessible”. Therefore it is evident that organizations will benefit from ICT through its adoption and be able to, as a result, reap benefits. According to Dasgupta [2], ICT adoption is “defined as the decision to accept, or invest in a technology”. In light of this the use of the terms acceptance and adoption are interchangeable in this article as supported by other literatures [3].

Botswana is currently experiencing the exponential growth of the ICT sector, yet it is still lagging behind in enjoying the ICT benefits compared to other countries dominated by ICT usage [13]. This has been attributed in part to financial constraints and the uneasiness of the Botswana society to adopt ICT, due to lack and or limited knowledge on it [13]. The literature revealed that ICT adoption is not an easy task in Botswana due to countless challenges which have hampered its growth and adoption in the education sector [13], consistent with other researches [4]. Therefore, it is important that the target that Botswana has set of having an informed and educated nation [5] can be reviewed as ICT also plays a key role in achieving such. Moreover, adoption of Information Communication Technology by institutions is one of the indicators that needs to be measured during the course of Vision 2016 implementation [13].

Closely related projects to this paper exists with their emphasis and scope being different. They have focused on identifying challenges of poor ICT distribution in Botswana’s Primary and or Secondary education levels, ICT training needs for teachers, challenges faced in implementing ICT in Botswana’s education system, or, ICT literacy in specific target groups in Botswana [6] [13] [15] [16]. Therefore, the adoption and acceptance part has been largely ignored, hence this research is a first of its kind conducted to measure ICT adoption and acceptance for educational purposes at tertiary level in Botswana with emphasis on Botho University.

II. RESEARCH OBJECTIVE

To present views of Botho University staff and students towards ICT as an eliciting may lead to and/or justifies its adoption or non-adoption.

III. LITERATURE REVIEW

A number of conducted studies have revealed that investing in ICT is beneficial for performance and productivity [8] [9] in diverse sectors of a country’s economy. Educational sector is one area which has benefited a lot from ICT and many countries that tend to do well in education have robust ICT education system [13] [14]. An ICT oriented education system provides varying benefits to both learners and educators alike, through impacting positively in improving quality, delivery and results [7] [13]. However, it is the challenges associated

with ICT which have hampered its adoption and acceptance at various organizations [10] [13].

Many countries across the world, especially developed countries, are doing everything possible to address the ICT adoption challenges at grass root level, with others introducing initiatives of giving laptops to student and teachers to enhance their ICT skills, introduction of ICT at early levels such as Kindergarten and or providing training to teaching personnel at all levels at low costs or for free [11] [13]. In Africa, and Botswana to be precise, this is just a dream[13], since the initiatives are difficult to implement due to diverse challenges such as financial deficiency, limited ICT trained personnel and electricity distribution amongst the so many

A. ICT in Botswana

The government of Botswana adopted ICT policy in 2004, known as Maitlamo [13] and it was also revised in 2007. The goals of this policy are to; “create enabling environment, Universal service and access to information and communications facilities, to make Botswana a Regional ICT Hub”. The country also have Vision 2016 which was approved in 1997, which articulates the role of ICT in an “Informed and Educated Nation” pillar [5]. However, although the government is concerned about ICT adoption, there are challenges that are encountered. According to Cummins [12], “There is also considerable difference in terms of urban and rural access to ICT services. Challenges include the relatively high cost of computers, absence of electricity in a number rural areas, and expensive Internet usage”. Many Universities and colleges in Botswana have computer laboratories, but the major challenge is the computer to student ratio, many students do not have or own computers [13].

Through the ICT policy [13] [14] various areas are being addressed and Botswana is channeling her efforts to address the areas identified in the policy. One such, is the education area where efforts to integrate ICT in education is being spearheaded. There is a national e-learning committee tasked with formulating and promotion of e-learning in Botswana [14]. This initiatives are tailored to meet the shortfalls identified in the findings and recommendations contained in the Botswana ICT policy [13]. The ministry of education is trying to encourage its partners to look at e-learning as one of the possible teaching modes in Botswana with government and other stakeholders working towards achieving that vision. The ICT policy, known as Maitlamo, has noted that the government is working on broadband connection to schools, refurbishment of computers then delivered to schools and training of teachers and administrators on e-learning. Botho University is one educational stakeholder who has put forth various blended learning technologies which among includes e-learning and other interactive learning applications.

IV. RESEARCH DESIGN AND METHODOLOGY

A. Research Design

This research was conducted at Botho University. Participants involved students, lecturers and administrative staff. The research utilized a qualitative approach and employed primary data collection through the questionnaire.

The questionnaire comprised of two sections namely the demographic details of participants and ICT Opinion questions. The questionnaire comprised of 18 questions spread into sub-categories with each categories focused at obtaining data from key variables such as the level of ICT anxiety, perceived ease of use of ICT, perceived effectiveness of ICT, and perceived level of ICT acceptance. These variables are described as follow.

1) *ICT Anxiety: This variable focused on finding opinions of respondents to indicate their feel and behavior in relation to ICT utilization.*

2) *Perceived Ease of Use of ICT: Perceived ease of use of a computer has been defined by Venkatesh (2000) as an individual's trust that using computers does not require too much effort.*

3) *Perceived Effectiveness of ICT: This is the opinion of respondents that indicate the level to which they are certain of the fact that ICT can be utilized to improve their job performance or daily routines.*

4) *ICT Acceptance: This is the opinion to measure the respondents' willingness to participate in ICT related ventures.*

B. Methodology

Sampling Procedure

These questions were based in the Likert scale ranging from strongly disagree to strongly agree. This questionnaire was used to explain and calculate the perception of tertiary schools towards ICT adoption and random sampling technique was utilized.

The sample size of this research is not fairly large as the research was only based in Botho University to capture the overall ICT perspective in the Botswana Higher Education sector, thus limiting the sample size. Therefore, research of similar and better magnitude would be employed to measure such with involvement of all sample sizes and demographics.

A pilot study was done to gauge the accuracy and correctness of the questionnaire. This pilot study was done on a sample size of 10 respondents. Based on the respondents' feedback, some changes were effected to improve their accuracy and correctness. The final questionnaire was then adopted and utilized for this research.

Research Instrument

This qualitative research instrument is formulated to answer two major questions which will then be further broken down into simpler and specific sub questions. In this way, the researchers hope to generate an instrument that can easily be analyzed and interpreted. These major questions were:

Main Question 1:

What are the perceptions of the respondents versus the adoption and acceptance of ICT?

1) Sub Question

Do Botho University staff and students have any reservations or concerns about ICT adoption and acceptance?

2) Sub Question

What are the staff and student's perceptions on the effectiveness of ICT in their school (Botho University)?

3) Sub Question

Do staff and students intend to adopt and use ICT in Botho University?

Main Question 2:

What the gaps in ICT adoption and acceptance are as identified from the questionnaire?

Assumption

The assumption tested was: a positive feeling towards the effectiveness and usefulness of ICT in tertiary schools is a sign of computer technology adoption.

V. RESEARCH ANALYSIS AND RESULTS

The study used qualitative data to examine ICT adoption and acceptance in tertiary institutions. These results are based on the analysis of the participants' views depending on the research variables advanced herein.

Qualitative data summaries were obtained and Microsoft Excel used to generate relevant charts for the data. Data summaries of demographic data was collected to provide percentages of describing characteristics of the population that participated in giving responses. There were two variables used for demographic data, namely; age and qualification. Fig.1 below shows the results of the age demographic variable.

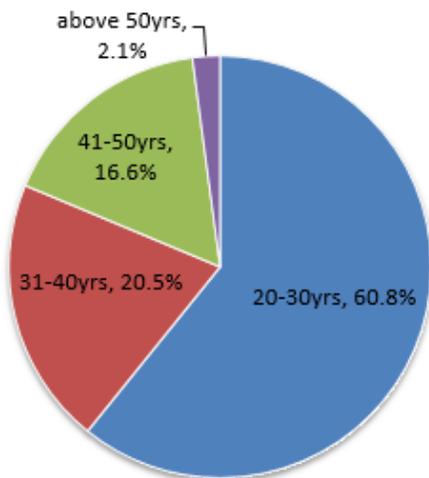


Fig. 1. Age distribution

The age of the participants resulted in 60.8% for the range of 20-30, 20.5% for the range of 31-40, 16.6% for the range 41-50, and 2.1% for participants above 50.

From the results, the majority of the respondents were in the range of 20 – 30 years, which is a true reflection of Botho University when looking at the age of students and staff members herein. Thus, the 20 – 30 years age group is the dominant of all the groups.

Fig. 2 below shows the results of the qualification demographic variable.

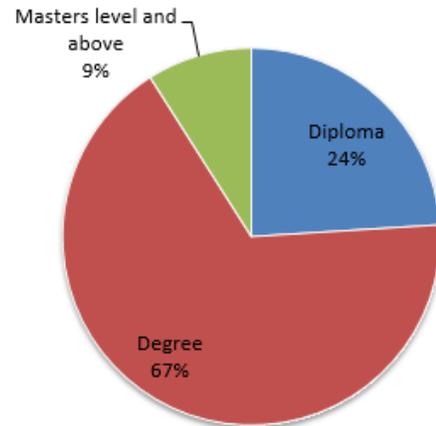


Fig. 2. Qualification distribution

Another demographic variable of qualification resulted in 24% of the participants having Diplomas, 67% having Degree qualification and 9% at masters' level and above.

From the results, it can be concluded that a fair distribution of the questionnaire was done and the majority of the respondents were of the Degree qualification, which is the dominant group in Botho University.

Data summaries were also done for the ICT Opinion questions and analyzed to give solution to the research questions as follows;

Question (a)

Do Botho University staff and students have any reservations or concerns about ICT adoption and acceptance?

The questions on how the participant's feel on the element of worry about ICT adoption had diverse responses. Therefore, this research does not have statistical clear data that can be used to measure the readiness of Botho University adopting and using ICT. Authors recommend further studies through interviews on computer anxiety or worry.

1) Question (b)

What are the staff and student's perceptions on the effectiveness of ICT in their school (Botho University)?

In this question, all participants (100%) strongly agreed that they found computers or ICT useful and effective in the daily delivery of their jobs and or studies.

From the results, it can be concluded that ICT plays a vital role in helping or realizing job delivery at Botho University. Thus having respondents fully agreeing to its role in their daily activities also influences its adoption and acceptance.

2) Sub Question(c)

Do staff and students intend to adopt and use ICT in Botho University?

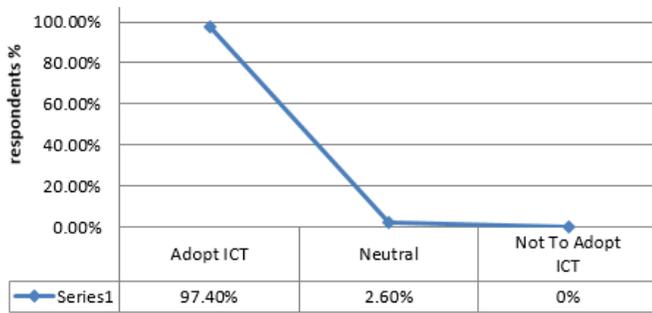


Fig. 3. ICT adoption and acceptance at Botho University

The results for this question produced mixed answers with the scales tipping to the majority who favor ICT adoption. Majority of the participants intend to adopt ICT (97.4%), just a few are neutral (2.6%) and no one does not totally intend to adopt ICT at Botho University. From the results, it can be concluded that majority of the respondents are eager to adopt and embrace the utilization of ICT at Botho University.

VI. CONCLUSION

From this research, it can be concluded that most people accept the role of ICT in their daily activities and tasks. It has also been concluded that they are still people who are hesitant to adopting and accepting the role of ICT in their tasks (2.6%) whilst majority (97.4%) acknowledges its role. It is not clear as to why some people cannot fully accept or adopt ICT, therefore, authors have concluded that further research needs to be conducted to answer and close this gap.

Finally, with the tested assumption, all participants had a positive feeling towards effectiveness and usefulness of ICT. Therefore, it can be concluded that this positive feeling towards the effectiveness and usefulness of ICT in tertiary schools is a sign of computer technology adoption and acceptance by the participants. It has been determined that Botho University staff and students adopt and accept ICT.

VII. RECOMMENDATION

- More research needs to be done

More research is needed on this area, with emphasis on the participant's feel and behavior towards ICT adoption, utilization and acceptance. Studies with focus on such would give a clear indication on the ICT adoption and acceptance trend in Botswana Higher Education Sector, and also provide a platform on policy making and implementation.

- Training people on ICT

ICT literacy is one hurdle impeding the adoption and acceptance of ICT in Institutions of Botswana, thus more training is needed in this area to have majority of people with basic to intermediary skills in the ICT area. This would help people to have relevant ICT skills on their Institutions.

- Early Exposure to ICT

Many people in Botswana tend to be exposed to ICT technologies at later stages such as during employment or at tertiary levels, which also contributes to delaying ICT adoption and acceptance due to limited and or lack of ICT skills. Thus, early exposure in the form of trainings at work entry level or high schools could help in fuelling ICT adoption at higher education sector [13].

REFERENCES

- [1] Unwin. "Information and communication technology for development. Cambridge": Cambridge University Press, 2009.
- [2] Dasgupta, S., Granger, M., & McGarry, N. User acceptance of e-collaboration technology: An extension of the Technology Acceptance Model. *Group Decision and Negotiation*, 11, 87-100, 2003
- [3] Shavo, "Information technology adoption", 2003
- [4] Russell, (2004). People and information technology in the supply chain; social and organizational influences on adoption. *International Journal of Physical Distribution and Logistics Management*, 34 (2), 102-122.
- [5] Long Term Vision 2016: Towards prosperity for all. (1997). From http://vision2016.co.bw/tempimg/media/mediac_103.pdf
- [6] Kanos M, (2013). "Challenges faced in Implementing ICT in Higher Learning Institutions. A Botswana perspective". *International Journal for Infonomics (IJI)*, Volume 6, Issues 1/2, 2013
- [7] Yusuf, M.O (2005) Information and Communication Technologies and Education: Analyzing the Nigerian National Policy for Information Technology. *International Education Journal*, 6(3), 316-321.
- [8] Bharadwaj, A., S. Bharadwaj and B. Konsynski (1999) "Information Technology Effects on Firm Performance as Measured by Tobin's q", *Management Science* (45)7, pp. 1008-1024.
- [9] Hitt, L. and E. Brynjolfsson (1996) "Productivity, Business Profitability, and Consumer Surplus: Three Different Measures of Information Technology Value", *MIS Quarterly* (20)2, pp.
- [10] Van den Ven, A. H. (1986) "Central Problems in the Management of Innovations", *Management Science* (32)5, pp. 590-607.
- [11] <http://laptop.org/en/laptop/hardware/index.shtml> (16 March 2012)
- [12] Cummins, J. (1996) *Negotiating Identities: Education and empowerment in a diverse society*. Ontario: California Association for Bilingual Education
- [13] Maitlamo Botswana's National ICT Policy. (2004). Legislative framework and Change report. Available at http://www.bits.org.bw/downloads/MAITLAMO_NATIONAL_ICT_POLICY.pdf (Retrieved 19-07-2014)
- [14] Botswana National e-government strategy 2011-2016. (2012). Available at <http://www.gov.bw/Global/Portal%20Team/eGovStrategy.pdf> (Retrieved from 18 July 2014)
- [15] Totolo, A. (2005). Information technology adoption in Botswana secondary schools and implications on leadership and school libraries in the digital age. In S. Lee, P. Warning, D. Singh, E. Howe, L. Farmer and S. Hughes (Eds.), *IASL Reports 2005: Information leadership in a culture of change* (p. 78). Erie, PA: International Association of School Librarianship.
- [16] Bose, K. (2004). Computer Training Programme for primary school teachers in teacher training institutions of the southern region of Botswana. Research in Post Compulsory Education.

Improving TCP Throughput Using Modified Packet Reordering Technique (MPRT) Over Manets

Prakash B. Khelage

Asst. Professor, Information Technology,
UMIT, SNDT Women's University,
Mumbai - 400049, India.

Dr. Uttam D. Kolekar

Principal
Smt. Indira Gandhi College of Engineering,
Navi Mumbai - 400709, India.

Abstract—at the beginning of development of network technology TCP transport agent were designed assuming that communication is using wired network, but recently there is huge demand and use of wireless networks for communication. Those TCP variants which are successful in wired networks are neither able to detect exact causes of packet losses nor unnecessary transmission delays over wireless networks. The biggest challenge over MANET is design of robust and reliable TCP variant which should give best performance in different network scenarios. Till date more than dozens of TCP variants designed and modified by researcher and scientist communities even though the level of TCP performance have to be optimum in different scenarios, Such as congestion, link failure, signal loss and interferences. Over rod, grid and bulk network model also. As some of TCP-variant performs well in particular network scenarios but degrades in other scenarios. The objective of this research work, to modify packet reordering technique based TCP variant, implement and compare its performance with other variants. Validation of basic and main network model done using network simulator (NS2) and calculated throughput, delay and packet drop by processing trace files. The simulated result shows that, proposed technique performs outstanding almost in all network scenarios with minimum packet losses and minimum delay.

Keywords—TCP; MANET; RTT; RTO; Congestion; Network validation model

I. INTRODUCTION

Instrumental in developing today's Internet. In particular, TCP has been successful due to its robustness in reacting dynamically to changing network traffic conditions and providing reliability on an end-to-end basis [4]. This Wide acceptance has driven the development of many TCP applications [2] [8], motivating the extension of this protocol to wireless networks. These networks pose some critical challenges to TCP since it was not originally designed to work in such complex Environments, where the level of bit error rate (BER) is not negligible due to the physical medium. High mobility may further degrade the end-to-end performance because TCP reduces its transmission rate whenever it perceives a dropped packet. Mobile ad hoc network is a collection of mobile nodes that offers different opportunities to TCP [6]. Reduction in deployment cost due to absence of fixed infrastructure and elimination of administration cost since it is self-configurable. However, MANET consists of unstable wireless communication links in compare to the wired network [7]. This instability is mainly due to mobility of nodes. Because TCP is originally invented for wired network, it ignores non-congestion loss

which occurs rarely in this environment. Thus, TCP in present form cannot address frequent link breakage in MANET and suffers from performance degradation [4]. TCP is responsible for providing reliability of connection by retransmitting lost packet. Congestion control is the most controversial parts of TCP which degrades performance in front of packet loss. Congestion control as its name appears, assumes all packet loss induced by congestion When link failure lasts greater than RTO (Retransmission timeout) [5].

Retransmission timer expires and TCP interprets packet loss as a congestion loss. Then congestion control executes back-off algorithm to grow RTO exponentially and retransmit packet [3]. After a few successive back-off executions, RTO becomes too long. Hence when route recovered, sender resumes data transmission with long RTO which forces sender remains idle unnecessary in case of probable next losses[10]. Thus, packet loss classification helps TCP to identify link failure loss from congestion loss and consequently triggers appropriate reaction instead of invoking congestion control [4]. Link failure needs TCP to explore how much new route is congested in compare to the broken one. Traffic characteristics can affect queuing delay and processing delay of intermediate nodes that consequently influences Round Trip Time (RTT). If discovered route suffers heavier traffic than old one, retransmission timer must wait more to receive acknowledgment and RTO should be increased [6].

This paper presents the description and implementation of RTT based proposal TCP-MPR variant, Attached as transport agent and validated, calculated performance parameters in different networks model. The rest of the paper is organized as follows. Section 2 gives brief explanations, main concept of MPRT and its implementation. In Section 3 presents in brief about tools, techniques and research methodology. Section 4 and 5 describes different validation network models to be validated for accurate performance measures and finally paper concludes with conclusion and future work in section 6.

II. MODIFIED PROPOSAL FOR THROUGHPUT IMPROVEMENT

The objective of the Modified Packet Reordering (MPR) is to increase throughput with minimum number of drop packets and minimum delay [1] [3]. Technique has used two lists: 1. To-be-ack 2. To-be-sent. In figure 1 to-be-ack list there are sequence numbers of packets which are to be acknowledged and another list is to-be-sent which contains the sequence numbers of packets which are waiting to be sent. When the congestion window allows it, the packet is sent to the receiver and moved to the to-be-ack list. When an ACK for that packet arrives from the

receiver, it is removed from the to-be-ack list (under cumulative ACKs, many packets will be simultaneously removed from to-be-ack). Alternatively, when it is detected that a packet was dropped, it is moved from the to-be-ack list back into the to-be-sent list. Drops are always detected through timers. To this effect, whenever a packet is sent to the receiver and placed in the to-be-ack list, a timestamp is saved. When a packet remains in the to-be-ack list more than a certain amount of time it is assumed dropped. In particular, it assumed that a packet was dropped at time when exceeds the packet's timestamp in the to-be-ack list plus an estimated maximum possible round-trip time $mxrtt$.

As data packets are sent and ACKs received, the estimate $mxrtt$ of the maximum possible round-trip time is continuously updated.

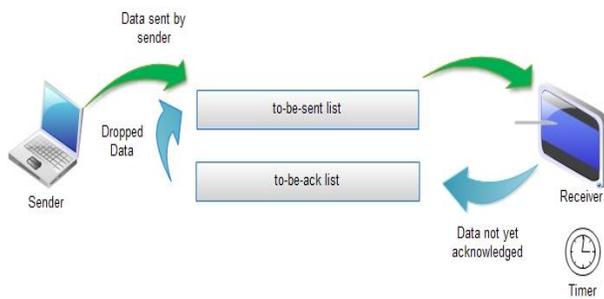


Fig. 1. Conceptual diagram for modified packet reordering (MPR)

Pseudo code for modified packet reordering (MPR) algorithm

Initialization:

- 1) $mode=slow-start$
- 2) $cwnd=1$
- 3) $memorize=0$
- 4) $alpha=0.99$
- 5) $beta=3.0$
- 6) $mxrtt=0$
- 7) $srtt=6.0$
- 8) $remove(to-be-ack, n)$
- 9) $add(to-be-sent, n)$
- 10) *if not is-in(memorize, n) then /*new drop*/*
- 11) $memorize=to-be-ack$
- 12) $cwnd = cwnd(n)/2$
- 13) *else*
- 14) $remove(memorize, n)$

New ack received:

- 15) *if (cwnd < ssthresh)*
- 16) */* Slow Start*/*
- 17) $cwnd = cwnd + 1;$
- 18) *else*
- 19) */* Congestion Avoidance */*
- 20) $cwnd = cwnd + 1/cwnd$

Timeout:

- 21) */* Multiplicative decrease */*
- 22) $ssthresh = cwnd/2;$
- 23) $cwnd = 1;$

III. RESEARCH METHODOLOGY

1) *NS2 (Network simulator version 2)*

NS2 is a discrete event simulator targeted at networking research. Ns provides substantial support for simulation of TCP, routing, and multicast protocols over wired and wireless (local and satellite) networks[12]. It is primarily UNIX based. It uses TCL as its scripting language.



Fig. 2. NS2 Implementation

2) *GNUPLOT*

Gnu plot is a command-line program that can generate two- and three-dimensional plots of functions, data, and data files. It is frequently used for publication-quality graphics as well as education [8]. The program runs on all major computers and operating systems (GNU/Linux, Unix, Microsoft Windows, Mac OS X, and others). gnu plot can produce output directly on screen, or in many formats of graphics files, including Portable Network Graphics (PNG), Encapsulated PostScript (EPS), Scalable Vector Graphics (SVG), JPEG and many others. The program can be used both interactively and in batch mode using scripts. The program is well supported and documented.

3) *AWK SCRIPTING LANGUAGE*

The AWK utility is an interpreted programming language typically used as a data extraction and reporting tool. It is a standard feature of most Unix-like operating systems. AWK is a language for processing text files. A file is treated as a sequence of records, and by default each line is a record. Each line is broken up into a sequence of fields, so we can think of the first word in a line as the first field, the second word as the second field, and so on. An AWK program is of a sequence of pattern-action statements [8]. AWK reads the input a line at a time. A line is scanned for each pattern in the program, and for each pattern that matches, the associated action is executed.

4) *EDRAW MAX*

This enables students, teachers and business professionals to reliably create and publish kinds of diagrams to represent any ideas. Edraw Max is an all-in-one diagram software that makes it simple to create professional-looking flowcharts, organizational charts, network diagrams, business presentations, building plans, mind maps, science illustration, fashion designs, UML diagrams, workflows, program structures, web design diagrams, electrical engineering diagrams, directional maps, database diagrams and more. The best thing about Edraw Max is its flexibility and even you can link diagrams to underlying data to provide even more detailed information to your audience.

5) SIMULATION PARAMETERS VALUE

The required parameters value from table 1 is used to set up wireless mobile ad-hoc network for simulation and validating results for different TCP variants.

TABLE I.

Parameter	Values
Channel Type	Wireless channel
Radio Propagation Model	Two Ray ground
Queue type	Droptail/PriQue
Max. packet(buffer size)	50
Network interface	Wirelessphy
MAC Protocol	802.11
Data Rate	1 Mbps
Transmission Radius	250
Interference Radius	550
Packet size	1000 bytes
Routing protocol	AODV, DSDV
Simulation Time	150 s
Value x	700
Value y	500
Agent trace	ON
Mac trace	OFF
Router trace	ON
Movement trace	ON

IV. BASIC VALIDATION NETWORK MODEL

1) Congestion Network Model

This scenario create a congested node at the middle of a five-node topology by generating three TCP data traffic flows that must pass by this intermediate node to reach the other communicating end [8][9]. One should also note that, different levels of data congestion can be generated by controlling the number of TCP data flows crossing this particular network node at a certain time. Fig. 3 referred as congestion network model to validate and corresponding values noted in table 2.

2) Link failure Network Model

In this model it has been forced to TCP agent to change its communication path by shutting down one intermediate node between the communicating end points. In addition, it is implied routes with different number of hops. Thus, each time TCP changes the communication route, the characteristics of the path between the communicating nodes changes [9]. It is obvious that the choice and the establishment delay of the new route will be dependent on the implemented ad hoc routing protocol. Packet losses and delay changes will also be implied by the link loss and the new chosen route. It is noticed that the effect of such networks nodes' mobility can be represented by the link failure scenario shown in fig. 4 as it is the most direct consequence of mobility, corresponding obtained parameters values depicted in table 3.

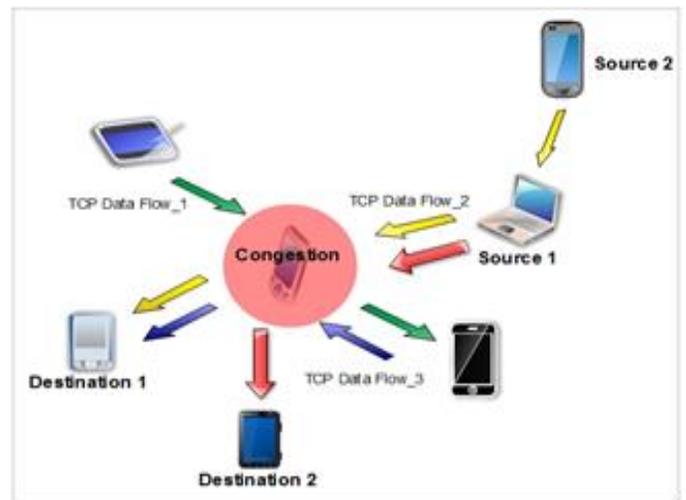


Fig. 3. Congestion Network Model [8]

TABLE II. VALUES OF CONGESTION NETWORK MODEL

Variant	Throughput	Delay	Drop Packet
TCP	371.54	491.632	600
Reno	371.25	482.392	572
Newreno	370.96	489.4	631
Westwood	371.70	506.764	594
WestwoodNR	371.70	456.783	562
Vegas	192.35	122.535	562
Sack	367.49	451.994	588
Fack	362.81	445.309	555
MPR	374.69	414.11	405

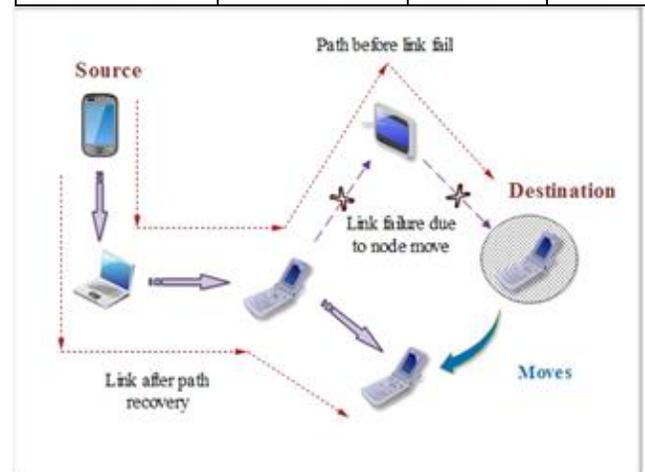


Fig. 4. Link failure Network Model [8]

TABLE III. VALUES OF LINK FAILURE NETWORK MODEL

Variant	Throughput	Delay	Drop Packet
TCP	374.09	226.1	17
Reno	374.09	226.1	17
Newreno	374.09	226.1	17
Westwood	374.09	226.1	17
WestwoodNR	374.09	226.1	17
Vegas	193.36	39.093	1
Sack	374.09	226.1	17
Fack	374.09	226.1	17
MPR	374.68	62.824	2

3) signal loss Network Model

This scenario illustrates the situation where the wireless signal is not stable. The communicating nodes loose the connection due to signal loss and resume the communication when the signal comes back shown in fig. 5. Signal losses are generated by moving one of the intermediate nodes out of the radio range of its connection neighbors [8]. This scenario created using three nodes end node acts as sender and receiver and intermediate node as router Transmission of ftp traffic source flow through intermediate node. Intermediate node moves away for few second so signal loss occurs between source and destination, after few second intermediate node moves at original place and again retransmission starts. Table 4 contains values of different TCP variants validated in Signal Loss Network Model.

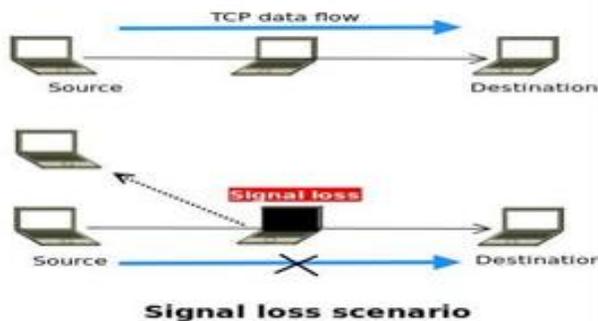


Fig. 5. Signal Loss Network Model [9]

TABLE IV. VALUES OF SIGNAL LOSS NETWORK MODEL

Variant	Throughput	Delay	Drop Packet
TCP	126.00	242.009	42
Reno	127.52	229.244	42
Newreno	48.99	242.241	27
Westwood	124.90	243.186	45
WestwoodNR	125.61	232.455	41
Vegas	93.94	55.8684	20
Sack	126.20	229.576	43
Fack	146.16	217.699	40
MPR	365.71	65.162	5

4) Interference Network Model

In this scenario, two TCP connections are established parallel indicated in fig. 6. The main TCP connection is disturbed by the interferences generated by the second TCP connection. Indeed, the node acting as forwarder for the main TCP connection is placed within the interference range of the second TCP connection sender. So, this situation creates interference and thus data packet losses.

Interference scenario in wireless environment created using two traffic sources Transmission of second traffic source will interfere to the first traffic source [8]. Referring with the obtained values from table 5 it is found that, TCP MRP gives improved throughput, reduced delay and drop packets compared with other TCP- variants.

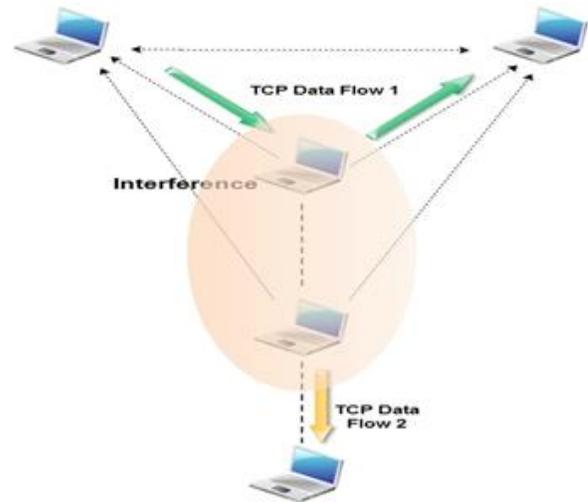


Fig. 6. Interference Network Model [8]

TABLE V. VALUES OF INTERFERENCE NETWORK MODEL

Variant	Throughput	Delay	Drop Packet
TCP	736.51	236.228	1861
Reno	736.51	236.228	1861
Newreno	736.51	236.228	1861
Westwood	736.51	236.228	1861
WestwoodNR	736.51	236.228	1861
Vegas	380.24	38.713	1905
Sack	736.51	236.228	1861
Fack	736.51	236.228	1861
MPR	737.78	136.494	1778

V. MAIN VALIDATION NETWORK MODEL

1) Chain multi-hop Network Model

There is only one route to travel from one node to another. So communication became faster between two nodes. It is very easy to implement. The network consists of variable length chain of static nodes, placed at a distance of 200m from one another. FTP traffic is transferred between the first and last node of the chain [9][11]. During the simulation we will keep one FTP connection active at a time. Sequential TCP connection are initiated and terminated.

The solid-line circle denotes a node's valid transmission range. The dotted-line circle denotes a node's interference range. Node 4's transmission will interfere with node 1's transmissions to node 2. In this scenario also TCP-MPR gives maximum throughput with minimum packet loss and delays comparatively.

Fig. 8 shows a static grid network as experiment topology with 4x4 nodes. The distance between two adjacent nodes is set to be 200 m, and the transmission and interference radii are set to 250 and 550 m, respectively. In a regular grid topology, each node in the network is connected with two neighbors along one or more dimensions. If the network is one-dimensional, and the chain of nodes is connected to form a circular loop, the resulting topology is known as a ring. Network systems such as FDDI use two counter-rotating token-passing rings to achieve high reliability and performance [8].

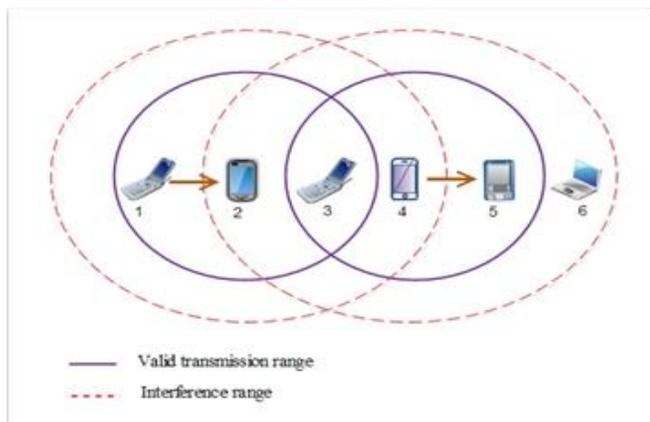


Fig. 7. Chain multi-hop Network Model [9]

TABLE VI. VALUES OF CHAIN MULTI-HOP NETWORK MODEL

Variant	Throughput	Delay	Drop Packet
TCP	253.03	333.966	173
Reno	253.03	333.966	243
Newreno	253.03	333.966	243
Westwood	253.03	333.966	243
WestwoodNR	253.03	333.966	243
Vegas	130.86	57.637	149
Sack	253.03	333.966	243
Fack	253.03	333.966	243
MPR	253.60	135.096	173

2) Grid multi-hop Network Model

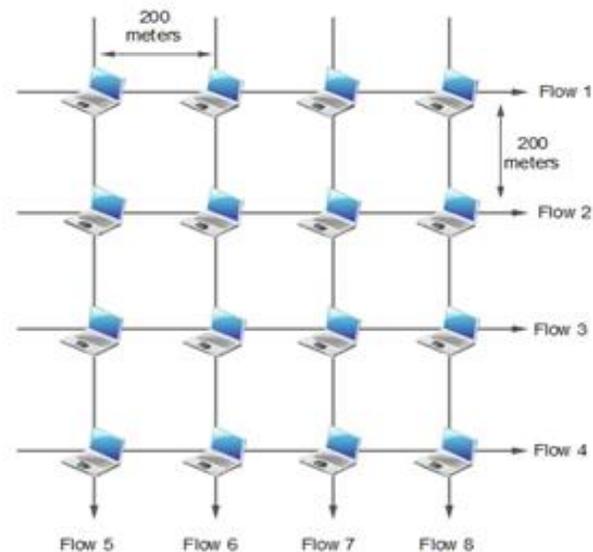


Fig. 8. Grid multi-hop Network Model [9]

In general, when an n-dimensional grid network is connected circularly in more than one dimension, the resulting network topology is a torus, and the network is called "toroidal". When the number of nodes along each dimension of a toroidal network is 2, the resulting network is called a hypercube [11]. The simulated results noted in table 7.

With reference to the graph shown in fig. 9 and corresponding histograms it is observed that, the proposed solution gives improved and optimal performance in all aspects.

Fig. 10-15 shows histogram plot corresponding to the congestion, link failure, signal loss, interference, rod and grid multi-hop network models. It is drawn with referring the tables values of parameters such as throughput, delay and packet drops along with vertical axis whereas different TCP variant taken along with the horizontal axis.

TABLE VII. VALUES OF GRID MULTI-HOP NETWORK MODEL

Variant	Throughput	Delay	Drop Packet
TCP	488.35	1345.12	5
Reno	488.35	1345.12	5
Newreno	488.35	1345.12	5
Westwood	495.59	1339.02	12
WestwoodNR	478.16	1285.47	4
Vegas	222.78	171.355	0
Sack	497.10	1113.29	63
Fack	504.30	1350.66	14
MPR	617.56	588.534	2

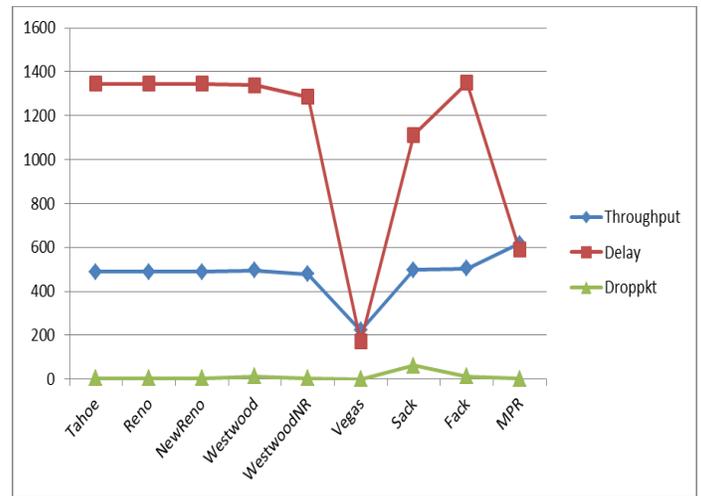


Fig. 9. Graph for Grid multi-hop Network

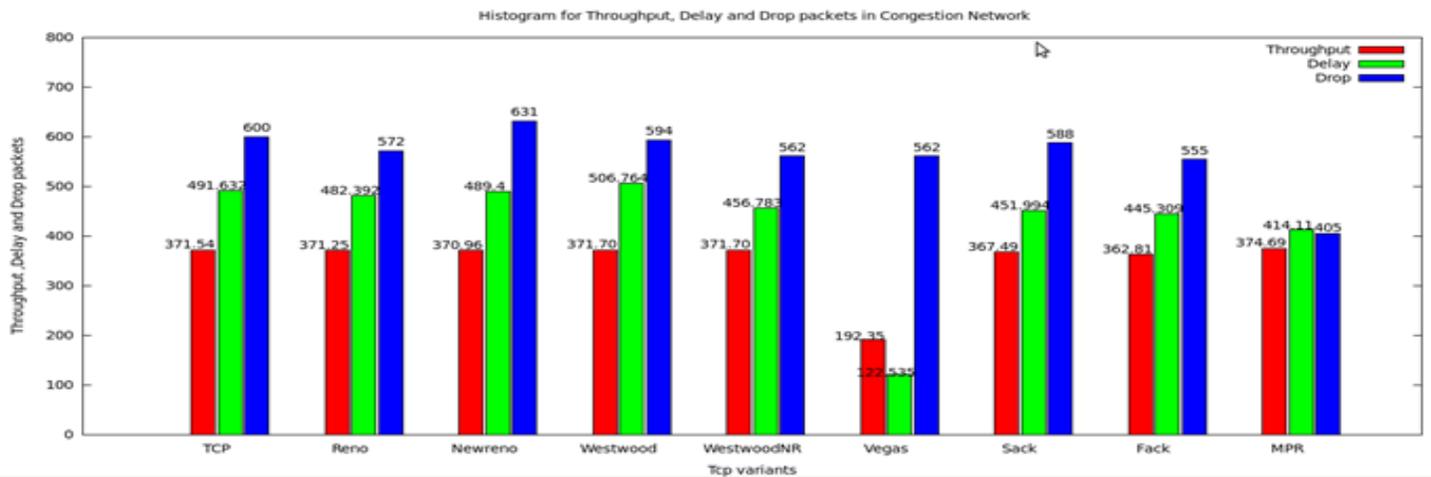


Fig. 10. Histogram Plot for Congestion Network Model

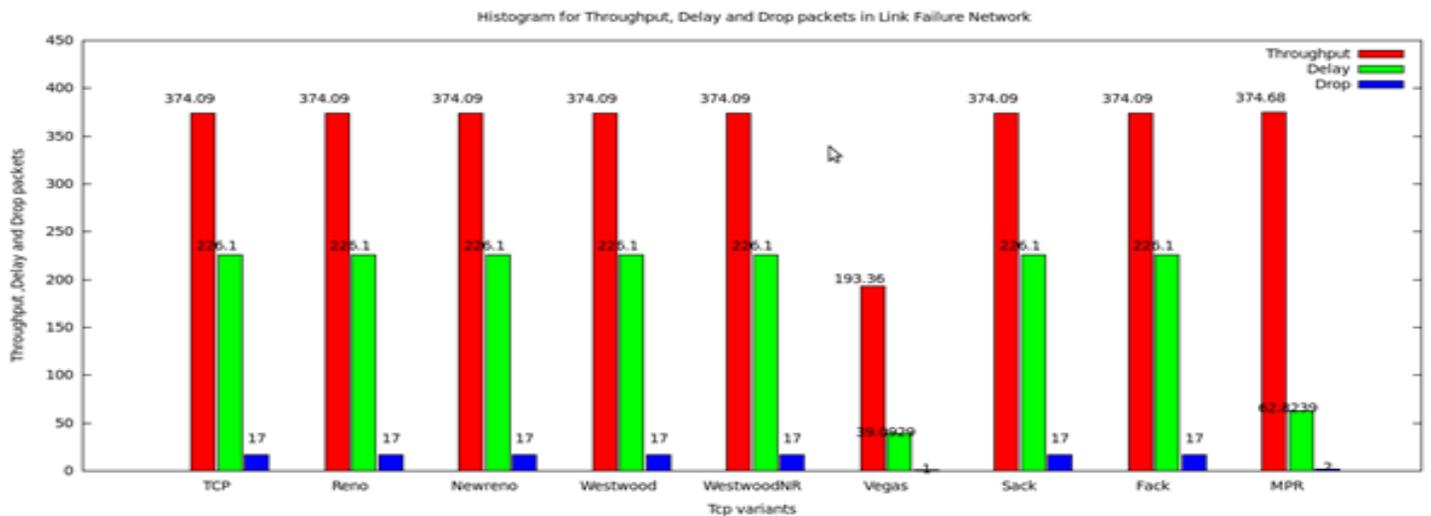


Fig. 11. Histogram Plot for Link failure Network Model

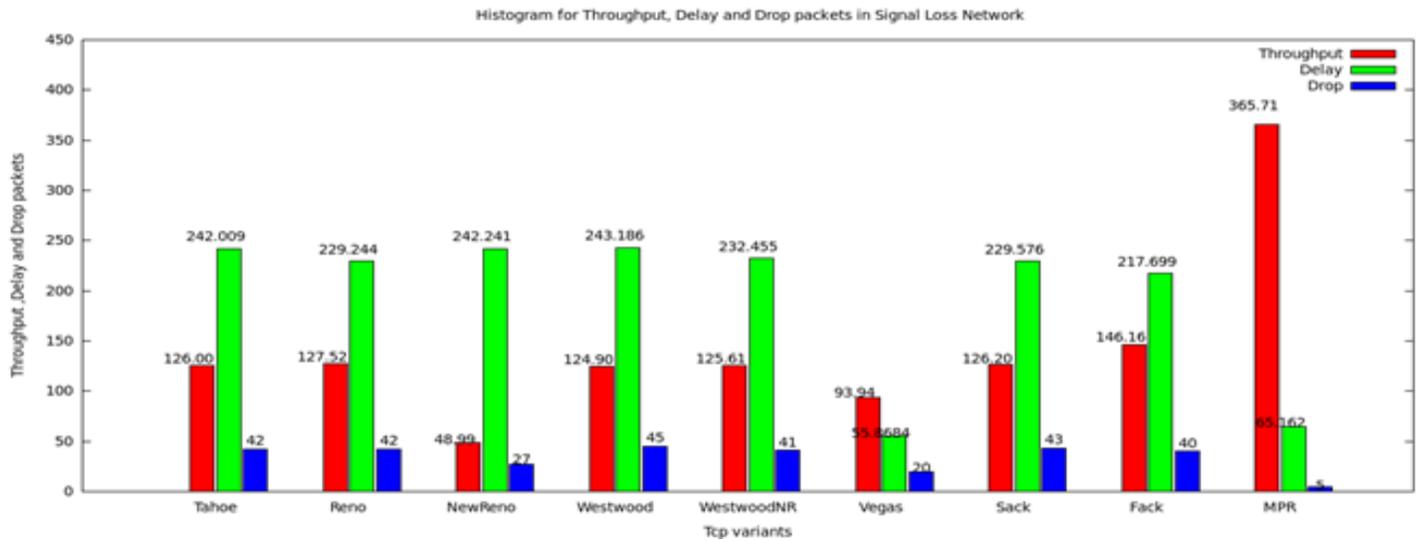


Fig. 12. Histogram Plot for Signal loss Network Model

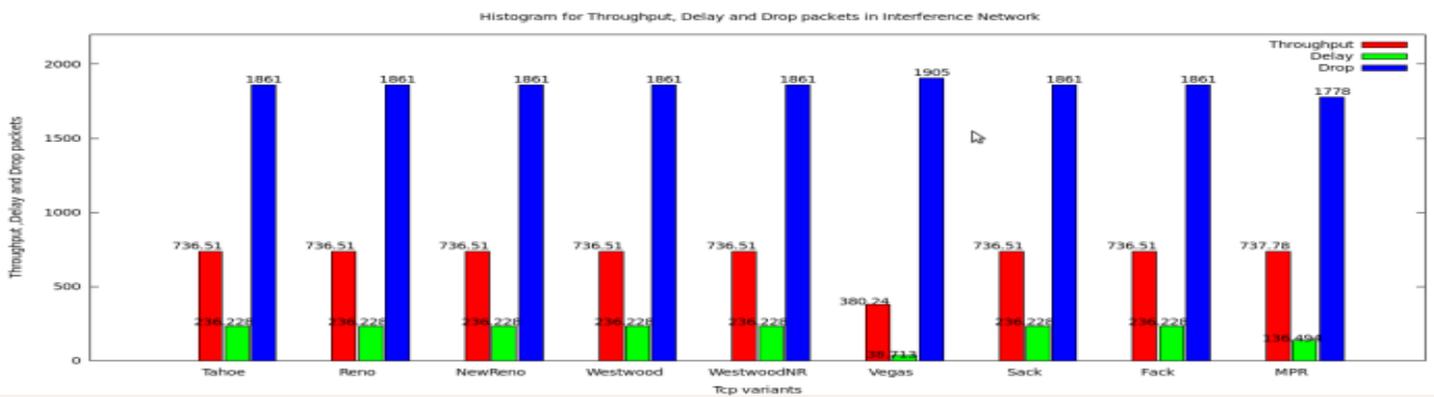


Fig. 13. Histogram Plot for Interference Network Model

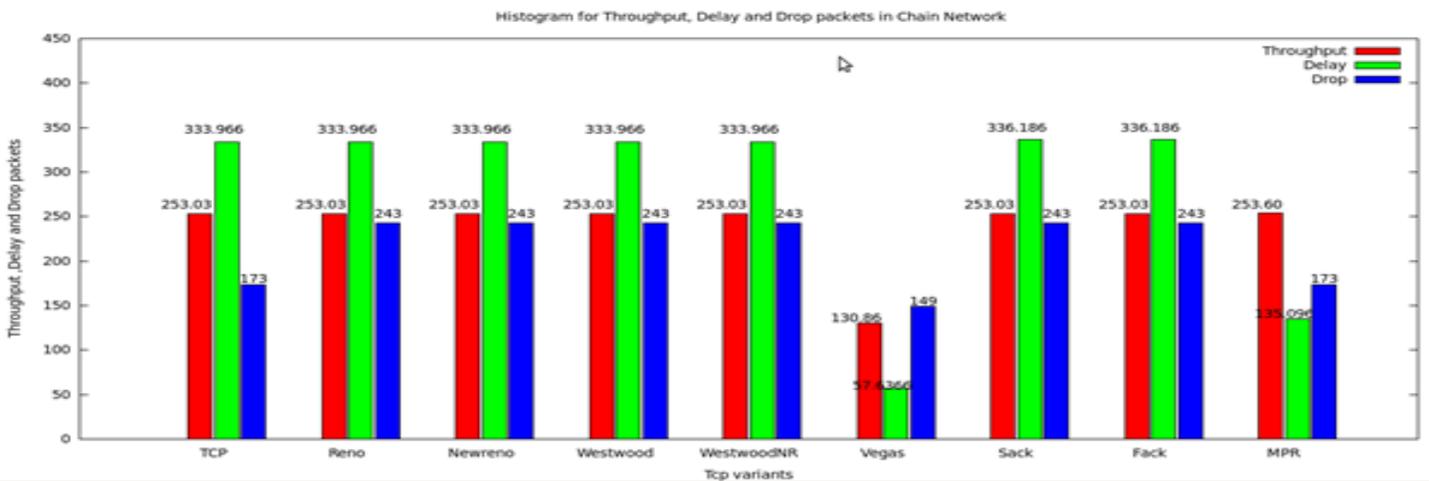


Fig. 14. Histogram Plot for Chain multi-hop Network Model

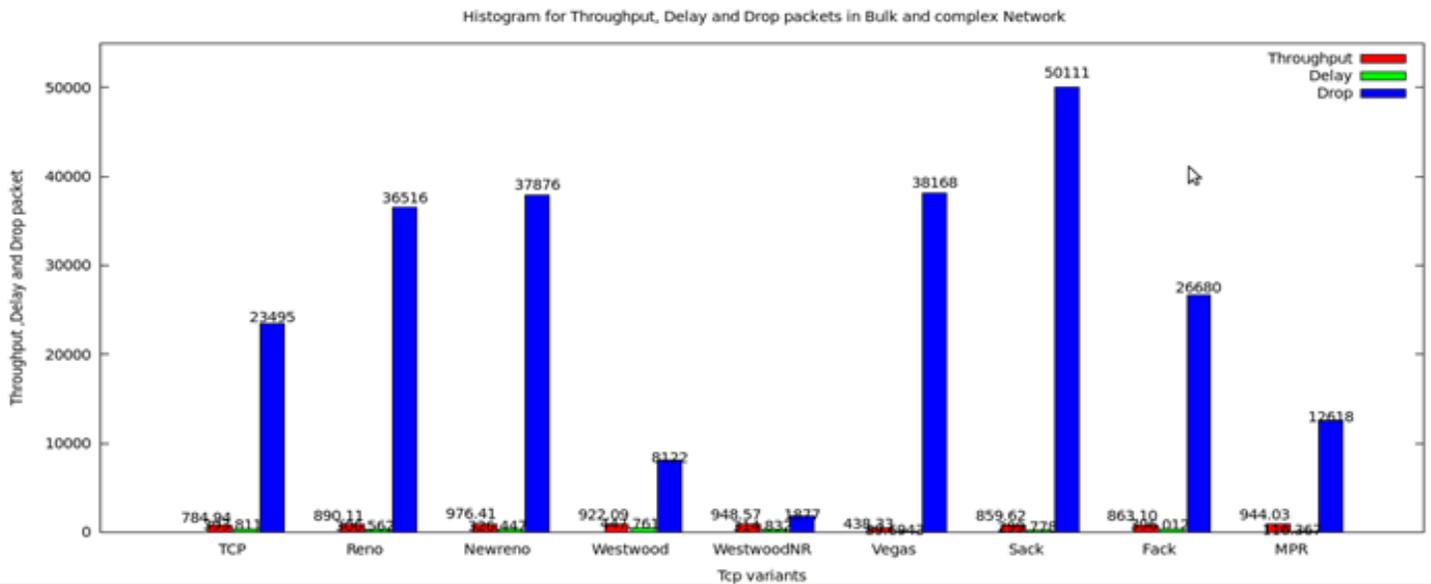


Fig. 15. Histogram Plot for Grid multi-hop Network Model

VI. CONCLUSION AND FUTURE WORK

In this paper we successfully implemented our research proposal (TCP-MPRT) using ns2. The proposal tested and validated in different network model scenarios. We also investigate throughput, delay and packet drop using other TCP-variants in same scenario. Based upon processed results and analysis it shows that, TCP-MPR is giving highest throughput, minimum delay and packet drop compared to all variants. In future the proposal has to be tested and validated in bulk and complex networks with different speed of node mobility.

ACKNOWLEDGMENT

I am deeply thankful to my M. Tech. Mentor Dr. Vijay Raisinghani who has boosted moral, confidence and taught how to do research, my Dear Friend Prof. Sanjay Sange for motivation, helps for best quality work, constantly encouragement and support.

REFERENCES

- [1] Ka-Cheong Leung, et al., "An Overview of Packet Reordering in Transmission Control Protocol (TCP): Problems, Solutions, and Challenges," in IEEE transactions on parallel and distributed systems, vol. 18, no. 4, april 2007
- [2] Ahmad Al Hanbali et al., "A survey of TCP over Ad-hoc Networks," IEEE Communications Surveys & Tutorials, Third Quarter 2005.
- [3] Hanaa Torkey, Gamal Attiya and Ibrahim Z. Morsi, "Modified Fast Recovery Algorithm for Performance Enhancement of TCP-New Reno" International Journal of Computer Applications (0975 – 8887) Volume 40–No.12, February 2012.
- [4] A. Seddik-Ghaleb, Y. Ghamri-Doudane, and S. M. Senouci, "TCP WELCOME TCP Variant for Wireless Environment, Link losses, and COngestion packet loss ModElS," in First International Communication Systems and Networks and Workshops, COMSNETS 2009.
- [5] P Bhaskar Sardar and Debashis Saha, "A Survey of TCP Enhancements for Last-Hop Wireless Networks" in IEEE Communications Surveys & Tutorials, 2006 3rd Quarter 2006, Volume 8, No 3.
- [6] Haifa Touati, Ilhem Lengliz and Farouk Kamoun, "TCP Adaptive RTO to Improve TCP performance in mobile ad hoc networks," in The Sixth Annual Mediterranean Ad Hoc Networking Workshop, Corfu, Greece, June 12-15, 2007.

- [7] H. M. El-Sayed, "Performance Evaluation of TCP in Mobile Ad-Hoc Networks," in the Second International Conference on Innovations in Information Technology (IIT) 2005.
- [8] Prakash B. Khelage, Uttam D. Kolekar, " Survey and Simulation based Performance Analysis of TCP-Variants in terms of Throughput, Delay and drop Packets over MANETs" International Journal of Scientific and Engineering Research(775-786) Volume 5, Issue 1 January 2014.
- [9] Prakash B. Khelage, Uttam D. Kolekar, " TCP- CostCO Reno: New Variant by Improving Bandwidth Estimation to adapt over MANETs" International Journal of advanced computer science and applications Vol. 05, No. 02, February 2014.
- [10] R. Oliveira and T. Braun, "A Dynamic Adaptive Acknowledgment Strategy for TCP over Multi-hop Wireless Networks," Proc. IEEE INFOCOM, Mar. 2005.
- [11] Z. Fu, P. Zerfos, H. Luo, S. Lu, L. Zhang, and M. Gerla, "The Impact of Multi-hop Wireless Channel on TCP Throughput and Loss," Proc. INFOCOM '03, Apr. 2003.
- [12] Network Simulator Ns2, <http://www.isi.edu/nsnam/ns>.

AUTHORS PROFILE



Prakash B. Khelage received his B.E. in Electronics and Telecommunication Engineering from Dr. Babasaheb Ambedkar Marathwada University Aurangabad, M.Tech in information Technology from NMIMS University Mumbai, Maharashtra, India. He is currently working as Assistant Professor with UMIT, SNDT Women's University. He has 13 years of experience in industrial as well as educational field; His research interest includes Ad Hoc Networks, Mobile Computing, Wireless Networks, Co-operative Communication Networks and Network Security. He has also interest in Computer Architecture design, Cloud Computing and Data Mining. He has published 2 International journal papers.



Uttam D. Kolekar received his B.E. in Electronics and Telecommunication Engineering, M.E. in Electronics from Shivaji University, Kolhapur and Awarded Ph. D. in electronics from Bharati Vidhyapith Pune, Maharashtra, India. He is currently working as Principal with Smt. Indira Gandhi College of Engineering, Mumbai University. He has more than 20 years of experience in educational institution; His research interest includes Ad Hoc Networks, Mobile Computing, Wireless Networks, Neural Network and Co-operative Communication Networks. He has published over 30 National and International Journals & conferences various papers accros India and other countries.

Local and Semi-Global Feature-Correlative Techniques for Face Recognition

Asaad Noori Hashim
Dept. of Computer Science
University of Babylon
P.O.Box 4, Babylon,
Iraq

Zahir M. Hussain
Dept. of Computer Science
University of Kufa
P.O.Box 21, Kufa, Najaf, Iraq
Adjunct Professor, ECU, Australia

Abstract—Face recognition is an interesting field of computer vision with many commercial and scientific applications. It is considered as a very hot topic and challenging problem at the moment. Many methods and techniques have been proposed and applied for this purpose, such as neural networks, PCA, Gabor filtering, etc. Each approach has its weaknesses as well as its points of strength. This paper introduces a highly efficient method for the recognition of human faces in digital images using a new feature extraction method that combines the global and local information in different views (poses) of facial images. Feature extraction techniques are applied on the images (faces) based on Zernike moments and structural similarity measure (SSIM) with local and semi-global blocks. Pre-processing is carried out whenever needed, and numbers of measurements are derived. More specifically, instead of the usual approach for applying statistics or structural methods only, the proposed methodology integrates higher-order representation patterns extracted by Zernike moments with a modified version of SSIM (M-SSIM). Individual measurements and metrics resulted from mixed SSIM and Zernike-based approaches give a powerful recognition tool with great results. Experiments reveal that correlative Zernike vectors give a better discriminant compared with using 2D correlation of the image itself. The recognition rate using ORL Database of Faces reaches 98.75%, while using FEI (Brazilian) Face Database we got 96.57%. The proposed approach is robust against rotation and noise.

Keywords—Zernike Moments; Face Recognition; Structural Similarity

I. INTRODUCTION

Face recognition has become one of the most successful applications of image analysis and computer vision. Face recognition software has been incorporated in a wide variety of biometrics-based security systems for the purposes of identification, authentication and video surveillance. Face recognition includes three stages. The first stage is detecting the location of the face, which is a difficult task for the position, orientation, and scaling of the face are unknown in an arbitrary image. The second stage involves extraction of the pertinent features of the localized facial image obtained in the first stage. Finally, the third stage requires classification of facial images based on the derived feature vector obtained in the previous stage.

Unlike humans who have an outstanding capability of recognizing different patterns and faces in varying conditions,

machines are still dependent on ideal face images; their performance suffers when there are variations in illumination, background, pose angle, obstacles, etc. Therefore, the problem of automatic face recognition is a very complex and challenging task [1]. Conventionally, face recognition methods are classified in two categories. The first one is based on extracting structural facial features that are local features of face images, for example, the shapes of eyes, nose and mouth. The structure-based approaches deal with local information instead of global information. The second category is based on statistical approaches, wherein features are extracted from the whole image and thus use global information instead of local information. Since the global data of an image are used to determine the feature elements, data that are irrelevant to facial portion such as hair, glasses, shoulders and background may result in the creation of erroneous feature vectors that can affect the recognition results [2].

Statistical approaches for feature extraction based on moment invariants have been utilized for classification and recognition applications because of their invariance properties. An image feature is considered invariant if it remains neutral to changes in size (scale), position (translation), orientation (rotation), or/and reflection in an image. The most popular appearance-based face recognition algorithms are: Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Linear Discriminant Analysis (LDA). PCA finds a set of the most representative projection vectors such that the projected samples retain most of the information about original samples. ICA captures both second and higher-order statistics and projects the input data onto the basis vectors that are statistically independent as possible. LDA uses the class information and finds a set of vectors that maximize the between-class scatter while minimizing the within-class scatter [3].

Pan and Bolouri (1999) used the discrete cosine transform to reduce image information redundancy, because only a subset of the transform coefficients are necessary to preserve the most important facial features such as hair, eyes and mouth. The researchers state that when DCT coefficients are fed into back-propagation neural network for classification, high recognition rate can be achieved by using a very small proportion of transform coefficients [4]. Hafeed et al. (2001) introduced an accurate and robust face recognition system. This system exploits the feature extraction capabilities of the discrete cosine transform (DCT) and invokes certain

normalization techniques that increase robustness to variations in facial geometry and illumination [5]. Hazim et al. (2005) proposed a local appearance-based face recognition algorithm, where local information is extracted using block-based discrete cosine transformation; and obtained local features are combined both at the feature level and the decision level [6].

In [7], a novel algorithm was proposed by Osslan et al. (2009) to solve the problem of automatic face recognition is presented; where the novelty of the algorithm is the ability to combine the computer vision tasks with Particle Swarm Optimization (PSO) to improve the execution time and to obtain better recognition results. The crucial stage of a typical system of face recognition has been improved by using a fitness function to measure the similarity of the input face compared with a database of faces. Sharma et al. (2010) introduced simple but efficient novel H-eigenface (Hybrid-eigenface) method for pose-invariant face recognition, ranging from frontal to profile view. The proposed method is based on the fact that face samples of same person under different poses are similar in terms of the combination patterns of facial features [9]. Zhang and Li (2010) proposed discriminative K-SVD (D-KSVD) based on extending the K-SVD algorithm by incorporating the classification error into the objective function [10].

Lone et al. (2011) used a multi-algorithmic approach, where they developed a face recognition systems based on one combination of four individual techniques, namely, Principal Component Analysis (PCA), Discrete Cosine Transform (DCT), Template Matching using Correlation (Corr) and Partitioned Iterative Function System (PIFS). Researchers fuse the scores of all of these four techniques in a single face recognition system [11].

Invariants, especially Zernike moments, also adopted by researchers and became more attractive due to its robustness against rotation and noise. Shi et al. (2012) proposed a feature extraction method based on pseudo-Zernike moment followed by LDA for dimensionality reduction [12].

Singh et al. (2012) proposed a modified PCA algorithm by using some components of the LDA algorithm for face recognition. The algorithm is based on the measure of the principal components of the faces, then to find the shortest distance between them [13]. On the other hand, many researchers utilize local features embedded within the human face. Manchula and Arumugam (2013) presented a feature-based multimodality face recognition system to recognize the human individuals in an environment of known faces using features like shape of the eyes, nose and jaw [14].

We outline this paper as follows: the next section describes the proposed system which consists of a number of steps: Image Pre-Processing, Image Dividing and Feature Extraction (such as Zernike Moments), Modified Structural Similarity Index Measurements (M-SSIM), Features Selection, Measurement Performance (that illustrate a number of main and derived measurements for face recognition), and the final step of proposed system, which is Classification. Section 3 explains the standard data sets which used for face recognition.

The remaining sections are Contributions, Results, Conclusions, and Discussions.

II. LOCAL AND SEMI-GLOBAL FEATURE EXTRACTION

Different emotions and occlusions are represented by facial features with more emphasis on specific areas of the face than other areas (center of the face); also, changing lighting conditions (lighting direction and illumination condition) are considered. In this paper, a novel weighted patch moment array face representation and recognition approach is introduced. The overall framework of the algorithm is illustrated in the following steps. There are four main steps in the proposed algorithm:

- 1) *Pre-processing operations,*
- 2) *Image partitioning and feature extraction,*
- 3) *Measurements performance,*
- 4) *Classification.*

A. Image Pre-Processing

The following pre-processing operations are needed before applying the proposed algorithm:

- 1) *Modifying image scales: All images must be square and have even dimensions.*
- 2) *Specifying the order and repetition of Zernike to get a set of polynomials, using the following algorithm.*
- 3) *Preparing window parameters for SSIM.*
- 4) *Face detection: face detection is a necessary pre-processing stage. However, it is a research issue by itself, with major difficulties and challenges. Therefore, we propose our approach based on standard datasets, which provide suitable (pre-processed) images for recognition without the need to face detection stage.*

Algorithm (1): Zernike Order-Repetition Set

```
Initialization:  m=[ ], n=[ ]
Input:          Po //Minimum Order
                P  //Maximum Order
Output:         Zernike Order-Repetition Set
```

Zernike Set:

```
for h=Po to P
for f=0 to P
if (f<=h and mod(f-h,2)=0)
n =[n h]
m =[m f]
end if
end loop f
end loop h.
```

B. Image Dividing and Feature Extraction

Choosing an efficient feature extraction method is the most important factor to achieve a high recognition rate in face recognition. In the proposed algorithm, a human face image is divided into a set of equal-sized blocks in an overlapping manner. In this work, the dimension of each face image training image or test (reference) is $N*N=92*92$ pixels, the following algorithm is used to divide each image in to overlapping windows.

Algorithm (2): Dividing images y,x into overlapping blocks

Initialization:

wg=[62 42], sg=[15 25] //Global/Local windows
yw=[], xw=[] //yw sub-block of test face
//xw sub-block of training image

Input: y is reference (test) image with size N*N

x is training image with size N*N

Output: y and x are divided into sub-blocks {yw, xw}

Step One: Compute the length of wg: Lg=length(wg) ;

Step Two: Partition Images

For g=1 to Lg

w=wg(g) ; s=sg(g); k=round(N-w)/s

for s1=0 to k ; s2=s1*s

for w1=0 to k ; w2=w1*s

yw = y(s2+1:s2+w, w2+1:w2+w);

xw = x(s2+1:s2+W, w2+1:w2+W);

end loop w1

end loop s1

end loop g

1) Zernike Moments

Moments compute a numeric quantity at some distance from a reference point or axis. While Zernike polynomials are defined as a set of orthogonal polynomials defined on the unit disk, Zernike moment is the projection of the image function onto these orthogonal basis functions. Zernike moments have been proven to be more robust in the presence of noise. Since their moment functions are defined using polar coordinate representation of the image space, Zernike moments are commonly used in recognition tasks requiring rotation invariance. Zernike moments are a good feature representation and provide more information about facial image and reduce the dimension of the feature vector leading to improved results. Implementation of Zernike moments is detailed in [15]; while some formulas were corrected by Sun-Kyoo Hwang. The kernel of Zernike moments is a set of orthogonal Zernike polynomials defined over the polar coordinate space inside the unit circle. Zernike moments of order p with repetition q of an image with intensity f(r, θ) are defined as follows [16]:

$$Z_{pq} = \frac{p+1}{\pi} \int_0^1 \int_0^{2\pi} R_{pq}^*(r) e^{-jq\theta} f(r, \theta) r dr d\theta, \quad |r| \leq 1 \quad (1)$$

where the radial polynomial $R_{pq}(r)$ is given as follows:

$$R_{pq}(r) = \sum_{k=0}^{\frac{p-|q|}{2}} (-1)^k \frac{(p-k)!}{k! \left(\frac{p+|q|}{2} - k\right)! \left(\frac{p+|q|}{2} + 1 - k\right)!} \quad (2)$$

with $0 \leq |q| \leq p$ and $p - |q|$ is even.

Zernike moments utilize polar coordinates (r,θ) inside the unit circle $|r| \leq 1$. To approximate and compute them in discrete form we perform a linear transformation of the image Cartesian coordinates (i, j) from the inside of the square $i, j=0, 1, \dots, N-1$ to the inside of the unit circle $|r| \leq 1$ to get the discrete form:

$$Z_{pq} = \lambda_z(p, N) \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} R_{pq}(r_{ij}) e^{-jq\theta_{ij}} f(i, j) \quad (3)$$

where

$$\left. \begin{aligned} r_{ij} &= \sqrt{x_i^2 + y_j^2}; & \theta_{ij} &= \tan^{-1} \left\{ \frac{y_j}{x_i} \right\} \\ x_i &= \frac{2i}{N-1} - 1; & y_j &= \frac{2j}{N-1} - 1 \\ \lambda_p(p, N) &= \frac{p+1}{N-1} \end{aligned} \right\} \quad (4)$$

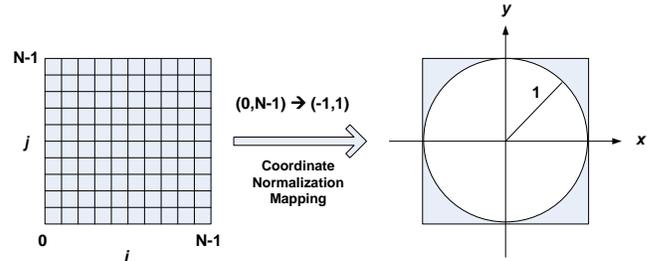


Fig. 1. Two Coordinate normalization schemes for radial and Zernike moments: (a) Discrete image coordinate space of size(N*N); (b) Coordinate normalization using map (0,N-1)→(-1,+1).

The above coordinate's transformation is shown in Figure 1. The figure illustrates that, the algorithm will focus on the center of an image, which includes the human face directly, this will increase the accuracy of recognition with small redundancy. The radial moments used here are complex in nature. Since their magnitude is invariant to rotation, so we are utilizing just their magnitude as a feature vector with several orders. Zernike moment is used as a feature extractor the value of order p, repetition q, which is varied to achieve the best classification performance.

2) Modified Structural Similarity Index Measurements (M-SSIM)

An objective image quality measure can have a significant role in image processing and its applications, where it can be used to monitor and adjust image quality. Also, a quality measure can be used to optimize algorithms and parameter settings of image processing systems. Machine evaluation of image and video quality is important for many image processing systems, for example, systems used for compression, restoration, enhancement, etc. The goal of quality assessment is to find robust techniques for objective evaluation of image quality in accord with subjective human assessment. Wang et al. (2004) [17] proposed a promising technique (SSIM) for distance covariance to measuring the structural similarity based on number of statistical measurements such as mean, standard deviation and they produced a new relation among these standards as n the following formula:

$$p(x,y) = \frac{(2\mu_x\mu_y + C_1)(\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (5)$$

where, $\rho(x,y)$ is the SSIM measure between two images x and y, μ_x and σ_x^2 are the statistical mean and variance of pixels in image x (μ_y, σ_y^2 are defined similarly), σ_{xy} is the statistical variance between pixels in images x and y, while the constants C1 and C2 are defined as $C1 = (K1L)^2$ and $C2 = (K2L)^2$,

with K1 and K2 are small constants and L = 255 (maximum pixel value). This approach gives a high level of similarity for noise free condition while it goes to zero when noise increase, in other words, it gives a similarity with two different images due to it dependent only the statistics features of images which may have some correlations. SSIM can't reveal all image structural properties, so we need to more specific measurements that are image-dependent. Pure SSIM gives a good results, but if it combined with Edge detection filters such as Canny, it will produce excellent results specially when the images are different from each other (in this case well return zero value for similarity) [18].

III. NEW MEASURES: A FEATURE-CORRELATIVE APPROACH

In previous works, we used either local or global features for recognition successfully[19-21]. In this work, global features proved ineffective, and we use mixed local and semi-global features. Seven measurements are investigated: First three are tested individually, while others are combined in some ways. Zernike moment is applied on equal size of overlapping blocks in a local manner (small window) and semi-global (large window). In general, the main measurements are Zernike correlation, Zernike errors, and SSIM. Each one of these measurements must be converted into one dimension before using the other measurements. In each measurement we tried to find first and second maximum values, where the first represent the required person with high probability while the second represent the person with probability less than the first. To satisfy above goals, many normalizations operations as well as dimensions reductions are applied. The measurements are explained later.

A. Zernike Correlation Measure

Correlation measures the linear relationship between any two variables, and if these variables are independent, the correlation will be approximately zero. In addition, correlation matrices give an overview of the pattern of relationships between variables.

With the proposed algorithm, Pearson correlation coefficients are computed to find the autocorrelation of Zernike moments of (windowed) reference image and cross correlation between the Zernike moments of (windowed) reference image and moments of (windowed) poses of database. Then the minimum distance between the two correlations over all possible windows will indicate a measure of similarity (recognition). The following formulas are used, with y_w indicating the w-th window of the reference image and x_w indicating the w-th window of the test pose:

$$e_c(x, y) = \min_w \left[\sum_{i=1}^{2L_z-1} \frac{\{R_{x_w y_w}(i) - R_{y_w}(i)\}^2}{M^2} \right] \quad (6)$$

where

$$R_{x_w y_w} = \text{xcorr}\{|Z_{p_o p}(x_w)|, |Z_{p_o p}(y_w)|\},$$

$$R_{y_w} = \text{xcorr}\{|Z_{p_o p}(y_w)|, |Z_{p_o p}(y_w)|\},$$

$$M = \max\{R_{y_w}\},$$

$$L_z = \text{length}\{Z_{p_o p}\}$$

$$= \text{length}\{Z_{pq} | P_o \leq p \leq P; 0 \leq q \leq p; \text{mod}(p - |q|, 2) = 0\}$$

$$\{p\} = \{P_o, \dots, P\} = \text{range of Zernike moments under test}$$

$$\begin{aligned} |Z_{p_o p}(y_w)| &= \{|Z_{pq}(y_w)| | P_o \leq p \leq P; 0 \leq q \leq p; \text{mod}(p - |q|, 2) = 0\} \\ |Z_{p_o p}(x_w)| &= \{|Z_{pq}(x_w)| | P_o \leq p \leq P; 0 \leq q \leq p; \text{mod}(p - |q|, 2) = 0\} \end{aligned} \quad (7)$$

noting that **xcorr** computes correlation of vectors (with equal length= $N=L_z$) giving a vector of $2L_z - 1$ length without normalization as follows:

$$R_{xy}(m) = \begin{cases} \sum_{n=0}^{N-m-1} x_{n+m} y_n & m \geq 0 \\ R_{yx}(-m) & m < 0 \end{cases} \quad (8)$$

where w ranges over corresponding blocks of reference image (which represents the test image) and different poses of each person which are stored in the face database which represent the training set.

The **Correlative Zernike Measure** is defined as:

$$E_z(y) = \arg[\max_x \{M_y [1 - e_c(x, y)]\}] \quad (9)$$

where

$$M_y = \max[\max_x \{e_c(x, y)\}, \max_x \{e_m(x, y)\}] \quad (10)$$

and the function $\arg[\cdot]$ indicates the (ordinal number of the) person in the database whose some specific pose gives maximal similarity with the **reference image y**. The above constant M_y is *defined jointly* with the **Min-Max Zernike error function** $e_m(x, y)$ as explained below.

B. Min-Max Zernike Measure

Applying the following formulas to find Zernike error function between the corresponding blocks:

$$e_m(x, y) = \min_w [\{M_{x_w} - M_{y_w}\}^2 + \{\mu_{x_w} - \mu_{y_w}\}^2] \quad (11)$$

where:

$$M_{x_w} = \max |Z_{p_o p}(x_w)| \quad ;$$

$$\mu_{x_w} = \min |Z_{p_o p}(x_w)|$$

We can define the **Min-Max Zernike Measure** as follows:

$$E_m(y) = \arg[\max_x \{M_y - e_m(x, y)\}] \quad (12)$$

C. Structural Similarity Measure

As we mentioned above, the SSIM measurement can be applied to corresponding blocks of reference image and different poses giving the similarity function:

$$S(x, y) = \max_w \text{SSIM}(x_w, y_w) \quad (13)$$

from which the **Structural Similarity Measure** is defined as follows:

$$E_s(y) = \arg[\max_x \{M_y \cdot S(x, y)\}] \quad (14)$$

D. Combined Similarity Measures

Based on the above three basic measures [Equations (9), (12), and (14)], we derive four **Combined Similarity Measures** as follows:

$$e_o(y) = \arg[\max_x \{E_1(x, y) \cdot E_2(x, y) \cdot S(x, y) / M_y\}] \quad (15)$$

$$e_h(y) = \arg[\max_x \{E_2(x, y) \cdot S(x, y)\}] \quad (16)$$

$$e_n(y) = \arg[\max_x \{E_1(x, y) \cdot S(x, y)\}] \quad (17)$$

$$e_v(y) = \arg[\max_x \{E_1(x, y) \cdot E_2(x, y) / M_y\}] \quad (18)$$

noting that:

$$E_1(x, y) = M_y - e_c(x, y) \quad (19)$$

$$E_z(x, y) = M_y - e_m(x, y) \quad (20)$$

where the function $\text{arg}[\cdot]$ indicates the ordinal number of the person in the database as stated before, hence, it is the recognition function.

E. Classification and Probability of Recognition

Many techniques may be used for classification stage such as K-means or Naïve Bayesian, which is considered as a probabilistic approach. In this work, a new threshold is derived based on using a set of seven measures as defined in Equations (9), (12), (14)-(18). A success ($D=1$ in our algorithms, which is the recognition of the face image as belonging to the data set) is reached when at least two measures recognize the reference face image from $C=14$ cases [seven measurements for local analysis with seven measurements for semi-global analysis]. So, the threshold of belonging (recognition) is $T = 2/C$.

Now we define our *confidence in this recognition* and call it *Probability of Recognition* (belonging), P_r . First, we find the second peak (maximum) in the above measures. Then we find the difference between the absolute maximum and the second maximum for each measure, which we call here the MM - difference. Normalization for these differences (by maximum difference) is necessary. The resulting quantity is the MM-difference for that measure. For example, $d_z(y)$ is the MM-difference for the Zernike Correlation Measure; $d_m(y)$ is the MM-difference for the Min-Max measure; and so on. Then, the probability of recognition is defined as follows:

$$P_r(y) = \frac{\text{length} \{d_i(y) \mid i = \text{successful measure with } d_i(y) > \frac{1}{C}\}}{C} \quad (21)$$

where "successful" means passing the Threshold $T = 2/C$. Hence, P_r calculates how much confidence we should put in this recognition, noting that $\frac{1}{C}$ means at least one measure recognizes the image.

F. Database Sets

1) *The AT&T: This face image database contains 10 different images (poses) of each person; the set consist of images for 40 persons taken at different illuminations, rotation and facial expressions and facial details like glasses. The size of each image is 92×112 pixels, with 256 grey levels per pixel.*

2) *FEI Face Database: The FEI face database is a Brazilian face database that contains a set of face images taken between June 2005 and March 2006 at the Artificial Intelligence Laboratory of FEI in Sao Bernardo do Campo, Sao Paulo, Brazil. There are 14 images for each of 200 individuals, a total of 2800 images. All images are colorful and taken against a white homogenous background in an upright frontal position with profile rotation of up to about 180 degrees. Scale might vary about 10% and the original size of each image is 640×480 pixels. All faces are mainly represented by students and staff at FEI, between 19 and 40 years old with distinct appearance, hairstyle, and adorns. The number of male and female subjects are exactly the same and equal to 100 [22]. See Fig(2).*



Fig. 2. Various face poses for a single person from The FEI Face Database.

IV. RESULTS AND DISCUSSION

The performance of the proposed methodology is compared with two different benchmark datasets as well as using images taking at unconditional environment (may be hard environments if we consider the high degree of rotation and complex emotions). The results illustrate the efficacy of Zernike moments for the face recognition problems. The algorithm has been tested using approaches:

- A. *Under verification branch where the test image represents person belong to the training dataset, then the program must be to back all poses related with that person. At this testing the recognition rate reach to 98.9% (see Table 1) with ORL database, but with FEI Face Database (Brazilian) degrade to 96.5% (see Table 2).*
- B. *Under classification branch where the tested image may be belong or not belong to the data set, the recognition rate may be reach to 95% based on the complexities of image.*

A Comparison: The following table illustrate a comparison among a number of existing face recognition algorithms under similar databases.

Proposed Algorithm vs. Existing Algorithms.

Paper	Algorithm(s)	Database	% Recog
Guang Dai et al. (2003)[23]	Support Vector Machines	AT&T	94.5
A. Nabatchian et al. (2008)[24]	pseudo Zernike moments	AT&T	95%
N. Farajzadeh et al. (2010)[25]	Zernike moments	FERET	94.3%
Zhan Shi et al. (2012)[26]	Pseudo-Zernike Moment	AT&T	89%
Sara Nazari et al. (2013)[27]	Global and Local Gabor Features	AT&T	91.8%
Raman Kumar et al. (2013)[28]	PCA	Indian	90%
Proposed Algorithm	Zerinke and SSIM	AT&T	98.75%,
Proposed Algorithm	Zerinke and SSIM	Brazilian	96.57%

Discussion:

When the query subject is in an unconstrained imaging environment, the true accept rate can fall from 99% to below 60% [29], so we can claim that our method attains higher accuracy than other method. Of course, it's difficult to determine the accurate rate, since the environment is un conditional (see Fig.3, where we used non-standard, self-made face images).

Experiments show that, in the image plane, when rotation exceeds 4° , the recognition rate drops rapidly, and the face image will be beyond recognition when rotation is more than 12° [30]. However, using the proposed algorithm, face recognition attains high degrees of probability in spite of such rotations, where we might reach 99%.

V. CONCLUSIONS

1) Zernike moments started with order $p=zero$ give acceptable recognition rate that reaches 94.5%. Low orders of Zernike moments are useful for face expression recognition.

2) Zernike moments started with $p=two$ give excellent recognition rate that reaches to 99%.

3) Any Zernike moments beyond five return the same results (this means that the best range for Zernike moment is order [2-5]).

4) Pure SSIM gives poor results, while modified SSIM, which combined edge detection methods with SSIM, gives better results.

5) Some of the above measurements are considered as weak measurements when they are used alone, but when combining these measurements with others excellent results are obtained.

6) Last conclusion is that: image blurring or noise does not affect these measures. In addition, they are very resistant against rotation.

REFERENCES

- [1] Arnold Wiliem, Vamsi Krishna Madasu, Wageeh Boles & Prasad Yarlagadda, "A face recognition approach using Zernike Moments for video surveillance", Queensland University of Technology website, <http://eprints.qut.edu.au>, 2007.
- [2] W. Zhao, R. Chellappa, and P. J. Phillips, "Face Recognition: A Literature Survey", ACM Computing Surveys, Vol. 35, No. 4, pp. 399-458, 2003.
- [3] Kresimir Delac, Mislav Grgic, and Sonja Grgic, "A comparative study of PCA, ICA AND LDA," International Journal of Imaging Systems and Technology, 2005.
- [4] Zhengjun Pan and Hamid Bolouri, "High Speed Face Recognition Based on Discrete Cosine Transform and Neural Networks", University of Hertfordshire, UK, 1999.
- [5] Ziad M. Hafid and Martin D. Levine, "Face Recognition Using the Discrete Cosine Transform", International Journal of Computer Vision 43(3), Kluwer Academic Publishers, 2001.
- [6] Hazim Ekenel and Rainer Stiefelwagen, "Local Appearance Based Face Recognition Using Discrete Cosine Transform", 13th European Signal Processing Conference (EUSIPCO), 2005.
- [7] Osslan Osiris Vergara Villegas, Mitzel Aviles Vianey Guadalupe Cruz Sánchez, and Humberto de, "A Novel Evolutionary Face Recognition Algorithm Using Particle Swarm Optimization", Fifth International Conference on Signal Image Technology and Internet Based Systems, 2009.
- [8] Lanzarini Laura, La Battaglia Juan, Maulini Juan, and Hasperué Waldo, "Face Recognition Using SIFT and Binary PSO Descriptors", Proceedings of the ITI 2010 32nd Int. Conf. on Information Technology Interfaces, June 21-24, Cavtat, Croatia, 2010.
- [9] Abhishek Sharma, Anamika Dubey, A. N. Jagannatha, R. S. Anand, "Pose invariant face recognition based on hybrid-global linear regression", Springer -Verlag London Limited, 2010.
- [10] Qiang Zhang and Baoxin Li, "Discriminative K-SVD for Dictionary Learning in Face Recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.
- [11] Manzoor Ahmad Lone, S. M. Zakariya, and Rashid Ali, "Automatic Face Recognition System by Combining Four Individual Algorithms", International Conference on Computational Intelligence and Communication Systems, 2011.
- [12] Zhan Shi, Guixiong Liu, and Minghui Du, "Rotary Face Recognition Based on Pseudo-Zernike Moment", ECICE 2012, AISC 146, pp. 641-646, Springer-Verlag Berlin Heidelberg, 2012.
- [13] Sukhvinder Singh, Meenakshi Sharma, N Suresh Rao, "Accurate Face Recognition Using PCA and LDA", IJMIE, Volume 2, Issue 4, 2012.
- [14] Manchula A., Arumugam S., "Robust Facial Data Recognition using multimodal Fusion Features in Multi-Variant Face Acquisition", International Journal of Computer Applications, Volume 64, No.11, 2013.
- [15] Chee-Way Chonga, P. Raveendranb, R.Mukundan, "Translation invariants of Zernike moments", Pattern Recognition 36, 2003.
- [16] Sun-Kyoo Hwang, Whoi-Yul Kim, "Anovel approach to the fast computation of Zernike moments", Pattern Recognition 39, 2006.
- [17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity". IEEE Trans. Image Proc., 13: 600-612. DOI: 10.1109/TIP.2003.819861, Simoncelli, 2004.
- [18] Asaad Noori Hashim and Zahir M. Hussain, "Novel Image Dependent Quality Assessment Measures", Journal of Computer Science 10 (9): 1548-1560, 2014.
- [19] Seyed Mehdi Lajevardi, Zahir M. Hussain, "Automatic facial expression recognition: feature extraction and selection," Signal, Image and Video Processing, v.6, no.1, pp.159-169, 2012.
- [20] Seyed Mehdi Lajevardi, Zahir M. Hussain, "Novel Higher-order Local Autocorrelation-like Feature Extraction Methodology for Facial Expression Recognition," IET Image Processing, vol. 4, no. 2, 2010.
- [21] Seyed Mehdi Lajevardi, Zahir M. Hussain, "Facial Expression Recognition Using Log-GaborFilters and Local Binary Pattern Operators," International Conference on Communication, Computer and Power (ICCCP'09), Muscat, Oman, 15-18 Feb. 2009.
- [22] Personal Homepage of Dr. Carlos Eduardo Thomaz (Caru), <http://fei.edu.br/~cet/facedatabase.html>.
- [23] Guang Dai and Changle Zhou, "Face Recognition Using Support Vector Machines with the Robust Feature", Proceedings of the 2003 IEEE International Workshop on Robot and Human interactive Communication Millbrae. Calimia, USA. Oct. 31 -Nov. 2, 2003.
- [24] A. Nabatchian, E. Abdel-Raheem and M. Ahmadi, "Human Face Recognition Using Different Moment Invariants: A Comparative Study", Congress on Image and Signal Processing. IEEE 2008.
- [25] N. Farajzadeh, K. Faez, and G. Pan, "Study on the performance of moments as invariant descriptors for practical face recognition systems", IET Comput. Vis., 2010.
- [26] Zhan Shi, Guixiong Liu and Minghui Du, "Rotary Face Recognition Based on Pseudo-Zernike Moment", Springer-Verlag Berlin Heidelberg 2012.
- [27] Sara Nazari and Mohammad-Shahram Moin, "Face Recognition Using Global and Local Gabor Features", IEEE 2013.
- [28] Raman Kumar and Satnam Singh, "Face Recognition using Principle Component Analysis for Biometric Security System", International Journal of Engineering Trends and Technology (IJETT) - Volume 4 Issue 9- Sep 2013
- [29] Joshua C. Klontz Anil K. Jain, "A Case Study on Unconstrained Facial Recognition Using the Boston Marathon Bombings Suspects", Columbia Southern University, Computer Society, November, 2013.
- [30] Shan, S.-G., Gao, W., "Curse of Miss-Alignment Problem in Face Recognition", Chinese Journal of Computers 05, 2005.

TABLE I. RESULTS OF APPLING PROPOSED ALGORITHM TO ORL DATA SET.

person	Successful poses		Failed poses		% Success		Notes
	Zernike Moment Ord 0-3	Zernike Moment Ord 2-5	Zernike Moment Ord 0-3	Zernike Moment Ord 2-5	Zernike Moment Ord 0-3	Zernike Moment Ord 2-5	
P1	10/10	10/10	-	-	100%	100%	1-Main problem with persons 4 and 10.
P2	10/10	10/10	-	-	100%	100%	
P3	09/10	10/10	6	-	90%	100%	
P4	05/10	10/10	1,3,4,7,8	-	50%	100%	
P5	10/10	09/10	-	9	100%	90%	With Zernike orders from 0-3
P6	09/10	10/10	2	-	90%	100%	
P7	10/10	10/10	-	-	100%	100%	
P8	10/10	10/10	-	-	100%	100%	
P9	10/10	9/10	-	1	100%	90%	
P10	07/10	10/10	4,5,10	-	70%	100%	
P11	10/10	10/10	-	-	100%	100%	
P12	06/10	10/10	1,4,7,8	-	60%	100%	1-Main problem with person 12 and 16.
P13	10/10	10/10	-	-	100%	100%	
P14	10/10	10/10	-	-	100%	100%	2-percentage of success is 92%.
P15	10/10	10/10	-	-	100%	100%	
P16	07/10	10/10	3,8,10	-	70%	100%	
P17	09/10	10/10	5	-	90%	100%	
P18	10/10	10/10	-	-	100%	100%	With Zernike orders from 0-3
P19	10/10	10/10	-	-	100%	100%	
P20	10/10	10/10	-	-	100%	100%	
P21	10/10	09/10	-	2	100%	90%	1-Ten poses of ten persons are succeed
P22	10/10	10/10	-	-	100%	100%	
P23	10/10	10/10	-	-	100%	100%	2-percentage of succeed is 100%
P24	10/10	10/10	-	-	100%	100%	
P25	10/10	9/10	-	8	100%	90%	
P26	10/10	10/10	-	-	100%	100%	
P27	10/10	10/10	-	-	100%	100%	With Zernike orders from 0-3
P28	10/10	10/10	-	-	100%	100%	
P29	10/10	10/10	-	-	100%	100%	
P30	10/10	10/10	-	-	100%	100%	
P31	10/10	10/10	-	-	100%	100%	1-Main problem with person 40
P32	10/10	10/10	-	-	100%	100%	
P33	10/10	10/10	-	-	100%	100%	2-percentage of succeed is 96%
P34	10/10	10/10	-	-	100%	100%	
P35	09/10	09/10	1	2	90%	90%	
P36	10/10	10/10	-	-	100%	100%	
P37	10/10	10/10	-	-	100%	100%	With Zernike orders from 0-3
P38	10/10	10/10	-	-	100%	100%	
P39	10/10	10/10	-	-	100%	100%	
P40	07/10	10/10	1,6,10	-	70%	100%	
Total	388/400	400/400	22/400	5/400	94.5%	98.75%	

TABLE II. RESULTS OF APPLING PROPOSED ALGORITHM TO BRAZILIAN DATA SET.

Person	Successful Poses	Failed Poses	Succeed %	Person	Successful Poses	Failed Poses	Succeed %
P1	14/14	-	100%	P26	12/14	11, 14	86%
P2	14/14	-	100%	P27	14/14	-	100%
P3	14/14	-	100%	P28	14/14	-	100%
P4	13/14	2	93%	P29	14/14	-	100%
P5	13/14	10	93%	P30	14/14	-	100%
P6	13/14	5	93%	P31	13/14	10	93%
P7	13/14	6	93%	P32	13/14	11	93%
P8	14/14	-	100%	P33	13/14	12	93%
P9	14/14	-	100%	P34	14/14	-	100%
P10	14/14	-	100%	P35	13/14	13	93%
P11	13/14	8	93%	P36	13/14	14	93%
P12	14/14	-	100%	P37	14/14	-	100%
P13	14/14	-	100%	P38	14/14	-	100%
P14	14/14	-	100%	P39	14/14	-	100%
P15	14/14	-	100%	P40	14/14	-	100%
P16	13/14	11	93%	P41	14/14	-	100%
P17	13/14	10	93%	P42	13/14	2	93%
P18	13/14	12	93%	P43	14/14	-	100%
P19	14/14	-	100%	P44	12/14	9,10	86%
P20	13/14	4	93%	P45	13/14	10	93%
P21	14/14	-	100%	P46	14/14	-	100%
P22	13/14	14	93%	P47	13/14	4	93%
P23	14/14	-	100%	P48	14/14	-	100%
P24	14/14	-	100%	P49	13/14	6	93%
P25	13/14	12	93%	P50	14/14	-	100%
				Total	676/700	24/700	%96.57

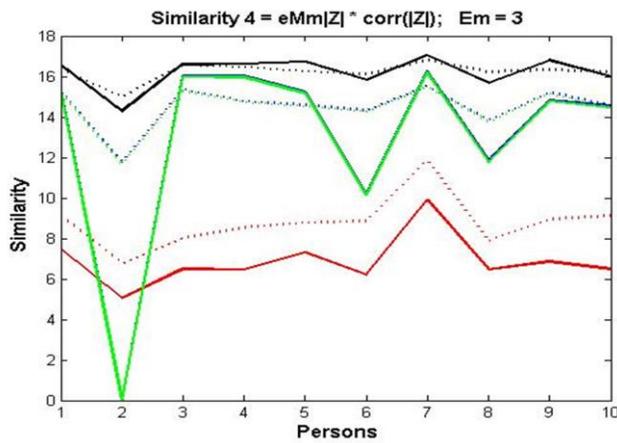
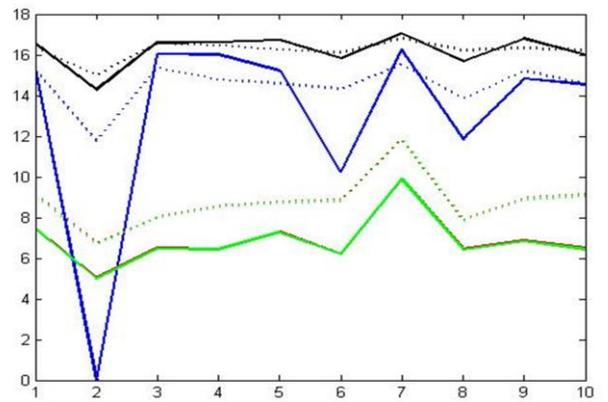
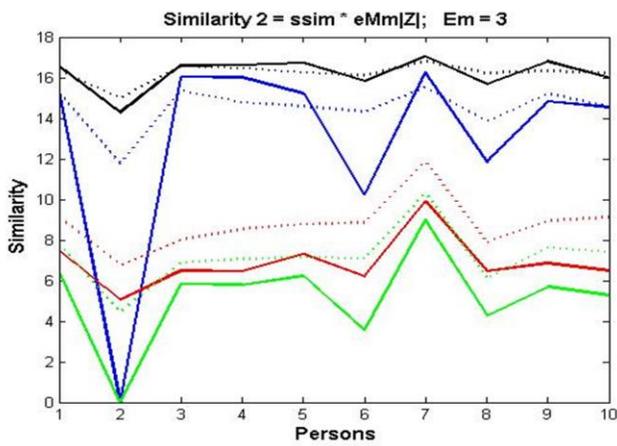
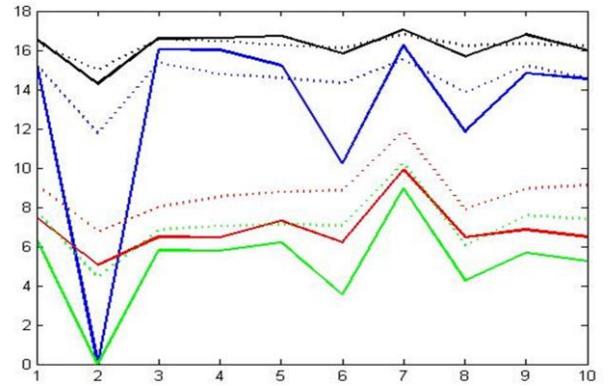
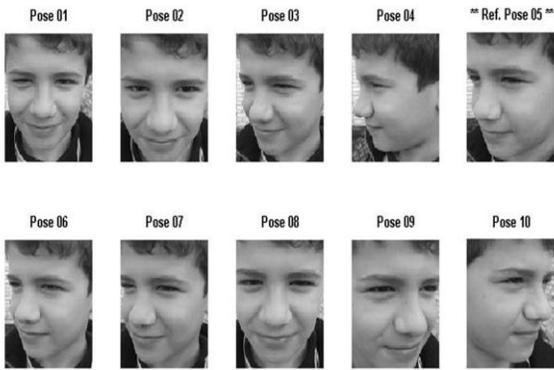


Fig. 3. Results with non-standard poses ($D = 1$, $Pe = 7$, $Pr = 0.5714$).

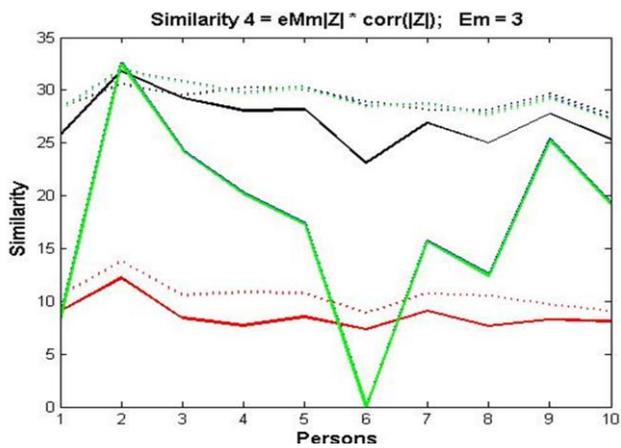
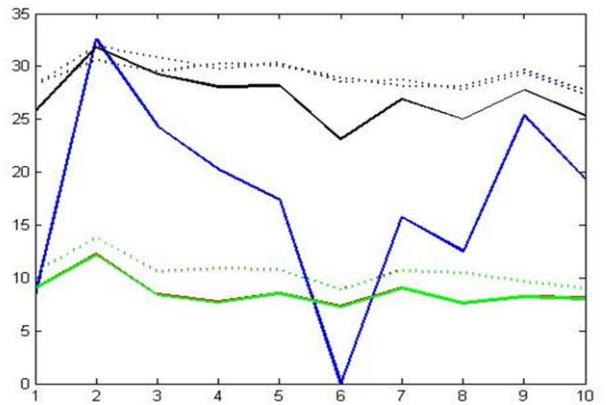
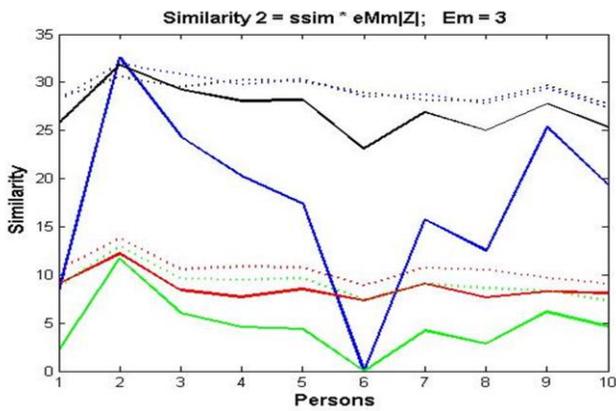
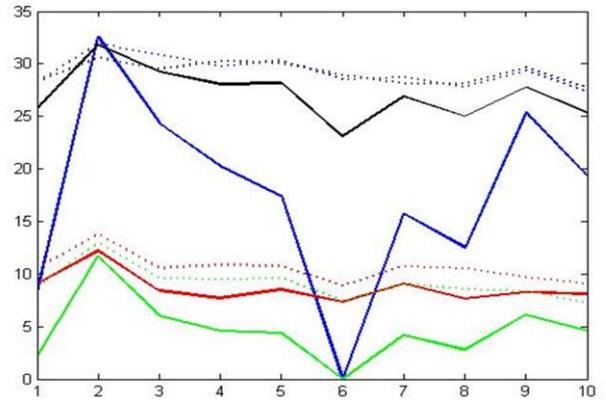
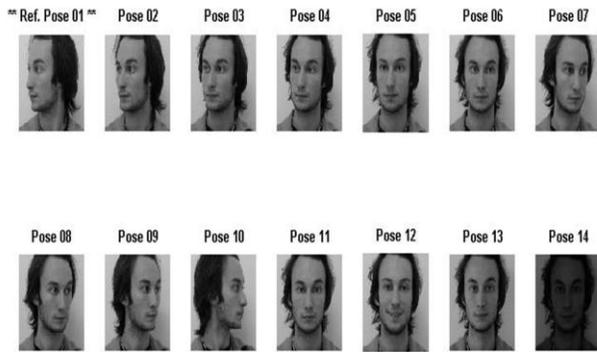


Fig. 4. Results with FEI Face Database(Brazilian) ($D = 1, Pe = 2, Pr = 0.785$)

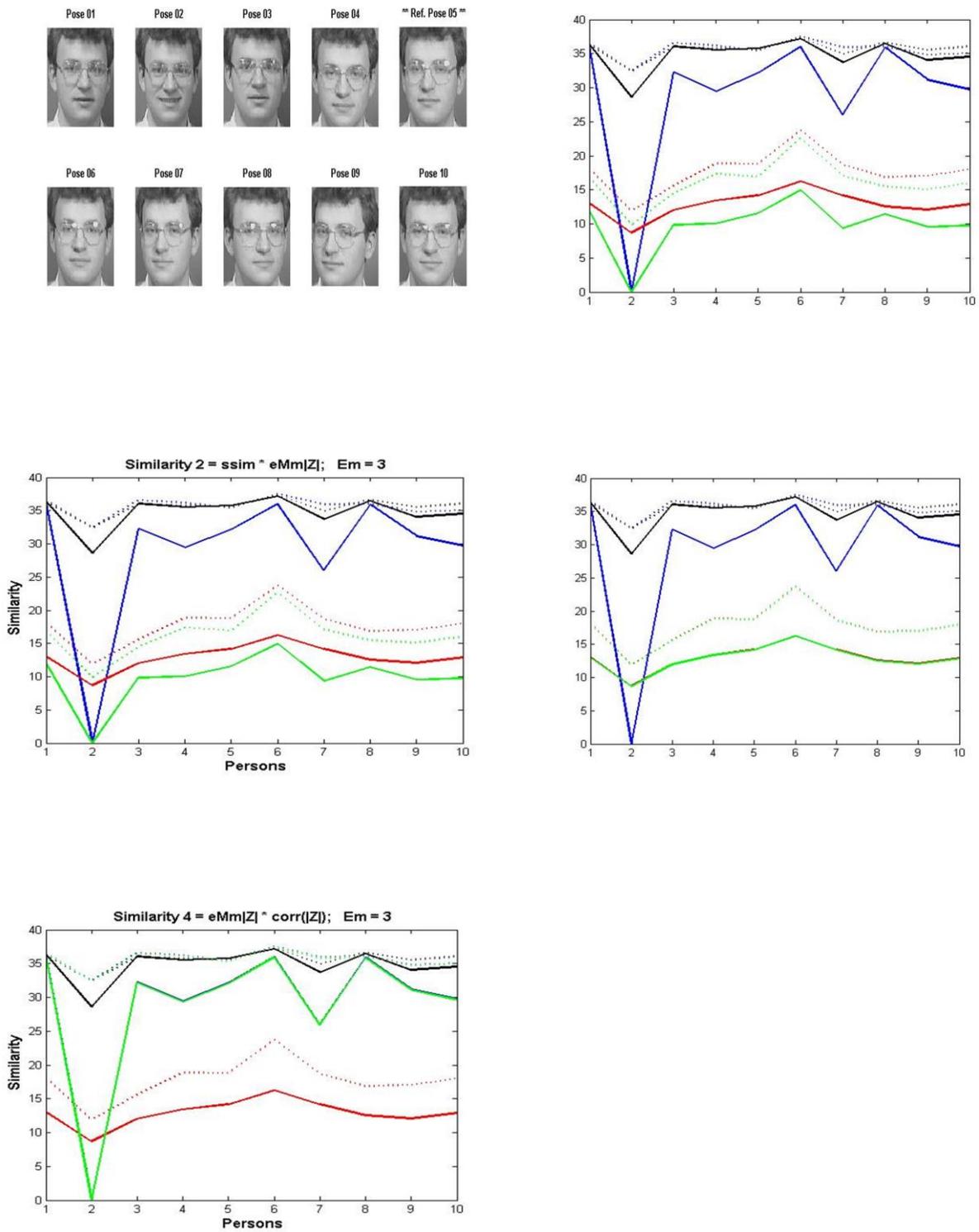


Fig. 5. Results with ORL Database ($D = 1$, $Pe = 6$, $Pr = 0.5714$)

Activity Based Learning Kits for Children in a Disadvantaged Community According to the Project “Vocational Teachers Teach Children to Create Virtuous Robots from Garbage”

Kuntida Thamwipat, Pornpapatsorn Princhankol, Thanakarn Khumphai and Vitsanu Sudsangket

Faculty of Industrial Education and Technology
King Mongkut’s University of Technology Thonburi
Bangkok, Thailand

Abstract—This research was aimed to develop and evaluate the activity based learning kits for children in a disadvantaged community according to the project “Vocational Teachers Teach Children to Create Virtuous Robots from Garbage”, to examine the learning achievement, to measure the satisfaction and to do an authentic assessment of children as regards the learning kits. The researchers chose the sampling group purposively out of children aged 4-14 years in the community under bridge zone 1 who could participate in the summer activity in the academic year 2013. The sampling group consisted of 40 children. Statistical tools in this research included mean and standard deviation. The results showed that the content quality was good ($\bar{x} = 4.40$, S.D. = 0.55), the presentation quality was good ($\bar{x} = 4.46$, S.D. = 0.68). After learning with the learning kits, the children in the disadvantaged community could achieve higher post-test score than pre-test score with statistical significance at a .05 level. The children expressed the highest level of satisfaction towards the learning kits ($\bar{x} = 4.57$, S.D. = 0.58). The authentic assessment of children as regards the learning kits was at a good level ($\bar{x} = 4.43$, S.D. = 0.51 and this complied with the hypotheses. Therefore, the activity based learning kits are useful and could be used in other nearby communities.

Keywords—Activity Based Learning Kits; Children in a Disadvantaged Community; Virtuous Robots; Garbage

I. BACKGROUND

In 1993, the cabinet decided to improve the living condition of people who live under the bridges in Bangkok and metropolitan areas because they are homeless. Bangkok Metropolitan Administration in collaboration with National Housing Authority provided over 700 families with home around the areas. The area called “Community under the Bridge Zone 1” is located around 10 kilometres away from King Mongkut’s University of Technology Thonburi. At the moment there are around 200 families and the majority or 70% of the population is itinerant junk buyers. According to an interview with the community chief Mr Chalorernsak Leewangsee [1], one of the concerns in this community is that during the time of work the parents do not have time to look after their children because they gather together to play with garbage which can be dangerous without awareness such as

firecrackers and fireworks, resulting in accidents. When the problem was taken into consideration, it was found that the main cause is that the children or the youth did not have space for recreation or activities and they lacked a good instructor. Therefore, they tended to use force or end a dispute through fights. They also did not appreciate the value of formal education, resulting in a low number of people who study beyond compulsory education.

According to the Public Welfare Education Development Plan for 5 Years (2012-2016), all disadvantaged children have a right and an opportunity to study compulsory education in accordance to their identity, maintain the quality of education through virtues and appreciation of Thai culture, and lead a life according to the self-sufficiency philosophy in the society for a happy life. According to the Office of the Education, Religion, Arts and Culture Commission, which have conducted many studies on behaviours among children and youths in relation to other people and communities [2], it was found that there is an increase in children and youth who display lack of virtues and moral conducts. Unless all the groups involved in solving the problem during the primary education, the nation could encounter serious problems when they grow up. Therefore, it is important to develop students with good character.

According to the problem and the significance mentioned above, the researchers decided to develop *the activity based learning kits* for children in a disadvantaged community according to the project “Vocational Teachers Teach Children to Create Virtuous Robots from Garbage” to mirror 8 kinds of virtues as follows: diligence, thrift, honesty, discipline, politeness, cleanliness, harmony, and generosity. Boonkuea Kuanhawet [3] says that learning kits are a kind of teaching material which is developed according to and consisted of topics, contents and experiences in each unit as part of a series for 3 stages: instruction, creation, promotion.

The researchers have been doing research in the community since 2011 and this year we were interested in the development of the activity based learning kits for children during the summer of the academic year 2013.

II. RESEARCH OBJECTIVES

- A. To develop and evaluate the activity based learning kits for children in a disadvantaged community according to the project “Vocational Teachers Teach Children to Create Virtuous Robots from Garbage”
- B. To examine the learning achievement of children who use the activity based learning kits for children in a disadvantaged community according to the project “Vocational Teachers Teach Children to Create Virtuous Robots from Garbage”
- C. To measure the satisfaction of children towards the activity based learning kits for children in a disadvantaged community according to the project “Vocational Teachers Teach Children to Create Virtuous Robots from Garbage”
- D. To do an authentic assessment of children who use the activity based learning kits for children in a disadvantaged community according to the project “Vocational Teachers Teach Children to Create Virtuous Robots from Garbage”

III. RESEARCH SCOPE

The institution responsible for the project consists of the researchers and students at Faculty of Industrial Education and Technology, King Mongkut’s University of Technology Thonburi

There were 2 periods of research as follows:

Preparation Period: from 1 December 2013 to 30 January 2014

Project Period: from 1 February 2014 to 30 May 2014

This project follows 3 stages of operation as in instruction-creation-promotion.

IV. POPULATION AND THE SAMPLING GROUP

The population in this research consisted of 110 children in the community under the bridge zone 1 at Pracha Uthid 76 Road who were 4-14 years of age [1]. The researchers chose the sampling group purposively out of those who could participate in the summer activity. There were 40 children in total.

V. RESEARCH HYPOTHESES

- A. The quality of the activity based learning kits for children in a disadvantaged community according to the project “Vocational Teachers Teach Children to Create Virtuous Robots from Garbage” would be at a good level.
- B. The learning achievement of the children who use the activity based learning kits for children in a disadvantaged community according to the project “Vocational Teachers Teach Children to Create Virtuous Robots from Garbage” would show higher post-test score than pre-test score with statistical significance at the .05 level.
- C. The satisfaction of the children towards the activity based learning kits for children in a disadvantaged community according to the project “Vocational Teachers

Teach Children to Create Virtuous Robots from Garbage” would be at a high level.

- D. The authentic assessment of the children who use the activity based learning kits for children in a disadvantaged community according to the project “Vocational Teachers Teach Children to Create Virtuous Robots from Garbage” would be at a good level.

VI. TOOLS FOR DATA COLLECTION

Tools in this research consisted of the quality evaluation form for the experts in contents and presentation learning activities for children, the learning achievement test, the children satisfaction questionnaire and the authentic assessment form.

VII. STATISTICAL METHODS USED

The statistical methods in this research were Mean and Standard Deviation.

VIII. RESEARCH RESULTS

The results from the development and the evaluation of the quality of the activity based learning kits for children in a disadvantaged community according to the project “Vocational Teachers Teach Children to Create Virtuous Robots from Garbage” could be seen as follows:

TABLE I. THE PROCESS OF DEVELOPMENT FOR THE ACTIVITY BASED LEARNING KITS FOR CHILDREN IN A DISADVANTAGED COMMUNITY ACCORDING TO THE PROJECT “VOCATIONAL TEACHERS TEACH CHILDREN TO CREATE VIRTUOUS ROBOTS FROM GARBAGE”

Step	Researchers	Outcome	Resource
Promoting the new project within the community	Creating pamphlets and visiting the community	Community was informed of the project and well-prepared for the activities	1. Instruction
Visiting the area and teaching the children in the disadvantaged community according to the activity based learning kits for 10 weeks	Visiting the area and implementing the research through video clips, plays, storytelling, painting, excursion and basic robot creation	Children in the community could create their robots which have 8 virtues as in diligence, thrift, honesty, discipline, politeness, cleanliness, harmony and generosity.	
Creating Virtuous Robots	Researchers, volunteers and children in the community create the robots which have 8 virtues for children in the nursery	Robots have 8 virtues	2. Creation

Step	Researchers	Outcome	Resource
Distributing the projects through TV, free TV channels and national newspaper	Presenting the virtuous robots created by children in the disadvantaged community	Children build up their self-esteem through sharing their experiences with the main media	3.Promotion

TABLE II. THE QUALITY EVALUATION OF THE ACTIVITY BASED LEARNING KITS ACCORDING TO THE EXPERT PANEL IN CONTENTS.

Item	\bar{x}	S.D.	Meaning
1.The contents about 8 virtues	4.33	0.80	Good
2. The presentation and activities	4.20	0.50	Good
3. The evaluation form	4.67	0.35	Very Good
Overall	4.40	0.55	Good

The mean score of the content quality was 4.40 on average with standard deviation of 0.55. When this value was compared to the criteria, it was at a good level, or supporting the hypothesis.

TABLE III. THE MEAN SCORE FOR THE QUALITY OF ACTIVITY BASED LEARNING KITS AS EVALUATED BY THE EXPERT PANEL IN PRESENTATION

Item	\bar{x}	S.D.	Meaning
1.Content and activity presentation	4.28	0.57	Good
2. Activity step 1: Instruction	4.44	0.85	Good
3. Activity step 2: Creation	4.65	0.50	Very Good
4. Activity step 3: Promotion	4.41	0.69	Good
5. Linking activities from each learning unit	4.52	0.81	Very Good
Overall	4.46	0.68	Good

The mean score of the presentation quality was 4.46 with standard deviation of 0.68. When the value was compared to the criteria, it was found to be at a good level, or supporting the hypothesis.

Photos below show the activity based learning kits for children in the disadvantaged community according to the project “Vocational Teachers Teach Children to Create Robots from Garbage” in relation to the 3 stages.

A. Instruction

Training for children in the disadvantaged community during the summer for 10 weeks



B. Creation

Robots with 8 virtues , 8 characters design





C. Promotion

Distribution of the projects through TV, free TV channels, and national newspaper



TABLE IV. THE LEARNING ACHIEVEMENT OF THE SAMPLING GROUP

Test	n	\bar{x}	S.D.	\bar{D}	S.D. _d	t
Pre-test	40	10.80	2.00	5.9	2.92	12.95*
Post-test	40	16.70	2.23			

* with statistical significance at the .05 level

According to the research results, it was found that the children who use the activity based learning kits for children in a disadvantaged community according to the project “Vocational Teachers Teach Children to Create Virtuous Robots from Garbage” showed higher post-test score than pre-test score with statistical significance at the .05 level, supporting the hypothesis.

TABLE V. THE CHILDREN'S SATISFACTION TOWARDS THE ACTIVITY BASED LEARNING KITS FOR CHILDREN IN A DISADVANTAGED COMMUNITY ACCORDING TO THE PROJECT "VOCATIONAL TEACHERS TEACH CHILDREN TO CREATE VIRTUOUS ROBOTS FROM GARBAGE"

Item	\bar{x}	S.D.	Meaning
1.Content and activity presentation	4.48	0.64	High
2. Activity step 1: Instruction	4.62	0.56	The Highest
3. Activity step 2: Creation	4.56	0.61	The Highest
4. Activity step 3: Promotion	4.62	0.56	The Highest
5. Benefits from activities in each learning unit	4.58	0.53	The Highest
Overall	4.57	0.58	The Highest

The mean score for the children's satisfaction towards the activity based learning kits was 4.57 with standard deviation of 0.58. When it was compared to the criteria, it was at the highest level.

TABLE VI. SHOWS THE MEAN SCORE OF THE AUTHENTIC ASSESSMENT OF CHILDREN AS REGARDS THE ACTIVITY BASED LEARNING KITS FOR CHILDREN IN A DISADVANTAGED COMMUNITY ACCORDING TO THE PROJECT "VOCATIONAL TEACHERS TEACH CHILDREN TO CREATE VIRTUOUS ROBOTS FROM GARBAGE"

Item	\bar{x}	S.D.	Meaning
Authentic assessment of children	4.43	0.51	Good

The mean score of the authentic assessment in children as regards the activity based learning kits was 4.43 with standard deviation of 0.51. When the value was compared with the criteria, it was at a good level, supporting the hypothesis.

IX. DISCUSSIONS

This research into the activity based learning kits for children in a disadvantaged community according to the project "Vocational Teachers Teach Children to Create Virtuous Robots from Garbage" was considered to be of good quality in terms of contents with the mean score of 4.40 and standard deviation of 0.55 and of good quality in terms of presentation with the mean score of 4.46 and standard deviation of 0.68. This was because the researchers adopted the principle proposed by Boonkuea Kuanhawet [3] along with the quality evaluation procedure for the learning kits as proposed by Wichai Wongyai [4] who put an emphasis on holistic thinking and attention by instructors for the development. The post-test score was higher than the pre-test score with statistical significance at the .04 level, supporting the hypothesis. This was in compliance with the research by Parinya Ubonkarn [5] who conducted a study into the development of games-based activities for mathematical preparation among primary school students at Thairat Wittaya 22 School (Tai Rom Yen) and showed similar level of learning achievement.

In terms of satisfaction, the children in this study expressed the highest level of satisfaction with the mean score of 4.57 and standard deviation of 0.58. The mean score of the authentic assessment for children as regards the activity based learning kits was 4.43 with standard deviation of 0.51, or at a good level and this supported the hypothesis. This was because the children in the study had many opportunities to learn and create their own works in accordance with the steps provided by the researchers in relation to the principle by Skinner [6] in that allowing learners to have freedom in thinking could lead to satisfaction towards learning.

X. SUGGESTIONS

- A. *The research into the activity based learning kits for children in a disadvantaged community according to the project "Vocational Teachers Teach Children to Create Virtuous Robots from Garbage" was of good quality. Therefore, this project could be expanded to other nearby communities such as the community behind Thonburi Rom Park.*
- B. *The research results showed that the learners showed the highest level of satisfaction towards the activity steps from instruction, creation to promotion. In this research, student volunteers worked closely with the researchers in each step. Therefore, in the future activities, there should be an announcement to ask for student volunteers to participate in activities to improve disadvantaged communities.*

ACKNOWLEDGMENT

The researchers would like to extend the gratitude to the funding by King Mongkut's University of Technology Thonburi and Red Bull U-Project.

REFERENCES

- [1] Chaloesmsak Leewangsee, Community Chief of Community under the Bridge Zone 1 at Pracha Utit 76 Road, Interview, 15 April 2013
- [2] Office of the Education, Religion, Arts and Culture Commission, 2005, Projects to Improve Virtues and Moral Conducts of Students. The Secretariat of the Senate Printing House, Bangkok, p. 35.
- [3] Boonkuea Kuanhawet. 1999. Educational Innovation . Chulalongkorn University Press, Bangkok, p. 40.
- [4] Wichai Wongyai. 1982. Curriculum Development and Instruction: A New Dimension. Suweeriyasarn. Bangkok, p. 123.
- [5] Parinya Ubonkarn. 2011 . The Development of Games-Based Activities for Mathematical Preparation among Primary School Students at Thairat Wittaya 22 School (Tai Rom Yen), a Master's dissertation in Industrial Education, King Mongkut's University of Technology Thonburi, p.174.
- [6] Skinner , BF..1971 . Beyond Freedom and Dignity. Toronto . A Bantam Vintage Boo, p. 127.

Reconsideration of Potential Problems of Applying EMIM for Text Analysis

D. Cai

School of Computing and Engineering
University of Huddersfield, HD1 3DH, UK
Email: d.cai@hud.ac.uk

Abstract—It seems that the term dependence methods developed using the expected mutual information measure (EMIM) have not achieved their potential in many areas of science, involving statistical text analysis or document processing. This study examines the reasons for the failure and highlights potential problems of applications. Several interesting questions are arisen, including, does a term provide any information if it occurs in all the sample documents? how the mutual information of two terms, under their status values, makes contribution to EMIM? are two terms highly dependent for their co-occurrence if they receive a high positive EMIM value? what may imply for dependence of two term pairs when they receive the same EMIM value? how can properly verify two terms to be high dependent for their co-occurrence? how can properly apply EMIM? does the size of the sample set matter? This study attempts to answer these questions in order to clarify confusions caused by the problems and/or suggest solutions to the problems. Some interesting examples are provided to clarify our viewpoints.

Index Terms—text analysis; term dependence; term co-occurrence; the expected mutual information measure (EMIM).

I. INTRODUCTION

The *expected mutual information measure* (EMIM) quantifies how much knowing one of two variables reduces our uncertainty about the other. The effectiveness of measuring the *mutual information of terms* (MIT) is an active research subject in many areas of science. This subject has been motivated by the concern: to developed a variety of techniques in order to assign a ‘dependence’ (‘relatedness’, ‘proximity’, ‘association’) value to each term pair, and then to make some decision based on those values. Many studies have used EMIM for a variety of tasks in, for instance, feature selection [1]–[4], document classification [5], face image clustering [6], noise and redundancy reduction [7], multi-modality image registration [8], information retrieval [9]–[13].

Despite the attractiveness of EMIM, however, it seems that the term dependence methods developed using EMIM have not achieved their potential. There may be two main issues for this. First, it is practically difficult to estimate the probability distributions required in EMIM. Second, different estimations conclude to different properties of EMIM and it is theoretically challenging to apply EMIM without clearly understanding the properties. This study focuses on the second issue.

There exist potential problems in applying EMIM. This study examines the reasons for the failure by analysing the

properties, particularly when considering the binary probability estimation, denoted by $P_{\Xi}(\delta_i)$ and $P_{\Xi}(\delta_i, \delta_j)$, widely used in many areas of science. We highlight eight problems through respective eight questions below: For two arbitrary distinct terms t_i and t_j (where $I(\delta_i; \delta_j)$ is EMIM and $emim(\delta_i; \delta_j)$ is a simplified form, which will be given in the next section),

- Q1: does t_i provide any information on t_j if it occurs in all the sample documents?
- Q2: what is a fact given from the relation between $I(\delta_i; \delta_j)$ and $emim(\delta_i; \delta_j)$?
- Q3: how the mutual information of t_i and t_j , under their status values, makes contribution to $I(\delta_i; \delta_j)$?
- Q4: are t_i and t_j highly dependent for their co-occurrence if (t_i, t_j) receives a high positive value of $I(\delta_i; \delta_j)$?
- Q5: what may imply for dependence of two term pairs (t_i, t_j) and (t'_i, t'_j) when $I(\delta_i; \delta_j) = I(\delta'_i; \delta'_j)$?
- Q6: how can properly verify t_i and t_j to be high dependent for their co-occurrence?
- Q7: how can properly apply $emim(\delta_i; \delta_j)$?
- Q8: does the size of the sample set matter?

This study attempts to answer the above questions in order to clarify confusions caused by the problems and/or suggest solutions to the problems. As it will be seen from this study, for instance, the occurrence of a term in all samples (which may be regarded as a good term in some applications) does not provide any information about the occurrence of other terms in the samples; two terms receiving a high positive EMIM value may not be necessarily high dependent for their co-occurrence; two term pairs receiving the same EMIM value may be dependent of each other in different implications; an inequality has to be verified, in order to properly apply EMIM or $emim$, to ensure two terms are high dependent for their co-occurrence. Some interesting examples are provided to clarify our viewpoints, and each question Q k is answered through a corresponding remark Remark- k ($k = 1, 2, \dots, 8$).

The remainder of the paper is organized as follows. Section 2 gives notation, the expressions of EMIM and $emim$. Section 3 considers the properties of EMIM and answers Q1 and Q2. Section 4 analyses the properties of four MIT measures, derived from EMIM, and answers Q3–Q7. Section 5 explains the sensitivity to the size of the sample set and answers Q8. Conclusions are drawn in Section 6 and detailed proofs of all the theorems given in this study are presented in Appendix.

II. BACKGROUND

This section gives notation, expressions of EMIM and its simplified form.

Let D be a collection of documents, $\Xi \subseteq D$ a sample set of documents interested, and V be a vocabulary of terms used to index individual documents in D . Denote $V_d \subseteq V$ as the set of terms occurring in document d , and $V_{\Xi} \subseteq V$ as the set of terms occurring in at least one of sample documents in Ξ .

In order to clarify our idea presented in this study, let us first give term state value distributions. A term is usually thought of having its state values *present* or *absent* in a document or a set of documents. For an arbitrary term $t \in V$, it will be convenient to introduce a variable δ taking values from set $\Omega = \{1, 0\}$, where $\delta = 1$ expresses that t is present and $\delta = 0$ expresses that t is absent. Denote $t^{\delta} = t, \bar{t}$ when $\delta = 1, 0$, respectively. We call $\Omega = \{1, 0\}$ a *state value space*, and each element in Ω a *state value*, of the term t . Thus, for a given term $t \in V_d$, its state distribution, denoted by $P_d(\delta) = P(t^{\delta}|d)$, is over Ω . Similar discussions can be given to $P_{\Xi}(\delta) = P(t^{\delta}|\Xi)$ over Ω for $t \in V_{\Xi}$, and to $P_{\Xi}(\delta_i, \delta_j) = P(t_i^{\delta_i}, t_j^{\delta_j}|\Xi)$ over $\Omega \times \Omega = \{(1, 1), (1, 0), (0, 1), (0, 0)\}$ for $(t_i, t_j) \in V_{\Xi} \times V_{\Xi}$ (where $i \neq j$).

There exists dependence between two terms if the state value of one of them provides mutual information about the probability of the state value of another. The study [14] also showed that there is a relationship between the frequencies (or probabilities) and the mutual information of terms. Therefore, term t taking some state value δ should be looked upon as complex because another state value of t , and state values of many other terms, may be dependent on this state value [?].

To enable to analyse and understand the properties of EMIM and its a simplified form, let us further denote $n_{\Xi}(t)$ as the number of samples in Ξ in which t occurs, and $n_{\Xi}(t_i, t_j)$ as the number of samples in Ξ in which t_i and t_j co-occur (where $i \neq j$). Then, under the binary assumption, using the statistics of the sample frequencies concerning the set Ξ , we can introduce the following two theorems, which are essential for estimating probability distributions required in EMIM.

Theorem 2.1 For an arbitrary term $t \in V$, the state value distribution, denoted by $P_{\Xi}(\delta)$, given by

$$\begin{aligned} P_{\Xi}(\delta = 1) &= P_{\Xi}(t) = \frac{n_{\Xi}(t)}{|\Xi|} \\ P_{\Xi}(\delta = 0) &= P_{\Xi}(\bar{t}) = 1 - \frac{n_{\Xi}(t)}{|\Xi|} \end{aligned} \quad (1)$$

is a probability distribution over Ω . For two arbitrary distinct terms $t_i, t_j \in V$, the state value distribution, denoted by $P_{\Xi}(\delta_i, \delta_j)$, given by

$$\begin{aligned} P_{\Xi}(\delta_i = 1, \delta_j = 1) &= P_{\Xi}(t_i, t_j) = \frac{n_{\Xi}(t_i, t_j)}{|\Xi|} \\ P_{\Xi}(\delta_i = 1, \delta_j = 0) &= P_{\Xi}(t_i, \bar{t}_j) = \frac{n_{\Xi}(t_i, t_j)}{|\Xi|} \\ P_{\Xi}(\delta_i = 0, \delta_j = 1) &= P_{\Xi}(\bar{t}_i, t_j) = \frac{n_{\Xi}(t_j) - n_{\Xi}(t_i, t_j)}{|\Xi|} \\ P_{\Xi}(\delta_i = 0, \delta_j = 0) &= P_{\Xi}(\bar{t}_i, \bar{t}_j) \\ &= \frac{|\Xi| - n_{\Xi}(t_i) - n_{\Xi}(t_j) + n_{\Xi}(t_i, t_j)}{|\Xi|} \end{aligned} \quad (2)$$

is a probability distribution over $\Omega \times \Omega$. And $P_{\Xi}(\delta_i)$ and $P_{\Xi}(\delta_j)$ are the marginal distributions of $P_{\Xi}(\delta_i, \delta_j)$.

Theorem 2.2 For two arbitrary distinct terms $t_i, t_j \in V$, suppose $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i, \delta_j)$ are given in Eq.(1) and Eq.(2), respectively. Then $P_{\Xi}(\delta_i, \delta_j)$ is absolutely continuous with respect to product $P_{\Xi}(\delta_i)P_{\Xi}(\delta_j)$ for $\delta_i, \delta_j = 1, 0$.

With Theorems 2.1 and 2.2, we can now substitute Eq.(1) and Eq.(2) into EMIM:

$$I_{\Xi}(\delta_i; \delta_j) = \sum_{\delta_i, \delta_j=0,1} P_{\Xi}(\delta_i, \delta_j) \ln \frac{P_{\Xi}(\delta_i, \delta_j)}{P_{\Xi}(\delta_i)P_{\Xi}(\delta_j)} \quad (3)$$

where \ln is the natural logarithm, which measures the amount of information that δ_j provides about δ_i , and vice versa.

In order to give a simplified form of EMIM, denoted by $emim_{\Xi}(\delta_i; \delta_j)$, let us adopt the notation given in [15]:

$$\begin{aligned} n_{1.} &= n_{\Xi}(t_i) \\ n_{.1} &= n_{\Xi}(t_j) \\ n_{11} &= n_{\Xi}(t_i, t_j) \\ n_{10} &= n_{\Xi}(t_i) - n_{\Xi}(t_i, t_j) \\ n_{01} &= n_{\Xi}(t_j) - n_{\Xi}(t_i, t_j) \\ n_{0.} &= |\Xi| - n_{\Xi}(t_i) \\ n_{.0} &= |\Xi| - n_{\Xi}(t_j) \\ n_{00} &= |\Xi| - n_{\Xi}(t_i) - n_{\Xi}(t_j) + n_{\Xi}(t_i, t_j) \end{aligned} \quad (4)$$

Then we can write

$$\begin{aligned} emim_{\Xi}(\delta_i; \delta_j) &= n_{11} \ln \frac{n_{11}}{n_{1.}n_{.1}} + n_{10} \ln \frac{n_{10}}{n_{1.}n_{.0}} + \\ & n_{01} \ln \frac{n_{01}}{n_{0.}n_{.1}} + n_{00} \ln \frac{n_{00}}{n_{0.}n_{.0}} \end{aligned} \quad (5)$$

which is well-known to many researchers, in particular, to information retrieval (IR) researchers. It was initially introduced by van Rijsbergen in his earlier book and papers [15], [16].

We will give the relation between EMIM and $emim$ and provide an example to illustrate the computation involved in EMIM and $emim$ in next section. In what follows, we will always assume, when mentioning two arbitrary terms $t_i, t_j \in V$, that they are distinct terms (i.e., $i \neq j$).

III. PROPERTIES OF EMIM

In order to enable us to gain an insight into $I_{\Xi}(\delta_i; \delta_j)$ and $emim_{\Xi}(\delta_i; \delta_j)$, this section introduces three theorems. These give interesting properties of EMIM and $emim$, and then give answers to questions Q1 and Q2.

Theorem 3.1 For two arbitrary terms $t_i, t_j \in V$, suppose $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i, \delta_j)$ are given in Eq.(1) and Eq.(2), respectively. Then $I_{\Xi}(\delta_i; \delta_j) = 0$ if $n_{\Xi}(t_i) = |\Xi|$ or $n_{\Xi}(t_j) = |\Xi|$.

Remark-1: Theorem 3.1 tells us, when EMIM is used with the estimation given Eq.(1) and Eq.(2), that the occurrence of t_i in all samples does not provide any information about the occurrence of t_j in the samples. Thus, t_i and t_j are statistically independent of one another with respect to Ξ . Consequently, in order to capture the dependence information of terms, we should always avoid many terms having $n_{\Xi}(t) = |\Xi|$ and take the sample set Ξ with a relatively larger size satisfying, for instance,

$$|\Xi| \geq \alpha + \beta \times \max\{n_{\Xi}(t) \mid t \in V_{\Xi}\}$$

where $\alpha, \beta \geq 1$ are integers. \diamond

Theorem 3.2 For two arbitrary terms $t_i, t_j \in V$, suppose $I_{\Xi}(\delta_i; \delta_j)$ and $emim_{\Xi}(\delta_i; \delta_j)$ are given in Eq.(3) and Eq.(5), respectively. Then

$$I_{\Xi}(\delta_i; \delta_j) = \frac{1}{n} \times emim_{\Xi}(\delta_i; \delta_j) + \ln(n) \quad (6)$$

where $n = |\Xi|$.

Remark-2: Theorem 3.2 gives the relation between $I_{\Xi}(\delta_i; \delta_j)$ and $emim_{\Xi}(\delta_i; \delta_j)$. Many applications use $emim_{\Xi}(\delta_i; \delta_j)$, rather than $I_{\Xi}(\delta_i; \delta_j)$, as a scale factor $\frac{1}{n}$ and a constant $\ln(n)$ are independent of all term pairs $(t_i, t_j) \in V \times V$, and thus they are eliminated for simplifying computation. It is clear that an essential difference between Eq.(3) and Eq.(5) is: the former is normalized by n but the latter is not. An important fact given by the above relation to notice is: $I_{\Xi}(\delta_i; \delta_j) \geq 0$ cannot infer $emim_{\Xi}(\delta_i; \delta_j) \geq 0$. Theorem 3.3 below is interesting. \diamond

Theorem 3.3 For two arbitrary terms $t_i, t_j \in V$, suppose $emim_{\Xi}(\delta_i; \delta_j)$ is given in Eq.(5). Then $emim_{\Xi}(\delta_i; \delta_j) \leq 0$.

Example 3.1 Suppose $\Xi = \{d_1, d_2, d_3\} \subseteq D$ is a sample set, $V_{d_1} = \{t_1, t_2, t_4, t_5, t_6, t_8\}$, $V_{d_2} = \{t_1, t_3, t_4, t_5, t_6, t_7\}$ and $V_{d_3} = \{t_2, t_4, t_6\}$. From $n_{\Xi}(t_1, t_2) = 1$, $n_{\Xi}(t_1) = 2$ and $n_{\Xi}(t_2) = 2$, we have

$$\begin{aligned} I_{\Xi}(\delta_1; \delta_2) &= \frac{1}{3} \ln \frac{\frac{1}{\frac{2}{3}}}{\frac{2}{3}} \\ &\quad + \frac{2-1}{3} \ln \frac{\frac{2-1}{3}}{\frac{2}{3}(1-\frac{2}{3})} \\ &\quad + \frac{2-1}{3} \ln \frac{\frac{2-1}{3}}{(1-\frac{2}{3})^2} \\ &\quad + \frac{3-2-2+1}{3} \ln \frac{\frac{3-2-2+1}{3}}{(1-\frac{2}{3})(1-\frac{2}{3})} \\ &= \frac{1}{3} \ln \frac{3}{4} + \frac{1}{3} \ln \frac{3}{2} + \frac{1}{3} \ln \frac{3}{2} + 0 \ln 0 \\ &\approx -0.0959 + 0.1352 + 0.1352 - 0.0000 \\ &= 0.1745 \\ emim_{\Xi}(\delta_1; \delta_2) &= 1 \times \ln \frac{1}{2 \times 2} \\ &\quad + (2-1) \ln \frac{2-1}{2 \times (3-2)} \\ &\quad + (2-1) \ln \frac{2-1}{(3-2) \times 2} \\ &\quad + (3-2-2+1) \ln \frac{3-2-2+1}{(3-2) \times (3-2)} \\ &= \ln \frac{1}{4} + \ln \frac{1}{2} + \ln \frac{1}{2} + 0 \ln \frac{0}{1} \\ &\approx -1.3863 - 0.6931 - 0.6931 - 0.0000 \\ &= -2.7725. \end{aligned}$$

Also, with the expression given in Eq.(6), we can see

$$\begin{aligned} &\frac{1}{3} \times emim_{\Xi}(\delta_1; \delta_2) + \ln(3) \\ &\approx \frac{1}{3} \times (-2.7725) + 1.0986 \\ &\approx 0.1745 = I_{\Xi}(\delta_1; \delta_2) \end{aligned}$$

which verifies the relation between $I_{\Xi}(\delta_1; \delta_2)$ and $emim_{\Xi}(\delta_1; \delta_2)$ for terms t_1 and t_2 . \triangle

IV. PROPERTIES OF MIT MEASURES

This section gives four measures of mutual information of terms (MIT), and then clarifies our viewpoints, which are used for answering questions Q3–Q7. The answers are essential for guiding practical applications.

Following the studies in [17] [18], we express EMIM given in Eq.(3) with the sum of four items,

$$\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j}) = P_{\Xi}(\delta_i, \delta_j) \ln \frac{P_{\Xi}(\delta_i, \delta_j)}{P_{\Xi}(\delta_i)P_{\Xi}(\delta_j)} \quad (7)$$

where $\delta_i, \delta_j = 0, 1$, each of which can be regarded as ‘mutual information of terms, t_i and t_j , in support of dependence rejecting independence under state value (δ_i, δ_j) . Thus, we can regard it as a general form of a MIT measure, computing the extent of the contributions made by t_i and t_j under the corresponding state values to $I_{\Xi}(\delta_i; \delta_j)$. The four MIT measures and example below enable a simple answer to the third question.

Example 4.1 Substituting the probability distributions given in Eq.(1) and Eq.(2) into the MIT measure in Eq.(7), we can write four concrete MIT measures for $\delta_i, \delta_j = 0, 1$. For instance, taking $\delta_i = 1$ and $\delta_j = 1$, we can write the first item of $I_{\Xi}(\delta_i; \delta_j)$:

$$\begin{aligned} \mathbf{mit}_{\Xi}(t_i, t_j) &= \mathbf{mit}_{\Xi}(t_i^{\delta_i=1}, t_j^{\delta_j=1}) \\ &= P_{\Xi}(t_i, t_j) \ln \frac{P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_i)P_{\Xi}(t_j)} \\ &= \frac{n_{\Xi}(t_i, t_j)}{|\Xi|} \ln \left(\frac{\frac{n_{\Xi}(t_i, t_j)}{|\Xi|}}{\frac{n_{\Xi}(t_i)}{|\Xi|} \frac{n_{\Xi}(t_j)}{|\Xi|}} \right) \end{aligned} \quad (8)$$

which is the MIT measure of terms t_i and t_j for their occurrence in Ξ . Also, if taking $\delta_i = 1$ but $\delta_j = 0$, then we have the second item of $I_{\Xi}(\delta_i; \delta_j)$:

$$\begin{aligned} \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) &= \mathbf{mit}_{\Xi}(t_i^{\delta_i=1}, t_j^{\delta_j=0}) \\ &= P_{\Xi}(t_i, \bar{t}_j) \ln \frac{P_{\Xi}(t_i, \bar{t}_j)}{P_{\Xi}(t_i)P_{\Xi}(\bar{t}_j)} \\ &= \frac{n_{\Xi}(t_i) - n_{\Xi}(t_i, t_j)}{|\Xi|} \ln \left(\frac{\frac{n_{\Xi}(t_i) - n_{\Xi}(t_i, t_j)}{|\Xi|}}{\frac{n_{\Xi}(t_i)}{|\Xi|} (1 - \frac{n_{\Xi}(t_j)}{|\Xi|})} \right) \end{aligned}$$

which is the MIT measure of term t_i occurring but term t_j not occurring in Ξ . \triangle

Remark-3: The expressions Eq.(3) and Eq.(7) tell us, in order to measure the term mutual information, we have to consider the mutual information under the individual state values. That is, we need to measure the extent of the contribution made by the respective four state value pairs (δ_i, δ_j) using the corresponding measure $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$, where $\delta_i, \delta_j = 0, 1$, to the expected mutual information. \diamond

Generally, each MIT measure, $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$, can be positive or negative (which can be seen in Example 3.1). The following theorem, which considers the relation between $\frac{n_{\Xi}(t_i, t_j)}{|\Xi|}$ and $\frac{n_{\Xi}(t_i)}{|\Xi|} \frac{n_{\Xi}(t_j)}{|\Xi|}$, is interesting.

Theorem 4.1 For two arbitrary terms $t_i, t_j \in V$, the four measures, $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$, where $\delta_i, \delta_j = 0, 1$, given in Eq.(7) have the following property.

- (1) if $\frac{n_{\Xi}(t_i, t_j)}{|\Xi|} = \frac{n_{\Xi}(t_i)}{|\Xi|} \frac{n_{\Xi}(t_j)}{|\Xi|}$ then
 $\mathbf{mit}_{\Xi}(t_i, t_j) = 0, \quad \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) = 0,$
 $\mathbf{mit}_{\Xi}(\bar{t}_i, t_j) = 0, \quad \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) = 0.$
- (2) if $\frac{n_{\Xi}(t_i, t_j)}{|\Xi|} > \frac{n_{\Xi}(t_i)}{|\Xi|} \frac{n_{\Xi}(t_j)}{|\Xi|}$ then
 $\mathbf{mit}_{\Xi}(t_i, t_j) > 0, \quad \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) \leq 0,$
 $\mathbf{mit}_{\Xi}(\bar{t}_i, t_j) \leq 0, \quad \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) > 0.$
- (3) if $\frac{n_{\Xi}(t_i, t_j)}{|\Xi|} < \frac{n_{\Xi}(t_i)}{|\Xi|} \frac{n_{\Xi}(t_j)}{|\Xi|}$ then
 $\mathbf{mit}_{\Xi}(t_i, t_j) < 0, \quad \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) \geq 0,$
 $\mathbf{mit}_{\Xi}(\bar{t}_i, t_j) \geq 0, \quad \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) < 0.$

Remark-4: By the property given in Theorem 4.1, it can be easily seen, when $\frac{n_{\Xi}(t_i, t_j)}{|\Xi|} < \frac{n_{\Xi}(t_i)}{|\Xi|} \frac{n_{\Xi}(t_j)}{|\Xi|}$, that the positive value $I_{\Xi}(\delta_i; \delta_j)$ is dominated by the positive quantities $\mathbf{mit}_{\Xi}(t_i, \bar{t}_j)$ and/or $\mathbf{mit}_{\Xi}(\bar{t}_i, t_j)$. Thus, the higher value the $I_{\Xi}(\delta_i; \delta_j)$ has, the larger quantities the $\mathbf{mit}_{\Xi}(t_i, \bar{t}_j)$ and/or $\mathbf{mit}_{\Xi}(\bar{t}_i, t_j)$ provide, and the more they indicate that t_i and t_j are highly dependent under state values (1, 0) and (0, 1), and that they should not co-occur in samples in Ξ . Consequently, a high positive value of $I_{\Xi}(\delta_i; \delta_j)$ may not indicate that t_i and t_j are highly dependent for their occurrence, namely, that the occurrence (absence) of term t_i accompanies the absence (occurrence) of term t_j . \diamond

The answer to the fourth question is now apparent. We can clarify our viewpoint by an example below, which can also help to answer the fifth and sixth questions.

Example 4.2 Suppose $\Xi = \{d_1, d_2, d_3\}$, $V_{d_1} = \{t_1, t_2, t_3, t_4, t_5\}$, $V_{d_2} = \{t_1, t_4, t_5, t_7\}$ and $V_{d_3} = \{t_4, t_7, t_8\}$. Then, it has $|\Xi| = 3$, $n_{\Xi}(t_1) = 2$, $n_{\Xi}(t_2) = 1$, $n_{\Xi}(t_1, t_2) = 1$, and

$$I_{\Xi}(\delta_1; \delta_2) = \frac{1}{3} \ln \frac{\frac{1}{3}}{\frac{2}{3} \cdot \frac{1}{3}} + \frac{1}{3} \ln \frac{\frac{1}{3}}{\frac{2}{3} \cdot \frac{1}{3}} + \frac{0}{3} \ln \frac{\frac{0}{3}}{\frac{1}{3} \cdot \frac{1}{3}} + \frac{1}{3} \ln \frac{\frac{1}{3}}{\frac{1}{3} \cdot \frac{2}{3}}$$

$$\approx 0.1352 - 0.0959 - 0.0000 + 0.1352 = 0.1745.$$

In this case, the value $I_{\Xi}(\delta_1; \delta_2)$ is dominated by both the quantities $\mathbf{mit}_{\Xi}(t_1, \bar{t}_2)$ and $\mathbf{mit}_{\Xi}(\bar{t}_1, t_2)$, and t_1 and t_2 are highly dependent for their co-occurrence in set Ξ . Also, from $n_{\Xi}(t_5) = 2$, $n_{\Xi}(t_7) = 2$ and $n_{\Xi}(t_5, t_7) = 1$, it has

$$I_{\Xi}(\delta_5; \delta_7) = \frac{1}{3} \ln \frac{\frac{1}{3}}{\frac{2}{3} \cdot \frac{2}{3}} + \frac{1}{3} \ln \frac{\frac{1}{3}}{\frac{2}{3} \cdot \frac{1}{3}} + \frac{1}{3} \ln \frac{\frac{1}{3}}{\frac{1}{3} \cdot \frac{2}{3}} + \frac{0}{3} \ln \frac{\frac{0}{3}}{\frac{1}{3} \cdot \frac{1}{3}}$$

$$\approx -0.0959 + 0.1352 + 0.1352 - 0.0000 = 0.1745.$$

In this case, the value $I_{\Xi}(\delta_5; \delta_7)$ is dominated by both the quantities $\mathbf{mit}_{\Xi}(t_5, \bar{t}_7)$ and $\mathbf{mit}_{\Xi}(\bar{t}_5, t_7)$, and t_5 and t_7 are highly dependent for their not-co-occurrence in set Ξ . \triangle

Remark-5: It can be seen, from Example 4.2, that two term pairs (t_1, t_2) and (t_5, t_7) receive the same value, $I_{\Xi}(\delta_1; \delta_2) = I_{\Xi}(\delta_5; \delta_7)$. However, the implications of the dependence information under the individual state values are entirely different: terms t_1 and t_2 provide the information highly supporting for either their co-occurrence or none of them occurrence (i.e., co-not-occurrence); whereas terms t_5 and t_7 provide the information highly supporting for one of them occurrence but another not occurrence (i.e., not-co-occurrence). \diamond

Remark-6: In a practical application, we normally concentrate on the statistics of co-occurrence of terms. That is, the

dependence under which we are really interested is state value $(\delta_i, \delta_j) = (1, 1)$. In this case, what we need is:

- to use the measure $\mathbf{mit}_{\Xi}(t_i, t_j)$ given in Eq.(8), and for every $(t_i, t_j) \in V \times V$, to verify an inequality,

$$\frac{n_{\Xi}(t_i, t_j)}{|\Xi|} > \frac{n_{\Xi}(t_i)}{|\Xi|} \times \frac{n_{\Xi}(t_j)}{|\Xi|} \quad (9)$$

- to select those term pairs (t_i, t_j) satisfying the above inequality as they guarantee both $\mathbf{mit}_{\Xi}(t_i, t_j) > 0$ (i.e., co-occurrence) and $\mathbf{mit}_{\Xi}(\bar{t}_1, \bar{t}_2) > 0$ (i.e., co-not-occurrence).

Then, we remove the term pairs not carrying the information supporting not-co-occurrence. \diamond

Example 4.3 (Example 4.2 continued). Consider terms t_1 and t_2 , we have

$$\frac{3}{9} = \frac{1}{3} = \frac{n_{\Xi}(t_1, t_2)}{|\Xi|} > \frac{n_{\Xi}(t_1)}{|\Xi|} \frac{n_{\Xi}(t_2)}{|\Xi|} = \frac{2}{3} \frac{1}{3} = \frac{2}{9}$$

From which we know that $\mathbf{mit}_{\Xi}(t_1, t_2) > 0$, $\mathbf{mit}_{\Xi}(t_1, \bar{t}_2) < 0$, $\mathbf{mit}_{\Xi}(\bar{t}_1, t_2) < 0$, $\mathbf{mit}_{\Xi}(\bar{t}_1, \bar{t}_2) > 0$, and that t_1 and t_2 are statistically dependent for their co-occurrence in Ξ . Also, if we consider terms t_5 and t_7 , then $n_{\Xi}(t_5) = 2$, $n_{\Xi}(t_7) = 2$, $n_{\Xi}(t_5, t_7) = 1$, and

$$\frac{3}{9} = \frac{1}{3} = \frac{n_{\Xi}(t_5, t_7)}{|\Xi|} < \frac{n_{\Xi}(t_5)}{|\Xi|} \frac{n_{\Xi}(t_7)}{|\Xi|} = \frac{2}{3} \frac{2}{3} = \frac{4}{9}$$

From which we know that $\mathbf{mit}_{\Xi}(t_5, t_7) < 0$, $\mathbf{mit}_{\Xi}(t_5, \bar{t}_7) > 0$, $\mathbf{mit}_{\Xi}(\bar{t}_5, t_7) > 0$, $\mathbf{mit}_{\Xi}(\bar{t}_5, \bar{t}_7) < 0$, and that t_5 and t_7 are highly dependent for their not co-occurrence in Ξ . \triangle

The following two Corollaries give properties of the MIT measures, that is, of the individual items of $I_{\Xi}(\delta_i; \delta_j)$ and $emim_{\Xi}(\delta_i; \delta_j)$. Their proofs are given in the proofs of Theorem 3.2 and Theorem 3.3, respectively.

Corollary 4.1 For two arbitrary terms $t_i, t_j \in V_{\Xi}$, if $n_{\Xi}(t_i) = |\Xi|$ or $n_{\Xi}(t_j) = |\Xi|$, then $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j}) = 0$ for $\delta_i, \delta_j = 0, 1$.

Corollary 4.2 For two arbitrary terms $t_i, t_j \in V_{\Xi}$, the individual items of $emim_{\Xi}(\delta_i; \delta_j)$ are always non-positive.

Remark-7: In order to apply $emim_{\Xi}(\delta_i; \delta_j)$ properly, let us compare the first item of $I_{\Xi}(\delta_i; \delta_j)$ given in Eq.(8) and the first item of $emim_{\Xi}(\delta_i; \delta_j)$ given in Eq.(5). Note that we have

$$\frac{n_{\Xi}(t_i, t_j)}{|\Xi|} = \frac{n_{11}}{n} \quad \text{and} \quad \frac{n_{\Xi}(t_i)}{|\Xi|} \frac{n_{\Xi}(t_j)}{|\Xi|} = \frac{n_{1.}}{n} \frac{n_{.1}}{n}$$

Thus, from the expressions in the respective \ln functions of the two first items:

- from the relation between $\frac{n_{\Xi}(t_i, t_j)}{|\Xi|}$ and $\frac{n_{\Xi}(t_i)}{|\Xi|} \frac{n_{\Xi}(t_j)}{|\Xi|}$ given in Theorem 4.1, we can infer all the signs of $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$ for $\delta_i, \delta_j = 1, 0$, and then determine whether term pair (t_i, t_j) is statistically dependent under the individual state values.
- however, the inference and determination cannot be made from the relation between n_{11} and $n_{1.}n_{.1}$; in fact, by Corollary 4.2, we know that the individual items of $emim_{\Xi}(\delta_i, \delta_j)$ are always non-positive.

Therefore, to solve the problem arisen by Q7, with Remark-6, we need to verify Eq.(9) or, equivalently, to verify a simpler inequality,

$$n_{11} = n_{\Xi}(t_i, t_j) > \frac{1}{|\Xi|} n_{\Xi}(t_i) n_{\Xi}(t_j) = \frac{1}{n} n_{1.} n_{.1} \quad (10)$$

which is a straightforward way to the solution. \diamond

TABLE I
THE DEPENDENCE VALUES AGAINST SIZES OF Ξ

$ \Xi $	$\text{mit}_{\Xi}(t_1, t_4)$	$\text{mit}_{\Xi}(t_1, t_4)$	$\text{mit}_{\Xi}(t_1, t_4)$	$\text{mit}_{\Xi}(t_1, t_4)$	$I_{\Xi}(\delta_1, \delta_4)$
3	0.0000	0.0000	0.0000	0.0000	0.0000
4	0.1438	0.0000	-0.1014	0.1733	0.2157
5	0.2043	0.0000	-0.1176	0.2043	0.2910
6	0.2310	0.0000	-0.1155	0.2027	0.3182
7	0.2421	0.0000	-0.1089	0.1923	0.3255
8	0.2452	0.0000	-0.1014	0.1798	0.3236
9	0.2441	0.0000	-0.0941	0.1675	0.3175
10	0.2408	0.0000	-0.0875	0.1562	0.3095
15	0.2146	0.0000	-0.0637	0.1145	0.2654
20	0.1897	0.0000	-0.0497	0.0896	0.2296
30	0.1535	0.0000	-0.0343	0.0621	0.1813
50	0.1125	0.0000	-0.0212	0.0384	0.1297
100	0.0701	0.0000	-0.0108	0.0196	0.0789
1000	0.0116	0.0000	-0.0011	0.0020	0.0125
10000	0.0016	0.0000	-0.0001	0.0002	0.0017

$$n_{\Xi}(t_1) = 2, n_{\Xi}(t_4) = 3, n_{\Xi}(t_1, t_4) = 2$$

V. SIZE OF SAMPLE SET

The binary estimation methods derive their importance from the fact that their simplicity of computation easily enables us to have an insight into the term dependence. However, the methods may be sensitive to the size of the sample set. This sections explains the sensitivity, using the probability estimation given in Eq.(1) and Eq.(2) as an example, and gives an answer to the last question Q8 through a simple example.

Example 5.1 (Example 4.2 continued) Suppose we have a sample set $\Xi \subseteq D = \{d_1, d_2, \dots, d_{10000}\}$. Consider two terms t_1 and t_4 with fixed numbers $n_{\Xi}(t_1, t_4) = 2, n_{\Xi}(t_1) = 2$ and $n_{\Xi}(t_4) = 3$. Then, when $|\Xi| = 3$, by Theorem 3.1,

$$I_{\Xi}(\delta_1; \delta_4) = \sum_{\delta_1, \delta_4=1,0} \text{mit}_{\Xi}(t_1^{\delta_1}, t_4^{\delta_4})$$

$$= 0.0000 - 0.0000 - 0.0000 + 0.0000 = 0.0000.$$

Next, taking $|\Xi| = 10$, then

$$I_{\Xi}(\delta_1; \delta_4) = \frac{2}{10} \ln \frac{\frac{2}{10} \frac{3}{10}}{\frac{2}{10} \frac{3}{10}}$$

$$+ \frac{2-2}{10} \ln \frac{\frac{2-2}{10}}{1 - \frac{3}{10}}$$

$$+ \frac{3-2}{10} \ln \frac{\frac{3-2}{10}}{(1 - \frac{2}{10}) \frac{3}{10}}$$

$$+ \frac{10-2-3+2}{10} \ln \frac{\frac{10-2-3+2}{10}}{(1 - \frac{2}{10})(1 - \frac{3}{10})}$$

$$= \frac{2}{10} \ln \frac{10}{3} + 0 \ln 0 + \frac{1}{10} \ln \frac{10}{24} + \frac{7}{10} \ln \frac{10}{8}$$

$$\approx 0.2408 - 0.0000 - 0.0875 + 0.1562 = 0.3095.$$

There are more dependence values of t_1 and t_4 against the increasing sizes of Ξ in Table I, in which, the numbers underlined are the maximum (in absolute values) for the corresponding EMIM and MIT measures. As it can be seen from Table I, the values vary as changing of $|\Xi|$ and the variation tells us about the behaviour of the individual measures. \triangle

The five different measures give us useful information; each indicates a different aspect about the dependence of terms and so should be interpreted in an appropriate way. Let us now

carefully examine Table I to look at what insight it can give regarding $|\Xi|$ for terms t_1 and t_4 .

- When $|\Xi| = 3$, it has $n_{\Xi}(t_4) = |\Xi|$, namely, t_4 occurs in all samples in Ξ . In this case, the occurrence of t_4 does not provide any information about the occurrence of t_1 in samples. Thus, t_1 and t_4 is statistically independent of each other, and $\text{mit}_{\Xi}(t_1^{\delta_1}, t_4^{\delta_4}) = 0$ for $\delta_1, \delta_4 = 1, 0$, so $I_{\Xi}(\delta_1; \delta_4) = 0$.
- As increasing of $|\Xi|$, the individual dependence values in each of the columns are increasing (in absolute values) till to the maximum. This is because if t_1 or t_4 occur in several (not many) samples, and also co-occur in some of these, then the values indicate that t_1 and t_4 are dependent to some extent.
- For larger and larger $|\Xi|$, t_1 and t_4 co-occur in less and less samples in Ξ (compared with $|\Xi|$) and they receive lower and lower dependence values. The values drop greatly when $|\Xi| = 100$ and almost are equal to zero when $|\Xi| = 10000 = |D|$.

Generally, when the numbers $n_{\Xi}(t_i, t_j), n_{\Xi}(t_i)$ and $n_{\Xi}(t_j)$ are fixed, we have $\text{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j}) \rightarrow 0$ (for $\delta_i, \delta_j = 1, 0$) and hence $I_{\Xi}(\delta_i; \delta_j) \rightarrow 0$, when $|\Xi| \rightarrow \infty$. The mathematical reason for this is simple. As it can be seen from the probability estimation given in Eq.(2) and the MIT measures given Eq.(7),

- except the last one, the individual probabilities $P_{\Xi}(\delta_i, \delta_j)$ approach 0, so the corresponding measures $\text{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$ approach $0 \times \ln(\alpha|\Xi|) = 0$ (where α is a constant), as $|\Xi| \rightarrow \infty$.
- the last probability $P_{\Xi}(\delta_i = 0, \delta_j = 0)$ approaches 1, so the measure $\text{mit}_{\Xi}(t_i^{\delta_i=0}, t_j^{\delta_j=0})$ approaches $1 \times \ln 1 = 0$, as $|\Xi| \rightarrow \infty$.

Remark-8: It worth mentioning that the binary estimation method given in Eq.(1) and Eq.(2) rely on statistics $n_{\Xi}(t), n_{\Xi}(t_i, t_j)$ and $|\Xi|$; it is thus sensitive to the sample size. A large sample size might overwhelm useful statistical information carried by those important terms having smaller statistics (or, concentrating in a few documents), thereby weaken and dilute the potential capability of EMIM and the MIT measures. \diamond

The sample size is an important feature of any empirical study, and generally a larger sample size leads to increased precision when estimating unknown (probability distribution) parameters. According to study given in [19], an appropriate sample size for a qualitative research depends on a number of factors, including: the quality of the data, the scope of the study, the nature of the topic, the amount of useful information obtained from the participants (samples), the qualitative method, experimental design and settings, and so on. It seems not clear at present how to determine an appropriate sample size against a set of term pairs in practical applications. It would be helpful to consider appropriateness of the sample size prior to determining some probability estimation method for applying EMIM in a specific application.

CONCLUSION

This study examined the reasons for the failure of applying EMIM and highlighted some potential problems of applications. We attempted to clarify confusions caused by the problems and/or suggest solutions to the problems by analysing a various of properties of $I_{\Xi}(\delta_i; \delta_j)$ and $emim_{\Xi}(\delta_i; \delta_j)$. The key points of this study were emphasised and formally discussed through a series of remarks, some of them are listed as follows.

- The occurrence of term t in all samples does not provide any information about the occurrence of other terms in the samples; in order to effectively capture the dependence information of terms, we should always avoid many terms having $n_{\Xi}(t) = |\Xi|$.
- It can be seen, from the relation given in Eq.(6), that $I_{\Xi}(\delta_i; \delta_j) \geq 0$ cannot infer $emim_{\Xi}(\delta_i; \delta_j) \geq 0$; in fact, we have $emim_{\Xi}(\delta_i; \delta_j) \leq 0$ for two arbitrary terms $t_i, t_j \in V$.
- Two term pairs, (t_i, t_j) and (t'_i, t'_j) , receiving the same EMIM value, $I_{\Xi}(\delta_i; \delta_j) = I_{\Xi}(\delta_{i'}; \delta_{j'})$, may be dependent of each other in entirely different implications under the individual state values.
- A high positive value of $I_{\Xi}(\delta_i; \delta_j)$ may not be necessary to indicate that t_i and t_j are highly dependent for their occurrence; we should always verify the inequality given in Eq.(9) to ensure $mit_{\Xi}(t_i, t_j) > 0$, and that terms are high dependent for their co-occurrence.
- In order to apply $emim(\delta_i; \delta_j)$ properly, we should always verify the inequality given in Eq.(10).
- The binary estimation method given in Eq.(1) and Eq.(2) is sensitive to the sample size; a large sample size might overwhelm useful statistical information carried by those terms concentrating in a small number of documents.

It is essential for this study to point out that different probability estimations may conclude to different properties of EMIM and the MIT measures, and therefore it is theoretically challenging to apply EMIM without clearly understanding the properties. A widely used binary estimation method is considered in this study as a good example to reveal practical application problems and to clarify our viewpoints. A more general discussion on this subject can be found in our another study [18]. Due to its generality, this study can be regarded as

a useful tool for many areas of science, involving statistical text analysis and document processing.

REFERENCES

- [1] A. Akadi, A. Abdeljalil El Ouardighi, and D. Aboutajdine, "A powerful feature selection approach based on mutual information," *International Journal of Computer Science and Network Security*, vol. 8, no. 4, pp. 116–121, 2008.
- [2] M. Ait Kerroum, A. Hammouch, and D. Aboutajdine, "Textural feature selection by joint mutual information based on Gaussian mixture model for multispectral image classification," *Pattern Recognition Letters*, vol. 31, no. 10, pp. 1168–1174, 2010.
- [3] H.-W. Liu, J.-G. Sun, L. Liu, and H.-J. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognition*, vol. 42, pp. 1330–1339, 2009.
- [4] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.
- [5] G. Wang, F. Lochovsky, and Q. Yang, "Feature selection with conditional mutual information maximin in text categorization," in *Proceedings of the 10th International Conference on Information and Knowledge Management*, 2004, pp. 342–349.
- [6] N. Vretos, V. Solachidis, and I. Pitas, "A mutual information based face clustering algorithm for movies," in *Proceedings of the 2006 IEEE International Conference on Multimedia and Expo (ICME'06)*, 2006, pp. 1013–1016.
- [7] X. Zhang, K. Liu, Z. Liu, B. Duval, J. Richer, X. Zhao, J. Hao, and L. Chen, "Narromi: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference," *Bioinformatics*, vol. 29, no. 1, p. 106113, 2013.
- [8] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.
- [9] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Journal of the American Society for Information Science*, vol. 16, no. 1, pp. 22–29, 1990.
- [10] H. Fang and C. X. Zhai, "Semantic term matching in axiomatic approaches to information retrieval," in *Proceedings of the 29th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 115–122.
- [11] S. Gauch, J. Wang, and S. M. Rachakonda, "A corpus analysis approach for automatic query expansion and its extension to multiple databases," *ACM Transactions on Information Systems*, vol. 17, no. 3, pp. 250–269, 1999.
- [12] M. Kim and K. Choi, "A comparison of collocation-based similarity measures in query expansion," *Information Processing & Management*, vol. 35, no. 1, pp. 19–30, 1999.
- [13] R. Mandala, T. Tokunaga, and H. Tanaka, "Query expansion using heterogeneous thesauri," *Information Processing & Management*, vol. 36, no. 3, pp. 361–378, 2000.
- [14] R. M. Losee, Jr., "Term dependence: A basis for Luhn and Zipf models," *Journal of the American Society for Information Science and Technology*, vol. 52, no. 12, pp. 1019–1025, 2001.
- [15] C. J. van Rijsbergen, *Information Retrieval*, 2nd ed. London: Butterworths, 1979.
- [16] —, "A theoretical basis for the use of co-occurrence data in information retrieval," *Journal of Documentation*, vol. 33, no. 2, pp. 106–119, 1977.
- [17] D. Cai and T. McCluskey, "A simple method for computing term mutual information," *Journal of Computing*, vol. 4, no. 6, pp. 1–6, 2012.
- [18] —, "A general framework of generating estimation functions for computing the mutual information of terms," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 11, pp. 198–207, 2013.
- [19] J. Morse, "Determining sample size," *Qualitative Health Research*, vol. 10, no. 1, p. 35, 2000.

APPENDIX

Theorem 2.1 Suppose $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i, \delta_j)$ are given in Eq.(1) and Eq.(2), respectively. Then $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i, \delta_j)$ are

probability distributions on Ω and $\Omega \times \Omega$, respectively; $P_{\Xi}(\delta_i)$ and $P_{\Xi}(\delta_j)$ are the marginal distributions of $P_{\Xi}(\delta_i, \delta_j)$.

Proof. For arbitrary terms $t, t_i, t_j \in V$ (where $i \neq j$), using the statistics of the document frequencies concerning the set Ξ , it is easy to estimate the probability distributions.

First, notice that the (total) number of documents in the sample set is $|\Xi|$. Thus, the probability that t occurs in some sample is $\frac{n_{\Xi}(t)}{|\Xi|}$ as the number of samples in which t occurs is $n_{\Xi}(t)$, and thus the probability that t does not occur is $1 - \frac{n_{\Xi}(t)}{|\Xi|}$. Therefore, we can write a probability distribution, $P_{\Xi}(\delta)$, over Ω as expressed by Eq.(2).

Second, with the size of the sample set, the probability that t_i and t_j co-occur is $\frac{n_{\Xi}(t_i, t_j)}{|\Xi|}$ as the number of samples in which t_i and t_j co-occur is $n_{\Xi}(t_i, t_j)$; the probability that t_i occurs but t_j does not occur is $\frac{n_{\Xi}(t_i) - n_{\Xi}(t_i, t_j)}{|\Xi|}$ as the number of samples in which t_i occurs but t_j does not occur is $n_{\Xi}(t_i) - n_{\Xi}(t_i, t_j)$; similarly, the probability that t_i does not occur but t_j occurs is $\frac{n_{\Xi}(t_j) - n_{\Xi}(t_i, t_j)}{|\Xi|}$; the probability that neither of t_i nor t_j occur is $\frac{n_{\Xi}(\bar{t}_i, \bar{t}_j)}{|\Xi|}$, where $n_{\Xi}(\bar{t}_i, \bar{t}_j) = |\Xi| - n_{\Xi}(t_i) - n_{\Xi}(t_j) + n_{\Xi}(t_i, t_j)$ is the number of samples in which none of t_i and t_j occur. Therefore, we can write a probability distribution, $P_{\Xi}(\delta_i, \delta_j)$, over $\Omega \times \Omega$ as expressed by Eq.(3).

Finally, it is easy to see: $P_{\Xi}(\delta_i = 1) = \sum_{\delta_j=1,0} P_{\Xi}(\delta_i = 1, \delta_j) = \frac{n_{\Xi}(t_i)}{|\Xi|}$ and $P_{\Xi}(\delta_i = 0) = \sum_{\delta_j=1,0} P_{\Xi}(\delta_i = 0, \delta_j) = 1 - \frac{n_{\Xi}(t_i)}{|\Xi|}$. Hence, $P_{\Xi}(\delta_i)$ is the marginal distributions of $P_{\Xi}(\delta_i, \delta_j)$. A similar discussion may be given for $P_{\Xi}(\delta_j)$. \square

An alternative way to derive $P_{\Xi}(\delta_i, \delta_j)$ is to use a conditional probability formula. The conditional probability of observing t_j occurs, given that t_i occurred, is $P_{\Xi}(\delta_j = 1 | \delta_i = 1) = \frac{n_{\Xi}(t_i, t_j)}{n_{\Xi}(t_i)}$, since before the observation there were $n_{\Xi}(t_i)$ documents in Ξ , in which t_i occurred. The conditional probability of observing t_j does not occur, given that t_i occurred, is $P_{\Xi}(\delta_j = 0 | \delta_i = 1) = 1 - \frac{n_{\Xi}(t_i, t_j)}{n_{\Xi}(t_i)}$, and similarly, we have $P_{\Xi}(\delta_i = 0 | \delta_j = 1) = 1 - \frac{n_{\Xi}(t_i, t_j)}{n_{\Xi}(t_j)}$. Then, we can immediately write the expressions:

$$\begin{aligned} P_{\Xi}(\delta_i = 1, \delta_j = 1) &= P_{\Xi}(\delta_i = 1)P_{\Xi}(\delta_j = 1 | \delta_i = 1) \\ &= \frac{n_{\Xi}(t_i)}{|\Xi|} \frac{n_{\Xi}(t_i, t_j)}{n_{\Xi}(t_i)} \\ &= \frac{n_{\Xi}(t_i, t_j)}{|\Xi|} \\ P_{\Xi}(\delta_i = 1, \delta_j = 0) &= P_{\Xi}(\delta_i = 1)P_{\Xi}(\delta_j = 0 | \delta_i = 1) \\ &= \frac{n_{\Xi}(t_i)}{|\Xi|} \left[1 - \frac{n_{\Xi}(t_i, t_j)}{n_{\Xi}(t_i)} \right] \\ &= \frac{n_{\Xi}(t_i)}{|\Xi|} - \frac{n_{\Xi}(t_i, t_j)}{|\Xi|} \\ P_{\Xi}(\delta_i = 0, \delta_j = 1) &= P_{\Xi}(\delta_j = 1)P_{\Xi}(\delta_i = 0 | \delta_j = 1) \\ &= \frac{n_{\Xi}(t_j)}{|\Xi|} \left[1 - \frac{n_{\Xi}(t_i, t_j)}{n_{\Xi}(t_j)} \right] \end{aligned}$$

$$\begin{aligned} &= \frac{n_{\Xi}(t_j)}{|\Xi|} - \frac{n_{\Xi}(t_i, t_j)}{|\Xi|} \\ P_{\Xi}(\delta_i = 0, \delta_j = 0) &= 1 - P_{\Xi}(\delta_i = 1, \delta_j = 1) \\ &\quad - P_{\Xi}(\delta_i = 1, \delta_j = 0) \\ &\quad - P_{\Xi}(\delta_i = 0, \delta_j = 1) \\ &= 1 - \frac{n_{\Xi}(t_i)}{|\Xi|} - \frac{n_{\Xi}(t_j)}{|\Xi|} + \frac{n_{\Xi}(t_i, t_j)}{|\Xi|}. \end{aligned}$$

The results are in agreement with one given in Eq.(2).

It worth mentioning that the reason why we give the detailed proofs of Theorem 2.1 is to interpret mathematical meaning of the estimation of the probability distributions. The proof may be greatly simplified by directly using the nature of the expressions given in Eq.(1) and Eq.(2), that is,

$$P_{\Xi}(\delta) \geq 0 \quad \text{and} \quad P_{\Xi}(\delta_i, \delta_j) \geq 0$$

for $\delta, \delta_i, \delta_j = 1, 0$, and

$$\sum_{\delta=1,0} P_{\Xi}(\delta) = 1 \quad \text{and} \quad \sum_{\delta_i, \delta_j=1,0} P_{\Xi}(\delta_i, \delta_j) = 1$$

Therefore, $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i, \delta_j)$ are probability distributions.

Theorem 2.2 Suppose $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i, \delta_j)$ are given in Eq.(1) and Eq.(2), respectively. Then $P_{\Xi}(\delta_i, \delta_j)$ is absolutely continuous with respect to product $P_{\Xi}(\delta_i)P_{\Xi}(\delta_j)$, denoted by $P_{\Xi}(\delta_i, \delta_j) \ll P_{\Xi}(\delta_i)P_{\Xi}(\delta_j)$, for $\delta_i, \delta_j = 1, 0$.

Proof. For two arbitrary terms $t_i, t_j \in V$, according to whether $P_{\Xi}(t_i) = 1$ and/or $P_{\Xi}(t_j) = 1$, there are four cases to be considered, that is,

- (C1) $0 < P_{\Xi}(t_i) < 1$ and $0 < P_{\Xi}(t_j) < 1$,
- (C2) $P_{\Xi}(t_i) = 1$ but $0 < P_{\Xi}(t_j) < 1$,
- (C3) $0 < P_{\Xi}(t_i) < 1$ but $P_{\Xi}(t_j) = 1$,
- (C4) $P_{\Xi}(t_i) = 1$ and $P_{\Xi}(t_j) = 1$.

We first prove (C1) and then prove (C2). Similar proofs can be given to (C3) and (C4).

In order to prove (C1), let us further consider four cases:

- (a) $t_i, t_j \in V_{\Xi}$;
- (b) $t_i \in V_{\Xi}$ but $t_j \notin V_{\Xi}$;
- (c) $t_i \notin V_{\Xi}$ but $t_j \in V_{\Xi}$;
- (d) $t_i, t_j \notin V_{\Xi}$.

Notice that, for (a), $P_{\Xi}(\delta_i, \delta_j) \ll P_{\Xi}(\delta_i)P_{\Xi}(\delta_j)$ as $0 < P_{\Xi}(\delta_i), P_{\Xi}(\delta_j) < 1$ for $\delta_i, \delta_j = 0, 1$ by Eq.(1). We now prove (b), and similar proofs can be given for (c) and (d). The proof is to verify four distinct state values, respectively.

On one hand, when $t_i \in V_{\Xi}$ but $t_j \notin V_{\Xi}$, it has $0 < P_{\Xi}(t_i) < 1$, $P_{\Xi}(t_j) = 0$, and $P_{\Xi}(t_i, t_j) = 0$ by Eq.(1). Thus, by Eq.(3),

$$\begin{aligned} P_{\Xi}(\delta_i = 1, \delta_j = 1) &= 0 \\ P_{\Xi}(\delta_i = 1, \delta_j = 0) &= P_{\Xi}(t_i) > 0 \\ P_{\Xi}(\delta_i = 0, \delta_j = 1) &= 0 \\ P_{\Xi}(\delta_i = 0, \delta_j = 0) &= 1 - P_{\Xi}(t_i) > 0 \end{aligned}$$

On the other hand, by Eq.(2), we have $0 < P_{\Xi}(\delta_i) < 1$ for $\delta_i = 1, 0$ when $t_i \in V_{\Xi}$; $P_{\Xi}(\delta_j = 1) = 0$ and $P_{\Xi}(\delta_j = 0) = 1$ when $t_j \notin V_{\Xi}$. Thus,

$$\begin{aligned} P_{\Xi}(\delta_i = 1)P_{\Xi}(\delta_j = 1) &= 0 \\ P_{\Xi}(\delta_i = 1)P_{\Xi}(\delta_j = 0) &= P_{\Xi}(\delta_i = 1) > 0 \\ P_{\Xi}(\delta_i = 0)P_{\Xi}(\delta_j = 1) &= 0 \\ P_{\Xi}(\delta_i = 0)P_{\Xi}(\delta_j = 0) &= P_{\Xi}(\delta_i = 0) > 0 \end{aligned}$$

Therefore, $P_{\Xi}(\delta_i, \delta_j) \ll P_{\Xi}(\delta_i)P_{\Xi}(\delta_j)$ for $\delta_i, \delta_j = 1, 0$.
In order to prove (C2), let us suppose we are given $t_i, t_j \in V_{\Xi}$ satisfying $n_{\Xi}(t_i) = |\Xi|$ and $n_{\Xi}(t_j) < |\Xi|$ (namely t_i occurs in all samples in Ξ , but t_j does not). In this case, it has $P_{\Xi}(t_i) = 1$ and $0 < P_{\Xi}(t_j) < 1$, and $n_{\Xi}(t_j) = n_{\Xi}(t_i, t_j)$. Thus,

- (a) $P_{\Xi}(\delta_i = 1) \cdot P_{\Xi}(\delta_j = 1) > 0$ since $P_{\Xi}(\delta_i = 1) = 1$, and $0 < P_{\Xi}(\delta_j = 1) < 1$. Thus, $P_{\Xi}(\delta_i = 1, \delta_j = 1) \ll P_{\Xi}(\delta_i = 1) \cdot P_{\Xi}(\delta_j = 1)$ for $(\delta_i, \delta_j) = (1, 1)$.
- (b) $P_{\Xi}(\delta_i = 1) \cdot P_{\Xi}(\delta_j = 0) > 0$ since $P_{\Xi}(\delta_i = 1) = 1$ and $0 < P_{\Xi}(\delta_j = 0) < 1$. Thus, $P_{\Xi}(\delta_i = 1, \delta_j = 0) \ll P_{\Xi}(\delta_i = 1) \cdot P_{\Xi}(\delta_j = 0)$ for $(\delta_i, \delta_j) = (1, 0)$.
- (c) $P_{\Xi}(\delta_i = 0) \cdot P_{\Xi}(\delta_j = 1) = 0$ since $P_{\Xi}(\delta_i = 0) = 0$ and $0 < P_{\Xi}(\delta_j = 1) < 1$. Also, $P_{\Xi}(\delta_i = 0, \delta_j = 1) = \frac{1}{|\Xi|} [n_{\cdot 1} - n_{11}] = 0$. Thus, $P_{\Xi}(\delta_i = 0, \delta_j = 1) \ll P_{\Xi}(\delta_i = 0) \cdot P_{\Xi}(\delta_j = 1)$ for $(\delta_i, \delta_j) = (0, 1)$.
- (d) $P_{\Xi}(\delta_i = 0) \cdot P_{\Xi}(\delta_j = 0) = 0$ since $P_{\Xi}(\delta_i = 0) = 0$ and $0 < P_{\Xi}(\delta_j = 0) < 1$. Also, $P_{\Xi}(\delta_i = 0, \delta_j = 0) = \frac{1}{|\Xi|} [|\Xi| - n_{\cdot 1} - n_{\cdot 1} + n_{11}] = \frac{1}{|\Xi|} [(|\Xi| - n_{\cdot 1}) - (n_{\cdot 1} - n_{11})] = 0$. Thus, $P_{\Xi}(\delta_i = 0, \delta_j = 0) \ll P_{\Xi}(\delta_i = 0) \cdot P_{\Xi}(\delta_j = 0)$ for $(\delta_i, \delta_j) = (0, 0)$.

Therefore, $P_{\Xi}(\delta_i, \delta_j) \ll P_{\Xi}(\delta_i) \cdot P_{\Xi}(\delta_j)$ for $\delta_i, \delta_j = 1, 0$. \square

Theorem 3.1 Suppose $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i, \delta_j)$ are given in Eq.(1) and Eq.(2), respectively. Then $I_{\Xi}(\delta_i; \delta_j) = 0$ if $n_{\Xi}(t_i) = |\Xi|$ or $n_{\Xi}(t_j) = |\Xi|$.

Proof. We prove that each item of $I_{\Xi}(\delta_i; \delta_j)$ is zero for $n_{\Xi}(t_i) = |\Xi|$. A similar proof can be given to $n_{\Xi}(t_j) = |\Xi|$. Notice that, we have $n_{\Xi}(t_j) = n_{\Xi}(t_i, t_j)$, Thus,

- 1) for $(\delta_i, \delta_j) = (1, 1)$, with $n_{11} = n_{\Xi}(t_i, t_j) = n_{\Xi}(t_j)$, it has

$$\begin{aligned} & \frac{n_{11}}{|\Xi|} \ln \left(\frac{n_{11}}{|\Xi|} / \frac{n_{\Xi}(t_i)}{|\Xi|} \frac{n_{\Xi}(t_j)}{|\Xi|} \right) \\ &= \frac{n_{11}}{|\Xi|} \ln \frac{n_{11}}{1 \times n_{\Xi}(t_j)} = \frac{n_{11}}{|\Xi|} \ln 1 = 0 \end{aligned}$$

- 2) for $(\delta_i, \delta_j) = (1, 0)$, with $n_{10} = n_{\Xi}(t_i) - n_{\Xi}(t_i, t_j) = |\Xi| - n_{\Xi}(t_j)$, it has

$$\begin{aligned} & \frac{n_{10}}{|\Xi|} \ln \left(\frac{n_{10}}{|\Xi|} / \frac{n_{\Xi}(t_i)}{|\Xi|} \left(1 - \frac{n_{\Xi}(t_j)}{|\Xi|}\right) \right) \\ &= \frac{n_{10}}{|\Xi|} \ln \frac{n_{10}}{1 \times (|\Xi| - n_{\Xi}(t_j))} = \frac{n_{10}}{|\Xi|} \ln 1 = 0 \end{aligned}$$

- 3) for $(\delta_i, \delta_j) = (0, 1)$, with $n_{01} = n_{\Xi}(t_j) - n_{\Xi}(t_i, t_j) = n_{\Xi}(t_j) - n_{\Xi}(t_j) = 0$, is has

$$\begin{aligned} & \frac{n_{01}}{|\Xi|} \ln \left(\frac{n_{01}}{|\Xi|} / \left(1 - \frac{n_{\Xi}(t_i)}{|\Xi|}\right) \frac{n_{\Xi}(t_j)}{|\Xi|} \right) \\ &= \frac{0}{|\Xi|} \ln \frac{0}{0 \times n_{\Xi}(t_j)} = 0 \ln \frac{0}{0} = 0 \end{aligned}$$

- 4) for $(\delta_i, \delta_j) = (0, 0)$, with $n_{00} = |\Xi| - n_{\Xi}(t_i) - n_{\Xi}(t_j) + n_{\Xi}(t_i, t_j) = |\Xi| - |\Xi| - n_{\Xi}(t_j) + n_{\Xi}(t_j) = 0$, it has

$$\begin{aligned} & \frac{n_{00}}{|\Xi|} \ln \left(\frac{n_{00}}{|\Xi|} / \left(1 - \frac{n_{\Xi}(t_i)}{|\Xi|}\right) \left(1 - \frac{n_{\Xi}(t_j)}{|\Xi|}\right) \right) \\ &= \frac{0}{|\Xi|} \ln \frac{0}{0 \times (|\Xi| - n_{\Xi}(t_j))} = 0 \ln \frac{0}{0} = 0 \end{aligned}$$

The proof is completed. \square

Theorem 3.2 Suppose $I_{\Xi}(\delta_i, \delta_j)$ and $emim_{\Xi}(\delta_i, \delta_j)$ are given in Eq.(3) and Eq.(5), respectively. Then

$$I_{\Xi}(\delta_i, \delta_j) = \frac{1}{n} \times emim_{\Xi}(\delta_i, \delta_j) + \ln(n)$$

where $n = |\Xi|$.

Proof. With the above notation n_{11} , $n_{\cdot 1}$, and $n_{\cdot 1}$ given in Eq.(5), we can write an alternative, but fully equivalent, expression:

$$\begin{aligned} I_{\Xi}(\delta_i; \delta_j) &= \frac{n_{11}}{n} \ln \left(\frac{n_{11}}{n_{\cdot 1} \cdot n_{\cdot 1}} n \right) \\ &+ \frac{n_{\cdot 1} - n_{11}}{n} \ln \left(\frac{n_{\cdot 1} - n_{11}}{n_{\cdot 1} \cdot (n - n_{\cdot 1})} n \right) \\ &+ \frac{n_{\cdot 1} - n_{11}}{n} \ln \left(\frac{n_{\cdot 1} - n_{11}}{(n - n_{\cdot 1}) n_{11}} n \right) \\ &+ \frac{n - n_{\cdot 1} - n_{\cdot 1} + n_{11}}{n} \times \\ &\quad \ln \left(\frac{n - n_{\cdot 1} - n_{\cdot 1} + n_{11}}{(n - n_{\cdot 1})(n - n_{\cdot 1})} n \right) \\ &= \left[\frac{n_{11}}{n} \ln \frac{n_{11}}{n_{\cdot 1} \cdot n_{\cdot 1}} \right. \\ &\quad + \frac{n_{\cdot 1} - n_{11}}{n} \ln \frac{n_{\cdot 1} - n_{11}}{n_{\cdot 1} \cdot (n - n_{\cdot 1})} \\ &\quad + \frac{n_{\cdot 1} - n_{11}}{n} \ln \frac{n_{\cdot 1} - n_{11}}{(n - n_{\cdot 1}) n_{11}} \\ &\quad + \frac{n - n_{\cdot 1} - n_{\cdot 1} + n_{11}}{n} \times \\ &\quad \quad \left. \ln \frac{n - n_{\cdot 1} - n_{\cdot 1} + n_{11}}{(n - n_{\cdot 1})(n - n_{\cdot 1})} \right] \\ &+ \left[\frac{n_{11}}{n} + \frac{n_{\cdot 1} - n_{11}}{n} + \frac{n_{\cdot 1} - n_{11}}{n} + \right. \\ &\quad \left. \frac{n - n_{\cdot 1} - n_{\cdot 1} + n_{11}}{n} \right] \times \ln(n) \\ &= \left[n_{11} \ln \frac{n_{11}}{n_{\cdot 1} \cdot n_{\cdot 1}} \right. \\ &\quad + (n_{\cdot 1} - n_{11}) \ln \frac{n_{\cdot 1} - n_{11}}{n_{\Xi}(t_i)(n - n_{\cdot 1})} \\ &\quad + (n_{\cdot 1} - n_{11}) \ln \frac{n_{\cdot 1} - n_{11}}{(n - n_{\cdot 1}) n_{11}} \\ &\quad + (n - n_{\cdot 1} - n_{\cdot 1} + n_{11}) \times \\ &\quad \quad \left. \ln \frac{n - n_{\cdot 1} - n_{\cdot 1} + n_{11}}{(n - n_{\cdot 1})(n - n_{\cdot 1})} \right] \\ &\quad \times \frac{1}{n} + \ln(n) \\ &= emim_{\Xi}(\delta_i; \delta_j) \times \frac{1}{n} + \ln(n) \end{aligned}$$

The proof is completed. \square

Theorem 3.3 Suppose $emim_{\Xi}(\delta_i, \delta_j)$ is given expression Eq.(5). Then $emim_{\Xi}(\delta_i, \delta_j) \leq 0$.

Proof. We prove each item of $emim_{\Xi}(\delta_i; \delta_j)$ non-positive. The proof is simple with an inequality $\frac{a}{a_1 a_2} \leq 1$ if $a \leq a_1$ and $a \leq a_2$.

- 1) we have $\frac{n_{11}}{n_{\cdot 1} \cdot n_{\cdot 1}} \leq 1$ since,

$$\begin{aligned} n_{11} &= n_{\Xi}(t_i, t_j) \leq n_{\Xi}(t_i) = n_{\cdot 1} \\ n_{11} &= n_{\Xi}(t_i, t_j) \leq n_{\Xi}(t_j) = n_{\cdot 1} \end{aligned}$$

2) we have $\frac{n_{10}}{n_{1.}n_{.0}} \leq 1$ since,

$$\begin{aligned} n_{10} &= n_{\Xi}(t_i) - n_{\Xi}(t_i, t_j) \leq n_{\Xi}(t_i) = n_{1.} \\ n_{10} &= n_{\Xi}(t_i) - n_{\Xi}(t_i, t_j) \leq |\Xi| - n_{\Xi}(t_i, t_j) \\ &\leq |\Xi| - n_{\Xi}(t_j) = n_{.0} \end{aligned}$$

3) the proof is similar to 2).

4) we have $\frac{n_{00}}{n_{0.}n_{.0}} \leq 1$ since,

$$\begin{aligned} n_{00} &= |\Xi| - n_{\Xi}(t_i) - n_{\Xi}(t_j) + n_{\Xi}(t_i, t_j) \\ &= |\Xi| - n_{\Xi}(t_i) - [n_{\Xi}(t_j) - n_{\Xi}(t_i, t_j)] \\ &\leq |\Xi| - n_{\Xi}(t_i) = n_{0.} \\ n_{00} &= |\Xi| - n_{\Xi}(t_j) - [n_{\Xi}(t_i) - n_{\Xi}(t_i, t_j)] \\ &\leq |\Xi| - n_{\Xi}(t_j) = n_{.0} \end{aligned}$$

The proof is completed. \square

Note that the fact that the individual items of $emim_{\Xi}(\delta_i, \delta_j)$ are non-positive can also be seen directly by the relations:

$$\begin{aligned} n_{1.} &= n_{11} + n_{10}, & n_{.0} &= n_{01} + n_{00}, \\ n_{.1} &= n_{11} + n_{01}, & n_{.0} &= n_{10} + n_{00}. \end{aligned}$$

Theorem 4.1 Suppose the four measures, $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$, where $\delta_i \delta_j = 0, 1$, are given in Eq.(7). Then we have the following property.

(1) If $\frac{n_{\Xi}(t_i, t_j)}{|\Xi|} = \frac{n_{\Xi}(t_i)}{|\Xi|} \frac{n_{\Xi}(t_j)}{|\Xi|}$ then

$$\begin{aligned} \mathbf{mit}_{\Xi}(t_i, t_j) &= 0, & \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) &= 0, \\ \mathbf{mit}_{\Xi}(\bar{t}_i, t_j) &= 0, & \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) &= 0. \end{aligned}$$

(2) If $\frac{n_{\Xi}(t_i, t_j)}{|\Xi|} > \frac{n_{\Xi}(t_i)}{|\Xi|} \frac{n_{\Xi}(t_j)}{|\Xi|}$ then

$$\begin{aligned} \mathbf{mit}_{\Xi}(t_i, t_j) &> 0, & \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) &\leq 0, \\ \mathbf{mit}_{\Xi}(\bar{t}_i, t_j) &\leq 0, & \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) &> 0. \end{aligned}$$

(3) If $\frac{n_{\Xi}(t_i, t_j)}{|\Xi|} < \frac{n_{\Xi}(t_i)}{|\Xi|} \frac{n_{\Xi}(t_j)}{|\Xi|}$ then

$$\begin{aligned} \mathbf{mit}_{\Xi}(t_i, t_j) &< 0, & \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) &\geq 0, \\ \mathbf{mit}_{\Xi}(\bar{t}_i, t_j) &\geq 0, & \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) &< 0. \end{aligned}$$

Proof. The proof of (1) is obvious. We only prove (2) here. A similar proof can be given to (3).

Now, substituting $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i, \delta_j)$ in Eq.(1) and Eq.(2) into $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$ in Eq.(7), we can rewrite the four MIT measures as follows (also see Example 4.1):

$$\begin{aligned} \mathbf{mit}_{\Xi}(t_i, t_j) &= P_{\Xi}(t_i, t_j) \ln \frac{P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_i)P_{\Xi}(t_j)} \\ \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) &= (P_{\Xi}(t_i) - P_{\Xi}(t_i, t_j)) \ln \frac{P_{\Xi}(t_i) - P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_i)(1 - P_{\Xi}(t_j))} \\ \mathbf{mit}_{\Xi}(\bar{t}_i, t_j) &= (P_{\Xi}(t_j) - P_{\Xi}(t_i, t_j)) \ln \frac{P_{\Xi}(t_j) - P_{\Xi}(t_i, t_j)}{(1 - P_{\Xi}(t_i))P_{\Xi}(t_j)} \\ \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) &= (1 - P_{\Xi}(t_i) - P_{\Xi}(t_j) + P_{\Xi}(t_i, t_j)) \times \\ &\quad \ln \frac{1 - P_{\Xi}(t_i) - P_{\Xi}(t_j) + P_{\Xi}(t_i, t_j)}{(1 - P_{\Xi}(t_i))(1 - P_{\Xi}(t_j))} \end{aligned}$$

Thus, on one hand, from $\frac{n_{\Xi}(t_i, t_j)}{|\Xi|} > \frac{n_{\Xi}(t_i)}{|\Xi|} \frac{n_{\Xi}(t_j)}{|\Xi|}$, we have

$$\begin{aligned} P_{\Xi}(t_i, t_j) &> P_{\Xi}(t_i)P_{\Xi}(t_j) \\ P_{\Xi}(t_i) - P_{\Xi}(t_i, t_j) &< P_{\Xi}(t_i) - P_{\Xi}(t_i)P_{\Xi}(t_j) \\ &= P_{\Xi}(t_i)(1 - P_{\Xi}(t_j)) \\ P_{\Xi}(t_j) - P_{\Xi}(t_i, t_j) &< P_{\Xi}(t_j) - P_{\Xi}(t_i)P_{\Xi}(t_j) \\ &= P_{\Xi}(t_j)(1 - P_{\Xi}(t_i)) \\ 1 - P_{\Xi}(t_i) - P_{\Xi}(t_j) + P_{\Xi}(t_i, t_j) &> 1 - P_{\Xi}(t_i) - P_{\Xi}(t_j) + P_{\Xi}(t_i)P_{\Xi}(t_j) \\ &= (1 - P_{\Xi}(t_i))(1 - P_{\Xi}(t_j)) \end{aligned}$$

which are equivalent respectively to

$$\begin{aligned} \frac{P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_i)P_{\Xi}(t_j)} &> 1 \\ \frac{P_{\Xi}(t_i) - P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_i)(1 - P_{\Xi}(t_j))} &< 1 \\ \frac{P_{\Xi}(t_j) - P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_j)(1 - P_{\Xi}(t_i))} &< 1 \\ \frac{1 - P_{\Xi}(t_i) - P_{\Xi}(t_j) + P_{\Xi}(t_i, t_j)}{(1 - P_{\Xi}(t_i))(1 - P_{\Xi}(t_j))} &> 1 \end{aligned}$$

then we obtain

$$\begin{aligned} \ln \frac{P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_i)P_{\Xi}(t_j)} &> 0 \\ \ln \frac{P_{\Xi}(t_i) - P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_i)(1 - P_{\Xi}(t_j))} &< 0 \\ \ln \frac{P_{\Xi}(t_j) - P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_j)(1 - P_{\Xi}(t_i))} &< 0 \\ \ln \frac{1 - P_{\Xi}(t_i) - P_{\Xi}(t_j) + P_{\Xi}(t_i, t_j)}{(1 - P_{\Xi}(t_i))(1 - P_{\Xi}(t_j))} &> 0 \end{aligned}$$

On the other hand, for $t, t_i, t_j \in V_{\Xi}$, from

$$\begin{aligned} 0 < P_{\Xi}(t) &= \frac{n_{\Xi}(t)}{|\Xi|} \leq 1 \\ 0 \leq 1 - P_{\Xi}(t) &< 1 \\ P_{\Xi}(t_i) &= \frac{n_{\Xi}(t_i)}{|\Xi|} \geq \frac{n_{\Xi}(t_i, t_j)}{|\Xi|} = P_{\Xi}(t_i, t_j) \\ P_{\Xi}(t_j) &= \frac{n_{\Xi}(t_j)}{|\Xi|} \geq \frac{n_{\Xi}(t_i, t_j)}{|\Xi|} = P_{\Xi}(t_i, t_j) \end{aligned}$$

we obtain

$$\begin{aligned} P_{\Xi}(t_i, t_j) &> P_{\Xi}(t_i)P_{\Xi}(t_j) > 0 \\ P_{\Xi}(t_i) - P_{\Xi}(t_i, t_j) &\geq 0 \\ P_{\Xi}(t_j) - P_{\Xi}(t_i, t_j) &\geq 0 \\ 1 - P_{\Xi}(t_i) - P_{\Xi}(t_j) + P_{\Xi}(t_i, t_j) &> (1 - P_{\Xi}(t_i))(1 - P_{\Xi}(t_j)) \geq 0 \end{aligned}$$

Hence, from the above four rewritten MIT measures, we can see that the four inequalities in (2) hold. \square

Multilabel Learning for Automatic Web Services Tagging

Mustapha AZNAG
Aix-Marseille University,
LSIS UMR 7296, France.
mustapha.aznag@univ-amu.fr

Mohamed QUAFARFOU
Aix-Marseille University,
LSIS UMR 7296, France.
mohamed.quafafou@univ-amu.fr

Zahi JARIR
University of Cadi Ayyad Marrakech,
LISI Laboratory, FSSM, Morocco.
jarir@uca.ma

Abstract—Recently, some web services portals and search engines as Biocatalogue and Seekda!, have allowed users to manually annotate Web services using tags. User Tags provide meaningful descriptions of services and allow users to index and organize their contents. Tagging technique is widely used to annotate objects in Web 2.0 applications. In this paper we propose a novel probabilistic topic model (which extends the CorrLDA model - Correspondence Latent Dirichlet Allocation-) to automatically tag web services according to existing manual tags. Our probabilistic topic model is a latent variable model that exploits local correlation labels. Indeed, exploiting label correlations is a challenging and crucial problem especially in multi-label learning context. Moreover, several existing systems can recommend tags for web services based on existing manual tags. In most cases, the manual tags have better quality. We also develop three strategies to automatically recommend the best tags for web services. We also propose, in this paper, WS-Portal; An Enriched Web Services Search Engine which contains 7063 providers, 115 sub-classes of category and 22236 web services crawled from the Internet. In WS-Portal, several technologies are employed to improve the effectiveness of web service discovery (i.e. web services clustering, tags recommendation, services rating and monitoring). Our experiments are performed out based on real-world web services. The comparisons of Precision@n, Normalised Discounted Cumulative Gain (NDCGn) values for our approach indicate that the method presented in this paper outperforms the method based on the CorrLDA in terms of ranking and quality of generated tags.

Keywords—Web services, Tags, Automatic, Recommendation, Machine Learning, Topic Models.

I. INTRODUCTION

The Service Oriented Architecture (SOA) is a model currently used to provide services on the Internet. The SOA follows the find-bind-execute paradigm in which service providers register their services in public or private registries, which clients use to locate web services. Web services¹ [27] are defined as software systems designed to support interoperable machine-to-machine interaction over a network. They are loosely coupled reusable software components that encapsulate discrete functionality and are distributed and programmatically accessible over the Internet. They are self contained, modular business applications that have open, internet-oriented and standards based interfaces [1]. Web services are autonomous software components widely used in various SOA applications according to their platform-independent nature. Different tasks like matching, ranking, discovery and composition have

been intensively studied to improve the general web services management process. Thus, the web services community has proposed different approaches and methods to deal with these tasks.

Recently, some web services portals and search engines as Biocatalogue² and Seekda!³ (Currently, the portal is no longer available.) and some other web services portals also support tags, have allowed users to manually annotate Web services using tags. User Tags provide meaningful descriptions of services and allow users to index and organize their contents. Tagging technique is widely used to annotate objects in Web 2.0 applications. This type of metadata provides a brief description of Web services and allows users to find appropriate services more easily. Tagging data provides meaningful descriptions, and is utilized as another information source for Web services.

Several web services tagging approaches have been proposed, for example the tagging system proposed in [14], [20]. However, most of them annotate web services manually. Moreover, several existing systems can recommend tags for web services based on existing manual tags [13], [9]. In most cases, the manual tags have better quality. In this paper we propose a novel approach based on our previous work on probabilistic topic models [23] to automatically tag web services according to existing manual tags. Our probabilistic topic model is a latent variable model that exploits local correlation labels. Indeed, exploiting label correlations is a challenging and crucial problem especially in Multi-Label learning context. We also develop three strategies to automatically recommend the best tags for web services. Our experiments are performed out based on real-world web services (i.e. Section IV). The experiment results show that the performance of our approach is affected by web services with or without user's tags. For this, we propose three strategies to learn the classifier before recommendation task.

The main contributions of this paper can be summarized as follows:

- 1) We propose an automatic tagging technique for web services, in which both the WSDL documents and service tags are effectively utilized. Our approach can work without existing tags, and works better when there exists manual tags.
- 2) We propose three tag recommendation strategies to improve the performance of our approach. We exploit

¹<http://www.w3.org/standards/webofservices>

²<https://www.biocatalogue.org/>

³<http://webservices.seekda.com/>

WSDL documents and related descriptions to extract the most important words and user's tags.

- 3) We generate tags for 22,236 real web services and these tags are published online in our developed Web Services Portal⁴

To validate the performance of our approach, a series of experiments are carried out. The comparisons of Precision@n, Normalised Discounted Cumulative Gain (NDCGn) values for our approach indicate that the method presented in this paper outperforms better when the selected tags from WSDL description are combined with the existing manual tags.

In this paper we propose also an enriched web service search engine called WS-Portal⁴ where we incorporate our research works to facilitate web services discovery task (see Section V) [6].

The rest of this paper is organized as follows. Section II analyzes some related work. In Section III, we describe in detail our web services tag recommendation approach. Section IV describes the experimental evaluation. Section V describes our developed web services search engine. Finally, the conclusion and future work can be found in Section VI.

II. RELATED WORK

Generally, every web service associates with a WSDL document that contains the description of the service. A lot of research efforts have been devoted in utilizing WSDL documents and Web service clustering [28], [19], [18], [12], [11] has been demonstrated as an effective mechanism to boost the performance of Web services discovery. Dong et al. [11] proposed the Web services search engine Woogle that is capable of providing Web services similarity search. However, their engine does not adequately consider data types, which usually reveal important information about the functionalities of Web services [18]. Liu and Wong [19] apply text mining techniques to extract features such as service content, context, host name, and service name, from Web service description files in order to cluster Web services. They proposed an integrated feature mining and clustering approach for Web services as a predecessor to discovery, hoping to help in building a search engine to crawl and cluster non-semantic Web services. Elgazzar et al. [12] proposed a similar approach which clusters WSDL documents to improve the non-semantic web service discovery. They take the elements in WSDL documents as their feature, and cluster web services into functionality based clusters. The clustering results can be used to improve the quality of web service search results.

Recently, tagging data provides meaningful descriptions, and is utilized as another information source for Web service. In this section, we briefly discuss some existing research works of tagging data related to different problems in web service. Meyer et al. use tags to annotate web services semantically [20]. Similarly this idea, Gawinecki et al. use structured collaborative tags to matchmake web services [14]. However, all these tags are generated manually and the authors spend 12

days to generate tags for just 50 services. Thus, manual tagging is very time-consuming and an automatic tagging system is needed for web services. To handle the problem of limited tags, Azmeh et al. [2] propose an automatic tagging system for web services which extracts tags from WSDL documents using machine learning technology and WordNet synsets. The system uses relevant synonyms in WordNet to enrich tags. Fang et al. [13] propose an approach to generate tags for web services automatically using two tagging strategies, tag enriching and tag extraction. In the first strategy, the system use clustering technique to enrich tags with existing manual tags. In the second strategy, recommended tags are extracted from WSDL documents and related descriptions. Liang et al. [10] propose a hybrid mechanism by using service-tag network information to compute the relevance scores of tags by employing semantic computation and HITS model, respectively.

In [9], the authors improve the performance of Web service clustering by introducing a novel approach based on the Author-Topic-Model [24] to explore the knowledge behind WSDL documents and tags and by proposing three tag pre-processing strategies to improve the performance of service clustering. But the system can't work if there is no manual tag in the system. Topic models are successfully used for a wide variety of applications including documents clustering and information retrieval [26], collaborative filtering [15], and visualization [16] as well as for modeling annotated data [8]. In our previous work [3], [4], we investigated the use of three probabilistic topic models PLSA, LDA and CTM to extract topics from semantically enriched service descriptions. These topics provide a model which represents any web service's description by a vector of terms. In our approach, we assumed all service descriptions were written in the WSDL and/or SAWSDL. The results obtained from comparing the three methods based on PLSA, LDA and CTM showed that the CTM model provides a scalable and interoperable solution for automated service discovery and ranking in large service repositories. The CTM model assumes that the concepts of each service arise from a mixture of topics, each of which is a distribution over the vocabulary. In this paper, we use CTM model to extract and select the candidates tag for a web services in the dataset. Then, we use the extracted tags from web service dataset to train our classifier using a latent variable model based on LocLDA (Local Correspondence Latent Dirichlet Allocation), which is a latent variable model that exploits local correlation labels [23]. LocLDA was built on Correspondence Latent Dirichlet Allocation (Corr-LDA) [8].

III. WEB SERVICES TAGS RECOMMENDATION SYSTEM

In this section, we describe the details of our web services tags recommendation approach. The overall process of our approach is divided into three phases:

- 1) Web Services Representation and Tags Extraction: We process the service descriptions and we use a probabilistic method to extract and select the candidates tag for a web services in the dataset (Section III-A).
- 2) Training Web Services Tags Recommendation Classifier: We use the extracted tags from web services dataset to train our classifier using a latent variable model (Section III-B).

⁴WS-Portal is available online:

- <http://wvmweb.esil.univ-mrs.fr/wsportal>
- <http://www.webvirtualmachine.fr/wsportal>
- <http://wsportal.aznag.net>

- 3) **Web Services Tags Recommendation:** Finally, we use the trained classifier to recommend the best tags for a new web service (Section III-C).

A. Web Services Representation and Tags Extraction

Web services are generally described with a standard Web Service Description Language (WSDL). The WSDL is an XML-based language, designed according to standards specified by the W3C, that provides a model for describing web services. It provides the specifications necessary to use the web service by describing the communication protocol, the message format required to communicate with the service, the operations that the client can invoke and the service location. To manage efficiently web service descriptions, we extract all features that describe a web service from the WSDL document (i.e. such as services, documentation, messages, types and operations).

As shown in Figure 1, our tags extraction process contains two main components, features extraction and tags selection. Before representing web services as TF-IDF (Text Frequency and Inverse Frequency) [25] vectors, we need some preprocessing. There are commonly several steps:

- **Features extraction** extracts all features that describe a web service from the WSDL document, such as service name and documentation, messages, types and operations.
- **Tokenization:** Some terms are composed by several words, which is a combination of simple terms (e.g., *get_ComedyFilm_MaxPrice_Quality*). We use therefore regular expression to extract these simple terms (e.g., *get, Comedy, Film, Max, Price, Quality*).
- **Stop words removal:** This step removes all HTML tags, CSS components, symbols (punctuation, etc.) and stop words, such as 'a', 'what', etc. The Stanford POS Tagger⁵ is then used to eliminate all the tags and stop words and only words tagged as nouns, verbs and adjectives are retained. We also remove the WSDL specific stop words, such as *host, url, http, ftp, soap, type, binding, endpoint, get, set, request, response*, etc.
- **Word stemming:** We need to stem the words to their origins, which means that we only consider the root form of words. In this step we use the Porter Stemmer Algorithm [22] to remove words which have the same stem. Words with the same stem will usually have the same meaning. For example, 'computer', 'computing' and 'compute' have the stem 'comput'. The Stemming process is more effective to identify the correlation between web services by representing them using these common stems (root forms).

After identifying all the functional terms, we calculate the frequency of these terms for all web services. We use the Vector Space Model (VSM) technique to represent each web service as a vector of these terms. In fact, it converts service description to vector form in order to facilitate the computational analysis of data. In information retrieval, VSM

is identified as the most widely used representation for documents and is a very useful method for analyzing service descriptions. The TF-IDF algorithm [25] is used to represent a dataset of WSDL documents and convert it to VSM form. We use this technique, to represent a services descriptions in the form of *Service Transaction Matrix (STM)*. In STM, each row represents a WSDL service description, each column represents a word from the whole text corpus (vocabulary) and each entry represents the TF-IDF weight of a word appearing in a WSDL document. TF-IDF gives a weight w_{ij} to every term j in a service description i using the following equation:

$$w_{ij} = tf_{ij} \cdot \log\left(\frac{n}{n_j}\right) \quad (1)$$

Where tf_{ij} is the frequency of term j in WSDL document i , n is the total number of WSDL documents in the dataset, and n_j is the number of services that contain term j .

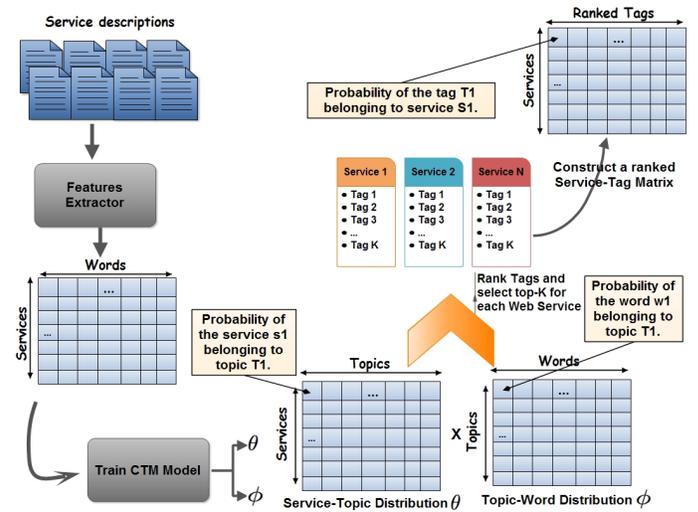


Fig. 1. An Overview of Web Services Tags Extraction Mechanism

Figure 1 presents an overview of our proposed Web Service Tag Extraction mechanism. For each Web Service, we generate top-K tags using our previous approach based on Correlated Topic Model (CTM) described in [3]. We utilized CTM to extract latent factors $z_f \in Z = \{z_1, z_2, \dots, z_k\}$ from web service descriptions (i.e., *STM*). In our work we use STM as training data for our implementation of CTM model. After the CTM model is trained, the distribution of words for each topic is known and all the services in the dataset can be described as a distribution of topics. Let

- 1) $\theta^{(s)} = P(z)$ refer to the multinomial distribution over topics in the service description s .
- 2) $\phi^{(j)} = P(w|z_j)$ refer to the multinomial distribution over words for the topic z_j .

Then, we use the extracted topics to rank the related tags for each web service. Each tag w in a service description s is generated by sampling a topic z from topic distribution (i.e. θ), and then sampling a word from topic-word distribution (i.e. ϕ). The probability of the i th tag occurring in a given service is given by Equation 2:

⁵<http://nlp.stanford.edu/software/tagger.shtml>

$$P(t_i|s) = \sum_{f=1}^k P(t_i|z_f)P(z_f|s) \quad (2)$$

Where z_f is a topic from which the i th word was drawn, $P(z_f|s)$ is the probability of topic z_f in the service s , and $P(t_i|z_f)$ is the probability of having tag t_i given the f th topic. The most relevant tags are the ones that maximize the probability $P(t_i|s)$ for a service s (See Algorithm 1)

Finally, we represent the output of this step by a matrix that contains for each web service the most related tags ranked in the descending order (i.e. $P(t_i|s)$). The main key of our approach is that the selected tags for a web service are not necessarily in its descriptions. In fact, we have represented all services in a topic space and tags are related to these topics. The result of this step will be used as input for the training classifier phase (Section III-B).

Algorithm 1 Web services tags extraction

Require:

- $S = \{s_1, \dots, s_D\}$ web services set. (D number of services).
- K Number of Topics.

Ensure: Ranked tags for each service.

- 1: Perform CTM on services set $S = \{s_1, \dots, s_D\}$.
- 2: **for** each service $s_i \in S = \{s_1, \dots, s_D\}$ **do**
- 3: **for** each word $w_m, m \in \{1, \dots, M\}$ **do**
- 4: Compute $P(w_m|s_i)$ (Equation 2)
- 5: **end for**
- 6: RankTags: The most relevant words are the ones that maximize the probability $P(w_m|s_i)$.
- 7: **end for**
- 8: **return** Set of K ranked tags for each service.

B. Training Web Services Tags Recommendation Classifier

In this step we have a dataset of service descriptions and extracted tags. From this training dataset, we first extract a list of candidate words using the probabilistic method based on CTM (Section III-A). Using this set of candidate tags, service transaction matrix and the original tags (manual tags) we train a classifier. We define our tags recommendation task as follows: given a set of web services, in which each service has not only a bag of words but also a bag of tags, our task is to learn a model using this dataset; and, when given an unseen service in which only the content words can be observed, we should predict a ranked list of tags based on the learned model and the observed words in the service. Our probabilistic approach based on LocLDA model (Local Correspondence Latent Dirichlet Allocation), which is a latent variable model that exploits local correlation labels [23]. LocLDA was built on Correspondence Latent Dirichlet Allocation (CorrLDA) [8]. More precisely, our model calculates dynamically the model structure depending on the data, and particularly, on the interaction between annotations. We originally developed the LocLDA model for generating captions for images. An image contains multiple regions, and each word in the image caption corresponds to one of the regions. The correspondence from words to regions is assumed to follow uniform distributions [23]. For our tags recommendation task, we adopt the LocLDA model for modeling the correspondence from the topic

Symbol	Description
D	Number of service in training set.
K	Number of topics.
M	Number of words related to a service.
T	Number of tags.
V	Neighborhood tags.
v	A neighbor tag: $v \in Index(Parents(tag))$.
θ	Multinomial distribution over topics: $\theta_i, i \in \{1, \dots, K\}$
z	Latent topic. $z_m^i = 1$ if z_m is the i th latent topic, else $z_m^i = 0$.
w	Word. $w_m^j = 1$ if w_m is the j th word else $w_m^j = 0$.
t	Tag. $t_n^j = 1$ if t_n is the j th tag else $t_n^j = 0$.
y	Discrete indexing variable.
W_s	Size of words vocabulary.
W_t	Size of tags vocabulary.
α	Dirichlet prior for θ : $\alpha_i, i \in \{1, \dots, K\}$
π	Multinomial: $\pi_{ij}, i \in \{1, \dots, K\}, j \in \{1, \dots, W_s\}$
β	Multinomial: $\beta_{ij}, i \in \{1, \dots, K\}, j \in \{1, \dots, W_t\}$
ϕ	Variational Multinomial: $\phi_{mi}, m \in \{1, \dots, M\}, i \in \{1, \dots, K\}$
γ	Variational Dirichlet: $\gamma_i, i \in \{1, \dots, K\}$
λ	Variational Multinomial: $\lambda_{nm}, n \in \{1, \dots, T\}, m \in \{1, \dots, M\}$
ψ	The digamma function, the first derivative of the log Gamma function.

TABLE I. NOTATIONS USED IN THIS PAPER

assignments for words and the topic assignments for tags of web services. We apply this model to the cases of automatic web services tagging. Given a service s with no tags, the task is to predict its missing tags.

Let $z = \{z_1, z_2, \dots, z_K\}$ be the latent factors that generate the web service, and $y = \{y_1, y_2, \dots, y_T\}$ be discrete indexing variables that take values from 1 to T with equal probability. Table I shows the notations used in this paper. Conditioned on T (i.e. Number of tags) and M (i.e. Number of words related to a web service), a K -topics (i.e. Number of topics), LocLDA model (Figure 2) assumes the following generative process for a pair service/tag (w, t) :

- 1) Find the parents of each tag.
- 2) Sample $\theta \sim Dirichlet(\theta|\alpha)$
- 3) For each word $w_m, m \in \{1, \dots, M\}$
 - Sample $z_m \sim Multinomial(\theta)$
 - Sample $w_m \sim p(w|z_m, \pi)$ from a multinomial distribution conditioned on z_m
- 4) For each tag $t_j, j \in \{1, \dots, T\}$
 - Sample $y_j \sim Uniform(1, \dots, T)$
 - Sample $t_j \sim p(t|y_j, y_v, z, \beta)$

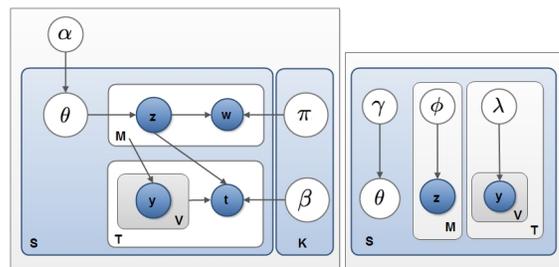


Fig. 2. (Left) Graphical model of LocLDA, (Right) Representation of variational distribution used to approximate the posterior in LocLDA.

In our model, the correspondence between a tag t_j and its associated service is obtained via a latent variable y_j . We consider that y_j is the parent of t_j and we note it by

$y_j = \text{Parent}(t_j)$. Thus, we would like to represent the interaction between different tags describing the same service. Let consider that the true caption of a given service \mathcal{S} is $\mathcal{T} = \{t_1, t_2, \dots, t_n\}$. We no longer consider that a tag $t_j \in \mathcal{T}$ is connected to the service \mathcal{S} via a single latent variable, but through a set of latent variables y_j which are parents of $t_j \in \mathcal{T} - t_j$ tags. To obtain the parents of a given tag, we first determine its neighbors by performing a multiple regression on each tag with respect to all other tags. Indeed, multiple regression is a statistical analysis that describes the relationships among variables [23]. Given a set of tags $\{t_1, t_2, \dots, t_n\}$, we seek to explain precisely the values taken by a single tag from all other tags. This process is performed for all tags. The theoretical model, formulated in terms of random variables, takes the form:

$$t_j = a_0 + a_1 t_1 + a_2 t_2 + \dots + a_n t_n + \epsilon_j$$

where ϵ is the model error that expresses the missing information in the explanation values of t_j from t_{-j} . t_{-j} represents all the tags not including the j th one. a_1, a_2, \dots, a_n are parameters to be estimated. By setting a threshold for the parameters a_i , we obtain the neighbors of a tag. We use the notation *Index*, which gives the indices of the parents of each tag (i.e. $v \in \text{Index}(\text{Parents}(\text{tag}))$ where v is a neighbor tag).

LocLDA model defines the joint distribution of the service description, tags and topics as follows:

$$P(\mathbf{w}, \mathbf{t}, \theta, \mathbf{z}, \mathbf{y} | \alpha, \pi, \beta) = P(\theta | \alpha) \prod_{m=1}^M P(z_m | \theta) P(w_m | z_m, \pi) \prod_{j=1}^T \prod_v P(y_j | M) P(y_v | M) P(t_j | y_j, y_v, z, \beta) \quad (3)$$

where α, π and β are the parameters to estimate.

The exact probabilistic inference is intractable for LocLDA, therefore, we turn to variational inference methods [17] to approximate the posterior distribution of the latent variables given a service/tag. We introduce a variational distribution q on the latent variables:

$$q(\theta, \mathbf{z}, \mathbf{y}) = q(\theta | \gamma) \prod_{m=1}^M q(z_m | \phi_m) \left(\prod_{n=1}^N q(y_n | \lambda_n) \prod_v q(y_v | \lambda_v) \right) \quad (4)$$

where γ, ϕ and λ are variational parameters.

The objective is to optimize the values of the variational parameters that make the variational distribution q close to the true posterior p by minimizing the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior. We bound the log-likelihood of a given service/tag using Jensen's inequality:

$$\begin{aligned} L(\gamma, \phi, \lambda; \alpha, \pi, \beta) &= E_q[\log P(\theta | \alpha)] + E_q[\log P(\mathbf{z} | \theta)] + E_q[\log P(\mathbf{w} | \mathbf{z}, \pi)] \\ &+ E_q[\log P(\mathbf{y} | M)] + E_q[\log P(\mathbf{t} | \mathbf{y} \in \text{parents}(\mathbf{t}), z, \beta)] \\ &- E_q[\log q(\theta | \gamma)] - E_q[\log q(\mathbf{z} | \phi)] - E_q[\log q(\mathbf{y} | \lambda)] \quad (5) \end{aligned}$$

Thus, by expanding each term of the equation 5 with respect to maximizing each variational parameter, we find the following updates rules:

- 1) Update the posterior Dirichlet parameters

$$\gamma_i = \alpha_i + \sum_{m=1}^M \phi_{mi} \quad (6)$$

- 2) For each service, update the posterior distribution over topics

$$\begin{aligned} \phi_{mi} &\propto \pi_{iw_m} \exp \left(\psi(\gamma_i) - \psi \left(\sum_{j=1}^K \gamma_j \right) \right. \\ &\left. + \sum_{n=1}^N \sum_v \lambda_{nm} \lambda_{vm} \log \beta_{it_n} \right) \quad (7) \end{aligned}$$

- 3) For each tag, update the posterior distribution over services

$$\lambda_{nm} \propto \exp \left(\sum_{i=1}^K \sum_v \phi_{mi} \lambda_{vm} \log \beta_{it_n} \right) \quad (8)$$

We maximize the lower bound with respect to the model parameters α, π, β . Given a training services set $D = \{(w_d, t_d)\}_{d=1}^D$, the objective is to find the maximum likelihood estimation for α, π, β . The corpus log-likelihood is bounded by :

$$L(D) = \sum_{d=1}^D \log P(w_d, t_d | \alpha, \pi, \beta) \geq \sum_{d=1}^D L(\gamma_d, \phi_d, \lambda_d; \alpha, \pi, \beta)$$

We then find α, π, β that maximize this lower bound:

$$\pi_{ij} \propto \sum_{d=1}^D \sum_{m=1}^{M_d} \phi_{dmi} w_{dm}^j \quad (9)$$

$$\beta_{ij} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} t_{dn}^j \sum_{m=1}^{M_d} \sum_v \phi_{dmi} \lambda_{dnm} \lambda_{dvm} \quad (10)$$

Finally, the Newton-Raphson algorithm [7] is used to estimate the Dirichlet α .

After obtaining parents of tags using the regression method, we present also in this section the variational EM algorithm [21] which performs iterative maximization of a lower bound of data in which some variables are unobserved. It maximizes a lower bound of the data log-likelihood with respect to the variational parameters, and then, for fixed values of the variational parameters, maximizes the lower bound with respect to the model parameters. Indeed, we have the following iterative algorithm:

- (E-Step) For each service, find the optimizing values of the variational parameters using equations (6), (7) and (8) with appropriate starting points for γ, ϕ_{mi} and λ_{nm} .
- (M-Step) Maximize the resulting lower bound on the log-likelihood with respect to the model parameters

for fixed values of the variational parameters, using equations (9), (10) and the Newton-Raphson algorithm.

These two steps are repeated until the lower bound on the log-likelihood converges.

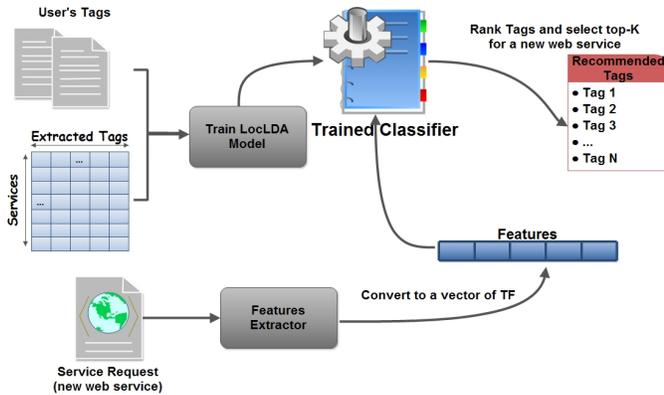


Fig. 3. An overview of Web Services Tags Recommendation mechanism

C. Web Services Tags Recommendation

Once our classifier has been trained, we can recommend tags for a new web service by performing the variational inference with fixed model parameters α , β and π . Thus, we can compute the conditional distributions of untagged service $p(t|s_{new})$ (Equation 11).

$$P(t|s_{new}) = \sum_{m=1}^M \sum_{z_m} P(z_m|\theta)P(t|z_m, \beta) \quad (11)$$

The most relevant tags are the ones that maximize the probability $P(t|s_{new})$ for a service s_{new} (See Algorithm 2)

Figure 3 presents an overview of our proposed Web Service Tag Recommendation mechanism.

Algorithm 2 Web services tags recommendation

Require: .

- $\mathcal{S} = \{s_1, \dots, s_D\}$ web services set. (D number of services).
- Set of extracted tags $\mathcal{E} = \{e_1, \dots, e_M\}$ (M number of extracted tags) (Algorithm 1).
- Set of original tags (manual tags) $\mathcal{T} = \{t_1, \dots, t_T\}$ (T number of original tags).
- K Number of Topics.
- Given service s_{new} .

Ensure: R Ranked tags for a given service.

- 1: Perform LocLDA on services datasets \mathcal{S} , \mathcal{E} and \mathcal{T}
 - 2: **for** each tag $t \in \mathcal{T}$ **do**
 - 3: Compute $P(t|s_{new})$ (Equation 11)
 - 4: **end for**
 - 5: RankTags: The most relevant tags are the ones that maximize the probability $P(t|s_{new})$.
 - 6: **return** Set of R ranked tags for a new web service s_{new} .
-

IV. EVALUATION

A. Web Services Corpus

Our experiments are performed out based on real-world web services that we collected from the web since 2011. We have considered different web service sources like Web-servicesX.net⁶, xMethods.net⁷, Seekda!⁸, Service-Finder!⁹ and Biocatalogue¹⁰. We have collected 22,236 real web services. For each Web service, we get the WSDL document and related tags if they exist. We generate tags for all web services in the dataset and these tags are published online in our Web Services search engine⁴.

Before applying the proposed approach, we process the WSDL corpus. The objective of this pre-processing is to identify the textual words of services, which describe the semantics of their functionalities. WSDL corpus processing consists of several steps: *Features extraction*, *Tokenization*, *Stop words removal*, *Word stemming* and *Service Transaction Matrix construction*. The observed words are represented in a Service Transaction Matrix (STM). In our work we use service transaction matrix as training data for our models.

To evaluate our method, we select 633 web services from our dataset. all these services have manual tags. We selected only the services having 3 to 10 manual tags, and there are totally 739 manual tags belonging to them. Then we use different approaches to tag these web services:

- 1) Original tags: In this approach, we just use the 633 web services and their tags to train our classifier and recommend tags for new web services.
- 2) Extracted tags: In this approach, we just generate extracted tags and select top-k extracted tags as final results (III-A).
- 3) Original tags + Extracted tags: In this approach, we mix original tags with extracted tags from WSDL documents.

The proposed approach is evaluated using the *Precision at n* ($Precision@n$) and the *Normalised Discounted Cumulative Gain* ($NDCG_n$) for the generated tags obtained for each of the service in the test set.

All experiments were performed on a Dell 64-bit Server with Intel®Xeon(R) CPU X5560 @ 2.80GHz x 16 and 16 Go of RAM.

B. Metrics Evaluation

In order to evaluate the accuracy of our approach, we compute two standard measures used in *Information Retrieval*: *Precision at n* ($Precision@n$) and *Normalised Discounted Cumulative Gain* ($NDCG_n$). $Precision@n$ and $NDCG_n$ are widely accepted as the metrics for ranking evaluation in IR. Formally, the previous metrics are defined as follows:

⁶<http://www.webservices.net/ws/default.aspx>

⁷<http://www.xmethods.net/ve2/index.po>

⁸<http://www.webservices.seekda.com>

⁹<http://demo.service-finder.eu/search>

¹⁰<https://www.biocatalogue.org/>

1) **Precision@n**: In our context, $Precision@n$ is a measure of the precision of the service tag recommendation and ranking system taking into account the first n retrieved tags. The $precision@n$ for a list of retrieved tags is given by Equation 12:

$$Precision@n = \frac{|RelevantTags \cap RetrievedTags|}{|RetrievedTags|} \quad (12)$$

Where the list of relevant tags to a given service is the ground truth tags related to the service.

2) **Normalised Discounted Cumulative Gain**: $NDCG_n$ uses a graded relevance scale of each retrieved tag from the result set to evaluate the gain, or usefulness, of a tag based on its position in the result list. This measure is particularly useful in Information Retrieval for evaluating ranking results. The $NDCG_n$ for n retrieved tags is given by Equation 13.

$$NDCG_n = \frac{DCG_n}{IDCG_n} \quad (13)$$

Where DCG_n is the Discounted Cumulative Gain and $IDCG_n$ is the Ideal Discounted Cumulative Gain. The $IDCG_n$ is found by calculating the DCG_n of the ideal first n generated tags for a given service. The DCG_n is given by Equation 14

$$DCG_n = \sum_{i=1}^n \frac{2^{relevance(i)} - 1}{\log_2(1 + i)} \quad (14)$$

Where n is the number of tags retrieved and $relevance(s)$ is the graded relevance of the tag in the i th position in the ranked list. The $NDCG_n$ values for all tags can be averaged to obtain a measure of the average performance of a ranking algorithm. $NDCG_n$ values vary from 0 to 1.

In Information retrieval, $NDCG_n$ gives higher scores to systems which rank a result list with higher relevance first and penalizes systems which return tags with low relevance.

3) **Caption Perplexity**: We compute the perplexity of the given tags under $P(t|s)$ for each service s in the test set to measure the tags quality of the models. In computational linguistics, the measure of perplexity has been proposed to assess generalizability of text models. The perplexity is algebraically equivalent to the inverse of the geometric mean per-word likelihood [8]. A lower perplexity score indicates better generalization performance. Assume we have D web services as a held-out dataset D_{test} and each web service s contains N_d tags. More formally, the perplexity for a dataset D_{test} is defined by:

$$Perplexity = \exp \left(- \sum_{d=1}^D \sum_{n=1}^{N_d} \frac{\log P(t_n|s_d)}{\sum_{d=1}^M N_d} \right) \quad (15)$$

Where $P(t_n|s_d)$ is the probability of having tag t_n given the d -th. service.

C. Results and Discussion

The choice of the number of topics corresponding to the original dataset has an impact on the interpretability of the results. In LocLDA and CorrLDA model the number of topics must be decided before training phase. There are several

methods to choose the number of topics that lead to best general performance [26]. We evaluated the performance of our system using $AveragePrecision$ for increasing numbers of topics and the results peak at $K = 70$ (where K is the number of topics) before the performance starts to decrease. These evaluation results are shown in Figures 4 and 5. As observed from these figures, the better performance is obtained for the approach when the extracted and original tags are used to learn our models. We also evaluated the performance of our system by computing the perplexity of LocLDA and CorrLDA according to the three strategies described previously. Figures 6 and 7 show the perplexity of the dataset test for each model by varying the number of topics (lower numbers are better). The results show that LocLDA and CorrLDA models achieve best performance when we mix the original tags and extracted tags.

As the manual creation of ground truth costs a lot of work, we use the 10% service of the dataset test and we generate top-10 tags using the probabilistic method based on CTM (Section III-A). The generated tags are considered as the true labels to evaluate the performance of our Web services tags ranking system. In addition, for each service in the dataset test, each of its tags is labeled as one of the five levels $relevance(s) \in \{1, 2, 3, 4, 5\}$ where 5 denotes *Most Relevant*, 4 denotes *Relevant*, 3 denotes *Partially Relevant*, 2 denotes *Weakly Relevant*, and 1 denotes *Irrelevant*.

The averaged $Precision@n$ and $NDCG_n$ were measured for up to the first ten generated tags from the complete list of results. These evaluation results are respectively shown in Figures 8 and 9. The results show that our approach performs better than the method based on CorrLDA model.

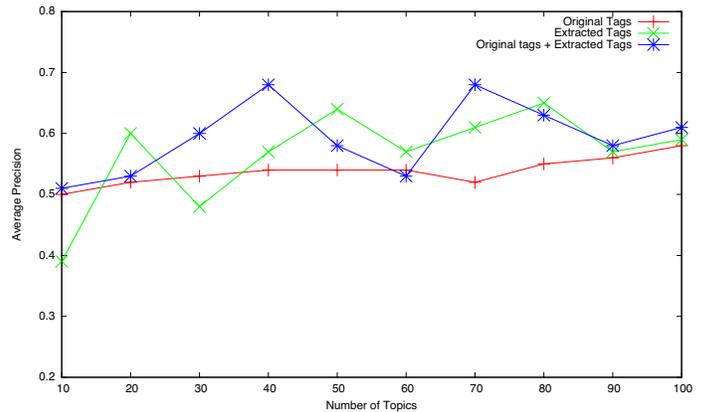


Fig. 4. Comparison of average $Precision$ values over all dataset test for CorrLDA.

V. WS-PORTAL; AN ENRICHED WEB SERVICES SEARCH ENGINE

In this section, we describe some functionalities for our web services search engine where we incorporate our research works to facilitate web service discovery task. Our WS-Portal⁴ contains 7063 providers, 115 sub-classes of category and 22236 web services crawled from the Internet [6]. In WS-Portal, several technologies, i.e., web services clustering, tags recommendation, services rating and monitoring are employed

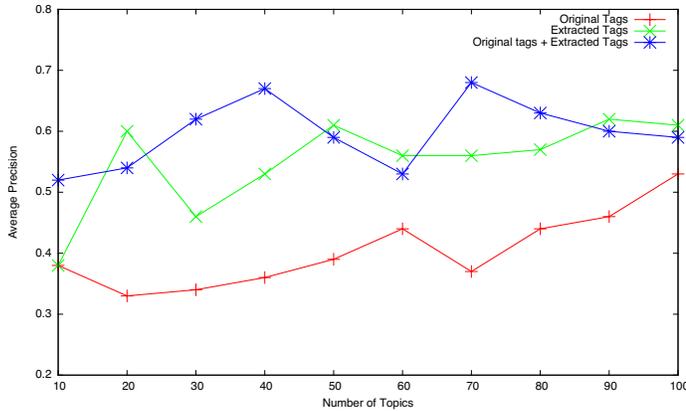


Fig. 5. Comparison of average *Precision* values over all dataset test for LocLDA.

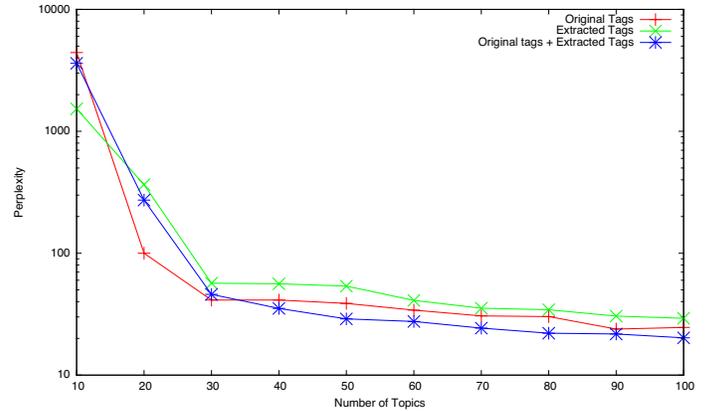


Fig. 7. Perplexity values obtained for learned LocLDA model.

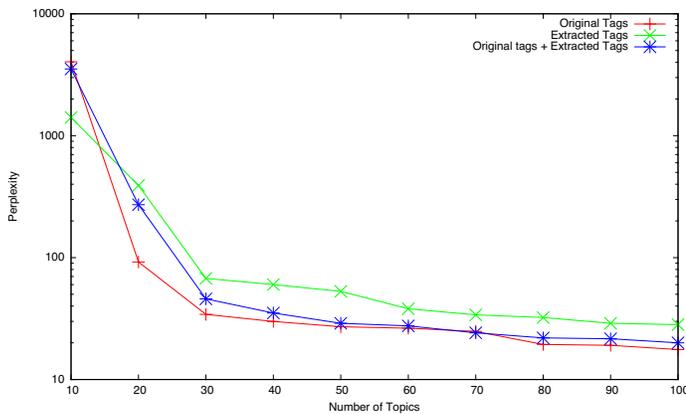


Fig. 6. Perplexity values obtained for learned CorrLDA model.

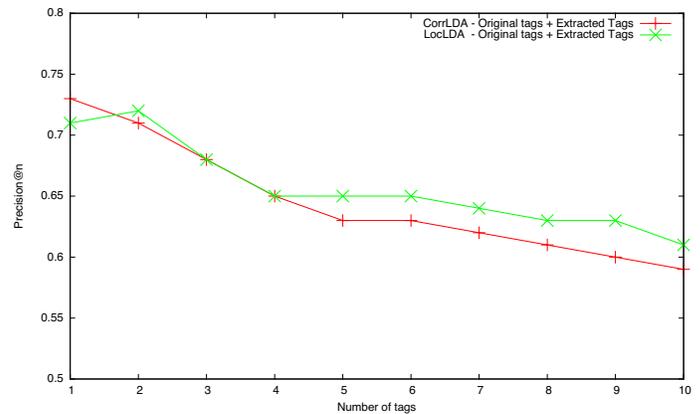


Fig. 8. Comparison of average *Precision@n* values for CorrLDA and LocLDA over the third dataset test (Original Tags and Extracted Tags).

to improve the effectiveness of web services discovery. Specifically, probabilistic topics models are utilized for clustering, services/topics and tags recommendation [3], [4], [5]. We use probabilistic topic models to extract topic from semantic service descriptions and search for services in a topics space where heterogeneous service descriptions are all represented as a probability distribution over topics.

A. Service Clustering

By organizing service descriptions into clusters, services become easier and therefore faster to discover and recommend. Web services are described as a distribution of topics [4]. A distribution over topics for a given service s is used to determine which topic best describes the service s . K clusters are created where K is the number of generated topics.

B. Service Discovery

Service Discovery and Selection aim to find web services with user required functionalities. A user query represented by a set of words is represented as a distribution over topics [3], [4], [5]. The service discovery is based on computing the similarity between retrieved topic's services and a user's query. We use the topics browsing technique as another method search to discover the web services that match with users requirements. Users can select the related topic to the their

query and our system gives automatically the topic's services that match with user's query.

C. Tags recommendation

We use the automatic tagging technique proposed in this paper to recommend automatically the tags for all published services in our repository.

D. Availability and performance monitoring

WS-Portal monitor all registered services. In addition, after registering a service in our service registry its availability will be monitored automatically. Our system measures the availability by calling the service endpoints periodically.

E. Services rating and comments posting

Our system allows users to rate and post comments to enrich the service descriptions.

F. Dynamic service invocation

Our system allows users to invoke the selected service using the html form generated automatically from the associated WSDL document for each service operations.

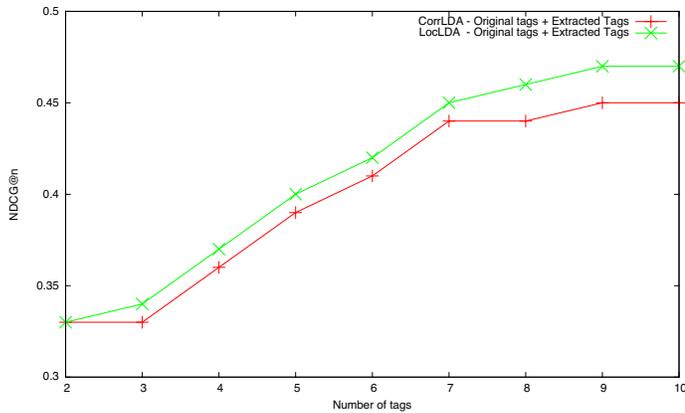


Fig. 9. Comparison of average $NDCG@n$ values for CorrLDA and LocLDA over the third dataset test (Original Tags and Extracted Tags).

G. User Interface

Our web services search engine is available online⁴ and consumers can use it to discover, register or annotate web services. Figure 10 shows the site home page of our Web Services Search Engine. When users submit the search form, our system gives a list of services that match with user's query and each search result entity show a brief service description:

- 1) Web service name,
- 2) Service description,
- 3) Tags given by users,
- 4) Service category,
- 5) Service provider,
- 6) Average rating score given by users,
- 7) Service availability.

In addition our system select automatically a top five related topics to the user's query. When users select a disered service, WS-Portal gives more details for selected service such as service name, wsdl url, service documentation, provider, categories, country, availability, rating score, user's tags, recommended tags and WSDL cache. Our system gives also more details for service monitoring (availabilty and response time values for each service endpoints). In addition, users can rate, annotate the selected service and post comments. Finally, users can invoke the selected service using the html form generated automatically from WSDL document for each service operations. Our system gives also two others important informations such as similar services and the related topics to the selected service. Indeed, we use the extracted topics from services descriptions to calculate the similarity between the selected service and others web services in our repository. For this, we compute the similarity score, using some probability metrics such as *Cosine Similarity* and *Symmetric KL Divergence* [5], between the vectors containing the service's distribution over topics. Finally, similar services are ranked in order of their similarity score to the selected service. Thus, we obtain automatically an efficient ranking of the services retrieved.

VI. CONCLUSION

In this paper, we propose a novel approach based on probabilistic topic model to tag web services automatically.

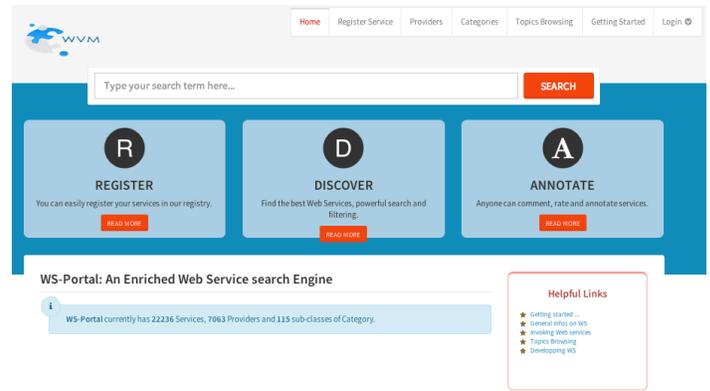


Fig. 10. Site Home Page of Our Web Services Search Engine.

Three tags recommendation strategies are also developed to improve the system performance. Our system performs better when we mix the original tags and extracted tags from WSDL documents. A series experiments prove that our method is very effective. The comparisons of Precision@n, Normalised Discounted Cumulative Gain ($NDCG_n$) values for our approach indicate that the method presented in this paper outperforms the method based on the CorrLDA in terms of ranking and quality of generated tags. We have presented also in this paper the Web Services Search engine developed to facilitate the service discovery process. In the future, we will focus our research on how to automatically tag RESTful services.

REFERENCES

- [1] Alonso, G., Casati, F., Kuno, H., Machiraju, V.: Web Services - Concepts, Architectures and Applications. Springer Verlag, Berlin Heidelberg, 2004.
- [2] Azmeh, Z.; Falleri, J.-R.; Huchard, M. and Tibermacine, C.: Automatic Web Service Tagging Using Machine Learning and WordNet Synsets, in International Conference on Web Information Systems and Technologies (WEBIST 2010).
- [3] Aznag, M., Quafafou, M. and Jarir, Z.: Correlated Topic Model for Web Services Ranking. In International Journal of Advanced Computer Science and Applications (IJACSA), vol. 4, no. 6, pp. 283–291, July 2013.
- [4] Aznag, M., Quafafou, M., Rochd, El M., and Jarir, Z.: Probabilistic Topic Models for Web Services Clustering and Discovery. In the European Conference on Service-Oriented and Cloud Computing (ESOC'2013), Springer LNCS 8135, pages 19-33, 11 September 2013.
- [5] Aznag, M., Quafafou, M. and Jarir, Z.: Leveraging Formal Concept Analysis with Topic Correlation for Service Clustering and Discovery. In 21th IEEE International Conference on Web Services (ICWS 2014). Alaska, USA.
- [6] Aznag, M., Quafafou, M. and Jarir, Z.: WS-Portal: An Enriched Web Services Search Engine. In 12th International Conference on Service Oriented Computing (ICSOC 2014), Paris, France.
- [7] Blei, D., Ng, A. Y. and Jordan, M. I.: Latent dirichlet allocation. Journal of Machine Learning Research, vol. 3:993-1022, 2003.
- [8] Blei, D., and Jordan, M.: Modeling annotated data. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, page 127-134. ACM Press, (August 2003)
- [9] Chen, L.; Wang, Y.; Yu, Q.; Zheng, Z. and Wu, J.: WT-LDA: User Tagging Augmented LDA for Web Service Clustering., in 12th International Conference on Service Oriented Computing (ICSOC'2013), Springer LNCS 8274, pages 162-176.

- [10] Chen, L., Wu, J., Zheng, Z., Lyu, M. R., Wu, Z.: Modeling and Exploiting Tag Relevance for Web Service Mining. in Knowledge and Information Systems Vol. 39, No. 1, pp 153-173. April 2014.
- [11] Dong, X., Halevy, A., Madhavan, J., Nemes, E., Zhang, J.: Similarity Search for Web Services. In VLDB Conference, Toronto, Canada, pp. 372-383, 2004.
- [12] Elgazzar, K., Hassan A., Martin, P.: Clustering WSDL Documents to Bootstrap the Discovery of Web Services. In IEEE International Conference on Web Services (ICWS'2010), pp. 147-154.
- [13] Fang, L.; Wang, L.; Li, M.; Zhao, J.; Zou, Y. and Shao, L.: Towards Automatic Tagging for Web Services., in IEEE International Conference on Web Services (ICWS 2012).
- [14] Gawinecki, M.; Cabri, G.; Paprzycki, M. and Ganzha, M., WSColab: Structured Collaborative Tagging for Web Service Matchmaking., in International Conference on Web Information Systems and Technologies (WEBIST 2010), pages 70-77.
- [15] Hofmann, T.: Collaborative filtering via Gaussian probabilistic latent semantic analysis. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 259-266. ACM Press, 2003.
- [16] Iwata, T., Yamada, T., and Ueda, N.: Probabilistic latent semantic visualization: topic model for visualizing documents. In KDD'2008: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 363-371. ACM, 2008.
- [17] Jordan, M.I., Ghahramani, Z., Jaakkola, T.S. and Saul, L.K: An introduction to variational methods for graphical models. In *Machine Learning*, 37:183–233, 1999.
- [18] Kokash, N.: A Comparison of Web Service Interface Similarity Measures. *Frontiers in Artificial Intelligence and Applications*, Vol. 142, pp.220-231, 2006.
- [19] Liu, Wei., Wong, W.: Web service clustering using text mining techniques. In *International Journal of Agent-Oriented Software Engineering (IJAOSE'2009)*, Vol. 3, No. 1, pp. 6-26.
- [20] Meyer, H. and Weske, M.: Light-Weight Semantic Service Annotations Through Tagging., in *International Conference on Service Oriented Computing (ICSOC'2006)*, Springer LNCS 4294, pages. 465-470.
- [21] Neal, R.M. and Hinton, G.E. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *M.I. Jordan, editor, Learning in Graphical Models*, pages 355-368. Kluwer, 1998.
- [22] Porter, M. F.: An Algorithm for Suffix Stripping, In: *Program* 1980, Vol. 14, No. 3, pp. 130-137.
- [23] Rochd, E. M., Quafafou, M.; Aznag, M.: Encoding Local Correspondence in Topic Models. In *IEEE 25th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pp.602,609, Washington, 4-6 Nov 2013.
- [24] Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In *20th Conference on Uncertainty in Artificial Intelligence*. pp. 487-494 (2004)
- [25] Salton, G.: *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA (1989).
- [26] Steyvers, M. and Griffiths, T.: Probabilistic topic models. In *Latent Semantic Analysis: A Road to Meaning*, T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, Eds. Laurence Erlbaum, 2007.
- [27] W3C (2004). Web services architecture. Technical report, W3C Working Group Note 11 February 2004.
- [28] Zheng, Z., Ma, H., Lyu, M.R., King, I.: QoS-aware Web service recommendation by collaborative filtering. *IEEE Transactions on Service Computing* 4(2), 140-152 (2011)

cFireworks: a Tool for Measuring the Communication Costs in Collective I/O

Kwangho Cha

National Institute of Supercomputing and Networking,
Korea Institute of Science and Technology Information,
Daejeon, KOREA
Email: khocha@kisti.re.kr

Abstract—Nowadays, many HPC systems use the multi-core system as a computational node. Predicting the communication performance of multi-core cluster systems is complicated job, but finding out it is important to use multi-core system efficiently. In the previous study, we introduced the simple linear regression models for predicting the communication costs in collective I/O. In the models, however, because it is important to get the communication characteristics of the given system, we designed *cFireworks*, an MPI application to measure the communication costs of HPC systems. In this paper, we explain the detail concept and experimental results of *cFireworks*. The performance evaluation showed that the expected communication costs with the linear regression models generated by using the output of *cFireworks* are reasonable to use.

Keywords—Collective I/O; Communication Costs; Parallel Computing; Parallel I/O

I. INTRODUCTION

Because modern HPC systems consist of multi-core computational nodes, the systems frequently issue the complex intra-node and inter-node communications. In such systems, predicting the communication performance is difficult, but it is an important process to use HPC systems efficiently.

Collective I/O is the specialized I/O which provides the functions of single-file based parallel I/O. As the number of processes and the size of a problem increase, the importance of collective I/O is also emphasized. The most well known parallel programming library, the message passing interface (MPI), also supports collective I/O and it follows the two-phase I/O scheme in order to improve the collective I/O performance[1], [2], [3], [4]. The two-phase I/O consists of data exchange phase and I/O phase. In terms of data exchange phase, it has to generate a number of complicated communication operations and they become some parts of collective I/O overheads.

In the previous study[5], we have shown it is possible to improve the performance of collective I/O by reducing the communication costs. Furthermore, we also have demonstrated that finding out the expected communication costs before launching an application is important to reduce the communication costs in collective I/O. We used the linear regression models for predicting the communication costs and it was important to understand the communication characteristics of given systems in order to get the reasonable linear regression model. For this reason, we considered making *cFireworks*, an MPI application to measure the communication characteristics of multi-core cluster systems and partially introduced the

basic concept of *cFireworks* in the previous work[5]. In this paper, we explain the more detail and improved concept of *cFireworks* and draw the experimental results with different kinds of multi-core cluster systems.

This paper is organized as follows. The previous research on communication model is summarized in Section II. Section III presents the main concept of *cFireworks*. The results of performance evaluations are described in Section IV. Finally, the conclusions are presented in Section V.

II. COMMUNICATION MODEL

When someone want to understand the process of communications or communication costs, it is helpful to use a valid communication model. In this section, we explain some communication models, such as the classical one and the linear regression model for collective I/O communications.

The *LogP* model is very well-known communication model which uses four parameters: L , o , g , and P stand for latency, overhead, bandwidth, and processors respectively[6][7]. It assumes a message passing procedure in distributed memory system and is intended for short messages. Many variants of *LogP* have been introduced as the system environments change[8][9].

Nowadays, many HPC systems use the multi-core system as a computational node. Communications in multi-core cluster systems are classified into two groups: intra-node and inter-node communications. In those multi-core cluster systems, because each core can communicate simultaneously, the communication media should be shared. Vienne et al.[10] suggested a predictive model for concurrent communication in multi-core systems. It sets several elementary sections of conflict parts and gets the communication time by predicting the cost of each section.

In some case, such as collective I/O, it is possible to expect the communication costs involving all processors by obtaining the communication time in the bottlenecked computational node[5]. Especially, data exchange time in collective I/O is proportional to the communication time in the hot-spot node. The simple linear model which uses the number of intra- and inter-node communications was introduced in order to expect the communication time in a node. The primary role of the prediction function in the study was predicting the relative performance of a given node set rather than obtaining accurate performance of the set. For this reason, they used a simple and

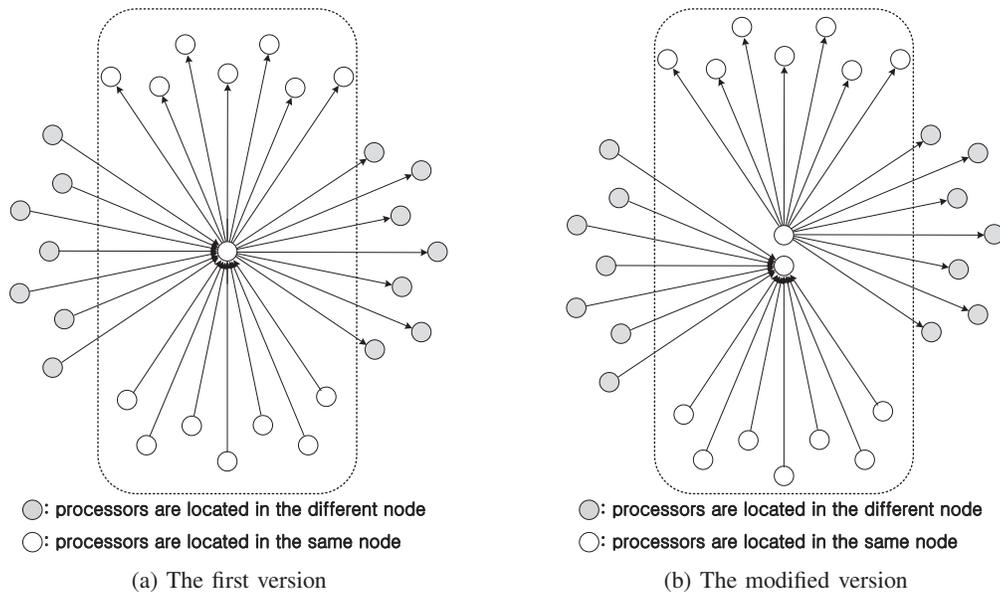


Fig. 1: Basic concept of *cFireworks*. The dotted lines represent a node; the circle in the center indicates the root process. *cFireworks* iterates to measure the communication time as an increase in the number of intra-node and inter-node communications

intuitive approach. The data exchange time in node n_i can be described as:

$$T_{n_i}(ca_i, ce_i) = \alpha \cdot ca_i + \beta \cdot ce_i + \gamma \quad (1)$$

where ca_i is the number of intra-node communications within n_i and ce_i is the number of inter-node communications of n_i .

III. *cFireworks*

In the previous study, we discovered that the data exchange time of collective I/O was determined by the communication time of the most overloaded node. Furthermore the communication time is represented by α , β and γ in equation (1). Because these values are related with the characteristics of the given system and communication procedures, it is necessary to identify the communication characteristics of the given system. For this reason we created a test program called *cFireworks*, in order to measure the appropriate communication parameters for the system.

Figure 1 shows the basic concept of the *cFireworks* test. In the first version of *cFireworks*, a process acts as a hot spot. In the real world, however, some processes in the same node can concurrently participate in the intra- and inter-node communications. For this reason, we designed the second version of *cFireworks* reflecting this situation. In the modified version, *cFireworks* has multiple hot spot processes. The processes are assigned to sub-groups and the processes send or receive data to their hot spot process in the sub-group. In this way, the program generates multiple concurrent communications in a node.

Algorithm 1 explains the pseudo code of *cFireworks*. It measures the communication time of a node by varying the number of intra- and inter-node communications. There is a simple double loop for increasing the number of intra- and inter-node communications (line 2, 3, 16, and 17)

and the communication times with each number of communication pair are measured in every iteration.

There are two kinds of procedures to post asynchronous communications. In case of the first procedure intra-node communications are posted first (line 5 and 9), while the second procedure issues inter-node communications first instead of the intra-node ones (line 19 and 23). In other words, in the first measurement method, it generates the intra-node communications and then launches the inter-node communications; whereas in the second method, the inter-node communications are called first instead of the intra-node communications. In many cases, calling the intra-node communications first shows slightly better performance.

IV. PERFORMANCE EVALUATION

All experiments in this study were performed with Tachyon cluster systems¹. Table I describes the specifications of Tachyon I and II system. A computational node of Tachyon I has four quad core CPUs, AMD's Barcelona. Each CPU is equipped with 2 Mbytes L3 cache memory, DDR2 memory controllers and HyperTransport controller. Tachyon II is equipped with Intel's Nehalem CPU which has an 8 Mbytes shared cache memory and DDR3 memory controllers.

A. Results of the *cFireworks* tests

Figures 2, 3, and 4 show the results of the *cFireworks* in the Tachyon I and II cluster system with a message size of 4 Mbytes. In order to reduce the number of iterations, *cFireworks* measures the communication time with a pair of intra- and

¹They are KISTI's fourth supercomputers and the phase I system is ranked at 130 in the list of TOP500 most powerful supercomputers published in June 2008, and the phase II system is ranked at 14 in the list released in November 2009[11].

TABLE I: Specifications of KISTI Tachyon cluster systems

	Hardware		Software	
	Tachyon I	Tachyon II	Tachyon I	Tachyon II
CPU	AMD Opteron 2.3GHz	Intel Xeon 2.93GHz	OS	CentOS 4.6 RedHat Enterprise 5.3
No. of nodes	188	3,176	MPI	MVAPICH2 1.4
No. of CPU cores	3008	25,408	File System	Lustre 1.6.6 Lustre 1.8.1.1
No. of CPU cores/node	16	8	Queue Scheduler	SGE 6.1u5 SGE 6.2u5
No. of CPU sockets/node	4	2		
Socket to socket bandwidth	8GB/s	25.6GB/s		
Memory	32GB/node	24GB/node		
Interconnection network	InfiniBand 4× DDR	InfiniBand 8× QDR		

Algorithm 1 cFireworks algorithm

```

1: procedure INTRA_FIRST           ▷ Intra-node communication first
2:   for x = 0; x < half_star; x++ do
3:     for y = 0; y < half_star; y++ do
4:       ...
5:       for z = 0; z < numprocs; z++ do
6:         MPI_irecv(recv_buff,...);
7:       end for
8:       ...
9:       for z = 0; z < numprocs; z++ do
10:        MPI_isend(send_buff,...);
11:      end for
12:    end for
13:  end for
14: end procedure

15: procedure INTER_FIRST          ▷ Inter-node communication first
16:   for x = 0; x < half_star; x++ do
17:     for y = 0; y < half_star; y++ do
18:       ...
19:       for z = numprocs - 1; z ≥ 0; z-- do
20:         MPI_irecv(recv_buff,...);
21:       end for
22:       ...
23:       for z = numprocs - 1; z ≥ 0; z-- do
24:         MPI_isend(send_buff,...);
25:       end for
26:     end for
27:   end for
28: end procedure

```

inter-node communications. That is, the hot spot process in Fig. 1 has the same number of ingress links and egress links for intra- or inter-node communications, respectively. For this reason, we’ve used a linear regression model obtained from the measured data considering equation (1) in order to cover every possible number of communications in a node. Figure 2a, 3a, and 4 illustrate the regression models derived from the data: the values of their coefficient of determination, R^2 , are approximately 0.98s.

In case of Tachyon I, Figs. 2 and 3 show that the increasing rates of the communication time had altered when there were more than two pairs of intra-node communications. That is, when the number of intra-node communications is in the range of 2 and 7, the graph shows the rapid increases in communication time unlike the results between 0 and 2. We checked the system throughput with the measured data and

could find that when the number of intra-node communications was less than 2, the throughput of the node still increased. If, however, it was more than two, the throughput remained steady and didn’t increase further. Consequently, the condition of that the number of intra-node communications reaches two is a criterion to determine whether the throughput of a node is saturated or not. For this reason, we’ve split the linear regression model into two variants: one for when throughput of the node is not saturated and another for when the throughput is saturated. By subdividing the regression model, the correctness of the model is improved. For example, when the number of intra-node communications is in the range of 2 and 7, R^2 s are approximately 0.99s.

B. Validation test for cFireworks

In this section, we introduce the results of validation tests. The results of cFireworks were used for predicting the communication costs of collective I/O. In order to generate collective I/O workload, we used the MPI-Tile-IO benchmark[12] and validated whether the linear regression models can provide a good indicator or not by comparing the execution time of MPI-Tile-IO and the results of *cFireworks*. In the test, a 4×4 array was distributed to 16 processes, which wrote and read an 1 GB file. If the selected nodes have the different number of processes, the communication times in collective I/O are different according to the sequence of the nodes[5]. The performance was measured using four types of node sets that had 16 processes from the eight nodes as described in Table II and Figure 5.

Figure 6 shows the communication cost of the MPI-Tile-IO and the expected values obtained by the linear regression models. In order to focus on the data exchange phase itself, the execution time without the file I/O phase was measured². In terms of collective I/O, if the size of I/O request is larger than the collective buffer size, collective I/O iterates the data exchange and I/O phases multiple times. We assumed that the data exchange time for a single iteration is proportional to the entire data exchange time and the linear regression models are used for predict the time for a single iteration. This is the reason why there is a gap between the measured data and the predicted ones in those figures.

²In most of MPI library, the write and read operations have the same communication workloads in the data exchange phase; however, unlike the read operation, the write operation has additional routines for *post write* and *read modify write*. Therefore, this causes the write operation to use more time than the read operation.

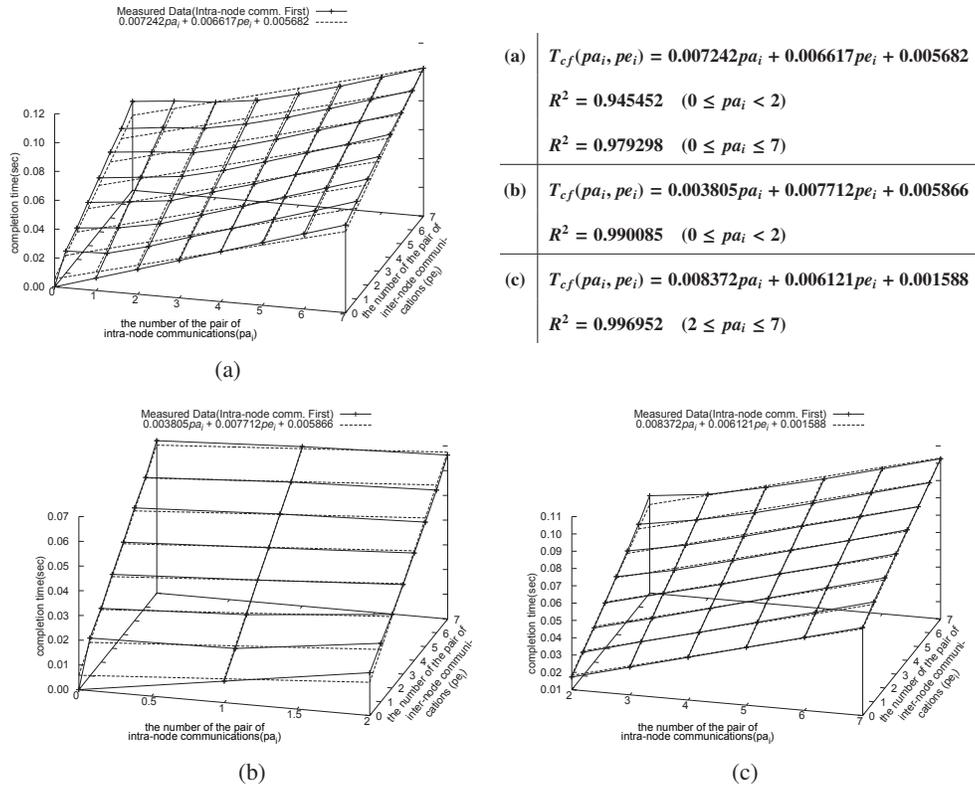


Fig. 2: Results of the *cFireworks* and their linear regression models (Tachyon I, intra-node communication first)

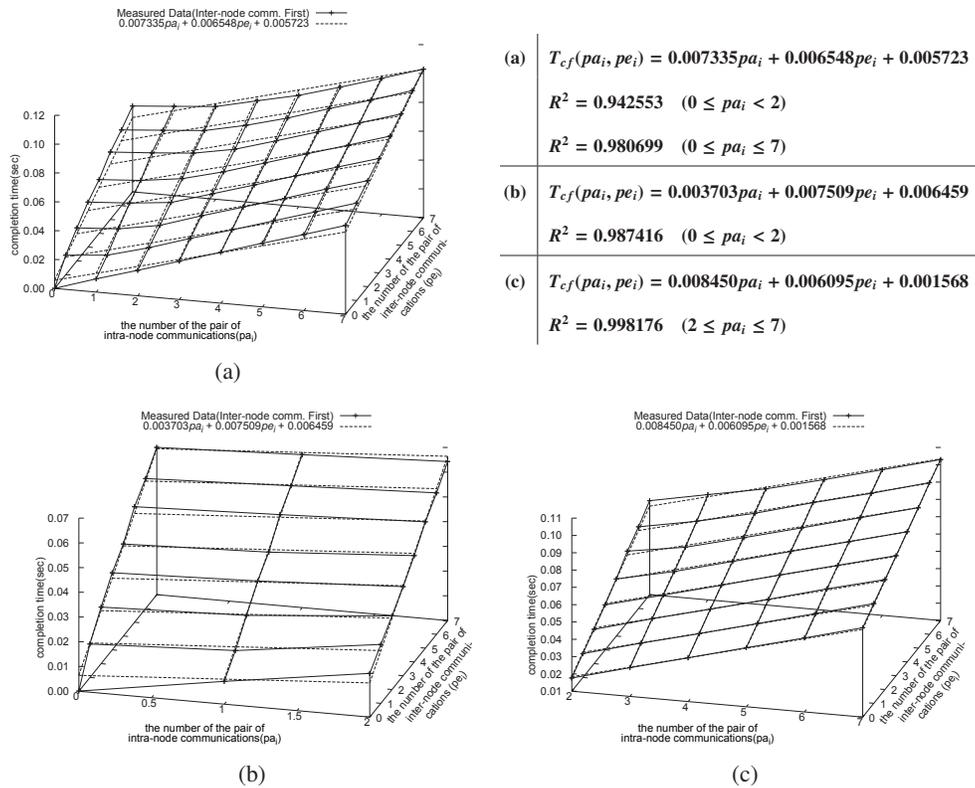


Fig. 3: Results of the *cFireworks* and their linear regression models (Tachyon I, inter-node communication first)

TABLE II: Test cases for the evaluation of the prediction functions

Tests	Node set	Expected Communication Costs			
		Tachyon I		Tachyon II	
		Intra-node comm. first	Inter-node comm. first	Intra-node comm. first	Inter-node comm. first
T16-01	{4,4,2,2,1,1,1,1}	0.052138	0.051513	0.015699	0.016514
T16-02	{1,1,1,1,2,4,4,2}	0.040519	0.040198	0.013773	0.014291
T16-03	{1,1,2,2,1,1,4,4}	0.052138	0.051513	0.015699	0.016514
T16-04	{1,1,1,4,4,2,2,1}	0.034710	0.034541	0.012810	0.013180

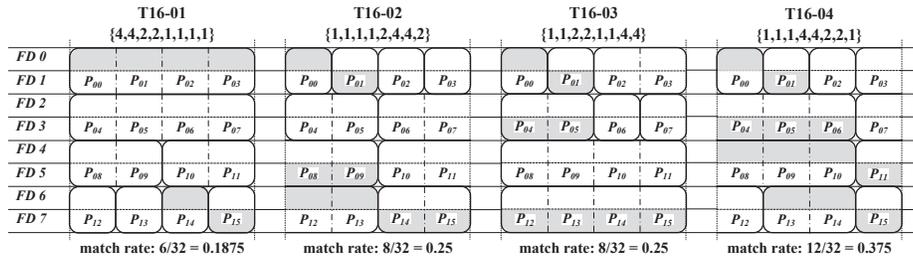
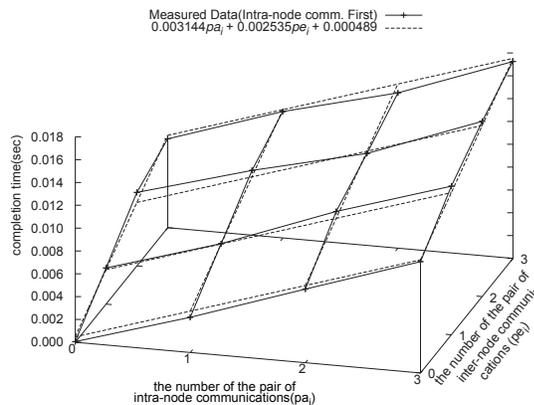
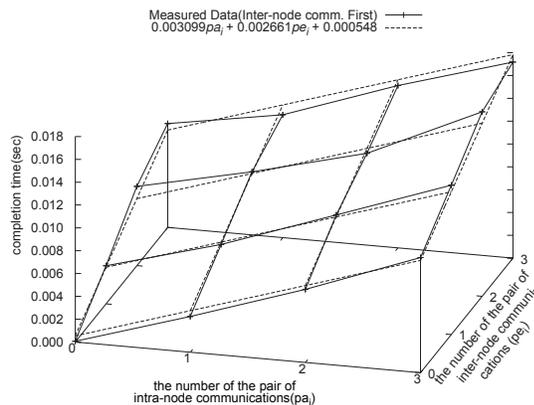


Fig. 5: Data distribution of each test cases in Table II



(a) intra-node comm. first, $R^2 = 0.989794$



(b) inter-node comm. first, $R^2 = 0.984673$

Fig. 4: Results of the *cFireworks* and their linear regression models (Tachyon II)

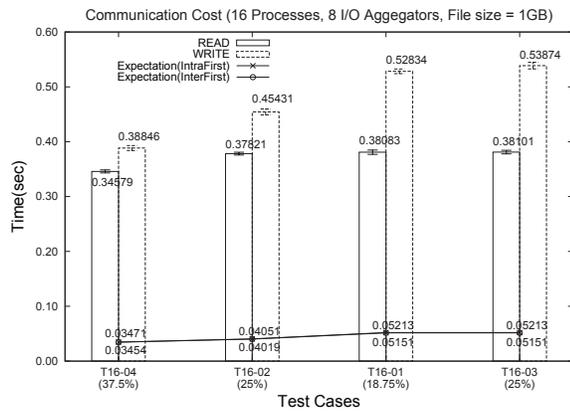
As seen in Table II and Fig. 6, the prediction values and measured date of Tachyon II are much less than those of Tachyon I. That is, the communication costs of Tachyon II are lower than those of Tachyon I because the communication performance of Tachyon II is much higher.

The result of the experiment also demonstrates that the regression model can provide reasonable predictions in general. As seen in Table II, we used four kinds of test sets for the experiments. Because each node set has the different order of nodes communication patterns in collective I/O are also changed. In other words, each test case has the different number of intra- and inter-node communications in a hot spot node and this hot spot node determines the communication time of collective I/O. We input the number of communications in hot spot node of each test into our regression model and compared the results with the measured data.

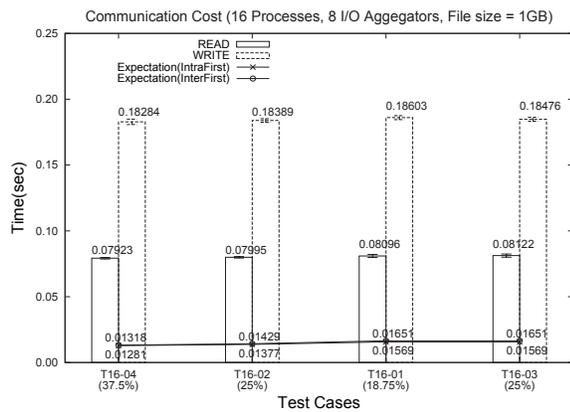
The experimental results in Fig. 6 showed that our regression model could generate the reasonable prediction values. Because the predicted values are proportional to the real measured data in a greater or less degree, it is possible to use our regression model as a prediction model which can find a good node set without MPI execution. The performance differences among node sets in Tachyon II are not significant but the linear regression model still can tell the expected communication performance of Tachyon II.

V. CONCLUSION

Although predicting the communication performance of multi-core cluster systems is troublesome task, finding out the expected communication performance is important. In this study, we introduced *cFireworks*, an MPI application to measure the communication costs of HPC systems and the outputs of *cFireworks* were used for generating the linear regression models for predicting the communication costs. The results of performance evaluation showed that the expected communication costs with the linear regression models are reasonable to use. Furthermore, they also proved that *cFireworks*



(a) Tachyon I



(b) Tachyon II

Fig. 6: Expected values and real data exchange times (Tachyon I and Tachyon II)

is simple and intuitive to use and helpful to generate the linear regression models.

REFERENCES

- [1] Rajeev Thakur, William Gropp, and Ewing Lusk, "Data Sieving and Collective I/O in ROMIO," in Proc. of the 7th Symposium on the Frontiers of Massively Parallel Computation, pp. 182-189, 1999.
- [2] Kwangho Cha, "An Efficient I/O Aggregator Assignment Scheme for Multi-core Cluster Systems," IEICE Transactions on Information and Systems, vol. E96-D, no. 2, pp. 259-269, 2013.
- [3] Kwangho Cha, and Seungryoul Maeng, "An Efficient I/O Aggregator Assignment Scheme for Collective I/O Considering Processor Affinity," in Proc. of the International Conference on Parallel Processing Workshops 2011 (SRMPDS 2011), pp. 380-388, Sep. 2011, Taipei, Taiwan
- [4] Kwangho Cha, Taeyoung Hong, and Jeongwoo Hong, "The Subgroup Method for Collective I/O," in Proc. of the 5th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT 2004), LNCS 3320, pp. 301-304, Dec. 2004.
- [5] Kwangho Cha, and Seungryoul Maeng, "Reducing Communication Costs in Collective I/O in Multi-core Cluster Systems with Non-exclusive Scheduling," The Journal of Supercomputing, vol. 61, no. 3, pp.966-996, 2012.
- [6] David Culler, Richard Karp, David Patterson, Abhijit Sahay, Klaus Erik Schauer, Eunice Santos, Ramesh Subramonian, Thorsten von Eicken, "LogP: towards a realistic model of parallel computation," in Proc. of the fourth ACM SIGPLAN symposium on Principles and practice of parallel programming (PPOPP), pp. 1-12, 1993.

- [7] David E. Culler, Richard M. Karp, David Patterson, Abhijit Sahay, Eunice E. Santos, Klaus Erik Schauer, Ramesh Subramonian, Thorsten von Eicken, "LogP: a practical model of parallel computation," Communications of the ACM, vol. 39, no. 11, pp. 78-85, 1996.
- [8] Thilo Kielmann, Henri E. Bal, Kees Verstoep, "Fast Measurement of LogP Parameters for Message Passing Platforms," Lecture Notes in Computer Science (15 IPDPS 2000 Workshops), vol. 1800, pp. 1176-1183, 2000.
- [9] Torsten Hoefler, Torsten Mehlan, Frank Mietke, Wolfgang Rehm, "LogfP - A Model for small Messages in InfiniBand," in Proc. of the 20th International Parallel and Distributed Processing Symposium(IPDPS), 2006.
- [10] Jérôme Vienne, Maxime Martinasso, Jean-Marc Vincent, Jean-François Méhaut, "Predictive models for bandwidth sharing in high performance clusters," in Proc. of the IEEE International Conference on Cluster Computing, 286-291, 2008.
- [11] TOP 500 Supercomputer Sites, <http://www.top500.org>, Accessed 14 August 2014
- [12] Parallel I/O Benchmarking Consortium, <http://www.mcs.anl.gov/research/projects/pio-benchmark>, Accessed 14 August 2014

Measuring Term Specificity Information for Assessing Sentiment Orientation of Documents in a Bayesian Learning Framework

D. Cai

School of Computing and Engineering
University of Huddersfield, HD1 3DH, UK
Email: d.cai@hud.ac.uk

Abstract—The assessment of document sentiment orientation using term specificity information is advocated in this study. An interpretation of the mathematical meaning of term specificity information is given based on Shannon's entropy. A general form of a specificity measure is introduced in terms of the interpretation. Sentiment classification using the specificity measures is proposed within a Bayesian learning framework, and some potential problems are clarified and solutions are suggested when the specificity measures are applied to estimation of posterior probabilities for the NB classifier. A novel method is proposed which allows each document to have multiple representations, each of which corresponds to a sentiment class. Our experimental results show, while both the proposed method and IR techniques can produce high performance for sentiment classification, that our method outperforms the IR techniques.

Index Terms—term specificity information; specificity measure; naive Bayes classifier; sentiment classification.

I. INTRODUCTION

The proliferation of web-centred social interaction has led to increasing quantities of opinion-dense text. The availability of this data, and the range of scientific, commercial, social and political uses to which it may be put, has reinforced interest in opinions as objects of analysis and fuelled the growth of text Sentiment Classification (SC). Sentiment analysis draws from, and contributes to, broad areas of text analytics, natural language processing and computational linguistics. The basic task for the analysis is to classify the polarity of a given text: whether the opinion expressed is positive or negative. Early studies at the whole document level such as [1], [2] used several methods to classify the polarity of product reviews and movie reviews, respectively. Classifying document polarity on n -ary scales, rather than just positive and negative, can be found, for instance, in [3]–[5]. Good reviews of SC methods can be found, for instance, in [5]–[7].

Generally, three main issues need to be considered in statistical methods of SC: i) methodologies to *identify* sentiment-bearing terms; ii) models to *represent* documents with the identified terms; iii) classifiers to *classify* each document by predicting a class that is most likely to generate the document representation. This study focuses on the second issue: design method to represent documents using *Term Specificity Information (TSI)* for accurate and reliable SC.

Several classical classifiers, such as Naive Bayes (NB),

k-Nearest Neighbours (kNN), Maximum Entropy (ME) and Support Vector Machine (SVM) have been developed further for SC. Studies have shown NB and SVM to be superior methods for SC [8]–[12]. Studies [13], [14] have experimentally shown performance benefits of representing documents using *TSI* along with SVM for SC. Our experimental results (not discussed in this paper) obtained from *TSI* with SVM, also support these conclusions.

In order to develop SC classifiers with a predictive capability, we need to know the explicit representation of the opinionated documents. That is, we have to design a *weighting* function to generate the document representation corresponding to the individual sentiment classes (each class is treated as a sub-collection). The weights of terms may be expected to enhance the likelihood of correctly predicting document sentiment orientation. This stage is crucial for SC, in particular, for estimating the posterior probability required by the NB classifier. There have been extensive studies on document representation in other areas, such as IR, in which a controlled vocabulary is constructed and the weights of carefully selected terms are used to represent the content of documents over the whole collection.

Specificity information measurement can be naturally and conveniently utilized to estimate posterior probabilities required in the NB classifier. Therefore, this study concentrates on SC in a Bayesian learning framework (rather than in SVM), in which, document representation using *TSI* is essential. The NB classifier is surprisingly effective in practice since its classification decision can be correct even if its probability estimates are inaccurate [15], [16], and it often competes well with more sophisticated classifiers [16], [17]. There are theoretical reasons [17] for the apparently unreasonable efficacy of the NB classifier. However, there has been no systematic discussion on how to use *TSI* to represent documents for SC and there exist some potential problems in applying specificity measures to the NB classifier for SC.

It is worth mentioning, rather than considering all terms in documents, that [18] attempts to determine the specificity of nouns. One possible indicator of specificity is how often the noun is modified: a very specific noun is rarely modified, while a very general noun is often modified. There are three categories of the modifiers: (prenominal) adjectives, verbs, or

other nouns. Their study uses the probability that the noun is modified by any of the modifiers and the probability that the noun is modified by each specific category. Their work considers also how much the modifiers vary: a noun with a great variety of modifiers suggests that it is general, whereas a noun rarely modified or modified by only a few different ways is likely specific. Clearly, their work is entirely different from methods presented in this paper. It is evident that the method given in [13] is a special case of one of our methods.

There are three main concerns in this study. First, we interpret the mathematical meaning of a basic concept on specificity information conveyed by a given term based on Shannon's entropy, and introduce a formal definition of a specificity measure in terms of the interpretation. Second, we propose a general method to represent the statistical importance of terms pertaining to individual documents with estimation of posterior probabilities using term weights obtained from *TSI* for the NB classifier. Third, we clarify some potential problems inherent in applying the specificity measures in a Bayesian learning framework and, then suggest solutions that are easy to apply in practice. Our methods allow each document to have multiple representations, each of which corresponds to a specific sentiment class, which we believe is of benefit to SC tasks. In addition, we present some experimental results, evaluating performance against a standard collection, MovieReviews [19], to verify that both *TSI* and the difference of *TSIs* over the individual sentiment classes may be regarded as appropriate measures for SC.

The remainder of the paper is organized as follows. Section 2 focuses on the mathematical interpretation and formal definition of *TSI*. Section 3 proposes a general form of the NB classifier with posterior probability estimation using *TSI*. Section 4 clarifies problems of applying *TSI* and suggests solutions. Some experimental results of our method are presented in Section 5 and conclusions are drawn in Section 6.

II. TERM SPECIFICITY INFORMATION (*TSI*)

This section gives a mathematical interpretation and formal definition of specificity information of terms.

To begin, let us introduce the notation. Let C be a *collection* of documents and $d \in C$ be a document. Let \mathcal{C} be the *classification* of documents over C and $X \in \mathcal{C}$ (or, $X \subseteq C$) be a *class*. Let V be a *vocabulary* of all the terms used to index individual documents. Let $V_X \subseteq V$ be the sub-vocabulary consisting of those terms appearing in at least one document $d \in X$ and $V_d \subseteq V$ be the set of terms appearing in document $d \in C$.

For simplicity, all our discussions are set to the situation where $|C| = 2$. Such a setting can be easily generalized to any finite number of classes. Thus, we have $\mathcal{C} = \{X, \bar{X}\}$, where $X = C_P$ (or, $X = C_N$) is a possible sentiment class consisting of all positive (or, negative) documents. Generally, $V_X \cap V_{\bar{X}} \neq \emptyset$, as terms often occur in both positive and negative documents.

A. A General Form of a *TSI* Measure

Intuitively, a term is said to contain *specificity* information if it tends to be capable of isolating the few documents of interest from many others.

Consider a conditional probability distribution $P_x(d|t)$ satisfying: $P_x(d|t) \geq 0$ and $\sum_{d \in X} P_x(d|t) = 1$. The entropy function (Shannon's entropy) of $P_x(d|t)$ is

$$H(P_x(d|t)) = - \sum_{d \in X} P_x(d|t) \log P_x(d|t)$$

where $P_x(d|t)$ is called the document frequency distribution (over X) of term $t \in V_X$. We here adopt the notational convention: $y \log(y) = 0$ if $y = 0$.

Note that, from the properties of the entropy function, if term t is uniformly distributed over X :

$$P_x(d|t) = \frac{1}{|X|} \quad \text{for every } d \in X$$

where $|X|$ is the cardinality of X , then the entropy of t arrives at the maximum:

$$H(P_x(d|t)) = - \sum_{d \in X} \frac{1}{|X|} \log \frac{1}{|X|} = \log(|X|) = H_{max}$$

which is called the maximum entropy of term t . Clearly, we have $H_{max} \geq 0$ as $|X| \geq 1$. $H(P_x(d|t))$ can be regarded as a measure of the degree of uncertainty based on what we know about t concerning X . Thus, t is said to be more informative than $t' \in V_X$ if $H(P_x(d|t)) < H(P_x(d|t'))$ as t reduces uncertainty. The reduction based on $H(P_x(d|t))$ essentially amounts to specificity information of t .

The above statements may already mathematically interpret what it is meant by the basic concept of specificity information conveyed by term t . Thus, we can now introduce a formal definition as follows.

Definition 2.1 For a given class $X \in \mathcal{C}$ and an arbitrary term $t \in V_X$, suppose $P_x(d|t)$ be the conditional probability distribution over X . A general form of a *term specificity information* measure, denoted by $tsi_x(t)$, is defined by

$$tsi_x(t) = \begin{cases} H_{max} - H(P_x(d|t)) & t \in V_X \\ \text{undefined} & t \in V - V_X \end{cases} \quad (1)$$

which measures the extent of uncertainty reduction caused by, or the amount of specificity information of, t concerning X .

Clearly, we have $tsi_x(t) \geq 0$ for every $t \in V_X \subseteq V$. A basic idea for $tsi_x(t)$ is: if term t has a skewed document frequency distribution, $P_x(d|t)$, over X , then t may be expected to be a good discriminator for distinguishing the few documents of interest from many others in X .

If we accept the assumption that the importance of a term in representing each document is dependent significantly, if not completely, on its specificity over the individual classes, the problem is then reduced to choosing a suitable specificity measure. With Definition 2.1, we discuss below two concrete specificity measures to clarify ideas involved in the general form given in Eq.(1).

B. Example TSI Measures

Two well-known specificity measures, $idf_x(t)$ and $int_x(t)$, as examples, are reconsidered to illustrate the general form, and the relationship between the two specificity measures are established based on the general form.

Example 2.1 Perhaps the most well-known measure capturing the specificity information of term t concerning some class X is the *inverse document frequency* [20]:

$$idf_x(t) = \begin{cases} \log \frac{|X|}{n_x(t)} & t \in V_X \\ \text{undefined} & t \in V - V_X \end{cases} \quad (2)$$

where $n_x(t)$, called the *document frequency* of t in X , is the number of documents in X in which term t occurs.

In order to interpret $idf_x(t)$ in terms of the entropy function as given in Eq.(1), let us consider documents represented by binary vectors. Note that t appears in at least one document of X , so $n_x(t) \neq 0$, for every $t \in V_X$. Then, for a given $t \in V_X$, the document frequency distribution for the binary representation is:

$$P_x(d|t) = \begin{cases} \frac{1}{n_x(t)} & d \in X_t \\ 0 & d \in X - X_t \\ \text{undefined} & d \in D - X \end{cases} \quad (3)$$

where $X_t \subseteq X$ is the set of document(s) in which t appears. Thus, we obtain

$$H(P_x(d|t)) = - \sum_{d \in X} \frac{1}{n_x(t)} \log \frac{1}{n_x(t)} = \log(n_x(t))$$

Hence, from Eq.(2) and Eq.(3), we have

$$idf_x(t) = \log(|X|) - \log(n_x(t)) = H_{max} - H(P_x(d|t))$$

which is the exact expression given in Eq.(1) when $t \in V_X$.

The measure $idf_x(t)$ states that the specificity of term $t \in V_X$ is inversely proportional to the document frequency over X . Therefore, it assigns higher values to more specific terms that tend to be capable of isolating few documents from the many others. However, $idf_x(t)$ does not take into consideration term frequency within documents, and terms with the same document frequency will be treated equally by assigning the same weights. \square

Example 2.2 A more accurate indication of term importance may be obtained by incorporating term frequency information into the document frequency distribution, which is *noise* of a term [21], it may be used to capture the unspecificity information of term $t \in V_X$ concerning some class X :

$$noise_x(t) = H(P_x(d|t)) = - \sum_{d \in X} \frac{f_d(t)}{f_x(t)} \log \frac{f_d(t)}{f_x(t)}$$

which is the entropy of the document frequency distribution:

$$P_x(d|t) = \begin{cases} \frac{f_d(t)}{f_x(t)} & t \in V_X \\ \text{undefined} & t \in V - V_X \end{cases} \quad (4)$$

where $f_d(t)$ is the frequency of t in d , $f_x(t) = \sum_{d \in X} f_d(t)$ is the total frequency of t in X . In other words, $noise_x(t)$

measures the extent of the lack of concentration of occurrence of t ; it emphasizes the uselessness of those terms that are in agreement with $P_x(d|t)$ for all the documents in X .

Note that the specificity of term t is in inverse relation to its noise. Thus, the specificity of t may be computed, for instance, by

$$int_x(t) = \begin{cases} H_{max} - noise_x(t) & t \in V_X \\ \text{undefined} & t \in V - V_X \end{cases} \quad (5)$$

which, called the *inverse noise* of t , is the same expression given in Eq.(1).

Because the measure $int_x(t)$ assigns low values to those terms that are not concentrated in a few particular documents, but instead are prevalent in X , it should be an appropriate measure of term specificity. \square

It is worth mentioning that there are two statistical concepts [22] used widely to test the performance of a binary classification: *sensitivity* of a test is the proportion of actual positives which are correctly predicted; *specificity* of a test is the proportion of negatives which are correctly predicted. Clearly, they are entirely different from our above discussion (i.e., the specificity of a term, rather than a test) and used in different contexts: sensitivity and specificity estimate the ability of the tests to predict positive and negative results, respectively.

III. SENTIMENT CLASSIFICATION USING TSI

So far, we have given a formal account of *TSI*. We are now in a position to see how the NB classifier, along with estimation of posterior probabilities using term weights obtained from *TSI*, can be used for effective SC.

A. The NB Classifier

The NB classifier is a learning method that requires an estimate of the posterior probability that a document belongs to some sentiment class, and then it classifies the document into the class with the highest posterior probability.

More specifically, the NB classifier is constructed based on Bayes' theorem with a strong conditional independence assumption. That is, for a possible sentiment class $X = C_P$ (or $X = C_N$), it computes the posterior probability, $p(X|d)$, that document $d \in C$ belongs to X :

$$p(X|d) = \frac{p(X)}{p(d)} \cdot p(d|X) \propto p(X) \cdot \prod_{t \in V_d} p(t|X) \quad (6)$$

where $p(t|X)$ is the conditional probability of term t occurring in some document of class X , $p(d)$ is the probability that a randomly picked document is d , and $p(X)$ is the probability that a randomly picked document belongs to class X . Note that $p(d)$ in Eq.(6) can be omitted as it is a scaling factor dependent only on terms, and that $p(t|X)$ may be interpreted as a measure of evidence of how much contribution t makes to support class X . Taking logarithms of probabilities on both sides of Eq.(6), we can write the NB Classifier by:

$$\Gamma(d, X) = \log(p(X)) + \sum_{t \in V_d} \log(p(t|X)) \quad (7)$$

given $0 < p(X) < 1$ and $p(t|X) > 0$ (where $t \in V_d$). Then document d is classified into class X^* if it has the highest posterior probability or, equivalently, it satisfies:

$$\Gamma(d, X^*) = \max \{ \Gamma(d, X), \Gamma(d, \bar{X}) \}$$

The parameters given in Eq.(7), such as, *a priori* probability $p(X)$ and the *posterior* probability $p(t|X)$ may be estimated by

$$p(X) = \frac{|X|}{|C|} \quad (8)$$

$$p(t|X) = \begin{cases} \frac{\varpi_x(t)}{\sum_{t \in V} \varpi_x(t)} & t \in V_X \\ \text{undefined} & t \in V - V_X \end{cases} \quad (9)$$

where $\varpi_x(t)$ is a *weighting function* estimating the importance of term $t \in V_X$ in representing class X .

Estimation of $p(X)$ is normally straightforward, it may be, for instance, the ratio of class cardinalities of X and C as given in Eq.(8). Estimation of $p(t|X)$ is however the main concern of studies and our discussion below is based on using term weights obtained from *TSI* as discussed in the last section.

B. Estimation of Posterior Probabilities

It can be seen, from Eq.(9), that estimation of $p(t|X)$ is uniquely determined by its argument $\varpi_x(t)$. Generally, we can express

$$\varpi_x(t) = \begin{cases} \sum_{d \in X} \pi(d) \cdot w_{d|X}(t) & t \in V_X \\ \text{undefined} & t \in V - V_X \end{cases} \quad (10)$$

in which, $w_{d|X}(t)$ is a *weighting function* estimating the importance of t in representing document $d \in X$; $\pi(d)$ is a function indicating the presumed importance of d in X . Thus, $\varpi_x(t)$ is the sum of weights, multiplied by the importance of the corresponding d , of term $t \in V_d$ over all documents $d \in X$.

It is now clear, with the general expression given in Eq.(10), that estimation of $p(t|X)$ is reduced to estimation of two components, $w_{d|X}(t)$ and $\pi(d)$, of $\varpi_x(t)$.

1) Estimation of $w_{d|X}(t)$

As we know, document representation, $w_{d|X}(t)$, plays an essential role in determining SC effectiveness. The issue of accuracy and validity of representation has long been a crucial and open problem. It is beyond the scope of this paper to discuss the issue in greater detail. A detailed discussion about representation techniques may be found, for instance, in [23].

Our concern is with applying *TSI* for the estimation of posterior probability required in the NB classifier. Therefore, in order to give a general expression of $w_{d|X}(t)$ incorporating term specificity information, we need to introduce a further piece of notation—we need to define the intuitive concept of specificity strength of terms over the classification.

Definition 3.1 Suppose we have a classification $\mathcal{C} = \{X, \bar{X}\}$. The *specificity strength* of term t in support of X against \bar{X}

is defined by

$$\Delta tsi_{(X:\bar{X})}(t) = \begin{cases} tsi_X(t) - tsi_{\bar{X}}(t) & t \in V_X \cap V_{\bar{X}} \\ \text{undefined} & t \in V - V_X \cap V_{\bar{X}} \end{cases} \quad (11)$$

where $tsi_X(t)$ is the *TSI* measure given in Eq.(1).

Obviously, the larger the *difference* is, the more specificity information term t conveys in support of X against \bar{X} . Thus, $\Delta tsi_{(X:\bar{X})}(t)$ may be regarded as the specificity strength of t over \mathcal{C} and as an appropriate measure for SC. Clearly, unlike $tsi_X(t)$, $\Delta tsi_{(X:\bar{X})}(t) \geq 0$ may or may not hold for every $t \in V_X \cap V_{\bar{X}}$.

Now we are ready to formally write $w_{d|X}(t)$. Suppose each document $d \in X$ can be represented as a $1 \times n$ matrix $M_{d|X} = [w_{d|X}(t)]$. With Definitions 2.1 and 3.1, a general expression of a weighting function incorporating term specificity information is defined as follows.

Definition 3.2 Suppose we have a classification $\mathcal{C} = \{X, \bar{X}\}$. A general form of the *weight* of term t in representing document $d \in X$ is defined by

$$w_{d|X}(t) = \begin{cases} w_{d|X}(f_d(t), \mathfrak{S}_{(X:\bar{X})}(t)) & t \in V_X \cap V_{\bar{X}} \\ \text{undefined} & t \in V - V_X \cap V_{\bar{X}} \end{cases} \quad (12)$$

where $\mathfrak{S}_{(X:\bar{X})}(t)$ is the *TSI* measure given in either Eq.(1) or Eq.(11).

It is worth emphasizing that we here express the weighting function by $w_{d|X}(t)$ rather than by $w_d(t)$. That is, our method facilitates SC with the NB classifier: it allows each document to have multiple representations, each of which corresponds to a specific sentiment class X . Estimation of term weights has been extensively studied in the area of IR. However, in traditional IR, document d is represented by a single weighting function $w_d(t)$ corresponding to the whole collection C .

Example 3.1 We may write a number of weighting functions. Eight weighting functions, derived immediately from Eq.(2) and Eq.(5), are given in Table I. The eight functions, and their variations, are widely applied in many applications (and they will be used in our experiments presented in Section 5). \square

TABLE I
EIGHT WEIGHTING FUNCTIONS $w_{d|X}(t)$

Symbols	Descriptions
idf	$idf_X(t)$
tf-idf	$f_d(t) \cdot idf_X(t)$
int	$int_X(t)$
tf-int	$f_d(t) \cdot int_X(t)$
Δidf	$idf_X(t) - idf_{\bar{X}}(t) = \log \frac{p(X)}{1-p(X)} - \log \frac{n_X(t)}{n_{\bar{X}}(t)}$
tf- Δidf	$f_d(t) \cdot [idf_X(t) - idf_{\bar{X}}(t)]$
Δint	$int_X(t) - int_{\bar{X}}(t) = \frac{H(P_{\bar{X}}(d t))}{\log(X)} - \frac{H(P_X(d t))}{\log(X)}$
tf- Δint	$f_d(t) \cdot [int_X(t) - int_{\bar{X}}(t)]$

We point out that [13] showed good performance using measure $\Delta tfidf = \log \frac{p(X)}{1-p(X)}$, along with SVM, for SC. It

is now clear that their measure is a special case of $\text{tf} \cdot \Delta \text{idf}$ (i.e., when $|X| = |\bar{X}|$).

2) Estimation of $\pi(d)$

There may be many ways to construct function $\pi(d)$. Two functions given in the example below indicate how they can be applied in practice.

Example 3.2 Let $\mathcal{V} = \{\mathcal{V}_X, \mathcal{V}_{\bar{X}}\} \subset V$ be the set of all sentiment-bearing terms selected, in which, \mathcal{V}_X is the subset consisting of all positive (or, negative) terms. Generally, we have $\mathcal{V}_X \cap \mathcal{V}_{\bar{X}} = \emptyset$, but $\mathcal{V}_X \cap V_{\bar{X}} \neq \emptyset$ (or, $\mathcal{V}_{\bar{X}} \cap V_X \neq \emptyset$), that is, a strong positive (or, negative) sentiment-bearing term may also occur in a negative (or, positive) document. Thus, for a given document $d \in C$, we may write a function:

$$\pi_1(d) = \begin{cases} \mu \cdot \left[1 + \frac{|V_d \cap \mathcal{V}_X|}{L_d}\right] & d \in X, V_d \cap \mathcal{V}_{\bar{X}} = \emptyset \\ \mu_1 & d \in X, V_d \cap \mathcal{V}_{\bar{X}} \neq \emptyset \\ \mu_2 & d \in \bar{X} \end{cases}$$

In particular, when $\mu = 0$, we have

$$\pi_2(d) = \begin{cases} \mu_1 & d \in X \\ \mu_2 & d \in \bar{X} \end{cases}$$

where $\mu, \mu_1, \mu_2 \geq 0$ are constants and $L_d = \sum_{t \in V_d} f_d(t)$ is the length of d . \square

The function $\pi_1(d)$ may involve SC using a small set of strong *sentiment-bearing* terms. For instance, two lists of strong positive and negative terms may be

$\mathcal{V}_X = \{\text{admirable, beautiful, creative, delicious, excellent, ...}\}$

$\mathcal{V}_{\bar{X}} = \{\text{aggravated, bored, confused, depressed, enraged, ...}\}$

respectively. The terms in the lists may be obtained in manual term selection and, hence, they may or may not be relevant to domains of interest or to training topics. Clearly, when taking $\mu \geq \mu_i$ ($i = 1, 2$), $\pi_1(d)$ assigns a relatively higher value to those documents that contain many strong sentiment-bearing terms in \mathcal{V}_X but contain no strong sentiment-bearing term in $\mathcal{V}_{\bar{X}}$; $\pi_1(d)$ is normally needed for applications where a set of good samples for learning is essential.

The function $\pi_2(d)$ is a special case of $\pi_1(d)$: there is no a set of strong sentiment-bearing terms and, thus it assigns the same value to all documents in X (or, \bar{X}). $\pi_2(d)$ is simple and may be the most commonly used function in practice: it indicates that all documents within X (or, \bar{X}) are treated as equally important; $\pi_2(d)$ may be needed when one has no particular reason to emphasize any document in X (or \bar{X}).

IV. PROBLEMS APPLYING TSI FOR SC

It seems that our method is a straightforward application of TSI, but it has some potential pitfalls. This section reveals problems and suggests solutions when applying the TSI measures for estimation of posterior probabilities for the NB classifier.

A. Problems

Let us first consider a simple example below, in which, the document frequency distributions are derived by expressions Eq.(3) and Eq.(4) and the values of term specificity information are computed by measures given in Eq.(2) and Eq.(5).

Example 4.1 Suppose we are given $C = \{d_1, \dots, d_7\}$, $C_P = \{d_1, \dots, d_4\}$, $C_N = \{d_5, d_6, d_7\}$ and $V = \{t_1, \dots, t_6\}$. Then we have $V_{C_P} = \{t_1, t_2, t_3, t_4, t_6\}$, $V_{C_N} = \{t_1, t_4, t_5, t_6\}$, and $V_{C_P} \cap V_{C_N} = \{t_1, t_4, t_6\}$. Thus, the term occurrence frequencies and the document frequency distributions are shown in Tables II and III, respectively, and the values of term specificity information computed by $tsi_x(t) = idf_X(t)$ and $tsi_x(t) = int_X(t)$ are given in Table IV. For instance, for $t_1 \in V_{C_P}$, we have

$$\begin{aligned} noise_{C_P}(t_1) &= - \sum_{d \in C_P} \frac{f_d(t_1)}{f_{C_P}(t_1)} \log \frac{f_d(t_1)}{f_{C_P}(t_1)} \\ &= - \left[\frac{1}{7} \log \frac{1}{7} + \frac{3}{7} \log \frac{3}{7} + \frac{1}{7} \log \frac{1}{7} + \frac{2}{7} \log \frac{2}{7} \right] \\ &= - \frac{1}{7} \cdot \log \frac{3^3 \times 2^2}{7^7} \end{aligned}$$

with expression Eq.(5) and $H_{max} = \log(|C_P|)$, we obtain

$$\begin{aligned} int_{C_P}(t_1) &= \log(|C_P|) - noise_{C_P}(t_1) \\ &= 4 - \left[-\frac{1}{7} \cdot \log \frac{108}{7^7} \right] \end{aligned}$$

Note that, in the above computation, we adopt the notational conventions: $y \log(y) = 0$ if $y = 0$. \square

TABLE II
TERM OCCURRENCE FREQUENCIES

	$f_d(t_1)$	$f_d(t_2)$	$f_d(t_3)$	$f_d(t_4)$	$f_d(t_5)$	$f_d(t_6)$
d_1	1	2	0	1	0	0
d_2	3	0	2	0	0	1
d_3	1	0	3	2	0	0
d_4	2	3	1	0	0	0
d_5	0	0	0	1	2	3
d_6	1	0	0	0	2	2
d_7	0	0	0	2	3	2

Some problems arise from the above example. First, for a given class X , the specificity measures $tsi_x(t)$ and $\Delta tsi_{(X:\bar{X})}(t)$ are meaningless for every $t \in V - V_X$ and for every $t \in V - (V_X \cap V_{\bar{X}})$, respectively. That is, some terms may have no TSI values. For instance, from Table IV, we can see that there is no specificity information concerning C_P for $t_5 \notin V_{C_P}$, concerning C_N for $t_2, t_3 \notin V_{C_N}$, concerning both C_P and C_N for $t_2, t_3, t_5 \in V - (V_{C_P} \cap V_{C_N})$.

Secondly, as mentioned previously, $\Delta tsi_{(X:\bar{X})}(t) \geq 0$ may not hold for every $t \in V_X \cap V_{\bar{X}}$. That is, it may assign negative weights to some terms that occur in both X and \bar{X} . For instance, from Table IV, we can see $\Delta tsi_{(V_{C_P}:V_{C_N})}(t) < 0$ for $idf_X(t)$ when $t_1 \in V_{C_P} \cap V_{C_N}$ and for $int_X(t)$ when $t_6 \in V_{C_P} \cap V_{C_N}$. Therefore, when $\Delta tsi_{(X:\bar{X})}(t)$ is applied, function $w_{d|X}(t)$ given in Eq.(12) may assign a negative weight to some terms and $\varpi_x(t)$ given in Eq.(10) cannot be

TABLE III
DOCUMENT FREQUENCY DISTRIBUTIONS $P_X(d|t)$

	t_1	t_2	t_3	t_4	t_5	t_6
for calculating $tsi_X(t) = idf_X(t)$						
$P_{CP}(d_1 t)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	-	$\frac{1}{1}$
$P_{CP}(d_2 t)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{2}$	-	$\frac{1}{1}$
$P_{CP}(d_3 t)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{2}$	-	$\frac{1}{1}$
$P_{CP}(d_4 t)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{2}$	-	$\frac{1}{1}$
$P_{CN}(d_5 t)$	$\frac{3}{1}$	-	-	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{3}{3}$
$P_{CN}(d_6 t)$	$\frac{3}{1}$	-	-	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{3}{3}$
$P_{CN}(d_7 t)$	$\frac{3}{1}$	-	-	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{3}{3}$
for calculating $tsi_X(t) = int_X(t)$						
$P_{CP}(d_1 t)$	$\frac{1}{7}$	$\frac{2}{5}$	$\frac{0}{6}$	$\frac{1}{3}$	-	$\frac{0}{1}$
$P_{CP}(d_2 t)$	$\frac{3}{7}$	$\frac{0}{5}$	$\frac{2}{6}$	$\frac{0}{3}$	-	$\frac{1}{1}$
$P_{CP}(d_3 t)$	$\frac{1}{7}$	$\frac{0}{5}$	$\frac{0}{6}$	$\frac{0}{3}$	-	$\frac{0}{1}$
$P_{CP}(d_4 t)$	$\frac{2}{7}$	$\frac{3}{5}$	$\frac{1}{6}$	$\frac{1}{3}$	-	$\frac{0}{1}$
$P_{CN}(d_5 t)$	$\frac{0}{1}$	-	-	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{3}{3}$
$P_{CN}(d_6 t)$	$\frac{1}{1}$	-	-	$\frac{0}{3}$	$\frac{0}{2}$	$\frac{0}{3}$
$P_{CN}(d_7 t)$	$\frac{0}{1}$	-	-	$\frac{0}{3}$	$\frac{0}{2}$	$\frac{0}{3}$

TABLE IV
TERM SPECIFICITY INFORMATION

	$tsi_{CP}(t)$	$tsi_{CN}(t)$	$\Delta tsi_{(CP:CN)}(t)$
obtained from $tsi_X(t) = idf_X(t)$			
t_1	$\log \frac{4}{4}$	$\log \frac{3}{1}$	$-\log 3$
t_2	$\log \frac{4}{3}$	-	-
t_3	$\log \frac{4}{3}$	-	-
t_4	$\log \frac{4}{2}$	$\log \frac{3}{3}$	$\log 2 - \log \frac{3}{2}$
t_5	-	$\log \frac{3}{3}$	-
t_6	$\log \frac{4}{1}$	$\log \frac{3}{3}$	$\log 4$
obtained from $tsi_X(t) = int_X(t)$			
t_1	$4 + \frac{1}{7} \cdot \log \frac{108}{77}$	0	$4 + \frac{1}{7} \cdot \log \frac{108}{77}$
t_2	$4 + \frac{1}{5} \cdot \log \frac{108}{55}$	-	-
t_3	$4 + \frac{1}{6} \cdot \log \frac{36}{65}$	-	-
t_4	$4 + \frac{1}{3} \cdot \log \frac{4}{33}$	$3 + \frac{1}{3} \cdot \log \frac{4}{33}$	1
t_5	-	$3 + \frac{1}{7} \cdot \log \frac{432}{77}$	-
t_6	0	$3 + \frac{1}{7} \cdot \log \frac{432}{77}$	$-(3 + \frac{1}{7} \cdot \log \frac{432}{77})$

expected to produce non-negative values for every $t \in V_X \cap V_{\bar{X}}$ and, thus $p(t|X)$ given in Eq.(9) may be non-positive. The negative weights may cause a problem in estimating the posterior probability for the NB classifier.

Thirdly, the estimation of the posterior probabilities are normally the maximum likelihood estimate which are given by weights $\varpi_X(t)$ and, thus $p(t|X) = 0$ if $\varpi_X(t) = 0$. This is problematic: it wipes out all information conveyed by other terms with non-zero probabilities when they are multiplied (see Eq.(6)); it also makes $\Gamma(d, X)$ given in Eq.(7) meaningless.

B. Solutions

There may be many ways to solve the above three problems. We here suggest some simple ways which are easy to apply in practice.

1) Terms Having No TSI Value

To solve the first problem, for each $t \in V (\supseteq V_{\bar{X}})$, let us

redefine the specificity measure $tsi_X(t)$ given in Eq.(1) to

$$tsi'_X(t) = \begin{cases} tsi_X(t) \geq 0 & t \in V_X \\ \varepsilon_1 & t \in V - V_X \end{cases} \quad (13)$$

where ε_1 , called a pseudo weight, is assigned to every $t \in V - V_X$ (i.e., to those terms occurring in only \bar{X}). A similar discussion can be given to $tsi'_{\bar{X}}(t)$ with a pseudo weight ε_2 assigned to terms occurring in only X . Generally, we have

$$0 \leq \varepsilon_1, \varepsilon_2 \leq \min \{tsi'_X(t), tsi'_{\bar{X}}(t'); t \in V_X, t' \in V_{\bar{X}}\}$$

Note that V can be partitioned into three disjoint sets:

$$V = (V_X - V_{\bar{X}}) \cup (V_X \cap V_{\bar{X}}) \cup (V - V_X)$$

Thus, in the same manner, we may redefine $\Delta tsi_{(X:\bar{X})}(t)$ given in Eq.(11) to

$$\Delta tsi'_{(X:\bar{X})}(t) = \begin{cases} tsi'_X(t) - \varepsilon_2 \geq 0 & t \in V_X - V_{\bar{X}} \\ tsi'_X(t) - tsi'_{\bar{X}}(t) & t \in V_X \cap V_{\bar{X}} \\ \varepsilon_1 & t \in V - V_X \end{cases} \quad (14)$$

where $tsi'_X(t)$ is given in Eq.(13) and ε_1 and ε_2 are the above pseudo weights. A similar discussion can be given to $\Delta tsi'_{(\bar{X}:X)}(t)$.

Clearly, both $tsi'_X(t)$ and $\Delta tsi'_{(X:\bar{X})}(t)$ are meaningful over V . According to the results given in Table IV, we may simply take, for instance, $\varepsilon_1 = \varepsilon_2 = 0$ as

$$\min \{tsi'_X(t), tsi'_{\bar{X}}(t'); t \in V_X, t' \in V_{\bar{X}}\} = 0$$

for $tsi_X(t) = int_X(t)$. Thus, the results given in Table V are examples of term specificity information obtained from the redefined specificity measures.

TABLE V
MODIFIED TERM SPECIFICITY INFORMATION

	$tsi'_{CP}(t)$	$tsi'_{CN}(t)$	$\Delta tsi'_{(CP:CN)}(t)$
obtained from $tsi_X(t) = idf_X(t)$			
t_1	$\log \frac{4}{4}$	$\log \frac{3}{1}$	$-\log 3$
t_2	$\log \frac{4}{3}$	$\varepsilon_2 = 0$	$\log 2$
t_3	$\log \frac{4}{3}$	$\varepsilon_2 = 0$	$\log \frac{4}{3}$
t_4	$\log \frac{4}{2}$	$\log \frac{3}{3}$	$\log 2 - \log \frac{3}{2}$
t_5	$\varepsilon_1 = 0$	$\log \frac{3}{3}$	$\varepsilon_1 = 0$
t_6	$\log \frac{4}{1}$	$\log \frac{3}{3}$	$2 \log 2$
obtained from $tsi_X(t) = int_X(t)$			
t_1	$4 + \frac{1}{7} \cdot \log \frac{108}{77}$	0	$4 + \frac{1}{7} \cdot \log \frac{108}{77}$
t_2	$4 + \frac{1}{5} \cdot \log \frac{108}{55}$	$\varepsilon_2 = 0$	$4 + \frac{1}{5} \cdot \log \frac{108}{55}$
t_3	$4 + \frac{1}{6} \cdot \log \frac{36}{65}$	$\varepsilon_2 = 0$	$4 + \frac{1}{6} \cdot \log \frac{36}{65}$
t_4	$4 + \frac{1}{3} \cdot \log \frac{4}{33}$	$3 + \frac{1}{3} \cdot \log \frac{4}{33}$	1
t_5	$\varepsilon_1 = 0$	$3 + \frac{1}{7} \cdot \log \frac{432}{77}$	$\varepsilon_1 = 0$
t_6	0	$3 + \frac{1}{7} \cdot \log \frac{432}{77}$	$-(3 + \frac{1}{7} \cdot \log \frac{432}{77})$

2) Terms Assigned a Negative TSI Value

To solve the second problem that $\Delta tsi_{(X:\bar{X})}(t) < 0$ may hold for some $t \in V_X \cap V_{\bar{X}}$, for each $t \in V (\supseteq V_X \cap V_{\bar{X}})$, we need to further redefine $\Delta tsi'_{(X:\bar{X})}(t)$ given in Eq.(14) to:

$$\Delta tsi_{(X:\bar{X})}^*(t) = \begin{cases} (tsi'_x(t) - \varepsilon_2) + \tau_4 & t \in V_X - V_{\bar{X}} \\ \Delta tsi'_{(X:\bar{X})}(t) + \tau_3 & t \in V_X \cap V_{\bar{X}}, \Delta tsi'_{(X:\bar{X})}(t) > 0 \\ \tau_2 & t \in V_X \cap V_{\bar{X}}, \Delta tsi'_{(X:\bar{X})}(t) = 0 \\ \tau_1 & t \in V_X \cap V_{\bar{X}}, \Delta tsi'_{(X:\bar{X})}(t) < 0 \\ \varepsilon_1 & t \in V - V_X \end{cases}$$

where $0 \leq \varepsilon_1 < \tau_1 < \tau_2 \leq \tau_3 \leq \tau_4$ are called *modifying parameters* (e.g., $\tau_1 = 0.5$, $\tau_2 = 1.0$, $\tau_3 = 1.5$, $\tau_4 = 2.0$, $\varepsilon_1 = \varepsilon_2 = 0$ were used in our experiments).

Clearly, $\Delta tsi_{(X:\bar{X})}^*(t) \geq 0$ for all $t \in V$. The basic idea of taking the above modifying parameters is simple. First, we assign τ_1 to those terms having negative weight and τ_2 to those terms having zero weight; the reason $\tau_2 > \tau_1$ is because we believe that terms having negative weight are worse than terms having zero weight. To avoid losing the importance of terms representing d caused by adding τ_1 and τ_2 , τ_3 and τ_4 are also added to those terms having positive weight; the reason $\tau_4 > \tau_3$ is because we regard terms occurring in X alone as being more important in representing $d \in X$ than terms occurring in both X and \bar{X} . Finally, $\varepsilon_1 < \tau_1$ for those terms occurring in only $V_{\bar{X}}$.

3) Terms with a Zero Posterior Probability

To solve the third problem that $p(t|X) = 0$ if $\varpi_x(t) = 0$, a smoothing method may be required to assign a non-zero probability mass to those terms with $\varpi_x(t) = 0$. For instance, with the *additive smoothing* method,

$$\varpi'_x(t) = \varpi_x(t) + \theta$$

where $\theta > 0$ is a smoothing parameter (for instance, $\theta = 0.5$ was used in our experiments), the posterior probability can be rewritten by

$$\hat{p}(t|X) = \frac{\varpi'_x(t)}{\Psi'} = \frac{\varpi_x(t)}{\Psi'} + \frac{\theta}{\Psi'}$$

and Ψ' is a *normalization factor* after smoothing:

$$\Psi' = \sum_{t \in V} \varpi'_x(t) = \Psi + \theta \cdot |V|$$

where, according to Eq.(9),

$$\Psi = \sum_{t \in V_X} \varpi_x(t) = \sum_{d \in X} \sum_{t \in V_X} \pi(d) \cdot w_{d|X}(t)$$

is a normalization factor before smoothing. Thus, all the terms with $\varpi_x(t) = 0$ are then assigned to an equal non-zero probability mass $\frac{\theta}{\Psi'}$.

4) Alternative

An alternative way, which can solve both the second and third problems together and may thus be the simplest one, is:

$$\varpi_x^*(t) = \begin{cases} \varpi_x(t) + \theta_1 & \varpi_x(t) > 0 \\ \theta_2 & \varpi_x(t) = 0 \\ \theta_3 & \varpi_x(t) < 0 \end{cases}$$

where $\theta_1 \geq \theta_2 \geq \theta_3 > 0$ are smoothing parameters. Clearly, $\varpi'_x(t)$ adds an equal value θ to all terms regardless of whether $\varpi_x(t)$ is zero or negative or not. That is, $\varpi'_x(t) = \varpi_x^*(t)$ when $\theta_1 = \theta_2 = \theta_3 = \theta$ and, therefore, it is a special case of $\varpi_x^*(t)$.

V. EXPERIMENTS

This section presents some results from three sets of experiments carried out in order to verify SC effectiveness of our methods. As this study focuses on introducing a general form of a specificity measure and clarifying some potential problems of applications and suggesting solutions, rather than an extensive experimental investigation into the measure, the readers interested in empirical evidence drawn from a number of performance experiments and comparisons are referred to those papers referenced.

Our experiments used a collection from the movie review domain [19], first used in [12] and widely used in SC research. There are 2000 labelled documents in the full collection, consisting of 1000 positive and 1000 negative documents. Before using our formulae, we removed stop words and very high frequency terms (occurring in more than 60% of documents), and used a stemming algorithm [24]. We disregarded the position of terms in documents. Each document was treated as a 'bag-of-words'. Only term frequencies were considered. In our experiments, 10-fold cross-validation and the standard measures *recall* and *precision* were used for evaluation.

The first set of experiments compared the performance obtained from eleven weighting functions: eight are listed in Table I (in Example 3.1) and another three below were used as benchmarks:

$$\begin{aligned} w_d^{(F)}(t) &= f_d(t) \\ w_d^{(O)}(t) &= \frac{(a+1) \cdot f_d(t)}{a \cdot [(1-b) + b \cdot \beta(d, C)] + f_d(t)} \\ w_d^{(S)}(t) &= \frac{[1 + \ln(1 + \ln(f_d(t)))] \cdot \log\left(\frac{|C|+1}{L_d}\right)}{(1-c) + c \cdot \beta(d, C)} \end{aligned}$$

where parameters $a = 1.2$, $b = 0.75$ and $c = 0.2$, and $\beta(d, C)$ is given in Eq.(15) (see the last set of experiments below). Past experimental studies emphasised that a weighting function using just term frequency information can produce good performance for SC [1], and the Okapi (BM25) [25] and Smart [26] weighting functions have widely been recognized to produce excellent retrieval performances in IR. Table VI displays our experimental results using the eleven weighting functions, and the best results are given in square brackets in bold face.

From the results in Table VI it can be seen: Classifications obtained from (i) all the eleven weighting functions achieved good performance (above 90% recall/precision) at most evaluation points; (ii) idf, tf-idf, int, tf-int achieved consistently better performance than from tf, Okapi and Smart functions; the improvements were shown at all the evaluation points, which verifies *TSSs* are appropriate measures for SC; (iii) Δ idf, tf- Δ idf, Δ int and tf- Δ int showed a bias towards C_P ,

which resulted in a relatively low precision for C_P but a very high precision for C_N ; the cause of the bias is an interesting question, and extensive experiments may need to be carried out to train the parameters of $\Delta tsi^*_{(x:\bar{x})}(t)$; (iv) int, tf-int, Δ int and tf- Δ int were consistently better than from idf, tf-idf, Δ idf and tf- Δ idf, respectively; the improvements were shown at all the evaluation points (the reason for the improvements was explained at the beginning of Section 2.2); (v) tf-idf, tf- Δ idf, tf-int and tf- Δ int seem not to achieve the anticipated performance improvements compared with from idf, Δ idf, int and Δ int, respectively; this indicates that term specificity information may dominate the classifier performance. In addition, our experimental results bear out past experimental studies that tf can produce good performance for SC.

TABLE VI
PERFORMANCE WITH 11 WEIGHTING FUNCTIONS

	Negative Class C_N		Positive Class C_P	
$w_d(t)$	recall	precision	recall	precision
tf	0.9280	0.9460	0.9470	0.9297
Okapi	0.9350	0.9482	0.9490	0.9363
Smart	0.9290	0.9411	0.9420	0.9303
$w_{d X}(t)$	recall	precision	recall	precision
idf	0.9420	0.9684	0.9690	0.9433
tf-idf	0.9350	0.9672	0.9680	0.9367
int	[0.9560]	[0.9747]	[0.9750]	[0.9566]
tf-int	0.9440	0.9705	0.9710	0.9452
Δ idf	0.8790	0.9921	0.9949	0.8927
tf- Δ idf	0.8420	0.9903	0.9900	0.8618
Δ int	0.8860	0.9988	0.9990	0.9190
tf- Δ int	0.8630	0.9947	0.9978	0.8954

The second set of experiments considered the issue of dimension reduction of term space. Dimension reduction is an important issue in document classification, IR, NLP, and many related areas. It is generally the process of reducing the number of random variables under consideration. In our case, it is the process of identification of *informative* terms and, then, documents are represented by all the identified terms. The identified informative terms pertaining to the positive (or, negative) class are regarded as positive (or, negative) *sentiment-bearing* terms. The directed divergence measure [27] was used for the identification:

$$I(P_x(t); P_c(t)) = P_x(t) \log \frac{P_x(t)}{P_c(t)}$$

in which, $P_x(t) = P(t|X)$ (where $t \in V_X$) may be estimated using expressions given in Eq.(9), Eq.(10) and Eq.(12). Dimension reduction enables sentiment analysis to be performed in the reduced space more accurately and reliably than in the original space. A detailed discussion on informative term identification can be found in [28].

We experimentally studied classification performance using the identified informative terms to represent documents. There were 25259 distinct terms in V after stop word removal. The top δ terms of a ranked list were selected as the informative terms. We iteratively evaluated the eleven weighting functions using the δ terms, with $\delta = 14000$ to $\delta = 4000$ stepping -2000. The best results with $\delta = 10000$ are given in Table VII.

From the results in Table VII it can be seen: Classifications obtained from (i) all the eleven weighting functions achieved consistently good performance at all the evaluation points; (ii) idf and tf-idf showed better performance than using tf, Okapi or Smart at most evaluation points; (iii) tf, Okapi and Smart showed better performance compared with the corresponding performance without using the informative terms at most evaluation points (see Table VI). In addition, our experimental results (not given in this paper) showed that if the number of identified terms is reduced to less than 40% of the original size of the vocabulary, it would not be possible to improve classification performance.

TABLE VII
PERFORMANCE USING 10000 INFORMATIVE TERMS

	Negative Class C_N		Positive Class C_P	
$w_d(t)$	recall	precision	recall	precision
tf	0.9340	0.9459	0.9460	0.9351
Okapi	0.9420	0.9529	0.9530	0.9430
Smart	0.9420	0.9536	0.9540	0.9435
$w_{d X}(t)$	recall	precision	recall	precision
idf	[0.9550]	0.9539	0.9530	[0.9641]
tf-idf	0.9520	[0.9553]	[0.9550]	0.9517
int	0.9960	0.9233	0.9010	0.9952
tf-int	0.9910	0.9300	0.9170	0.9904
Δ idf	0.9210	0.9472	0.9490	0.9243
tf- Δ idf	0.9150	0.9488	0.9510	0.9192
Δ int	0.9240	0.9455	0.9470	0.9268
tf- Δ int	0.9210	0.9472	0.9490	0.9243

The last set of experiments involved the construction of the normalization factor, denoted by $\psi(d, X)$, according to the individual documents. There are many ways to construct ψ . One way is to consider a linear combination (with a parameter $\lambda > 0$):

$$\psi(d, X) = (1 - \lambda) + \lambda \cdot \beta(d, X) \quad (15)$$

where $\beta(d, X) = \frac{L_d}{ave(X)}$ is a *length moderation factor* and $ave(X) = \frac{1}{|X|} \sum_{d \in X} L_d$ is the average length of all $d \in X$. The $\beta(d, X)$ may be used to further moderate the effect of the document length across the individual classes and thus be used to construct ψ . The $\psi(d, X)$ given in Eq.(16) is used in both Okapi and Smart weighting functions.

It is thus interesting to test the usefulness of a combination with the form:

$$w_{d|X}^*(t) = \frac{w_{d|X}(t)}{\psi(d, X)}$$

where $w_{d|X}(t)$ is one of the eight weighting functions listed in Table 1, and λ is set to 0.2, 0.5, 0.8 and 1.0.

The results (not given in this paper) showed that $w_{d|X}^*(t)$ did not provide performance improvement compared with $w_{d|X}(t)$ itself whether using the informative terms or not. The reason for the worse performance is not yet clear. However, we conjecture that it may be because tf, Okapi and Smart are basically term frequency-based weighting functions, and are therefore sensitive to the document length normalization. In contrast, our methods provide term specificity-based weighting functions, thus a skewed document frequency distribution over a class plays a key role in determining SC performance.

Note that the normalization factor Ψ given in Section 4.2 serves for $\varpi_x(t)$, whereas the normalization factor ψ here serves for $w_{d|x}(t)$. That is, the former is used for the weighting function regarding classes, the latter is used for the weighting function according to the individual documents.

VI. CONCLUSIONS

This study has advocated the use of *TSI* to assess document sentiment orientation. We discussed the mathematical concept of specificity information conveyed by a given term based on Shannon's entropy, and then introduced a general form of a specificity measure in terms of the concept. Two well-known specificity measures were considered, as examples, to illustrate the general form and their relationship was established based on the general form. We introduced an intuitive concept on specificity strength of terms over the classification and, then proposed a general method to represent the statistical importance of terms pertaining to individual documents with estimation of posterior probabilities using term weights obtained from *TSI* for the NB classifier. We clarified some potential problems inherent in applying the *TSI* measures in a Bayesian learning framework and, then suggest solutions that are easy to apply in practice. We proposed a novel multiple representation method, where each term is assigned multiple weights against individual sentiment classes, and explored a method of applying existing advanced single representation IR techniques to SC. We presented some experimental results and showed that the proposed method outperforms existing advanced IR single representation techniques. We attributed this to the capacity of the proposed method to capture aspects of term behaviour beyond a single representation. Our experimental results also verified that *TSI* may be regarded as an appropriate measure for effective SC. In ongoing work we are exploring reasons why using specificity information may result in a classification bias. Due to its generality, our method can be expected to be a useful tool for a variety of tasks of document classification, IR, NLP, and many related areas.

REFERENCES

- [1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*, 2002, pp. 79–86.
- [2] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, 2002, pp. 417–424.
- [3] B. Snyder and R. Barzilay, "Multiple aspect ranking using the Good Grief algorithm," in *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)*, 2007, pp. 300–307.
- [4] T. Wilson, J. Wiebe, and R. Hwa, "Just how mad are you? Finding strong and weak opinion clauses," in *Proceedings of the 21st Conference of the American Association for Artificial Intelligence (AAAI'04)*, 2004.
- [5] M. Thelwall, K. Buckley, G. Paltoglou, C. D., and A. Kappas, "An information theoretic foundation for the measurement of discrimination information," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, 2010.
- [6] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 3, pp. 121–144, 2008.
- [7] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 1, no. 1-2, pp. 1–135, 2008.
- [8] A. Aue and M. Gamon, "Customizing sentiment classifiers to new domains: A case study," in *Proceedings of RANLP'05*, 2005.
- [9] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Computational Intelligence*, vol. 22, no. 2, pp. 110–125, 2006.
- [10] S. Matsumoto, H. Takamura, and M. Okumura, "Sentiment classification using word sub-sequences and dependency sub-trees," in *Proceedings of PAKDD'05*, 2005, pp. 301–311.
- [11] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources," in *Proceedings of EMNLP'04*, 2004, pp. 412–418.
- [12] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of 42nd ACL*, 2004, pp. 271–278.
- [13] J. Martineau and T. Finin, "Delta tfidf: An improved feature space for sentiment analysis," in *Proceedings of the Third AAAI International Conference on Weblogs and Social Media*, 2009.
- [14] J. Martineau, T. Finin, A. Joshi, and S. Patel, "Improving binary classification on text problems using differential word features," in *Proceedings of 18th ACM Conference on Information and Knowledge Management (CIKM'09)*, 2009, pp. 2019–2023.
- [15] A. McCallum and K. Nigam, "A comparison of event models for naive Bayes text classification," in *AAAI-98 Workshop on Learning for Text Categorization*, 1998.
- [16] I. Rish, "An empirical study of the naive Bayes classifier," in *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 2001.
- [17] H. Zhang, "The optimality of naive Bayes," in *The 17th International FLAIRS Conference*, 2004.
- [18] S. Caraballo and E. Charniak, "Determining the specificity of nouns from text," in *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999, pp. 63–70.
- [19] "Movie Review Data: <http://www.cs.cornell.edu/people/pabo/movie-review-data/> (2004)."
- [20] K. Sparck Jones, "A statistical interpretation of term specificity and its application to retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.
- [21] G. Salton and M. H. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.
- [22] A. Akobeng, "Understanding diagnostic tests 1: sensitivity, specificity and predictive values," *Foundation Acta Paediatrica/Acta Paediatrica*, vol. 96, pp. 338–341, 2006.
- [23] D. Cai, "Determining semantic relatedness through the measurement of discrimination information using Jensen difference," *International Journal of Intelligent Systems*, vol. 24, no. 5, pp. 477–503, 2009.
- [24] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.
- [25] S. E. Robertson and S. Walker, "Okapi/Keenbow at TREC-8," in *The 8th Text REtrieval Conference (TREC-8)*. NIST Special Publication, 1999, pp. 151–161.
- [26] A. Singhal, J. Choi, D. Hindle, D. Lewis, and F. Pereira, "AT&T at TREC-7," in *The 7th Text REtrieval Conference (TREC-7)*. NIST Special Publication, 1999, pp. 239–252.
- [27] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.
- [28] D. Cai, "An information theoretic foundation for the measurement of discrimination information," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 9, pp. 1262–1273, 2010.